# Explainable AI for Mental Disorder Detection via Social Media: A survey and outlook

Yusif Ibrahimov, Tarique Anwar, Tommy Yuan

**Abstract**—Mental health constitutes a complex and pervasive global challenge, affecting millions of lives and often leading to severe consequences. In this paper, we conduct a thorough survey to explore the intersection of data science, artificial intelligence, and mental healthcare, focusing on the recent developments of mental disorder detection through online social media (OSM). A significant portion of the population actively engages in OSM platforms, creating a vast repository of personal data that holds immense potential for mental health analytics. The paper navigates through traditional diagnostic methods, state-of-the-art data- and AI-driven research studies, and the emergence of explainable AI (XAI) models for mental healthcare. We review state-of-the-art machine learning methods, particularly those based on modern deep learning, while emphasising the need for explainability in healthcare AI models. The experimental design section provides insights into prevalent practices, including available datasets and evaluation approaches. We also identify key issues and challenges in the field and propose promising future research directions. As mental health decisions demand transparency, interpretability, and ethical considerations, this paper contributes to the ongoing discourse on advancing XAI in mental healthcare through social media. The comprehensive overview presented here aims to guide researchers, practitioners, and policymakers in developing the area of mental disorder detection.

**Index Terms**—Explainable AI, Mental health, Social data mining, Interpretable deep learning, Natural language processing



## 1 Introduction

Mental health (MH) is a complex phenomenon that significantly influences individuals' psychological well-being, affective states, and behavioral patterns. Globally, mental health issues pose a significant challenge, affecting approximately 12.5% of individuals at some point in their lifespan, thereby contributing to a substantial burden [1]. Depression, alongside anxiety, bipolar disorder, and schizophrenia, is a leading mental condition [2]. The repercussions of mental disorders are profound, impacting both mental and physical well-being and often leading to improper behaviors [3]. This global challenge extends to an estimated 270 million individuals [4]. These conditions are identified by the World Health Organization (WHO) as a leading cause of suicide, claiming around 800,000 lives annually [1].

The delay or avoidance of seeking professional help for mental health issues is a common phenomenon, often influenced by societal stigma and various circumstantial factors [5]. This behavior exacerbates critical situations, particularly in advanced stages of mental disorders, where individuals experience a significant decline in their ability to perform routine activities. This can manifest as impaired cognitive functioning, self-harming behaviors, and, in extreme cases, culminate in suicidal acts. The challenges in accessibility faced by healthcare practitioners during critical situations, such as pandemics and armed conflicts, add further complexity [6], [7]. Hence, there is an urgent need to explore novel paradigms and develop innovative methodologies for the early identification and management of these mental health challenges. Doing so not only alleviates the burden on individuals but also supports healthcare professionals in addressing mental health crises. In recent years, the intersection of data science, artificial intelligence (AI) and mental healthcare has witnessed remarkable advancements [8]. Researchers are making rapid progress in developing novel AI- and data-driven solutions for mental healthcare, including applications such as early diagnostics [9], AI-driven therapy [10], and monitoring through regular screening [11].

A significant portion of our society (59.9% approx. [12]) actively engages in online social media (OSM) platforms [13]. Particularly among the youth, there is a notable inclination to discuss sensitive topics on online platforms rather than in-person interactions [14]. Individuals share their emotions, daily routines, problems, ideas, and health conditions on OSM, generating vast amounts of personal data that hold potential for mental health analytics [7], [15], [16]. Recent studies propose OSM as a futuristic solution for continuous mental health care [17]. Significant advancements are being made in deep learning for the development of models like DepressionNet [18] and EDNet [6] for detecting mental disorders. However, while deep learning models are valuable, their adoption in healthcare demands explainability [19]. Many of these models operate as black boxes, rendering the reasoning behind their decisions unclear. Given the critical nature of healthcare decisions, reliance on black-box models raises safety concerns, as they cannot guarantee 100% accuracy [20]. Explainable AI (XAI) models offer a solution by shedding light on the decision-making mechanisms of AI models. Therefore, investigating deep learning models at a low level is crucial for achieving explainability and developing robust XAI models for detecting mental disorders.

The existing survey and review papers extensively examine the use of AI on social media for mental disorder detection; however, they generally focus on text mining tools on social media data and ignore the nature of social

media and the importance of social interactions among social media users, as do many researchers. Furthermore, the survey and review publications show little concern for explainability and interpretability [8], [21], [22], [23], [24], [25].

In summary, this paper explores the recent advancements in the field of Explainable AI in the context of mental disorder detection (MDD) through social media. To provide a comprehensive understanding of the domain, we commence with an overview of traditional diagnostic methods for treating mental disorders in Section 2. Subsequently, we delve into recent data- and AI-driven research studies in Section 3. This section covers the state-of-the-art (SOTA) machine learning (ML) methods (modern deep learning (DL) specifically) for MDD, outlining their respective advantages and disadvantages, and offering our perspectives on their applicability. Notably, research on XAI for mental healthcare, as well as general healthcare, is in its early stages. Section 4 presents a review of existing research on explainability, accompanied by our insights that contribute to the development of XAI solutions for mental healthcare. To facilitate a deeper understanding of the experimental setup, Section 5 outlines prevalent practices in experimental evaluation and existing datasets. This encompasses the selection of experimental datasets and evaluation approaches. After presenting the existing research, we delve deeper in Section 6 to present a comprehensive outlook on this research area. Section 6.1 identifies and discusses the key issues and challenges inherent in this research. Looking towards the future, Section 6.2 outlines promising research directions and our views for the continued advancement of XAI in mental healthcare via social media. Finally, Section 7 concludes the paper.

## 2 Traditional Diagnostic Methods

Mental health conditions are typically diagnosed by healthcare professionals (e.g. therapists and psychologists) through face-to-face interviews with patients. Additionally, the patients may be required to respond to pre-defined standard questionnaires to enhance the quality of diagnosis, with each type of mental disorder often having its own set of self-questionnaires [26], [27], [28], [29]. This comprehensive diagnostic process is guided by individuals' experiences, with international standards such as the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) employed for diagnosis [7], [30]. DSM-5 [30] is a manual listing all mental disorders, along with their standard diagnostic methods. Main symptoms of various mental disorders are identified through these methods. For example, depression is characterised by a loss of interest in previously enjoyed activities, feelings of hopelessness, reduced motivation, concentration, energy and libido, and changes in cognition [30]. Anxiety manifests through symptoms like tachycardia, dizziness, shortness of breath, trembling, worry, anger, and fear [30]. Eating disorders are identified through signs such as problematic eating habits, abnormal changes in weight, changes in mood and unhappiness with body shape [30]. Several standardised questionnaires play a pivotal role in assessing mental conditions. The depression inventory by Beck et al. [26] and the CES-D Scale [27], each comprising 21 and 20 questions, respectively, assess the mental status for depression diagnosis. Gold standards for estimating depression severity include the Beck Depression Index-II (BDI-II) [31], Hamilton Depression Rating Scale (HDRS) [32], and Patient Health Questionnaire-9 (PHQ-9) [33], focusing on key depression symptoms such as sadness, pessimism, loss of interest, and fatigue. Recognised questionnaires for anxiety detection include the Generalised Anxiety Disorder 7 (GAD-7) [34] and Beck Anxiety Disorder [35]. For the assessment and diagnosis of eating disorders (EDs), the Eating Disorder Examination Questionnaire (EDE-Q) [28] and the Eating Attitudes Test (EAT-26) [29] are employed.

Despite the well-established area of mental healthcare relying on traditional diagnostic methods, there are inherent limitations in the current approaches [36], [37]. $(i)$ The *frequency of monitoring sessions* is typically limited to once per week. It hinders the ability to track the course of mental conditions in real-time. $(ii)$ The *subjective approach* of clinicians may occasionally lead to instances where professionals employ subjective judgement, introducing potential inaccuracies in diagnosis . $(iii)$ *Accessibility to therapy sessions* may be negatively affected by factors such as financial constraints and societal circumstances, particularly during crisis like pandemics and wars. $(iv)$ *The phenomenon of patient prejudice* is seen quite often, as patients' responses to self-questionnaires may exhibit bias due to individual interpretations, contextual factors, and potential misunderstandings of the questions, including intentional concealment of information. $(v)$ Additionally, *humanity barriers*, including cultural diversity, can pose challenges for patients in effectively communicating their needs due to language and traditional limitations.

For a timely detection of mental disorders while addressing the existing limitations, it is crucial to explore widely accessible, impartial, consistently monitored, and immediately available alternatives. As mentioned earlier, the ubiquitous OSM platforms are used by a vast majority of individuals, and therefore serve as a repository of substantial volumes of individual data. To address the limitations observed in conventional diagnostic approaches, this study aims to examine preliminary diagnostic methods for MDD using data-driven techniques. Additionally, we will conduct a thorough evaluation of the strengths and weaknesses of current data-driven methods, followed by providing our views on potential strategies to mitigate identified limitations.

## 3 Data-driven Methods for Detection

Healthcare approaches are rapidly evolving with technological advancements and the integration of data- and AI-driven techniques, contributing to enhanced outcomes [6], [8], [18]. A particularly promising area where technology can play a pivotal role is the timely detection of mental disorders. Recent research progresses suggest a foreseeable future in this direction. While various individual data sources can contribute to MDD, our primary focus in this paper centers on text-based OSM platforms, such as 𝕏 (formerly Twitter) and Reddit. These platforms, known for their widespread usage and expressive nature, contain valuable content such as users' posts, online behaviors, and social

TABLE 1: Some common mental disorders, their definitions, and indicative examples of OSM posts

| Mental disorder | Definition | Example of OSM post |
|---|---|---|
| Depression | Negative changes marked by affect, cognition, mood, and neurovegetative functions lasting at least 2 weeks [7], [30]. | *Can Someone Cheer Me Up? - I'm diagnosed with depression. I will sometimes out of no where...* |
| Anxiety | Conditions characterised by excessive fear and anxiety, along with associated mental and behavioral disorders [30], [38]. | *Anxiety is seriously affecting my life and it is ruining my life and I don't know how I'm going to provide for myself in the future* |
| Eating Disorder | A persistent disturbance in eating behavior that leads to altered food consumption and significantly impairs mental and behavioral functioning [6], [30]. | *Getting treatment for an eating disorder is awful and I wish I never started it. (Vent / cry session haha)* |
| Bipolar Disorder | Abnormally and persistently elevated, expansive, or irritable mood, along with increased activity or energy [30], [37]. | *Finally after years long struggle I considered seeking a psychiatrist. And just diagnosed bipolar disorder. She prescribed me with two medicines* |
| Schizophrenia | Abnormalities in delusions, hallucinations, disorganised thinking, and grossly disorganised or abnormal motor behavior [30], [39]. | *Hi. I'm **** and I'm a ******** with schizophrenia and I take a medication. I also have evil hallucinations and it scares me. Is it my schizophrenia or is it really the Devil?* |
| Suicide Ideation | Thoughts or contemplation of taking one's own life or self-inflicted harm [30], [40]. | *I'm having a really hard time with a lot of things in my life right now. My father started to verbally abuse me and my sisters since I was 6 or 7. I want to commit suicide.* |
| Obsessive-Compulsive Disorder (OCD) | Characterised by the presence of recurrent and persistent thoughts, urges, or images, along with repetitive behaviors or mental acts [30], [41]. | *I've been going to a doctor to see if I had autism and while there they also diagnosed me with OCD.* |

network interactions. Table 1 summarises the most frequent mental disorders along with their definition and example.

In our exploration of current methodologies for MDD via OSM, we observe a common workflow, depicted in Figure 1. This standard process involves several key steps: MH data collection, preprocessing, MH feature extraction, MDD modelling and evaluation, and explainability. The MH data collection involves sourcing necessary data from various OSM platforms. Following this, data pre-processing cleans and refines the raw data through basic cleaning and denoising procedures. MH feature extraction transforms it into a modeling-friendly representation of MH-related features using various techniques (Section 3.1). The MDD modelling step generates a model that is able to extract interesting patterns and insights specific to MH, providing a basis for accurate assessments for MDD. This modelling may follow supervised, unsupervised, and semi-supervised approaches. However, the literature has found supervised models as the most promising (Section 3.2). The subsequent explainability step generates insights into the decision-making process of MDD models, enhancing their overall interpretability (Section 4). Figure 2 shows a detailed pipeline of XAI methods for MDD from a technical perspective. All the techniques shown in the pipeline are discussed with sufficient details in the following sections.

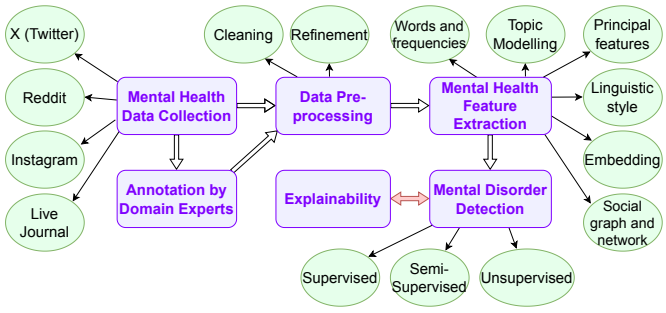



Fig. 1: Workflow of data-driven methods for MDD

### 3.1 Mental health feature extraction

The OSM data used for MDD contain lots of information in various formats, including predominantly textual contents in unstructured form, associated metadata, user behavioural data, and social communication networks. Transforming this raw data into a set of relevant features or a numeric representation is essential for ML models to process and make sense of them. An effective feature extraction and representation enables a deeper understanding of the underlying patterns and trends within the OSM data.

**Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and N-gram** [42], [43], [44], [45], [46], [47], [48], [49], [50] are fundamental feature extraction techniques in NLP. BoW emphasises word existence, TF-IDF considers word frequency within documents, and n-gram extracts contiguous sequences of $n$ tokens for vectorisation.

**Latent Dirichlet Allocation (LDA)** [51], [52] is an iterative probabilistic model that assigns topics to documents, which makes it useful for uncovering latent thematic structures within textual data. The extracted topics can be used as features for ML modelling.

**Principal Component Analysis (PCA)** [53], [54] is a dimensionality reduction technique used to identify and retain the most important or principal features, while avoiding the curse of dimensionality from a dataset. These features can be used for ML modelling.

