IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 December 2024, accepted 11 January 2025, date of publication 16 January 2025, date of current version 23 January 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3530862

TOPICAL REVIEW

Depression Detection in Social Media: A Comprehensive Review of Machine Learning and Deep Learning Techniques

WALEED BIN TAHIR[1], SHAH KHALID[1], SULAIMAN ALMUTAIRI[2], MOHAMMED ABOHASHRH[3], SUFYAN ALI MEMON[4], AND JAWAD KHAN[5]

[1]School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad 44000, Pakistan
[2]Department of Health Informatics, College of Public Health and Health Informatics, Qassim University, Qassim 51452, Saudi Arabia
[3]Department of Basic Medical Sciences, College of Applied Medical Sciences, King Khalid University, Abha 62521, Saudi Arabia
[4]Department of Defense Systems Engineering, Sejong University, Seoul 05006, Republic of Korea
[5]School of Computing, Gachon University, Seongnam 13120, South Korea

Corresponding author: Shah Khalid (shah.khalid@seecs.edu.pk)

The authors extend their appreciation to the King Salman Centre for Disability Research for funding this work through Research Grant no KSRG-2023-560.

ABSTRACT Depression is a widespread mental health disorder that may remain undiagnosed by conventional clinical methods. The rapidly growing world of social media sites such as Twitter, Reddit, Facebook, Instagram, and Weibo has provided new avenues for depression detection using Machine Learning (ML) as well as Deep Learning (DL), which analyze user behavior patterns and linguistic cues for more accurate detection of depression. Many techniques have been developed for this aim over the years. Identifying relevant publications on this topic using current academic search systems is challenging due to the rapid growth of research publications, unclear or limited search terms, and the complexity of citation networks. Several review papers have been published to ease this task by summarizing the methodologies, key findings, and recommendations for future research. However, most current reviews often do not provide a clear overview of the evolution, latest techniques, and challenges. This paper aims to address that gap by providing a comprehensive review of ML and DL methodologies for detecting depression on social media. We propose a generic architecture for these systems and present a detailed analysis of methodologies and datasets used for evaluation in this field. In addition, we highlight key open research areas, providing a useful starting point for further research and development. By narrowing our focus to social media, this review contributes to advancing the understanding and application of cutting-edge methods for depression detection. While this review highlights advancements in social media-based depression detection, it excludes alternative approaches like graph-based systems and reinforcement learning, and its focus on social media may limit its applicability to other domains.

INDEX TERMS Deep learning, depression detection, machine learning, natural language processing, sentiment analysis, social media.

I. INTRODUCTION

Depression is a widespread mental illness that affects millions of people worldwide. Depression is characterized by loss of interest, prolonged sadness, and several emotional and physical problems that in turn interfere with daily functioning and quality of life. According to WHO [1], 1 in every 8 people in the world is living with some form of mental disorder in which the majority of the population does not have access to an effective cure. Early detection of depression can help avoid suicide and self-harm occurrences, which are more common among those who suffer from depression.

The techniques traditionally used to diagnose depression include clinical evaluations and questionnaires filled in by

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar.