**Linguistic Inquiry and Word Count (LIWC)** [47], [54], [55] is a text analysis tool that examines psychological and linguistic content by categorising words into various categories, such as emotions, social processes, cognitive processes and linguistic style. LIWC scores are values assigned to the words and can play the feature roles in data analysis.

**Word2Vec** [56] leverages neural networks to learn numeric representations of different words from a large corpus. It is widely used today for its simplicity and effectiveness. There exists two primary approaches: continuous bag of words (CBOW) and skip-gram. There are pre-trained versions of the Word2Vec model which can easily be used to vectorise the textual inputs for modelling [57], [58].

**Global Vectors for Word Representation (GloVe)** [59]

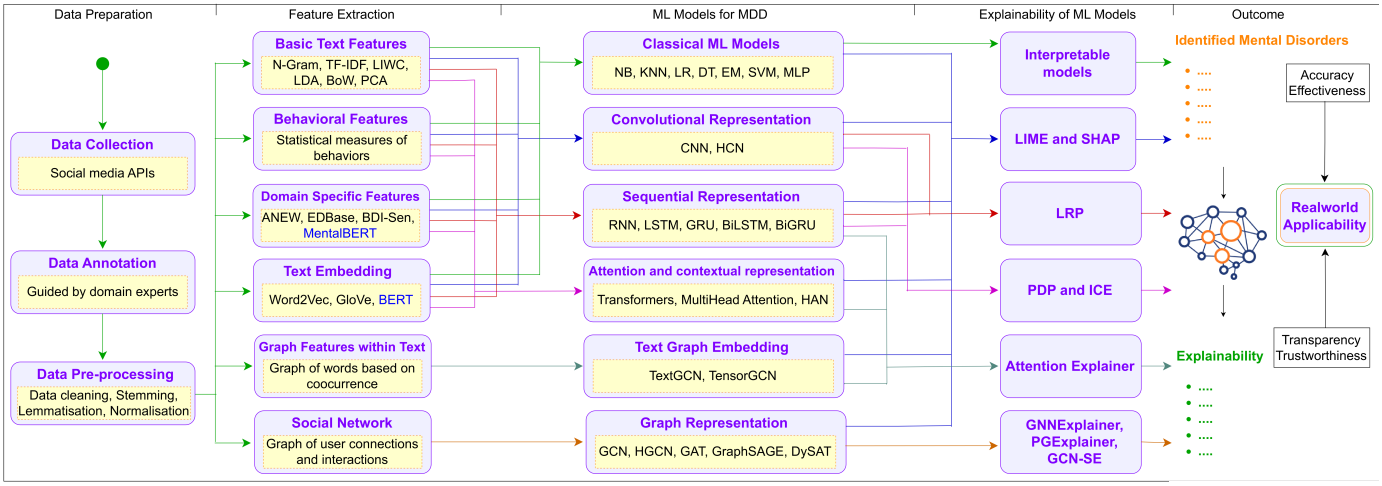



Fig. 2: Explainable AI pipeline for MDD via social media

is a statistical model used for text embedding similar to word2vec. It is an unsupervised learning algorithm that maps textual features into meaningful vector spaces. Like word2vec, GloVe is a pre-trained model as well, and can be used for textual vectorisation of textual mental health related documents [47], [57], [60], [61], [62].

**Bidirectional Encoder Representations from Transformers (BERT)** [63].

**Graph representation of text** [9]. Representing textual inputs as a graph structure enables to capture intricate structural patterns within words and documents. Examples like TextGCN [64] and TensorGCN [65] demonstrate effective graph-based representations that can be adopted for mental health analytics. These models analyse graphs constructed with words, sentences, and their relations using graph convolutional networks (GCN) [66] (detailed later).

**Social network representation** [67], [68]. A substantial portion of our population is actively engaged in OSM [12]. Research suggests a correlation between individuals and their friends with regards to the likelihood of experiencing mental disorders [69]. Thus, incorporating the social network of users/posts in addition to user-specific features is advantageous for MDD.

## 3.2 Machine learning on social data for MDD

The task of MDD from social data poses the challenges of dealing with high-dimensional, imbalanced, and non-linear data, along with the need to understand complex mental health related properties and social relationships. As classical ML techniques often rely on robust feature engineering, it is difficult for them to deal with such complicated data and the associated challenges. Therefore, recent advancements have seen a paradigm shift towards DL-based methods, surpassing those based on classical ML, in their abilities without the need for feature engineering [8]. Given that much of the OSM data are textual, current studies often employ deep neural networks (DNNs), such as recurrent neural networks (RNNs), and their variants like long short-term memory (LSTM) and gated recurrent unit (GRU) to learn sequential representations. Additionally, various other types of DNNs are utilised to capture different properties. In this section, we explore the different types of representation learning using these DNNs, and discuss the studies that leverage them for MDD.

### 3.2.1 *Convolution-based representation learning*

The convolution operation is a fundamental concept that represents the combination of two signals to produce a third signal. The power of convolution paved the way for convolutional neural networks (CNNs) [70]. They have found applications across various domains, including natural language processing. CNN models are capable of recognising patterns in textual input embeddings which can be beneficial to identify signs of the mental disorders in OSM posts. Several studies have harnessed their capabilities to explore and understand OSM texts related to mental health [71], [72], [73]. Gaur et al. [72] developed a dictionary of suicide risk severity using medical knowledge bases and suicide ontology to detect clues about suicidal thoughts and actions. They also generated a labelled dataset of suicide risk severity from Reddit. For their multiclass classification, they considered a simple CNN model. Similarly, Zogan et al. [71] employ CNNs along with other networks to detect depression on 𝕏 and analyse the impact of depression during the pandemic. The authors introduce an advanced model based on CNN and hierarchical attention incorporating both word- and tweet-level encodings to identify depressive tweets. Nevertheless, there are notable limitations to this study. The classification focuses solely on texts, treating depression detection as a binary classification task. However, assessing the severity or level of depression in a fine-grained manner would be more helpful. The social and communication network of users and textual posts are not considered in the study.

### 3.2.2 *Sequential representation learning*

In handling long-term sequential data generated on OSM and exhibiting effective parameter-sharing capabilities, RNNs outperform classical multilayer perceptrons (MLPs) and CNNs. Among the various RNN types, our focus is on three main types: vanilla RNN, LSTM, and GRU. The utilisation of sequential representation learning is crucial,

as representing textual posts as sequences allows for a more precise detection of mental disorders. This approach enables an accurate capture of relationships among words and sentences.

**Vanilla RNNs**. In these networks, the hidden layer and output at time $t$ are calculated as $y_t = f(W_{xh}x_t + W_{hh}y_{t-1})$ and $h_t = g(W_{hy}y_t)$, respectively, where $W_{xh}$, $W_{hh}$ and $W_{hy}$ are trainable parameters, $x_t$ is the input at time $t$, and $f(\cdot)$ and $g(\cdot)$ are differentiable activation functions. Vanilla RNNs are susceptible to the problem of vanishing gradients.

**Long Short-Term Memory (LSTM)** [74]. It is an RNN variant that addresses the vanishing gradient problem inherent in vanilla RNNs and handles long-term dependencies more effectively. An LSTM node comprises three gates - *input* gate, *forget* gate, and *output* gate - which collectively determine how to update the cell and hidden states. The *candidate cell state* and *cell state* are responsible for short-term and long-term memory storage, respectively. *The hidden state* is the output of the LSTM at each time step. In a regular (unidirectional) LSTM network, information flows from the first element of a sequence to its last element in a single direction. It may be limiting its potential in cases where the meaning of a word or sequence depends on both preceding and succeeding elements (e.g., OSM texts). BiLSTMs, on the other hand, process the input data in both forward and backward directions simultaneously, allowing them to capture contextual information from both sides.

**Gated recurrent unit (GRU)** [75]. It is another RNN variant that addresses the vanishing gradient problem with a simpler architecture than LSTM using *reset* and *update* gates to regulate the flow of information. In GRUs, there is no explicit memory cell with long-term memory features. Instead, a *candidate hidden state* is introduced to update the *hidden state*, which is the output of the GRU model. Similar to BiLSTMs, BiGRUs capture the contextual information from both forward and backward directions.

**MDD using RNNs**. Due to their capabilities of learning sequential representations from OSM data, RNNs have been widely employed in previous studies on MDD [7], [62], [76]. In their study, Ghosh and Anwar [7] approach MDD as a regression problem, aiming to provide a fine-grained assessment based on the disorder's intensity. They extract a variety of features and process them through an LSTM network with Swish activation function to predict the intensity of depression. The study utilised an existing $\mathbb{X}$ dataset labeled with binary classes related to depression diagnosis. Given its binary nature, the authors applied a weak-labeling approach to assign depression intensities using measures of sadness and semantic similarity with depression. They employed a self-supervised learning approach to predict depression severity. While the model yielded promising results, there are several aspects that *require further consideration*. *First*, the model employed an NLP-based relabeling technique without the involvement of mental health professionals and without a deep exploration of the psychological underpinnings. *Second*, for tweets known for their informal linguistic style, a more in-depth analysis of semantic meaning with advanced deep learning architectures and NLP techniques could enhance the model's performance *Finally*, the model does not account for user interactions

#### 3.2.3 Attention and contextual representation

OSM posts often exhibit clear signs of mental health problems, and assigning higher importance to these signs can be vital for accurate MDD. In the context of DL, emphasising "attention" to specific language patterns within OSM posts is likely to enhance the performance of MDD models [77]. The attention mechanism was first proposed by Bahdanau et al [78] to enhance sequence-to-sequence (Seq2Seq) performance by focusing on specific important parts of the input sequence. In 2017, Vaswani et al. [79] employed attention to introduce the transformer architecture. The attention mechanism involves three key matrices - query ($Q$), key ($K$), and value ($V$) - learned from the input data during the training process. The query represents a specific aspect of the input data that the attention mechanism wants to focus on. The key provides information that can be matched against the query. It helps in deciding which information is most relevant to the query. The value contains the actual information associated with the key, which is used to respond to the query. These matrices are defined as $Q = W^q \times \mathbf{X}$, $K = W^k \times \mathbf{X}$ and $V = W^v \times \mathbf{X}$, where $W^q$, $W^k$ and $W^v$ are trainable weight matrices corresponding to the query, key, and value components and $\mathbf{X}$ is the input data or the sequence of elements to be used. The attention values are computed using scaled dot product attention shown in Equation 1, where $d_k$ is the dimension of the query and key. The scaling by $d_k$ is applied to achieve gradient stability and mitigate vanishing/exploding gradient problems.

$$Attention(Q, K, V) = softmax(\frac{Q \times K^T}{\sqrt{d_k}}) \times V \quad (1)$$

By stacking the scaled dot product attention, the Transformer introduces a key component called multi-head attention shown in Equation 2, where $h_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$, and the parameter matrices vary for each head (self-attention layer).

$$MultiHead(Q, K, V) = Concat(h_1, h_2, ..., h_h)W^O \quad (2)$$

**Transformers and BERT**. After the introduction of Transformers [79], it sparked a revolution in NLP. Unlike traditional sequence-to-sequence models, Transformers rely on the powerful attention mechanism and do not require recurrent connections, making them highly effective for a wide range of NLP tasks. Transformers play an important role in capturing long-term dependencies and developing a contextualised understanding of OSM posts, which enables the detection of anomalies in users' mental health. The Transformer architecture comprises an encoder and a decoder. The encoder consists of six stacked layers, each containing multi-head attention, feed-forward, and layer normalisation units. This architecture allows the model to capture complex dependencies in the input data. The decoder, on the other hand, also consists of six identical layers but employs *masked* multi-head attention to prevent the model from attending to future tokens in the sequence, avoiding data leakage. The Transformer architecture marked a milestone in modern NLP, which led to the development of pretrained language models. BERT [63] is one of the most widely used pretrained language models for obtaining contextualised embeddings for words and sentences. It adopts a bidirectional multilayer transformer model, comprising a stack of

transformer encoders. There are two primary versions of BERT: $BERT_{BASE}$ and $BERT_{LARGE}$. The $BERT_{BASE}$ model consists of 12 transformer layers with a hidden size of 768 and 110 million trainable parameters, while $BERT_{LARGE}$ features 24 transformer layers with a hidden size of 1024 and 340 million parameters. The BERT model can be fine-tuned by adding task-specific layers, such as a fully connected layer. Then it can be used for a wide range of tasks, including mental health research with OSM texts. There has been some follow-up works on extending BERT with unique features and improvements, catering to different requirements in natural language understanding and processing. For example, *DistilBERT* [80] is a distilled, smaller and faster variant of BERT with fewer parameters. *RoBERTa* [81] is an optimised variant of BERT. It is trained on more data, omits the next sentence prediction task of BERT, utilises dynamic masking during training and improves the contextual understanding of texts.

**MDD using transformers**. The transformer-based models have had a significant impact on research in mental health analytics [6], [18], [82], [83], [84], [85]. Ragheb et al. [82] developed a comprehensive ensemble method for detecting at-risk users with transformer-based models. In their method, they introduce some noise into the textual feature representation generated by transformer-based contextual embedding models. Th resulting noise-induced representation is then used for classification to identify the at-risk users. The noisy learners help in generalising the model for multiple disorders such as depression, anorexia, self-harm and suicide, and reducing overfitting. However, the proposed method focuses only on post contents relying on clear expressions of disorders, which makes it less suitable for early detection. Furthermore, although their model is designed to detect three different disorders, there is no parameter sharing among them and each disorder is treated independently. In [83], Cao et al. adopted masked language structures to feed transformer-based contextual embedding models in order to detect latent signs of suicide ideation with an LSTM model. A limitation of their approach is the disregard of severity degrees associated with suicide ideation, a critical consideration, especially when assessing latent risk detection. Additionally, the method neglects the interactions among users and their social network friends, which often provide useful clues about the mental state of users. Utilising the 𝕏 dataset [86], Zogan et al. [18] developed DepressionNet. It leverages both user behavior and posts, and extracts a wide range of features including social network activities, emotional content, domain-specific information, and topic-related attributes. DepressionNet uses extractive-abstractive summarisation technique to summarise the historical posts of each user. This process begins by embedding each post with BERT. Subsequently, important posts are identified using K-Means clustering. Then BART [87], a denoising autoencoder for pre-training sequence-to-sequence models, is employed to generate abstractive summaries of the posts. These summarised posts are further embedded and passed through BiGRU and attention layers before being concatenated with the stack of BiGRU vectors representing user behaviors. While DepressionNet takes a step forward in depression detection, it has *two major limitations*. *First*, this model treats the problem of depression detection as a binary classification task, the classes being whether a person is depressed or not. Instead of binary classes, a more fine-grained severity levels or a severity measure will be a more informative outcome. *Second*, it utilises the textual posts and user behaviours for representation learning, but does not consider any kind of social or communication network among users and posts. The network contains some very useful information for MDD, given that the mental health related thoughts and experiences are propagated through such networks on OSM paltforms.

Abuhassan et al. [6] introduced EDNet, a multimodal deep learning model designed to identify the different types of OSM users engaged in Eating Disorders (EDs). This model integrates various data sources, including historical posts, user biographies, and online behaviors from 𝕏, and employs a multi-class classification approach to differentiate between distinct user types, such as ED-users, healthcare professionals, communicators, and non-ED users. EDNet consists of several deep neural layers, including an input layer, an embedding layer, a representation layer, a behavior modeling layer, and an output layer. This layered architecture facilitates a comprehensive understanding of diverse modalities within OSM data, encompassing users' historical tweets, biographies, and online behaviors. The authors employed the BERT model to generate contextual embeddings for textual tweets and user biographies, while incorporating variations of temporal convolutional layers and Bi-GRU layers to capture intricate patterns associated with EDs. However, the model specifically focuses on English language, which limits its effectiveness when applied to other languages. Another challenge arises when a user undergoes changes in their engagement type over time. For example, an ED user becoming non-ED a month later, or a communicator starting to experience ED. This kind of occasional behaviors leads to misclassifications, especially when users frequently use phrases associated with EDs in their posts. Furthermore, it does not give a fine-grained information about the severity of an ED user.

#### 3.2.4 Graph representation learning

Current research emphasises the significant impact of social connections on mental wellbeing [69]. There exists a correlation between an individual's mental health and the mental health status of their friends. Individuals are 93% more likely to experience depression if one of their friends is experiencing it [69]. To comprehensively understand users' mental wellbeing, capturing the graph representation of their social networks is essential. This approach discovers structural patterns, information flow dynamics, and community-level interactions. It aids in identifying their support systems, detecting isolation, and understanding the dissemination of mental health-related information within the network. Integrating this graph representation with textual information of posted contents provides a holistic view, revealing insights into users' contextual interactions. Additionally, the posts themselves also contain useful structural information that can be captured by a graph representation of their textual contents [64]. Recognising the capabilities of DNNs, it is imperative to explore DL models that are good at working with graph structures. It takes our attention to graph neural

networks (GNNs). They utilise message passing to gather information from neighboring nodes and edges, enabling the learning of node and edge representations based on complex structural information inherent in the input graph. There are several types of GNNs depending on the message passing and node/graph based representation, including GCN, GAT, GraphSAGE and DySAT.

**Graph Convolutional Network (GCN)**. An ordinary GNN aggregates representations from neighboring nodes, and therefore the nodes with large degrees tend to have a significant influence in their representations. On the other hand, those with smaller degrees have a minor influence. This can lead to a gradient explosion problem. This issue is addressed by GCNs [66] using fast approximate spectral graph convolution. They are able to capture long-term dependencies and complex representations over the graph effectively. The updated hidden layer of nodes (or activations) $H^{(l+1)}$ in the $(l+1)$-th layer of a GCN, after applying a graph convolution operation, is computed as $H^{(l+1)} = \sigma\left(\tilde{D}^{-0.5}\tilde{A}\tilde{D}^{-0.5}H^{(l)}W^{(l)}\right)$, where $H^{(l)}$ denotes the node activations in the $l$-th layer, $\sigma(\cdot)$ denotes the activation function, $\tilde{A} = A + I_N$ is an adjacency matrix with added self-connections, $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ is a diagonal matrix of $\tilde{A}$, $W^{(l)}$ is a layer-specific trainable weight matrix, and $H^0 = X$ is the initial node representations. As GCNs use normalised Laplacian matrix for convolution and aggregation of node features, they have to face scalability challenges, especially in large graphs. This limitation has led to the development of more scalable and flexible GNN architectures.

**Text graph embedding**. While textual contents are generally represented with text-based embeddings, they also contain some structural information. It enables them to be represented as graphs and be processed through GNNs. For example, TextGCN [64] and TensorGCN [65] adapt GCN on graph structures extracted from textual contents for text classification. TextGCN [64] constructs a graph based on the given set of documents and the words in them. A node is created corresponding to each word as well as document, and connections among them are established using pointwise mutual information (PMI) and TF-IDF. Similarly, TensorGCN [65] aims to enhance the comprehension of syntactic, semantic, and sequential relationships within text data. Operating on graph structures where nodes represent words and textual documents, TensorGCN establishes connections between words and documents using TF-IDF. The model integrates tensors representing three distinct graphs: a semantic-based graph, a syntactic-based graph, and a sequential-based graph. Semantic-based connections are determined by LSTM-generated embeddings' similarity, syntactic-based edges are extracted using the Stanford CoreNLP parser for grammatical dependencies, and sequential-based edges are formed through PMI. Following graph construction, TensorGCN employs inter-graph and intra-graph propagation processes. Intra-graph propagation gathers and consolidates information within a given graph, while inter-graph propagation facilitates information exchange between distinct graphs within the tensor. These graph creation and propagation methods enable the application of GCN for text classification. The scalability issues present in GCN are inherently present in both TextGCN and TensorGCN, making them difficult to work with long texts.

**Graph Attention Network (GAT)** [88]. GATs address the limitations of GCNs in terms of scalability by incorporating attention mechanism within the GNN. The attention enables the nodes to selectively aggregate information from their neighbors based on their importance. Each neighboring node is assigned an attention coefficient, which is used to make more contributions from more important nodes to the aggregation process. The attention coefficients $\alpha_{ij}$ between nodes $i$ and $j$ is computed using Equation 3, where $||$ represents concatenation, $h_i$ and $h_j$ are hidden representations of $i$ and $j$, and $\mathbf{a}$ and $\mathbf{W}$ are trainable parameters.

$$\alpha_{ij} = \frac{\exp(LeakyReLU(\mathbf{a}^T[\mathbf{W}h_i||\mathbf{W}h_j]))}{\sum_{k\in\mathcal{N}_i}\exp(LeakyReLU(\mathbf{a}^T[\mathbf{W}h_i||\mathbf{W}h_k]))} \tag{3}$$

The information from neighboring nodes $j \in \mathcal{N}_i$ of $i$ are then aggregated based on their attention coefficients $\alpha_{ij}$ and learnable parameters $\mathbf{W}$, and passed through a non-linear activation function $\sigma$, to generate its updated feature representation $h'_i$ from $h_i$, as shown in Equation 4. To stabilise the learning process of self-attention, the original GAT model employs multihead attention [79], with K = 3, using Equation 5, where $||$ represents concatenation, $\alpha^k_{ij}$ is the attention coefficient computed by the $k$-th attention mechanism, and $\mathbf{W}^k$ is the corresponding weight matrix. Using multiple heads allows the model to capture different aspects of node relationships.

$$h'_i = \sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}\mathbf{W}h_j\right) \tag{4}$$

$$h'_i = ||_{k=1}^{K}\sigma\left(\sum_{j\in\mathcal{N}_i}\alpha^k_{ij}\mathbf{W}^k h_j\right) \tag{5}$$

**GraphSAGE** [32]. Traditional GNNs like GCN and GAT require knowledge of all nodes in the graph during training, including those that will be part of the test set, leading to data leakage concerns. GraphSAGE tackles this problem by adopting a novel approach. The main idea behind GraphSAGE is to learn feature aggregation methods from local neighborhoods by utilising node sampling techniques. It significantly enhances the efficiency, flexibility, and scalability of GNNs, making them suitable for larger graphs. In a nutshell, GraphSAGE takes as input a graph $\mathcal{G}(V, E)$, node features, a depth parameter ($K$), trainable weight matrices, a non-linear activation function, and an aggregator function. The aggregator function plays a pivotal role. It aggregates information from a node's sampled neighbors and concatenates it with the node's own representation from the previous layer. Subsequently, these concatenated representations are normalised at each layer. The final embeddings are obtained from the last layer at depth $K$. Depending on the given context and specific requirements, different aggregation methods can be employed in GraphSAGE, including mean aggregators, pooling operations or aggregation based on neural networks.

**MDD using GNNs**. Some recent studies have adapted GNNs for MDD, in order to learn structural patterns within

data. This is done by utilising users' textual graphs generated from posts [9], [89] and social networks [67], [68], [90], [91], [92].

Depression severity assessment is a pressing concern within MDD. Naseem et al. [9] treat this problem as a fine-grained classification task, in which they consider four classes of varying severity from *minimal* to *severe*. Their method utilises several layers of DNNs and leverage a dataset extracted from Reddit. The method begins with the preprocessing of user posts, which are then processed through a representation layer with TextGCN. It generates their numeric embeddings by constructing a textual graph of posts (as nodes) based on their constituting words, and applying GCN on it to learn the structural post (node) representations. These are fed into a BiLSTM layer augmented with an attention mechanism to learn sequential properties. As the severity classes are ordinal in nature, the resultant vectors are passed into an ordinal classification layer to generate the output severity class. While this method exhibits promising results, it's important to note that textual data alone may fall short in identifying the depression severity level. Depressed individuals may not always explicitly post about their mental health status on OSM platforms. A more holistic approach that incorporates historical and behavioral data is important for a generalised model. Furthermore, leveraging social interactions and communications between OSM users can enhance the reliability and effectiveness of such models. Along a similar direction, GHAN [89] utilises textual posts extracted from Reddit to identify the various levels of suicide ideation, including *support*, *indicator*, *ideation*, *behavior*, and *attempt*. TensorGCN is employed to exploit semantic, syntactic, and contextual information for embedding, after which an attentive transformer encoder is used to capture temporal information across the user's posts. As the classes are ordinal, an ordinal classification layer is used to enhance the model's performance. While the model demonstrates great performance compared to the baselines, its generalisation may face challenges related to factors such as language and demographics. Although linguistic elements have proven beneficial, there are instances where they may be insufficient.

As discussed earlier, social interactions play a significant role in our mental health. Building upon this, Liu et al. [90] developed a heterogeneous network approach by taking into account an array of factors, encompassing social interactions, personality traits, social status, physical health, and overall wellbeing. A heterogeneous graph is constructed by amalgamating these diverse facets, in which each user is represented as a node and multiple types of links are established between different pairs of users. Thus, the MDD task is treated as a node-classification problem. The leveraged data is obtained from various sources including smartphones, social media accounts, and wearable devices. While this approach looks comprehensive and promising, one major challenge is the accessibility and fusion of individual data from different sources, posing a logistical challenge. Introduced by Pirayesh et al. [91], MentalSpot is another method incorporating social interactions. To enable this study, the authors firstly generated a dataset (PsycheNet) with users' social contagion network from 𝕏. The method begins by embedding user tweets with GloVe, followed by deploying 1D convolutional maps, which considered the top-k friends of each user. While the authors admirably considered the social contagion among users, they did not delve into GNNs. They use only the textual data of tweets and treat MDD as a binary classification task. To identify suicide ideation among OSM users, Sawhney et al. [92] developed a model (hyper-SOS) based on hyperbolic GCN (HGCN) [93]. Hyper-SOS considers the historical OSM posts of users and uses them together with social interactions through replies, comments, and quotes. The textual posts are embedded with BERT, and Hawkes temporal emotion aggregation (HEAT) mechanism is used to synthesise the posting history. This method adopts the Hawkes stochastic process for aggregating the historical posts using an exponential kernel. Once the posts are embedded and the connections are established, the generated graph is processed through an HGCN for detecting suicide ideation cases. HGCN is a variant of GCN that uses hyperbolic geometry to capture long range connections and achieves enhanced robustness, superior performance, and improved interpretability. The model, in its current form, is applied on static graphs, which is a hurdle when considering real-time dynamic data of evolving OSM.

Some studies exploit advanced graph structures. MentalNet [67] utilises graph-structured data by forming an ego network of users as a heterogeneous graph. Each user is denoted as a node, and connections are established based on interactions through replies, mentions, and quote tweets. The study extends the existing PsycheNet dataset [91] by incorporating interaction data to create PsycheNet-G. The problem is framed as heterogeneous graph classification. Node features are derived by applying an LSTM autoregressive sequence-to-sequence embedding once the graph is constructed. Given the adoption of heterogeneous graphs with repeated convolution, a normalisation process, known as doubly stochastic normalisation, is employed. The constructed graph is then processed through a stack of GCN and convolution layers. MentalNet's strength lies in its holistic approach of considering both tweet content and user interactions. However, it has a few limitations with possible areas of enhancement. It solely relies on textual posts for initial node features, without considering user-specific and behavioral features that can provide further insights into depression detection. One potential direction forward is to shift its focus from binary classification for detection to estimating the intensity level. Furthermore, model complexity is a concern. Lastly, the model relies on static graphs, but the dynamics of social contagion can shift rapidly in evolving OSM platforms. Incorporating dynamic graph structures and properties could bolster the model's real-world applicability. Kuo et al. [68] extend the MDD research into dynamic settings with their ContrastEgo network. They leverage the PsycheNet-G dataset [67] to construct a heterogeneous graph based on social interactions. Once the temporal graph is generated, interpersonal dynamics are established using GNN models, and a transformer encoder captures temporal dynamics. The final layer employs contrastive learning with binary cross-entropy loss and supervised contrastive loss to maximise user agreements. While ContrastEgo can capture the temporal evolution of communication patterns to some extent, it relies solely on textual posts for node features, ne-

glecting user-specific and behavioral attributes. The model performs binary classification for detection, but a fine-grained analysis of severity levels or intensity estimation is crucial for a comprehensive mental health assessment. Lastly, as ContrastEgo is a complex model, challenges arise in terms of its explainability.

In summary, DL models, as illustrated in Table 2, have demonstrated promising outcomes for MDD. However, it is noteworthy that much of the data used in these studies, particularly those achieving high accuracy, predominantly feature explicit patterns of disorders. This reliance on overt indicators may not offer a robust approach to MDD. To advance holistic methodologies, it is important to account for various factors such as historical posts and user interactions. Existing methods that encompass these features tend to be complex and less explainable. Striking the right balance is important, particularly in applications like healthcare, where explainability plays a crucial role.