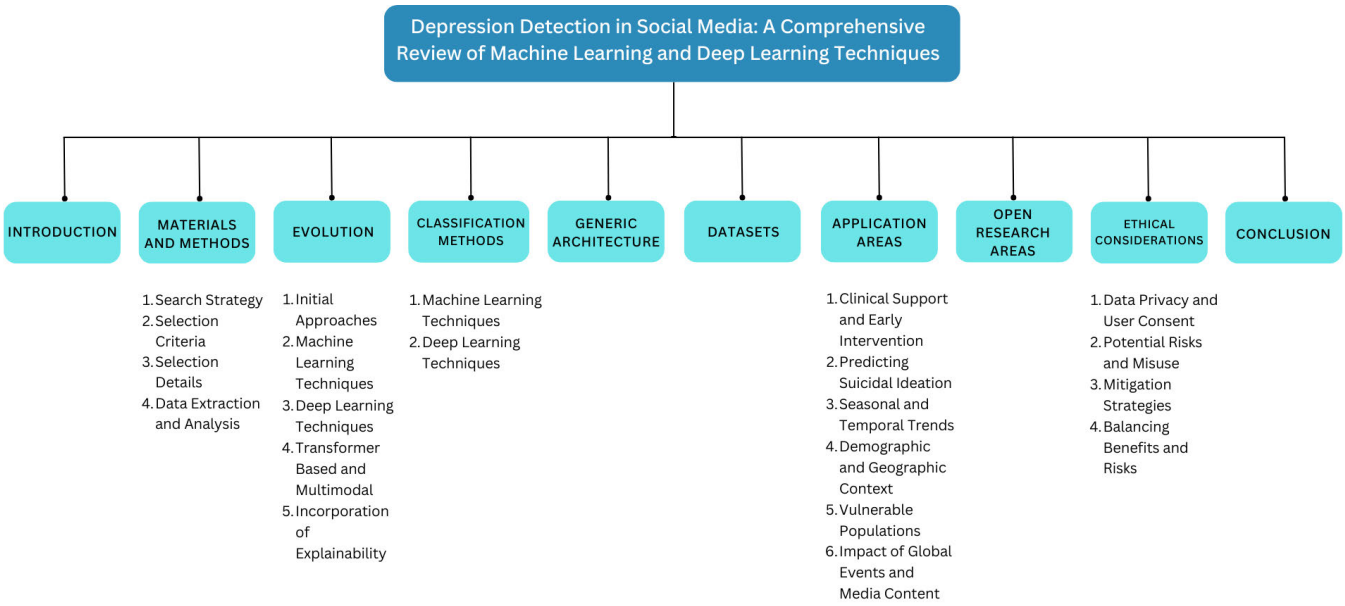

FIGURE 1. Paper structure.

patients, which are at times, not fully accurate, as the questions do not cover the whole range of individuals affected by depressive symptoms, the majority of whom do not look for help.

The emergence of numerous social media sites including X (formerly Twitter), Instagram, Reddit, and Facebook has opened up the road for finding out human behavior and mood in real time. People often share their ideas, feelings, and the things they do every day on these social media platforms, generating an extensive dataset that could possibly indicate mental illness information. Consequently, the idea of utilizing social media data to detect depression has gained popularity.

Techniques of ML and DL have been found to be very promising in domains such as Natural Language Processing (NLP), predictive analytics, and image recognition. These advanced computational techniques can analyze large unstructured data volumes and predict results with great accuracy by detecting hidden patterns. In the case of depression detection using social media posts, ML and DL models can be trained to identify the linguistic cues, behavioral patterns, and other depressive symptom indicators.

In early studies, the identification of depressive symptoms was mostly done by textual features from social media posts using classic ML models like Support Vector Machines (SVMs) [2], [3], [4], Decision Trees (DTs) [5], [6], [32], and Random Forest (RF) [7], [8], [9]. Although these early methods showed off ML's potential in this field, they were frequently limited by the variety and complexity of human speech.

Significant progress was made with the introduction of deep learning. To extract sequential and contextual information from text, researchers started using complex models like Recurrent Neural Networks (RNNs) [10], [11], [12] and Convolutional Neural Networks (CNNs) [13], [14], [15]. The introduction of models based on transformers, such as BERT [16], marked a significant leap, offering improved accuracy in understanding and processing natural language. These models facilitated more nuanced detection of depressive cues from social media content.

The field saw further enhancements with the integration of multimodal data [17], [18], [19], combining text, images, and user behavior to gain a more thorough understanding of a user's mental state. Advances in transfer learning and the development of pre-trained language models allowed for more efficient and effective depression detection systems.

The main goal of this paper is to streamline researchers' efforts by compiling comprehensive information into a single source while addressing critical gaps left by existing review papers. For instance, while [20] provides detailed coverage of selected studies and mentions datasets, it lacks an in-depth discussion of the datasets, application areas, and generic architecture. Similarly, [21] offers limited insight, as it neither details the reviewed studies nor discusses the datasets, application areas, or generic architecture. Furthermore, [22] focuses on a small subset of studies in detail but also fails to address the datasets, application areas, generic architecture, and open research areas. Additionally, [23] discusses methods and datasets in detail but does not cover application areas or generic architecture. Notably, none of these reviews examine the evolution of depression detection techniques in social media—a critical area given the field's rapid technological advancements. This paper addresses these gaps by offering a comprehensive analysis of key studies, datasets, and application areas. It also proposes a generic architecture and examines the field's evolution. These contributions are timely

due to the rising need for effective mental health detection in social media and the challenges posed by multimodal data.