### 3.3 Importance of domain knowledge

While data-driven methods rely heavily on machine learning for learning complex patterns, additional domain knowledge allows the method to further enrich and enhance the learning process according to the context. Taking proper consideration of domain knowledge offers a transparent view of decision-making processes with minimal reliance on computational methods. Nguyen et al. [105] conducted a study to explore the characteristics of depression among LiveJournal users by employing statistical learning methods. LiveJournal is an OSM platform that provides 132 predefined mood labels such as *depressed*, *happy*, *hungry*, and *cheerful*, and allows users to tag these labels in their posts. The researchers harness the Affective Norms for English Words (ANEW) [106], a lexicon of words with their affective measures in terms of valence, arousal, and dominance, for extracting sentiment and mental health related information. They extract Psycholinguistic features using LIWC, and topics using LDA. These information were found to be significantly different between clinical and control groups. Statistical tests confirmed substantial differences between the two groups. Notably, users from the clinical group were found to use negative emotions more frequently. The study observed an increased risk of clinical cases associated with the use of suicide-related words such as coffin, kill, and bury. Interestingly, the posts in the suicide and depression communities displayed similarities. These findings can serve as a foundation for creating guidelines to detect depressed users on social media platforms.

Our society has a social stigma that impedes the application of traditional methods for MDD. To this end, Perez et al. [4] introduced a computational approach for estimating depression intensity of an OSM user by automatically responding to the BDI-II questionnaire (Section 2) on users' behalf. The authors leveraged eRISK-2019 [107] and eRISK-2020 [108] datasets, which contain Reddit posts as well as answers to the BDI-II questionnaire for a sample of users. BDI-II has 21 questions, each having four options. For all these questions, a unique representative vector is learned with respect to each available option from the training data, resulting in a total of 84 vectors. The learning process utilised Word2Vec embeddings and averaging operations. To infer severity for a test user, the authors calculate Pearson correlation between text embeddings of the user and the representative option vectors. The option with the highest correlation is then assigned to the related question, thus automating the questionnaire filling process. Once all the questions are answered, they can be aggregated for severity assessment following traditional methods.

SOTA approaches and models for MDD exhibit limited capability for their generalisation and interpretation within the context of clinical settings. In order to achieve this objective, Perez et al. [14] developed a sentence dataset known as BDI-Sen, which is specifically designed to capture the clinical symptoms associated with depression. The dataset utilised in this study is derived from the eRISK-2019 dataset and encompasses a comprehensive range of symptoms associated with depression as assessed by the BDI-II questionnaire. In order to create this dataset, four distinct queries are formulated and the relevance is determined by calculating the cosine similarity of their sentence embeddings. Similarly, Anwar et al. [109] developed a lexicon for EDs called EDBase, comprising a comprehensive collection of ED-related terminology used in OSM and their ED relevance scores. The lexicon enables a domain-relevant interpretation and generalisation of ED-related OSM contents. It is particularly useful for domain-specific feature extraction, which can potentially enrich DL models with domain knowledge [6]. Transfer learning on domain-specific data is another essential component in enhancing detection performance. Within this context, deep embedding approaches pertaining to mental health hold significant importance. *MentalBERT* [110] is a pre-trained BERT model tailored for mental health. It is trained on a dataset of Reddit posts related to mental health, making it a valuable resource for analysing mental health related texts.

## 4 Explainability is what we need

Though modern DL models achieve remarkable performance, their architectures are often as complex as they are effective. Explainability is the key to unlocking the black box nature of a model, shedding light on the models' internal decision-making processes. This facet of understanding is of paramount importance, particularly for healthcare applications, where transparency can be a matter of life and death. There exists two different types of explainability [111], [112].

- *Local explainability* focuses on clarifying the rationale behind a single prediction made by the model.
- *Global explainability*, in contrast, provides insights into the entire decision-making process, with a holistic view of arriving at conclusions.

Some machine learning models are inherently able to explain themselves, referred to as *self-explainable* or *directly interpretable* [111], [112]. Decision trees, for instance, offer transparent decision paths that can be easily understood by humans. However, the scenario changes when it comes to deep learning models. Their intricate neural networks do not readily provide a comprehension. Instead, they necessitate a process known as *post-hoc explainability* [111], [112].

TABLE 2: Significant studies on ML for MDD from OSM (×: Not explainable, ✓: Can be explained but the paper does not demonstrate its explainability, ✓✓: Explainability demonstrated in the paper)

| Ref | Disorder | Methodology | Our comments | | XAI |
|---|---|---|---|---|---|
| | | | Pros | Cons | |
| [37] | Bipolar Disorder | Behaviour change graph, Stochastic gradient descent, Random forest, KNN, DT, LR, SVM | Mood changes are considered | Non-robust classifier, Unconsidered OSM post contents and social network. Focuses only on disorder existence | × |
| [92] | Suicide Ideation | BERT, Hawkes, Hyperbolic GCN | User interaction, Contextualised embedding | Static social network, only post content based node features. Focuses only on disorder existence | × |
| [6] | Eating Disorders | BERT, TCN, BiGRU | User-type identification, Multimodality, Contextual embedding | No social network | × |
| [18] | Depression | BERT, KMeans, BART, BiGRU, CNN | Multi-modality, Contextual embedding | No social network, Focuses only on disorder existence | × |
| [67] | Depression | GCN, BiLSTM, CNN | Augmented dataset, User interaction | Static social network, Only post contents as node features, Focuses only on disorder existence | × |
| [68] | Depression | RoBERTa, GCN, Transformer | Dynamic social network, User interactions, Contextual embedding | Only post contents as node features, Focuses only on disorder existence | × |
| [94] | Suicide Ideation | BERT, LSTM, Transformer | Contextual Embedding, Time-aware detection, Considers historical posts | No social network, Only post contents, Focuses only on disorder existence | × |
| [40] | Suicide Ideation | Knowledge graph, BERT, CNN, LSTM | Graphical representation, Contextual embedding, Multimodal data | No social network, Focuses only on disorder existence | × |
| [82] | Depression, Anorexia, Self-harm and Suicide | Transformer-based encoders, Negatively correlated noisy learners | Contextual embedding, Noisy learners, Multiple disorders | No social network, Only post contents, Focuses only on disorder existence | × |
| [83] | Suicide Ideation | LSTM, Attention, ResNet | Extracts latent information, Contextual embedding, Multimodal data | No social network, Focuses only on disorder existence | ✓ |
| [95] | Suicide Ideation | Longformer, BiLSTM, Attention | Ordinal classification, Contextualised embedding | No social network | ✓ |
| [9] | Depression | TextGCN, BiLSTM, Attention | Ordinal classification, Contextualised embedding | No social network, Only post contents | ✓ |
| [71] | Depression | CNN, MLP, Attention | Considers both word and sentence importance | No social network, Only post contents, Focuses only on disorder existence | ✓ |
| [96] | Depression | Mental RoBERTa, Transformer, Multi-head soft attention | Ordinal classification, Contextual embedding | No social network, Only post contents | ✓ |
| [97] | Bipolar disorder, Depression, Anxiety, ED, OCD, Schizophrenia | BERT, Attention | Multiple disorders, Contextual embedding | No social network, Only post contents, Focuses only on disorder existence | ✓ |
| [98] | Suicide Ideation | LIWC, LSTM, Attention | Considers multi-sources, Comprehensive textual analysis | No social network, Only post contents, No contextual embedding, Focuses only on disorder existence | ✓ |
| [99] | Depression | BERT, Transformer, Self attention | Real time detection, Contextual mood and content embedding | No social network, Only post contents, Focuses only on disorder existence | ✓ |
| [100] | Suicide Ideation | BERT, Bi-LSTM, MLP, Attention | Contextual embedding, Ordinal classification, Robust model | No social network, Only post contents | ✓ |
| [101] | Suicide Ideation | Doc2Vec, Multi-head attention, | Extracts emotions, Detects disorder couple of months earlier | No social network, Only post contents, No contextual embedding, Focuses only on disorder existence | ✓ |
| [102] | Eating Disorder (Anorexia Nervosa) | ELMo, CNN, Attention | Contextual embedding | No social network, Only post contents, Focuses only on disorder existence | ✓✓ |
| [103] | Depression | BERT, HAN | Adopts metaphor generation, Contextual embedding | No social network, Only post contents, Focuses only on disorder existence | ✓✓ |
| [19] | Depression | GloVe, MLP, HAN | Multi-modality | No social network, No contextual embedding, Focuses only on disorder existence | ✓✓ |
| [104] | Suicide Ideation | TensorGCN, Longfromer, Transformer | Ordinal classification, Contextualised embedding, Performs well on long sequences | No social network, Only post contents | ✓✓ |

In this post-processing phase, additional techniques are employed to investigate and elucidate the model's decisions.

In terms of their applicability, the explainability techniques can be categorised into two approaches [113], [114].

- *Model-agnostic* explainability techniques are versatile. They can be applied to unravel the mysteries of any machine learning model, regardless of its specific architecture.
- *Model-specific* explainability techniques, on the other hand, are customised to work with particular models, providing insights into their internal processes.

In the quest for explainability and transparency, researchers explore both approaches and employ diverse techniques to decipher advanced DL models.

Current research on MDD lacks sufficient attention to the crucial aspects of model explainability and interpretability, limiting the real-world applicability of these models. The complex and diverse nature of mental disorders poses a significant challenge in developing comprehensive explanatory models. Some existing studies use self-attention [102], hiearchical attention [19], [103] and multi-head attention [104] models for generating post-hoc explainability. We aim to bridge the existing gap between the black-box nature of current models and the demand for transparent mental health assessments. Through a comprehensive exploration of SOTA explainability techniques, we seek to advance research on explainable AI for MDD, providing practitioners and end-users with valuable insights into how MDD models arrive at their predictions. This, in turn, will empower informed decision-making processes and establish a foundation for responsible implementation in real-world scenarios.

### 4.1 Interpretable models

Some models, such as linear regression, logistic regression, and decision trees, stand out for their inherent interpretability. Their internal architectures make it easy to understand the rationale behind the models' decisions. Linear Regression and Logistic Regression provide transparency through the generation of feature weights. These weights assign significance to individual features, enabling us to recognise their importance in influencing predictions. Decision Trees, on the other hand, adopt a hierarchical structure of questions, forming a tree-like visualisation. This representation offers an intuitive way to grasp the decision logic. By following the branches of the tree, we can trace the series of questions and conditions that lead to a particular outcome.

### 4.2 LIME

LIME [115] is an acronym for local interpretable model-agnostic explanations. It is a widely-used model-agnostic method for providing local explanations. Its primary aim is to explain complex ML models by approximating their behavior with a simpler and more interpretable model, such as linear regression or logistic regression. The local neighborhood is created by generating a dataset through a perturbation method, which introduces controlled noise to the input data. Mathematically, given a prediction function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and a data point $x$, LIME seeks a simpler model $g(x)$ within a local neighborhood of $x$. This is done by minimising the loss between the original complex model and the simpler model using Equation 6, where $\mathcal{L}(f, g, \pi_x)$ represents the loss function shown in Equation 7, and $\Omega(g)$ is a regularization term applied to the simpler model $g$. The goal is to find the parameters for this simpler model that best approximate the behavior of the complex model within the local neighborhood. In Equation 7, $\mathcal{Z}$ denotes the set of points within the local neighborhood of $x$, $\pi_x(z)$ denotes the weight assigned to data point $z \in \mathcal{Z}$ determined using a perturbation method, and $f(z)$ and $g(z')$ are the predictions made by the complex and simpler models, respectively.

$$\xi(x) = \underset{g \in G}{\arg\min} \left[\mathcal{L}(f, g, \pi_x) + \Omega(g)\right] \tag{6}$$

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \cdot \left(f(z) - g(z')\right)^2 \tag{7}$$

LIME has proven valuable in various applications, including causal analysis of depression cases [116]. The study used LIME to compare model-generated explanations with human-generated explanations.

### 4.3 SHAP and KernelSHAP

SHAP [117], an acronym for SHapley Additive exPlanations, can be used to obtain the importance (called Shapley value) of individual features used in predictive models. The Shapley value $\phi_i(f, x)$ of a specific feature $i$ of a data point $x$ for a complex model $f$ is calculated using Equation 8, where $x'$ denotes the simplified input set (often a subset of the complete set of features), $M$ is the total number of available features, $|z|!$ calculates the factorial of the size of subset $z$, $f_x(z')$ denotes the model's prediction for $z'$ subset given as input, and $f_x(z' \backslash i)$ denotes the model's prediction after removing feature $i$ from the subset $z'$.

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z|!(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \backslash i)\right] \tag{8}$$

Shapley values tend to perform well with linear models, but may encounter limitations with complex models. An extension of SHAP, known as KernelSHAP, is tailored to handle complex models by breaking down Shapley values into smaller subsets of features and employing Monte Carlo sampling techniques to approximate these values.

### 4.4 LRP

LRP (or layer-wise relevance propagation) [118] is an explainability method used to analyse the feature contributions of complex models. It works by propagating relevance scores backward through the layers of a neural network model, determining the importance of each neuron all the way back to the input features. Let's consider two consecutive layers, $i$ and $j$. The propagation of relevance score, $R_j$, can be calculated as $R_j = \sum_k \frac{\theta_{jk} a_{jk}}{\sum_j a_{jk}} R_k$ where $a_{jk}$ is the activation vector of neuron $j$, and $\theta_{jk}$ is the weight vector of the connection from neuron $j$ to neuron $k$. $R_j$ quantifies the importance of neuron $j$ in model's predictions, and it simply aggregates the relevance scores $R_k$ from the subsequent layer. This process allows us to trace back and understand the contribution of each neuron and input feature in the model's decision-making.

TABLE 3: Overview of Explainability methods

| Explanation Type | | Scope | Model Type | Example | Applied for MDD? |
|---|---|---|---|---|---|
| Interpretable Models | | Global | Model Specific | Linear and Logistic Regression, Decision Trees | ✓ |
| Post-hoc | Feature relevance | Local | Model Agnostic | SHAP, KernelSHAP | × |
| | | Local | Model Specific | Layer-wise Relevance Propagation (LRP) | × |
| | | Local | Model Specific | A-Grad and RePAGrad | × |
| | By Approximation | Local | Model Agnostic | LIME | ✓ |
| | | Global | Model Agnostic | PGExplainer (for graphs) | × |
| | By Example | Local | Model Specific | GNNExplainer and GCN-SE (for graphs) | × |
| | | Local | Model Specific | Attention | ✓ |
| | By Visual Explanation | Local | Model Agnostic | Individual Conditional Expectation (ICE) | × |
| | | Global | Model Agnostic | Partial Dependence Plot (PDP) | × |

### 4.5 Explanation by Attention values

Attention values [78] provide a way to determine the importance of a feature. They are often employed to provide insights in various NLP tasks by generating word-level importance scores. Some studies [102], [104] leveraged the attention mechanism to detect suicide risk from Reddit data, while offering explanations based on attention values. Attention values are also helpful in demystifying the inner workings of transformer models, as transformers are complex and require further explanation for different tasks, including embedding [119], [120], [121]. However, the usage of attention values are not limited to word-level importance. They can also be applied to assess the importance of entire sentences with hierarchical attention networks (HAN) [122]. In this approach, word-level attention values are first computed, followed by the generation of sentence-level attention values. For example, MDHAN [19] detects depressed users while providing model explainability using HAN. It encodes user posts at both the tweet and word levels, providing explanations at both granularities. While MDHAN works well for classification as well as explainability, it focuses solely on textual posts and does not provide insights into user behavior features. Additionally, its explainability (word- and tweet-level attention) does not clarify whether a word or tweet has a positive or negative impact.

### 4.6 PDP and ICE

PDP (partial dependence plot) visualises the marginal effect of model features using a partial dependence method [123]. Based on this, Greenwell et al. [124] introduced *importance measure* ($IM$) to compute the importance of a feature. For continuous variable $x_i$ in a model $f()$, $IM(x_i)$ is calculated using Equation 9, where $k$ is the number of unique instances within the feature. For categorical variable $x_i$, $IM(x_i)$ is computed using Equation 10.

$$IM(x_i) = \sqrt{\frac{1}{k-1}\sum_{j=1}^{k}[f_i(x_{ij}) - \frac{1}{k}\sum_{j=1}^{k} f_i(x_{ij})]^2} \quad (9)$$

$$IM(x_i) = \frac{max_j(f_i(x_{ij})) - min_j(f_i(x_{ij}))}{4} \quad (10)$$

While PDP illustrates the average effect of a feature, it does not focus on the prediction changes based on individual instances. This limitation is addressed by ICE (individual conditional expectation) plots [125]. They depict one line for each instance and each feature by manipulating the feature of interest while keeping the other features fixed.

### 4.7 Explainability of GNNs

While GNNs are able to deliver great results for mental health analytics, achieving explainability in the context of MDD remains a challenge due to the inherent complexity of these models. While ongoing research is striving to enhance the explainability of GNNs, it is still an open problem [113]. Existing explainability techniques, including LIME, attention mechanisms, and LRP, have been applied to elucidate the decision-making processes of GNNs. Given the complex structure of GNNs, some non-traditional explainability methods have also been developed. Recent noteworthy contributions to GNN explainability include GNNExplainer [126], PGExplainer [127], GCN-SE [128].

**GNNExplainer** [126] is a powerful model-agnostic technique for explaining GNN predictions across diverse graph-related machine learning tasks, including node classification, link prediction, and graph classification. It furnishes explanations in the form of a concise subgraph of the input graph and a subset of node features that exercise the most significant influence on GNN predictions. GNNExplainer is adaptable to both single-instance and multi-instance scenarios. In single-instance scenarios, it elucidates a GNN's prediction for a specific instance, be it a node label, link, or graph-level label. For multi-instance scenarios, it provides a coherent explanation covering a set of instances, such as nodes belonging to a specific class. Given a trained GNN model $\Phi$ and its predicted label distribution $Y$, GNNExplainer seeks to identify a subgraph $G_S \subseteq G_c$ (computation graph) and the associated node features $X_S = \{x_j | v_j \in G_S\}$ that maximise mutual information with the GNN's prediction, as illustrated in Equations 11 and 12. Here, $H(Y)$ represents the entropy of the predicted label distribution $Y$, and $H(Y|G_S, X_S)$ is the conditional entropy. Equation 11 strives to minimise uncertainty when the GNN is confined to the explanation subgraph $G_S$. Equation 11 quantifies how much information about the GNN's prediction is present in the explanation subgraph and node features, ensuring that the identified subgraph $G_S$ maximises the probability of the GNN's prediction $\hat{y}$.

$$\max_{G_S} MI(Y, (G_S, X_S)) = H(Y) - H(Y|G_S, X_S) \quad (11)$$

$$H(Y|G_S, X_S) = -E_{Y|G_S,X_S}[\log P_\Phi(Y|G_S, X_S)] \quad (12)$$

Although GNNExplainer is effective, it suffers from several limitations that impact its applicability in the real-world. *Firstly*, the method primarily focuses on providing local interpretability by generating customised explanations for individual instances, like nodes or graphs, independently. This limits its effectiveness in the inductive setting,

as the explanations can not be generalised to other unexplained nodes. *Secondly*, GNNExplainer requires retraining for every single explanation, making it time-consuming and impractical when dealing with a large number of nodes. *Lastly*, as GNNExplainer was designed for interpreting individual instances, the explanatory motifs are not learned end-to-end with a holistic view of the entire GNN model. This lack of a comprehensive global perspective can lead to suboptimal generalisation performance.

**PGExplainer** [127] recognises the importance of collective and inductive explanations for GNN predictions, and addresses these limitations of GNNExplainer. Though PGExplainer works in a similar fashion as GNNExplainer maximising the mutual information between the original input graph and the subgraph, PGExplainer offers explanations that collectively cover multiple instances and provide a more global understanding. It leverages a parameterised generative model for graph data, and uncovers underlying structures crucial for GNN predictions. This generation process with a DNN enables collective explanations with shared parameters. Its improved generalisation enhances its practicality, allowing it to infer explanations for unexplained nodes in an inductive setting without the need for retraining. Both GNNExplainer and PGExplainer are able to explain GNN predictions when the graphs are static. They are unable to explain the decision-making process of dynamic GNNs.

**GCN-SE** [128] addresses the explainability in dynamic GNNs with the help of attention values. It treats graph snapshots at different times of dynamic graphs as different channels of data and attaches a set of learnable attention weights with them based on GCN [66] and SE-Net [129].

## 5 Evaluation Methods and Datasets

### 5.1 Evaluation measures

Evaluating an ML model on experimental datasets using specific evaluation metrics is important in order to assess its prediction performance. For MDD models, standard metrics like *Accuracy*, *Precision*, *Recall*, *F-Score*, and *ROC-AUC Score* are commonly employed. Moreover, there also exist problem-specific and domain-specific metrics tailored to consider relevant additional factors. Definition 1 is applied to evaluate the performance of a model working on early prediction, Definitions 2-5 are applied to evaluate models that predict answers of MDD questionnaires, and Definition 6 is applied to evaluate models that perform a multi-class classification of mental disorder severity.

**Definition 1: (Early Risk Detection Error (ERDE))** This measure introduces a penalty for late correct predictions by considering the number of posts seen before the mental disorder alert is issued by the model [130]. A lower ERDE value means a better model performance for early detection. For an instance, it is calculated using Equation 13, where $k$ denotes the delay (number of posts seen before decision), $o$, $c_{fp}$,$c_{fn}$ and $c_{tp}$ are external parameters, and $lc_o(k) = 1 - \frac{1}{1+exp(k-o)}$. The overall $ERDE$ measure for an MDD model is calculated by averaging the measures of all the users.

$$ERDE_o = \begin{cases} c_{fp} & \text{for False Positives (FP);} \\ c_{fn} & \text{for False Negatives (FN);} \\ lc_o(k) \times c_{tp} & \text{for True Positives (TP);} \\ 0 & \text{for True Negatives (TN);.} \end{cases} \quad (13)$$

**Definition 2: (Average Hit Rate (AHR))** Hit rate (HR) computes the ratio of items the model has estimated the same answer option as the user, and AHR is the average of HR of all users [4].

**Definition 3: (Average Closeness Rate (ACR))** Closeness rate (CR) is computed as CR $= mad - ad$, where where $mad$ and $ad$ represent the maximum absolute difference and absolute difference, respectively [4]. ACR is then computed as the average of closeness rates (CR) of all users.

**Definition 4: (Depression Category Hit Rate (DCHR))** It computes the fraction of users where the model prediction and ground truth score of the user fall in the same severity level according to BDI-II [4]. Although DCHR is used only for depression, it can be generally applied to all disorders with questionnaires.

**Definition 5: (Average Difference between Overall Depression Levels (ADODL))** The Difference between Overall Depression Levels (DODL) is computed as DODL $= \frac{63-ad}{63}$, where $ad$ is the absolute difference between model prediction and ground truth value, and 63 is the maximum possible absolute difference (21x3 = 63, for a total of 21 question, each with a maximum possible difference of 3 points) [4]. The ADODL is computed as the average of DODL of all users. ADODL can also be applied to other disorders with questionnaires.

**Definition 6: (Ordinal Regression)** For multi-class classification, the classes are assumed to be independent. However, the classes may have an order sometimes. For example, in case of a mental disorder severity classification with classes - *severe*, *moderate*, *mild*, and *minimum* - *severe* is closer to *moderate* than *mild* in terms of intensity. To model this ordinal relationship, Sawhney et al. [95] introduced the idea of ordinal regression (*aka* ordinal classification) that captures dependencies between the severity classes. Let $Y = \{r_i\}_{i=0}^{n}$ represent the $n$ ordinal severity levels. For given actual severity levels $r_t \in Y$, soft labels are calculated as probability distributions, $y = [y_0, y_1, ...., y_n]$. The probability $y_i$ of each severity level $r_i$ is determined using Equation 14, where $\phi(r_t r_i)$ is a cost function that penalises the difference between the actual severity level $r_t$ and the predicted severity level $r_i \in y$. As the difference between $r_t$ and $r_i$ increases, there is a decrease in the probability $y_i$ of the associated prediction $r_i$.

$$y_i = \frac{exp^{-\phi(r_t r_i)}}{\sum_{k=1}^{\lambda} exp^{-\phi(r_t r_k)}} \quad (14)$$

To evaluate the effectiveness of a severity estimation model, the measures of false negatives (FN) and false positives (FP) are modified using Equation 15, where $N_T$ is the size of the test data, $k^a$ is the actual severity level, and $k^p$ is the predicted severity level over the test data.

$$FN = \frac{\sum_{i=1}^{N_T} I(k_i^a > k_i^p)}{N_T}, FP = \frac{\sum_{i=1}^{N_T} I(k_i^a < k_i^p)}{N_T} \quad (15)$$

### 5.2 Experimental datasets

Datasets serve as fundamental components for data-driven decision-making systems. The dataset curated by Shen et al. [86] from 𝕏 stands out as a frequently utilised resource in research on depression detection. Another notable data source is the eRISK series [130], collected from Reddit, extensively used by researchers for detecting various mental disorders, including depression, EDs, self-harm, and pathological gambling. Dreaddit, compiled by Turcan et al. [131], is a Reddit-based dataset designed for binary depression classification, later enhanced by Naseem et al. [9] for multi-class severity classification. PsycheNet [91] is a social-contagion-driven dataset constructed from 𝕏, primarily for depression detection. Mihov et al. [67] improved this dataset, creating PsycheNet-G, by incorporating additional features like bidirectional replies, mentions, and quote-tweets to enhance the robustness of social network data. Further details about these and other relevant datasets are available in Table 4.

Data collected from OSM for MDD are sometimes annotated through automatic or semi-supervised methods. However, fair and unbiased diagnostics require annotation by field professionals. Many datasets often include anchor posts[1] with explicit hints and patterns of mental disorders, potentially leading to data leakage. To address this, there is a need for datasets that contain a detailed historical information of users including their posts (excluding the anchor post), online activities, and social network. It is crucial to highlight that the current volume of annotated data from field professionals is insufficient for robust data modeling in the context of MDD.

## 6 Research Horizons

### 6.1 Open issues and challenges

The development of accurate and explainable AI models for MDD faces several important issues and challenges. These obstacles impede progress in the field and demand careful consideration for the advancement of robust, reliable, and transparent mental health analytics systems.

**Dataset**. *Issue:* A critical issue in the domain of MDD is the scarcity and limitations of benchmark datasets. There exist very few datasets annotated by domain experts, due to which many existing datasets fail to comprehensively represent the complexity of mental health conditions [9], [96]. They also lack crucial associated data such as metadata and users' social network information. Furthermore, the diverse spectrum of mental disorders is not captured in these datasets. *Challenge:* Collaborative initiatives with mental health professionals are essential to ensure datasets are not only larger but also carefully annotated with a focus on severity levels and temporal dynamics. The datasets require to include comprehensive user metadata, social network information, and fine-grained severity labels. The limited access of OSM APIs makes data scraping challenging, which necessitates strategies to overcome these limitations.

1. An anchor post is the post of a user having clear indication of a disorder, based on which the user is labelled with the disorder. For example, "I have been diagnosed with depression".

**User-level vs post-level MDD**. *Issue:* The ultimate goal of MDD is to identify disorders experienced by users (user-level MDD). However, a significant portion of existing research focuses on a more straightforward approach, detecting disorders expressed within individual OSM posts (post-level MDD) [9], [67], [95]. While post-level MDD provides valuable insights into the mental states conveyed in isolated posts, the user-level MDD represents a more holistic perspective, offering a comprehensive understanding of an individual's mental health condition [6]. *Challenge:* Addressing user-level MDD necessitates a holistic examination of users' historical data. The challenge lies in developing methodologies that seamlessly integrate various types of historical data, including textual content, behavioral patterns, and social interactions, to construct a rich profile of an individual's mental health journey .