This review aims to achieve the following objectives:

- Identify and discuss the progression and evolution of techniques used for depression detection.
- Provide a detailed overview of the current status of research on depression detection in social media using machine and deep learning techniques.
- Outline a generic pipeline architecture commonly implemented in these detection methods.
- Discuss the datasets used to evaluate depression detection methods.
- Highlight the application areas of depression detection using social media.
- Identify research gaps and future directions in this field.
- Focus on ethical aspects like user consent, data privacy, and potential risks.

The structure of this paper is organized as follows: Section II outlines the methodology and selection criteria for the literature review. Section III outlines the progression of techniques utilized for depression detection. Section IV explores various machine and deep learning techniques utilized for depression detection on social networking platforms. Section V outlines the generic pipeline implemented for these methods. Section VI discusses the datasets used in assessing the effectiveness of these techniques. Section VII highlights the practical applications of depression detection in social media. Section VIII discusses current open research areas in the field. Section IX addresses ethical issues such as privacy, consent, and potential misuse of these technologies. Finally, Section X concludes the study.

The paper's structure is illustrated in Figure 1.

II. MATERIALS AND METHODS

This section outlines the methodology utilized to conduct literature review for depression detection in social media.

A. SEARCH STRATEGY

To conduct a comprehensive review of the literature on depression detection in social media using ML and DL, a systematic search strategy was employed. The following databases were utilized to gather relevant research articles: ScienceDirect, IEEE Xplore, ACM Digital Library, Google Scholar, Scopus, and DBLP. Keywords and phrases like ''depression detection'', ''social media'', ''deep learning'', ''machine learning'', ''sentiment analysis'', ''multi modal'', ''reddit'', and ''twitter'' were combined to search.

The search was restricted to articles published between January 2018 and July 2024 in order to guarantee the inclusion of recent and relevant studies. Additionally, a manual review of the reference lists of key papers was conducted to find further relevant studies.

B. SELECTION CRITERIA

The inclusion and exclusion criteria were carefully designed to ensure the review focuses on high-quality, relevant studies that contribute meaningfully to the understanding of depression detection using ML and DL techniques. These criteria aim to achieve two primary purposes: (1) to ensure that only studies addressing the review's central theme—depression detection in social media using ML and DL—are considered, and (2) to maintain the scientific rigor of the review by excluding irrelevant or low-quality sources.

Inclusion Criteria:

1) Studies published between January 2018 and July 2024 to focus on the most recent and relevant advancements in the field.
2) Articles written in English to ensure accessibility and consistency in interpretation.
3) Research focusing on depression detection using ML or DL techniques, aligning with the review's scope.
4) Studies utilizing social media data for analysis, reflecting the specific application context of the review.
5) Peer-reviewed journal articles, and conference papers to maintain a high standard of scientific reliability.

Exclusion Criteria:

1) Studies not focused on depression detection, to eliminate unrelated topics and maintain thematic consistency.
2) Articles not utilizing social media data, as the review emphasizes this specific data source.
3) Papers that did not employ ML or DL techniques, excluding studies outside the technical focus of the review.
4) Non-peer-reviewed articles, opinion pieces, and editorials, which are less likely to meet the methodological rigor required for reliable synthesis.

By methodically reviewing and synthesizing available literature, this methodology aims to provide a thorough understanding of the current status of depression detection in social media using machine learning and deep learning techniques.

C. SELECTION DETAILS

Using the inclusion and exclusion criteria, a total of 86 papers were selected, of which 27 utilized machine learning techniques and 59 employed deep learning techniques. Figure 2 depicts the PRISMA flowchart, which details the study selection process followed in this systematic review.