**Binary classification vs severity estimation**. *Issue:* The predominant focus on binary classification (indicative of a disorder or not) of a user or a post in existing studies on MDD limits the granularity of gained insights [68], [86]. It lacks the diverse spectrum of mental health conditions and their varying degrees of severity. *Challenge:* The primary challenge lies in transitioning from binary classification towards more sophisticated severity estimation models. While some recent studies have ventured into severity detection [7], [9], they remain limited in scope and lack comprehensive coverage across various mental disorders. Effectively addressing this challenge requires leveraging diverse data sources, establishing standardised severity scales for OSM extending the traditional severity scales, and the development of robust severity estimation models capable of discerning subtle variations in severity across different disorders.

**Multimodality and social graph**. *Issue:* Most existing studies exclusively rely on the textual content of user posts within their machine learning models for MDD [9], [97], [98], [103]. While text-based approaches offer valuable insights, they fall short of capturing the richness of multimodal data prevalent in OSM, such as behavioral patterns, contextual information, and the social network. They neglect a comprehensive understanding of mental health expressions in the online space. *Challenge:* The challenge lies in effectively capturing and representing multimodal and social graph data withing ML models for MDD. To adequately address this challenge, future research should explore the integration of behavioral patterns, visual content, and metadata associated with user posts [6], [18]. Additionally, harnessing the power of social network structures with the help of GNNs can provide valuable context and relationships [67], [68].

**Volume of OSM data**. *Issue:* The volume of OSM data (whether lengthy individual posts or large number of historical posts) sometimes become significantly large. In such cases, reasonable strategies need to be employed for properly utilising them. *Challenge:* Implementing strategies like summarisation and other advanced techniques becomes imperative to condense the data, maintaining a balance between compression and retaining crucial information [143].

**Explainability**. *Issue:* A critical concern in the current status of MDD models is the pervasive neglect of explainability. A majority of existing studies in this domain overlook the

TABLE 4: Notable experimental datasets (✓: publicly available, ⊙: available on request).

| Dataset | Level | Available | Platform | Disorder | Classes | Statistics |
|---|---|---|---|---|---|---|
| [86] | User | ✓ | 𝕏 | Depression | Binary | Depressed Users: 1,402, Non-Depressed Users: > 300 million, Depressed Tweets : 292,564, Non Depressed Tweets > 10 billion |
| Dreaddit [131] | Post | ✓ | Reddit | Depression | Binary | Depressed Posts: 1,857, Non-Depressed Posts: 1,698 |
| [9] | Post | ✓ | Reddit | Depression severity | Multi-class | Minimum depression level: 2,587, Mild depression level: 290, Moderate depression level: 394, Severe depression level: 282 |
| eRISK-2020 [108] | User | ⊙ | Reddit | Self-harm and depression | Binary and Multi-class | Submissions for Self-harm: 18,618, Submissions for non-self harm case: 254,642; Depression statistics: Not available, answers of BDI-II questionnaire |
| eRISK-2021 [132] | User | ⊙ | Reddit | Pathological Gambling, Self-harm and Depression | Binary and Multi-class | Gambling submissions: 54,674, Non-Gambling submissions: 1,073,883; Self-harm submissions: 69,722, Non self-harm submissions: 943,465; Depression statistics: Not available, answers of BDI-II questionnaire |
| eRISK-2022 [133] | User | ⊙ | Reddit | Pathological Gambling, Self-harm and depression | Binary and Multi-class | Gambling submissions: 69,301, Non-Gambling submissions: 2,087,210; Depressed submissions: 35,332, Non depressed submissions: 687,228; Eating Disorders statistics: Not available, answers of EDE-Q questionnaire |
| PsycheNet [91] | User | ⊙ | 𝕏 | Depression | Binary | Depressed users: 372, Non-Depressed users: 445 |
| PsycheNet-G [67] | User | ⊙ | 𝕏 | Depression | Binary | Depressed users: 242, Non-Depressed users: 349 |
| [134] | User | ✓ | Reddit | Suicide Ideation | Multi-class | 500 total users; Attempt: 45; Behavior: 77; Ideation: 171; Indicator: 99; Supportive: 108. |
| Twitter-STMHD [135] | User | ✓ | 𝕏 | ADHD, Bipolar, Anxiety, Depression, PTSD, OCD | Binary and Multi-class | 43269 tweets; 27003 users. |
| [136] | Post | ✓ | Reddit | Suicide and Depression | Binary | 1894 total posts, with 915 control and 980 diagnosed positive. |
| RSDD [137] | User | ⊙ | Reddit | Depression | Binary | 9210 depressed users and 107274 control users. |
| SMHD [138] | User | ⊙ | Reddit | Multiple (inc. Depression, Anxiety, Bipolar, OCD, Schizophrenia, ED) | Binary | Depression: 14,139 users and 1,272 posts; Anxiety: 8,783 users and 795K posts; Bipolar: 6,434 users and 575K posts; EDs: 598 users and 53K posts; OCD: 2,336 users and 203K posts; Schizophrenia: 1,331 users and 123K posts; Control: 335,952 users and 116M posts. |
| [139] | Post | ⊙ | 𝕏 | Suicide Ideation | Multi-class | 534 safe to ignore posts; 1029 Possibly concerning posts; 258 strongly concerning posts. |
| [39] | User | ⊙ | 𝕏 | Schizophrenia | Binary | 174 positively diagnosed users; 3200 posts. |
| [140] | User | ⊙ | 𝕏 | Multiple (inc. ADHD, Depression, Anxiety, OCD, ED, PTSD, Bipolar, Schizophrenia) | Binary | ADHD: 102 users and 384k posts; Anxiety: 216 users and 1591k posts; Bipolar 188 users and 720k posts; Depression: 393 users and 546k posts; EDs: 238 users and 724k posts; OCD: 100 users and 314k posts; PTSD: 403 users and 1251k posts; Schizophrenia: 172 users and 493k posts. |
| [141] | User | ⊙ | Tumblr | Recovery from Anorexia | Binary | 18,923 users and 55,334 posts. |
| [142] | User | ⊙ | Reddit | Bipolar Disorder | Binary | 3488 positively diagnosed users and 3931 control users. |

need of ensuring transparency and interpretability in their models [67], [68]. Particularly in the context of healthcare, where decisions based on these models can have acute implications, the lack of emphasis on explainability compromises the trustworthiness of these models [62], [144]. Hence, it limits their acceptance in real-world applications and clinical practices. *Challenge:* The main challenge lies in the dual pursuit of achieving high accuracy while concurrently ensuring explainability. The prevalent use of complex DL architectures, while effective in achieving accuracy, poses a barrier to interpretability. The intricate internal workings of these models make it challenging to uncover how they arrive at specific decisions, rendering them akin to "black boxes" .

**Temporal evolution of disorders over time**. *Issue:* A critical gap in the current status of research on MDD lies in its static nature. Existing studies generally treat it as a snapshot problem, neglecting the inherent temporal fluctuations that individuals experience [68]. In reality, mental health is a dynamic continuum, where individuals may transition through varying states of well-being and distress over different time intervals. *Challenge:* The challenge lies in developing models that transcend the static paradigm and incorporate temporal information to capture the evolving patterns of mental disorders over time. Addressing this challenge involves exploring innovative approaches to temporal modeling accommodating the transient and dynamic nature of these disorders, possibly with the help of dynamic DNNs.

**Simultaneous consideration of multiple disorders**. *Issue:* Existing studies predominantly focus on the detection of a single mental disorder [6], neglecting the reality where individuals often experience multiple interrelated disorders simultaneously [97]. This myopic approach fails to capture the complexity of mental health conditions, limiting the practical utility of these models in real-world scenarios.

*Challenge:* The challenge lies in developing models that can simultaneously consider multiple interrelated disorders with the help of mental health domain knowledge. The dynamic nature of these disorders over time adds another layer of complexity. The challenge is not merely to expand the scope of detection but to devise models that can adeptly capture the intricate link between multiple disorders and their evolving manifestations.

**Consideration of domain-specific knowledge**. *Issue:* The integration of domain-specific knowledge in MDD models is pivotal for enhancing interpretability, contextuality, and transparency in decision-making. Some existing studies have demonstrated the efficacy of incorporating domain knowledge from various sources, such as domain lexicons, psychological features, and clinical questionnaires, to enrich the understanding of mental health states within OSM platforms [4], [14], [62], [109]. However, a significant issue persists in their limited application. *Challenge:* The foremost challenge lies in fostering interdisciplinary collaborations between domain experts, data scientists and AI researchers to develop robust methodologies. This requires establishing effective communication channels, bridging the gap between mental health professionals and data/AI practitioners, and developing methodologies for extracting relevant domain information and insights. Moreover, creating transparent and interpretable models that leverage domain knowledge poses a unique challenge, especially when dealing with inhenerntly complex DL architectures.

### 6.2 Future research directions

The field of Explainable AI for MDD holds considerable promise, with several research directions well-positioned to advance transparency, interpretability, and real-world applicability. Addressing these directions can contribute significantly to the evolution of robust and resilient mental health analytics systems.

**Enhanced multi-modal data collection**. To enhance data comprehensiveness, researchers should develop advanced strategies for generating diverse and well-annotated multimodal datasets representing mental health conditions. It includes textual as well as multimedia posts, user interactions, behaviors, social networks and contextual information [40], [83]. Collaborative initiatives with mental health professionals are important to ensure the creation of larger datasets that are carefully annotated, encompassing severity levels and temporal dynamics [109].

**Transfer learning for multi-disorder detection**. To address the challenge of simultaneous detection of multiple mental disorders, researchers should explore transfer learning techniques [97]. It is crucial to investigate how knowledge gained from detecting one disorder can be effectively transferred to improve the accuracy of detecting other co-occurring disorders. This direction requires a comprehensive understanding and representation of multi-disorder patterns, paving the way for more robust and adaptable detection models .

**Advanced graph representation learning**. To detect (the severity of) multiple disorders leveraging social networks and other structural information within OSM contents, it is important to further graph representation learning [68], [128]. It involves developing novel approaches that incorporate temporal dynamics and evolving relationships in OSM data. The aim is to create models that effectively represent and analyse the structural components, providing deeper insights into the complex patterns of mental health expressions within online communities .

**Domain-aware fine-grained severity detection**. While existing research has made some progress in detecting mental disorders, a critical gap exists in fine-grained severity detection, particularly with the support and enrichment of domain-specific knowledge. The collaboration with mental health professionals becomes imperative to capture the symptoms of severity and align with domain-specific insights [14], [39], [45], [109]. The developed models should not only leverage domain knowledge but also offer transparent explanations for their decision-making processes.

**Realtime mental health analytics**. Mental health is a continuous and evolving process, which necessitates a paradigm shift from static analyses to real-time perspectives. This research direction advocates for developing real-time mental health analytics systems capable of tracking and predicting moment-by-moment mental health conditions and conducting longitudinal analysis. It is important to investigate how models can adapt and provide timely insights to address emerging mental health concerns, especially during significant global events such as warfares, economic recessions and pandemics. This requires the development of agile and adaptive models capable of monitoring and responding to shifts in online mental health expressions in real-time [68], [107], [133]. It will pave the way for proactive interventions and personalised support.

**Novel explainability methods**. Existing research on explainable MDD models is very limited. In order to enhance their realworld applicability, it is important to develop novel explainability techniques and ensure the transparency of the decision-making processes. In particular, GNNs and multimodal data are getting increasingly useful for MDD. The multi-modal data can be used as node/link features within GNNs. However, such models are highly complex. There is a need for development of dedicated explainability methods for enhancing the transparency of GNN-based MDD. Investigating techniques that effectively communicate the rationale behind predictions, while considering the diverse nature of features, connections, and historical data used in mental health analytics, is essential [126], [127], [128].

**Explanation with large language models**. Large language models (LLMs), such as ChatGPT, are a recent breakthrough in NLP, prominently built upon transformer-based architectures. Researchers are starting to use LLMs to explain insights of a wide range of things [145]. To enhance the generation of realistic and diverse explanations of MDD models, a promising research direction is to investigate the integration of these generative models with MDD models [146]. This integration has the potential to improve the explainability of MDD models, providing more meaningful insights into mental health predictions, and thereby contribute to a more transparent and understandable decision-making process.

**Ethical considerations and bias mitigation**. To maintain fairness, transparency, user privacy, and the responsible use of AI for mental health analytics, it is extremely important

to carefully identify and implement the ethical concerns surrounding the development and deployment of MDD models. This research direction focuses on mitigating biases, ensuring fairness, addressing potential ethical concerns related to user privacy and data sensitivity, and ensuring equitable deployment in real-world scenarios [147].

**Human-in-the-loop approaches**. Instead of complete reliance on AI models for MDD, a valuable research direction is to explore *human-in-the-loop* approaches that involve mental health professionals in model training and decision-making processes. It is important to investigate how the expertise of professionals can enhance model interpretability, reduce biases, and improve overall reliability through collaborative interdisciplinary approaches [148].

**Incorporating cultural variations**. To enable an MDD model to accurately understand user expressions on OSM, it is important to be mindful of the inherent diversities in cultural expressions of mental health around the globe. It requires collaboration with experts from diverse cultural backgrounds to ensure the construction of datasets that accurately represent the spectrum of cultural experiences. Additionally, the models should be developed to adapt and generalise across different cultural settings [149], [150].

## 7 Conclusion

Mental health challenges affect millions of people worldwide, and demand innovative solutions for early detection and intervention. The advent of deep learning models, such as DepressionNet and EDNet, brings powerful tools to the field, yet their black-box nature raises ethical concerns in healthcare applications. The recognition of this gap has led to the rise of XAI models, aiming to illuminate the decision-making processes of complex AI systems. In this survey paper, we explored and presented a summary of the traditional mental disorder diagnostic methods, the state-of-the-art research on data- and AI-driven mental disorder detection, and the rapidly developing field of XAI. It emphasises the need for XAI models to shed light on complex AI decision-making processes, especially in the context of mental health analytics. The paper calls for a balanced approach by aligning technological advancements with transparency and ethics. It is crucial to bridge the gap between deep learning efficacy and interpretability, in order to obtain meaningful insights that benefit users and healthcare professionals. The outlined research directions provide a roadmap for future endeavors, emphasising real-time analytics for enhanced mental health AI. Ultimately, the paper envisions a future where mental health AI not only detects disorders but also contributes positively to societal well-being.

## References


[1] Mental disorders. WHO Fact Sheets. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[2] T. Zhang, K. Yang, S. Ji, and S. Ananiadou, "Emotion fusion for mental illness detection from social media: A survey," *Inf. Fus.*, vol. 92, pp. 231–246, 2023.

[3] U. Ahmed, J. C.-W. Lin, and G. Srivastava, "Graph attention network for text classification and detection of mental disorder," *ACM Trans. Web*, vol. 17, no. 3, pp. 1–31, 2023.

[4] A. Pérez, J. Parapar, and Álvaro Barreiro, "Automatic depression score estimation with word embedding models," *AI in Med.*, vol. 132, p. 102380, 2022.

[5] A. Thompson, C. Hunt, and C. Issakidis, "Why wait? reasons for delay and prompts to seek help for mental health problems in an australian clinical sample," *Social Psychiatry and Psychiatric Epidem.*, vol. 39, no. 10, pp. 810–817, 2004.

[6] M. Abuhassan, T. Anwar, C. Liu, H. K. Jarman, and M. Fuller-Tyszkiewicz, "Ednet: Attention-based multimodal representation for classification of twitter users related to eating disorders," in *ACM Web Conf.*, 2023, p. 4065–4074.

[7] S. Ghosh and T. Anwar, "Depression intensity estimation via social media: A deep learning approach," *IEEE Trans. on Comput. Soc. Syst.*, vol. 8, no. 6, pp. 1465–1474, 2021.

[8] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: a narrative review," *npj Dig. Med.*, vol. 5, no. 1, p. 46, 2022.

[9] U. Naseem, A. G. Dunn, J. Kim, and M. Khushi, "Early identification of depression severity levels on reddit using ordinal classification," in *ACM Web Conf.*, 2022, p. 2563–2572.

[10] S. D'Alfonso, "Ai in mental health," *Current Opinion in Psychology*, vol. 36, pp. 112–117, 2020.

[11] M. Chen, K. Shen, R. Wang, Y. Miao, Y. Jiang, K. Hwang, Y. Hao, G. Tao, L. Hu, and Z. Liu, "Negative information measurement at ai edge: A new perspective for mental health monitoring," *ACM Trans. Internet Technol.*, vol. 22, no. 3, pp. 1–16, 2022.

[12] Social media users. DataReportal. [Online]. Available: https://datareportal.com/social-media-users

[13] T. Anwar, S. Nepal, C. Paris, J. Yang, J. Wu, and Q. Z. Sheng, "Tracking the evolution of clusters in social media streams," *IEEE T. on Big Data*, vol. 9, no. 2, pp. 701–715, 2023.

[14] A. Pérez, J. Parapar, A. Barreiro, and S. Lopez-Larrosa, "Bdi-sen: A sentence dataset for clinical symptoms of depression," in *ACM SIGIR*, 2023, p. 2996–3006.

[15] A.-M. Bucur, I. R. Podină, and L. P. Dinu, "A psychologically informed part-of-speech analysis of depression in social media," in *RANLP*, 2021, pp. 199–207.

[16] G. Coppersmith, K. Ngo, R. Leary, and A. Wood, "Exploratory analysis of social media prior to a suicide attempt," in *CLPsych*, 2016, pp. 106–117.

[17] C. Sanchez, A. Grzenda, A. Varias, A. S. Widge, L. L. Carpenter, W. M. McDonald, C. B. Nemeroff, N. H. Kalin, G. Martin, M. Tohen, M. Filippou-Frye, D. Ramsey, E. Linos, C. Mangurian, and C. I. Rodriguez, "Social media recruitment for mental health research: A systematic review," *Compr. Psychiatry*, vol. 103, p. 152197, 2020.

[18] H. Zogan, I. Razzak, S. Jameel, and G. Xu, "Depressionnet: Learning multi-modalities with user post summarization for depression detection on social media," in *ACM SIGIR*, 2021, p. 133–142.

[19] H. Zogan, I. Razzak, X. Wang, S. Jameel, and G. Xu, "Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media," *World Wide Web*, vol. 25, no. 1, pp. 281–304, 2022.

[20] Y. Jia, J. McDermid, T. Lawton, and I. Habli, "The role of explainability in assuring safety of machine learning in healthcare," *IEEE T. on Emerg. Topics in Comput.*, vol. 10, no. 04, pp. 1746–1760, 2022.

[21] C. Su, Z. Xu, J. Pathak, and F. Wang, "Deep learning in mental health outcome research: a scoping review," *Translational Psychiatry*, vol. 10, no. 1, p. 116, 2020.

[22] A. Le Glaz, Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T. C. Ryan, J. Marsh, J. Devylder, M. Walter, S. Berrouiguet *et al.*, "Machine learning and natural language processing in mental health: systematic review," *JMIR*, vol. 23, no. 5, p. e15708, 2021.

[23] A. Abd-Alrazaq, D. Alhuwail, J. Schneider, C. T. Toro, A. Ahmed, M. Alzubaidi, M. Alajlani, and M. Househ, "The performance of artificial intelligence-driven technologies in diagnosing mental disorders: an umbrella review," *Npj Digital Medicine*, vol. 5, no. 1, p. 87, 2022.

[24] S. Chancellor and M. De Choudhury, "Methods in predictive techniques for mental health status on social media: a critical review," *NPJ digital medicine*, vol. 3, no. 1, p. 43, 2020.

[25] V. Schønning, G. J. Hjetland, L. E. Aarø, and J. C. Skogen, "Social media use and mental health and well-being among adolescents–a scoping review," *Frontiers in psychology*, vol. 11, p. 542107, 2020.

[26] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh, "An inventory for measuring depression." *Archives of General Psychiatry*, vol. 4, pp. 561–571, 1961.

[27] L. S. Radloff, "The ces-d scale: A self-report depression scale for research in the general population," *App. Psycho. Meas.*, vol. 1, no. 3, pp. 385–401, 1977.

[28] Z. Cooper and C. Fairburn, "The eating disorder examination: A semi-structured interview for the assessment of the specific psychopathology of eating disorders," *Int. J. of Eat. Dis.*, vol. 6, no. 1, pp. 1–8, 1987.

[29] D. M. Garner and P. E. Garfinkel, "The eating attitudes test: an index of the symptoms of anorexia nervosa," *Psychological Med.*, vol. 9, no. 2, p. 273–279, 1979.

[30] A. P. Association, *Diagnostic and statistical manual of mental disorders: DSM-5™, 5th ed.* American Psychiatric Publishing, Inc., 2013.

[31] R. B. Aaron T. Beck, Robert A. Steer and W. F. Ranieri, "Comparison of beck depression inventories-ia and-ii in psychiatric outpatients," *Journal of Personality Assessment*, vol. 67, no. 3, pp. 588–597, 1996.

[32] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *NeurIPS*, 2017, p. 1025–1035.

[33] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The phq-9: Validity of a brief depression severity measure." *Journal of General Internal Med.*, vol. 16, no. 9, pp. 606–613, 2001.

[34] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7," *Archives of Internal Med.*, vol. 166, no. 10, pp. 1092–1097, 2006.

[35] R. A. Steer and A. T. Beck, "Beck anxiety inventory," in *Evaluating stress: A book of resources*. Scarecrow Education, 1997, pp. 23–40.

[36] S. Dhelim, L. Chen, S. K. Das, H. Ning, C. Nugent, G. Leavey, D. Pesch, E. Bantry-White, and D. Burns, "Detecting mental distresses using social behavior analysis in the context of covid-19: A survey," *ACM Comput. Surv.*, vol. 55, no. 14s, 2023.

[37] E. Kadkhoda, M. Khorasani, F. Pourgholamali, M. Kahani, and A. R. Ardani, "Bipolar disorder detection over social media," *Informatics in Med. Unlocked*, vol. 32, p. 101042, 2022.

[38] T. Richter, B. Fishbain, A. Markus, G. Richter-Levin, and H. Okon-Singer, "Using machine learning-based analysis for behavioral differentiation between anxiety and depression," *Scientific Reports*, vol. 10, no. 1, p. 16381, 2020.

[39] M. Mitchell, K. Hollingshead, and G. Coppersmith, "Quantifying the language of schizophrenia in social media," in *CLPsych*, 2015, pp. 11–20.

[40] L. Cao, H. Zhang, and L. Feng, "Building and using personal knowledge graph to improve suicidal ideation detection on social media," *IEEE Trans. on Multimedia*, vol. 24, pp. 87–102, 2022.

[41] R. Pavelko and J. G. Myrick, "Tweeting and trivializing: How the trivialization of obsessive–compulsive disorder via social media impacts user perceptions, emotions, and behaviors," *Imagination, Cognition and Personality*, vol. 36, no. 1, pp. 41–63, 2016.

[42] S. Bhatia, M. Hayat, and R. Goecke, "A multimodal system to characterise melancholia: Cascaded bag of words approach," in *ACM ICMI*, 2017, p. 274–280.

[43] C.-H. Chang, E. Saravia, and Y.-S. Chen, "Subconscious crowdsourcing: A feasible data collection mechanism for mental disorder detection on social media," in *IEEE/ACM ASONAM*, 2016, pp. 374–379.

[44] E. Saravia, C.-H. Chang, R. J. De Lorenzo, and Y.-S. Chen, "Midas: Mental illness detection and analysis via social media," in *IEEE/ACM ASONAM*, 2016, pp. 1418–1421.

[45] H. Yan, E. E. Fitzsimmons-Craft, M. Goodman, M. Krauss, S. Das, and P. Cavazos-Rehg, "Automatic detection of eating disorder-related social media posts that could benefit from a mental health intervention," *Int. J. of Eat. Dis.*, vol. 52, no. 10, pp. 1150–1156, 2019.

[46] B. Shickel and P. Rashidi, "Automatic triage of mental health forum posts," in *CLPsych*, 2016, pp. 188–192.

[47] R. Skaik and D. Inkpen, "Using twitter social media for depression detection in the canadian population," in *AICCC*, 2021, p. 109–114.

[48] Y. Yuan, K. Saha, B. Keller, E. T. Isometsä, and T. Aledavood, "Mental health coping stories on social media: A causal-inference study of papageno effect," in *ACM Web Conf.*, 2023, p. 2677–2685.

[49] Y. Q. Lim, M. J. Lee, and Y. L. Loo, "Towards a machine learning framework for suicide ideation detection in twitter," in *AiDAS*, 2022, pp. 153–157.

[50] M. E. Aragón, A. P. López-Monroy, L. C. González-Gurrola, and M. Montes-y Gómez, "Detecting mental disorders in social media through emotional patterns - the case of anorexia and depression," *IEEE Trans. on Affect. Comput.*, vol. 14, no. 1, pp. 211–222, 2023.

[51] J. Ma, L. Wang, Y.-R. Zhang, W. Yuan, and W. Guo, "An integrated latent dirichlet allocation and word2vec method for generating the topic evolution of mental models from global to local," *Expert Syst. with Applic.*, vol. 212, p. 118695, 2023.

[52] L. J. Hagg, S. S. Merkouris, G. A. O'Dea, L. M. Francis, C. J. Greenwood, M. Fuller-Tyszkiewicz, E. M. Westrupp, J. A. Macdonald, and G. J. Youssef, "Examining analytic practices in latent dirichlet allocation within psychological science: Scoping review," *J. Med. Internet Res.*, vol. 24, no. 11, p. e33166, 2022.

[53] S. A. Lee, "Coronavirus anxiety scale: A brief mental health screener for covid-19 related anxiety," *Death Studies*, vol. 44, no. 7, pp. 393–401, 2020.

[54] M. Tlachac and E. Rundensteiner, "Screening for depression with retrospectively harvested private versus public text," *IEEE J. of Biomed. and Health Inform.*, vol. 24, no. 11, pp. 3326–3332, 2020.

[55] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in *CLPsych*, 2014, pp. 51–60.

[56] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv:1301.3781*, 2013.

[57] I. Straw and C. Callison-Burch, "Ai in mental health and the biases of language based models," *PLOS ONE*, vol. 15, no. 12, pp. 1–19, 2020.

[58] Y. Liu, C. Xu, X. Kuai, H. Deng, K. Wang, and Q. Luo, "Analysis of the causes of inferiority feelings based on social media data with word2vec," *Scientific Reports*, vol. 12, no. 1, p. 5218, 2022.

[59] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.

[60] U. Ahmed, G. Srivastava, U. Yun, and J. C.-W. Lin, "Eandc: An explainable attention network based deep adaptive clustering model for mental health treatment," *Future Generation Computer Syst.*, vol. 130, pp. 106–113, 2022.

[61] K. Dheeraj and T. Ramakrishnudu, "Negative emotions detection on online mental-health related patients texts using the deep learning with mha-bcnn model," *Expert Syst. with Applications*, vol. 182, p. 115265, 2021.

[62] H. Song, J. You, J.-W. Chung, and J. C. Park, "Feature attention network: Interpretable depression detection from social media," in *PACLIC*, 2018.

[63] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019, pp. 4171–4186.

[64] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *AAAI*, vol. 33, no. 01, 2019, pp. 7370–7377.

[65] X. Liu, X. You, X. Zhang, J. Wu, and P. Lv, "Tensor graph convolutional networks for text classification," in *AAAI*, vol. 34, no. 05, 2020, pp. 8409–8416.

[66] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.

[67] I. Mihov, H. Chen, X. Qin, W.-S. Ku, D. Yan, and Y. Liu, "Mentalnet: Heterogeneous graph representation for early depression detection," in *IEEE ICDM*, 2022, pp. 1113–1118.

[68] A.-T. Kuo, H. Chen, Y.-H. Kuo, and W.-S. Ku, "Dynamic graph representation learning for depression screening with transformer," *arXiv:2305.06447*, 2023.

[69] J. N. Rosenquist, J. H. Fowler, and N. A. Christakis, "Social network determinants of depression," *Molecular Psychiatry*, vol. 16, no. 3, pp. 273–281, 2011.

[70] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[71] H. Zogan, I. Razzak, S. Jameel, and G. Xu, "Hierarchical convolutional attention network for depression detection on social media and its impact during pandemic," *IEEE J. of Biomed. and Health Inform.*, pp. 1–9, 2023.

[72] M. Gaur, A. Alambo, J. P. Sain, U. Kursuncu, K. Thirunarayan, R. Kavuluru, A. Sheth, R. Welton, and J. Pathak, "Knowledge-aware assessment of severity of suicide risk for early intervention," in *The Web Conf.*, 2019, p. 514–525.