D. DATA EXTRACTION AND ANALYSIS

From the selected articles, we systematically recorded the following data: authors, year of publication, journal/conference, research objectives, techniques used, features extracted, datasets, evaluation metrics, key findings, and challenges and limitations. An iterative process of thematic coding was utilized for analyzing the data in order to find recurrent patterns in approaches, datasets, and challenges. Trends were synthesized by examining the frequency and evolution of techniques, such as the increasing use of deep learning and multimodal approaches. By comparing these themes

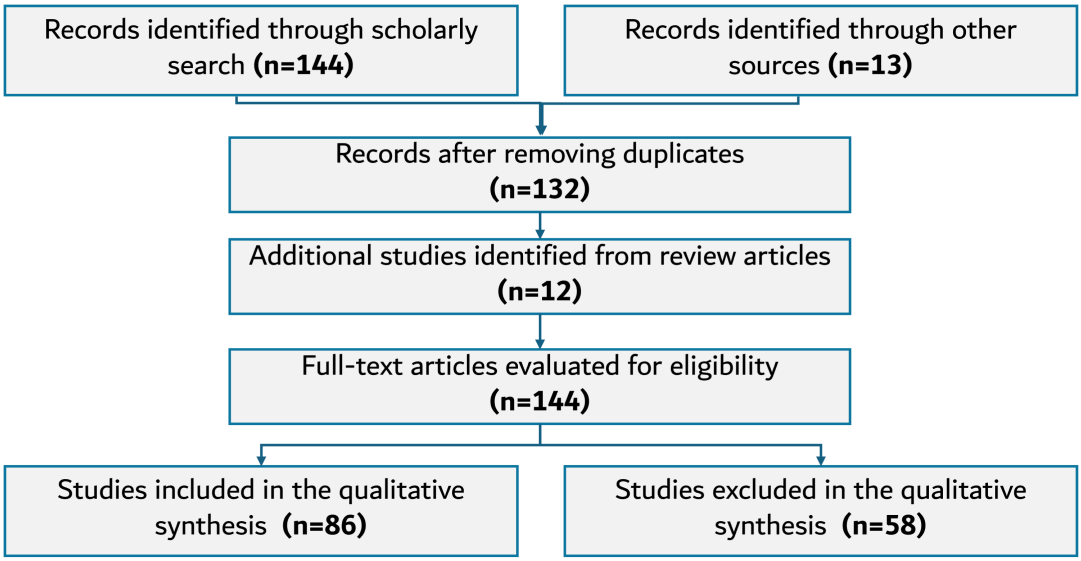

FIGURE 2. PRISMA Flowchart illustrating the inclusion and exclusion process for studies in the systematic review of depression detection from social media using machine learning and deep learning techniques.

and trends to stated research objectives and limitations, the analysis identified gaps in the literature, highlighting under-explored areas and methodological limitations. This rigorous approach ensured a comprehensive and insightful synthesis of the existing research.

III. EVOLUTION

This section outlines the key milestones, showcasing the progression from initial approaches to modern multimodal and explainable techniques. A summary timeline of these advances is provided in Figure 3, focusing on key moments in the development of the methods applied in this area.

A. INITIAL APPROACHES (PRE-2015)

The early stages of research in this domain focused on utilizing social media as a novel resource for gaining insights into mental health. Early works like [24] analyzed Twitter posts by examining linguistic and behavioral patterns to identify depressive symptoms. These efforts were primarily rule-based, relying on predefined linguistic rules to assign polarity scores, which were then used for depression classification [25]. During this time, statistical methods were also investigated, [26] used statistical techniques such as Fisher's exact test and chi-squared test to classify depression illnesses. Additionally, [27] used a model based on both node and linkage features for the classification of depression.

B. EMERGENCE OF MACHINE LEARNING TECHNIQUES (2015-2021)

A significant turning point toward the automation of depression detection occurred in the mid-2010s. Researchers started utilizing machine learning techniques for this purpose when large-scale datasets, such as the CLPsych dataset [28], became more readily available. Social media posts were classified using advanced NLP methods and ML models, like SVMs [29], [30], [31], K-Nearest Neighbors (kNN) [32], Logistic Regression (LR) [33] and DT [54]. During this time, initial attempts were also made to integrate ML models with more complex linguistic elements like syntactic patterns and n-grams. However, these models frequently had issues capturing long-range dependencies and complex contextual information in text, which limited their capacity to detect subtle signs of depression in social media posts.