[73] V. Tejaswini, K. S. Babu, and B. Sahoo, "Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model," *ACM TALLIP*, 2022.

[74] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[75] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning*, 2014.

[76] A. S. Uban, B. Chulvi, and P. Rosso, "Understanding patterns of anorexia manifestations in social media data with deep learning," in *CLPsych*, 2021, pp. 224–236.

[77] A.-S. Uban, B. Chulvi, and P. Rosso, "An emotion and cognitive based analysis of mental health disorders from social media data," *Future Generation Computer Syst.*, vol. 124, pp. 480–494, 2021.

[78] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473*, 2014.

[79] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[80] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv:1910.01108*, pp. arXiv–1910, 2019.

[81] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv:1907.11692*, 2019.

[82] W. Ragheb, J. Azé, S. Bringay, and M. Servajean, "Negatively correlated noisy learners for at-risk user detection on social networks: A study on depression, anorexia, self-harm, and suicide," *IEEE TKDE*, vol. 35, no. 1, pp. 770–783, 2023.

[83] L. Cao, H. Zhang, L. Feng, Z. Wei, X. Wang, N. Li, and X. He, "Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention," in *EMNLP-IJCNLP*, 2019, pp. 1718–1728.

[84] M. Matero, A. Idnani, Y. Son, S. Giorgi, H. Vu, M. Zamani, P. Limbachiya, S. C. Guntuku, and H. A. Schwartz, "Suicide risk assessment with multi-level dual-context language and BERT," in *CLPsych*, 2019, pp. 39–44.

[85] M. Abuhassan, T. Anwar, M. Fuller-Tyszkiewicz, H. K. Jarman, A. Shatte, C. Liu, and S. Sukunesan, "Classification of twitter users with eating disorder engagement: Learning from the biographies," *Computers in Human Behav.*, vol. 140, p. 107519, 2023.

[86] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu, "Depression detection via harvesting social media: A multimodal dictionary learning solution," in *IJCAI*, 2017, pp. 3838–3844.

[87] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *ACL*, 2020, pp. 7871–7880.

[88] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv:1710.10903*, 2017.

[89] U. Naseem, J. Kim, M. Khushi, and A. Dunn, "Graph-based hierarchical attention network for suicide risk detection on social media," in *ACM Web Conf.*, 2023, p. 995–1003.

[90] S. Liu, F. Vahedian, D. Hachen, O. Lizardo, C. Poellabauer, A. Striegel, and T. Milenković, "Heterogeneous network approach to predict individuals' mental health," *ACM Trans. Knowl. Discov. Data*, vol. 15, no. 2, 2021.

[91] J. Pirayesh, H. Chen, X. Qin, W.-S. Ku, and D. Yan, "Mentalspot: Effective early screening for depression based on social contagion," in *ACM CIKM*, 2021, p. 1437–1446.

[92] R. Sawhney, H. Joshi, R. R. Shah, and L. Flek, "Suicide ideation detection via social and temporal user representations using hyperbolic learning," in *NAACL*, 2021, pp. 2176–2190.

[93] I. Chami, Z. Ying, C. Ré, and J. Leskovec, "Hyperbolic graph convolutional neural networks," in *NeurIPS*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, 2019.

[94] R. Sawhney, H. Joshi, S. Gandhi, and R. R. Shah, "A time-aware transformer based model for suicide ideation detection on social media," in *EMNLP*, 2020, pp. 7685–7697.

[95] ——, "Towards ordinal suicide ideation detection on social media," in *ACM WSDM*, 2021, p. 22–30.

[96] T. Zhang, K. Yang, and S. Ananiadou, "Sentiment-guided transformer with severity-aware contrastive learning for depression detection on social media," in *BioNLP*, 2023, pp. 114–126.

[97] Z. Jiang, S. I. Levitan, J. Zomick, and J. Hirschberg, "Detection of mental health from Reddit via deep contextualized representations," in *LOUHI*. ACL, 2020, pp. 147–156.

[98] A. M. Schoene, A. P. Turner, G. De Mel, and N. Dethlefs, "Hierarchical multiscale recurrent neural networks for detecting suicide notes," *IEEE T. on Affect. Comput.*, vol. 14, no. 1, pp. 153–164, 2023.

[99] J. Wu, X. Wu, Y. Hua, S. Lin, Y. Zheng, and J. Yang, "Exploring social media for early detection of depression in covid-19 patients," in *ACM Web Conf.*, 2023, p. 3968–3977.

[100] R. Sawhney, A. Neerkaje, and M. Gaur, "A risk-averse mechanism for suicidality assessment on social media," in *ACL*, 2022, pp. 628–635.

[101] N. Wang, L. Fan, Y. Shivtare, V. Badal, K. Subbalakshmi, R. Chandramouli, and E. Lee, "Learning models for suicide prediction from social media posts," in *CLPsych*, 2021, pp. 87–92.

[102] H. Amini and L. Kosseim, "Towards explainability in using deep learning for the detection of anorexia in social media," in *NLDB*, vol. 12089, 2020, pp. 225–235.

[103] S. Han, R. Mao, and E. Cambria, "Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings," in *COLING*, 2022, pp. 94–104.

[104] U. Naseem, M. Khushi, J. Kim, and A. G. Dunn, "Hybrid text representation for explainable suicide risk identification on social media," *IEEE Trans. on Comput. Soc. Syst.*, pp. 1–10, 2022.

[105] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, and M. Berk, "Affective and content analysis of online depression communities," *IEEE Trans. on Affec. Comp.*, vol. 5, no. 3, pp. 217–226, 2014.

[106] M. M. Bradley and P. J. Lang, "Affective norms for english words (anew): Instruction manual and affective ratings," 1999.

[107] D. E. Losada, F. Crestani, and J. Parapar, "Overview of erisk 2019 early risk prediction on the internet," in *CLEF*, 2019, p. 340–357.

[108] ——, "erisk 2020: Self-harm and depression challenges," in *Adv. in Inf. Retr.*, 2020, pp. 557–563.

[109] T. Anwar, M. Fuller-Tyszkiewicz, H. K. Jarman, M. Abuhassan, A. Shatte, W. Team, and S. Sukunesan, "Edbase: Generating a lexicon base for eating disorders via social media," *IEEE J. of Biomed. and Health Inform.*, vol. 26, no. 12, pp. 6116–6125, 2022.

[110] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly available pretrained language models for mental healthcare," in *LREC*, 2022, pp. 7184–7190.

[111] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilovic *et al.*, "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques," in *INFORMS Ann. Meeting*, 2021.

[112] Y. Jia, J. McDermid, T. Lawton, and I. Habli, "The role of explainability in assuring safety of machine learning in healthcare," *IEEE T. on Emerging Topics in Computing*, vol. 10, no. 4, pp. 1746–1760, 2022.

[113] H. Yuan, H. Yu, S. Gui, and S. Ji, "Explainability in graph neural networks: A taxonomic survey," *IEEE T. PAMI*, vol. 45, no. 05, pp. 5782–5799, 2023.

[114] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, "A survey of the state of explainable AI for natural language processing," in *IJCNLP*, 2020, pp. 447–459.

[115] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *ACM SIGKDD*, 2016, p. 1135–1144.

[116] C. Saxena, M. Garg, and G. Ansari, "Explainable causal analysis of mental health on social media data," in *ICONIP*, 2023, p. 172–183.

[117] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *NeurIPS*, 2017, p. 4768–4777.

[118] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, pp. 1–46, 2015.

[119] B. Hoover, H. Strobelt, and S. Gehrmann, "exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models," in *ACL: Syst. Demo.*, 2020, pp. 187–196.

[120] B. v. Aken, B. Winter, A. Löser, and F. A. Gers, "Visbert: Hidden-state visualizations for transformers," in *ACM Web Conf.*, 2020, pp. 207–211.

[121] J. Vig, "Visualizing attention in transformer-based language representation models," *arXiv e-prints*, pp. arXiv–1904, 2019.

[122] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *NAACL*, 2016, pp. 1480–1489.

[123] J. H. Friedman, "Greedy function approximation: A gradient boosting machine." *The Annals of Statistics*, vol. 29, no. 5, pp. 1189 – 1232, 2001.

[124] B. M. Greenwell, B. C. Boehmke, and A. J. McCarthy, "A simple and effective model-based variable importance measure," *arXiv:1805.04755*, 2018.

[125] J. B. Alex Goldstein, Adam Kapelner and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *J. of Comput. and Graph. Stat.*, vol. 24, no. 1, pp. 44–65, 2015.

[126] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: generating explanations for graph neural networks," in *NeurIPS*, 2019, pp. 9244–9255.

[127] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang, "Parameterized explainer for graph neural network," in *NeurIPS*, vol. 33, 2020, pp. 19 620–19 631.

[128] Y. Fan, Y. Yao, and C. Joe-Wong, "Gcn-se: Attention as explainability for node classification in dynamic graphs," in *ICDM*, 2021, pp. 1060–1065.

[129] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE/CVF CVPR*, 2018, pp. 7132–7141.

[130] D. E. Losada and F. Crestani, "A test collection for research on depression and language use," in *CLEF*, 2016, pp. 28–39.

[131] E. Turcan and K. McKeown, "Dreaddit: A Reddit dataset for stress analysis in social media," in *LOUHI*. ACL, 2019, pp. 97–107.

[132] J. Parapar, P. Martín-Rodilla, D. E. Losada, and F. Crestani, "Overview of erisk 2023: Early risk prediction on the internet," in *CLEF*, 2023, pp. 294–315.

[133] ——, "Overview of erisk 2023: Early risk prediction on the internet," in *CLEF*. Springer, 2023, pp. 294–315.

[134] M. Gaur, A. Alambo, J. P. Sain, U. Kursuncu, K. Thirunarayan, R. Kavuluru, A. Sheth, R. Welton, and J. Pathak, "Knowledge-aware assessment of severity of suicide risk for early intervention," in *The Web Conf.*, 2019, p. 514–525.

[135] S. , A. K. Singh, U. Arora, S. Shrivastava, A. Singh, R. R. Shah, and P. Kumaraguru, "Twitter-stmhd: An extensive user-level database of multiple mental health disorders," *Proc. of the Internat. AAAI Conf. on Web and Soc. Med.*, vol. 16, no. 1, pp. 1182–1191, 2022.

[136] A. Haque, V. Reddi, and T. Giallanza, "Deep learning for suicide and depression identification with unsupervised label correction," in *ICANN*, 2021, pp. 436–447.

[137] A. Yates, A. Cohan, and N. Goharian, "Depression and self-harm risk assessment in online forums," in *EMNLP*, 2017, pp. 2968–2978.

[138] A. Cohan, B. Desmet, A. Yates, L. Soldaini, S. MacAvaney, and N. Goharian, "SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions," in *COLING*, 2018, pp. 1485–1497.

[139] B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen, "Detecting suicidality on twitter," *Internet Interventions*, vol. 2, no. 2, pp. 183–188, 2015.

[140] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, "From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses," in *CLPsych*, 2015, pp. 1–10.

[141] S. Chancellor, T. Mitra, and M. De Choudhury, "Recovery amid pro-anorexia: Analysis of recovery in social media," in *ACM CHI*, 2016, p. 2111–2123.

[142] I. Sekulic, M. Gjurković, and J. Šnajder, "Not just depressed: Bipolar disorder prediction on Reddit," in *WASSA*. ACL, 2018, pp. 72–78.

[143] U. Naseem and K. Musial, "Dice: Deep intelligent contextual embedding for twitter sentiment analysis," in *ICDAR*, 2019, pp. 953–958.

[144] H. Zogan, I. Razzak, X. Wang, S. Jameel, and G. Xu, "Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media," *World Wide Web*, vol. 25, no. 1, pp. 281–304, 2022.

[145] J. Shen, J. Liu, D. Finnie, N. Rahmati, M. Bendersky, and M. Najork, ""why is this misleading?": Detecting news headline hallucinations with explanations," in *ACM Web Conf.*, 2023, p. 1662–1672.

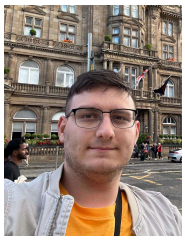
[146] K. Yang, S. Ji, T. Zhang, Q. Xie, Z. Kuang, and S. Ananiadou, "Towards interpretable mental health analysis with large language models," in *EMNLP*, 2023, pp. 6056–6077.

[147] J. Morley, C. C. Machado, C. Burr, J. Cowls, I. Joshi, M. Taddeo, and L. Floridi, "The ethics of ai in health care: A mapping review," *Social Science and Medicine*, vol. 260, p. 113172, 2020.

[148] S. Chancellor, E. P. S. Baumer, and M. De Choudhury, "Who is the "human" in human-centered machine learning: The case of predicting mental health from social media," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, 2019.

[149] S. R. Pendse, K. Niederhoffer, and A. Sharma, "Cross-cultural differences in the use of online mental health support forums," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, 2019.

[150] N. Gopalkrishnan, "Cultural diversity and mental health: Considerations for policy and practice," *Frontiers in public health*, vol. 6, p. 179, 2018.

**Yusif Ibrahimov** received the B.Sc.(Hons) degree in Computer Science from French-Azerbaijani University in Azerbaijan and University of Strasbourg in France. Currently, he is a full time PhD student and Teaching Assistant at the University of York in United Kingdom. His research interests include advanced deep learning, graph machine learning, social data mining and explainable artificial intelligence.

**Tarique Anwar** received the Masters and PhD degrees in Computer Science from Jamia Millia Islamia in India and Swinburne University of Technology in Australia, respectively. Currently, he is working as a Lecturer at the University of York in the United Kingdom. His research interests include data science, machine learning, social data mining and big data analytics.

**Tommy Yuan** received the B.Sc. degree in railway transport engineering from Southwest Jiaotong University, Chengdu, China, in 1993, and the M.Sc. degree in software engineering and the Ph.D. degree in computer science from Leeds Metropolitan University (currently, Leeds Beckett University), Leeds, U.K., in 2001 and 2004, respectively. He was a Railway Transport Engineer in China. After the completion of his Ph.D. degree, he started his academic career as an Assistant Professor in Iceland. In 2009, he joined the University of York, York, U.K., where he is a currently a Reader. His research interests include argument, dialogue, and their applications to computer system dependability, human–computer dialogue, and agent communication.