C. INTEGRATION OF DEEP LEARNING (2018-PRESENT)

The emergence of DL in the late 2010s addressed these limitations, marking a new era for depression detection systems. Models like Long Short-Term Memory (LSTM) [10], [34], [35] replaced conventional ML techniques due to better performance in processing sequential and contextual data. LSTMs were especially useful for evaluating text data from social media, where context and nuance are critical for identifying depressive symptoms. This is because of their ability to capture long-term dependencies. Furthermore, CNNs began to be used for text classification tasks, which improved the capability of automated systems.

D. RISE OF TRANSFORMER-BASED MODELS AND MULTIMODAL APPROACHES (2020-PRESENT)

The recent introduction of transformer-based models, like BERT, has greatly changed the field. These models facilitate more precise and contextually aware analysis of textual data. With their ability to capture complex linguistic nuances and contextual dependencies, these models have set new benchmarks in depression detection tasks [15], [36], [37]. The period also witnessed the rise of multimodal approaches [17], [18], [19], where researchers began combining text, images,

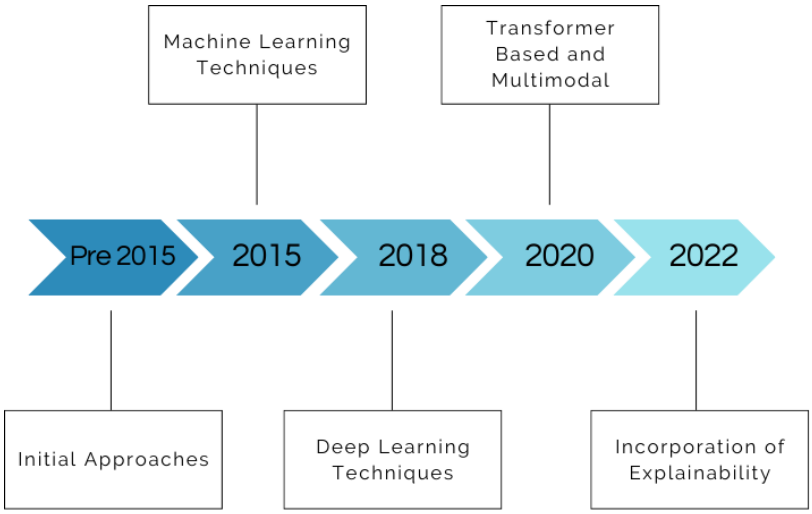

FIGURE 3. Evolution of depression detection techniques.

and even video data to enhance detection accuracy. These approaches have opened new avenues for integrating various data types, offering a more comprehensive understanding of a person's mental state.

E. INCORPORATION OF EXPLAINABILITY (2022-PRESENT)

As the field matures, there has been a growing emphasis on the explainability of models. Researchers are increasingly focused on making models interpretable, ensuring that their predictions can be understood and trusted by clinicians and users alike. Techniques such as attention scores [38], SHapley Additive exPlanation (SHAP), and Local Interpretable Model-agnostic Explanations (LIME) [39] are used to provide insights into model decisions. Explainable AI (XAI) frameworks ensure that predictions can be trusted and validated, addressing ethical concerns and enhancing the adoption of these technologies in real-world clinical settings.

This evolution underscores the rapid advancements in this field, moving from simple rule-based systems to sophisticated, interpretable, and multimodal approaches, reflecting a growing commitment to both technical excellence and societal impact.

IV. CLASSIFICATION METHODS

This section provides a detailed overview of ML and DL methods used for depression detection in social media. For each study, we discuss the dataset, preprocessing steps, feature extraction, classification methods, and results.

A. MACHINE LEARNING TECHNIQUES

A brief overview of select studies regarding depression detection through machine learning methods is provided in Table 1.

A novel approach for recognizing depression related content on Reddit was introduced by Pirina and Çöltekin [40]. Data gathered by Ramirez-Esparza et al. [41] were used in the study. They also collected data from the depression subreddit and used posts from a subreddit on breast cancer as a control set. Several classification techniques, including LR, SVMs, and RNNs, were applied. The best performance was obtained by SVMs using a combination of character and word n-grams of various sizes. The BM25 algorithm [42] was utilized to extract and weigh these n-grams from the texts. Random search was used to optimize the parameters for the maximum order of n-grams and the SVM margin parameter C. The BM25 parameters 'k1' and 'b' were fixed at 0.75 and 2.0, respectively. 5-fold cross-validation was employed for evaluation. High cross-validation performance was shown by the classifiers when both data sets were specific. However, the classifier's performance diminished when negative class texts came from less specific domains, and there was a notable performance drop between the cross-validation and test set results. Combining positive and negative instances from different settings resulted in a model that performed comparably but worse than more specific models due to less harmonized data sources.

A novel method for identifying depression with social network data was presented by Islam et al. [43]. The dataset used for this study consists of Facebook posts. Preprocessing techniques such as tokenization, stemming, and sentiment analysis were used. Ensemble, SVMs, DTs, kNN, and SVMs were used for classification. 10-fold cross-validation was employed for evaluation. The results demonstrated that the DT model outperformed other models, especially in recall and F-measure. Time series analysis indicated that the frequency of depressive posts was higher during the AM hours and there are seasonal trends in August and September.

Katchapakirin et al. [29] introduced a depression detection method for the Thai Facebook community. The system incorporates feature extraction using NLP tools, along with sentiment analysis, language translation, data collection, and attribute extraction from posts. ML algorithms, including

SVM, RF, and DL, were applied to classify the presence of depression based on these features. Majority Vote was used as the baseline and 8-fold cross-validation was employed for evaluation. Experimentation with the models demonstrated that all outperformed the baseline, with RF exhibiting superior performance over SVM, and DL achieving the highest accuracy, especially in identifying negative sentiment as a key indicator. The DL model highlighted that individuals who post negative sentiments without using emoticons and set their privacy to 'only me' are more likely to suffer from depression. Conversely, individuals who frequently share others' posts, post their memories, are active between 6 AM and 12 PM, and regularly tag friends are less likely to be depressed. Despite the study's potential, limitations such as small sample size and translation inaccuracies suggest the need for more extensive research and improved methods to enhance the detection algorithm.

Aragón et al. [30] introduced a novel methodology for depression detection in social media posts by utilizing fine-grained emotional data. Their method builds on datasets from the eRisk 2017 and 2018 evaluation tasks, which include user posts labeled according to confirmed depression diagnoses [44], [45]. The authors employed a two-step process: first, they used a lexicon of eight core emotions and sentiment labels [46], followed by AP clustering [47] on FastText sub-word embeddings [48] to create detailed emotional sub-groups. Text data was then transformed into these fine-grained emotional representations through a masking technique, where each word was replaced by its closest emotional label based on cosine similarity. The study formulated two distinct Bag of Sub Emotions (BoSE) models: unigrams, which represent emotion occurrence histograms, and bigrams, delineating emotion sequence histograms. For classification purposes, a SVM equipped with a linear kernel was employed. The results showed that the BoSE representation outperformed traditional Bag of Words (BoW) models. This highlights how crucial fine-grained emotional information is for detecting depression. When sentiment data was added to BoSE, the performance was improved even more. This suggests that people who are depressed communicate their feelings in more specific ways.

Tadesse et al. [31] introduced a comprehensive technique to detect depression posts on Reddit forums. The dataset was tokenized, preprocessed, and then stemmed. Preprocessing steps included the removal of stop words, URLs, and punctuation. Features were extracted by using Latent Dirichlet Allocation (LDA) for topic modeling, Term Frequency-Inverse Document Frequency (TF-IDF) vectorization with N-gram features to capture meaningful word patterns, and Linguistic Inquiry and Word Count (LIWC) dictionary to derive psycholinguistic features. LDA was limited to 70 topics and LIWC analysis extracted 68 out of 95 psycholinguistic features. Combining LIWC, LDA, and bigram features with an Multilayer Perceptrons (MLP) classifier achieved the highest overall accuracy of 91% and F1 score of 0.93, while bigram features with an SVM classifier achieved an accuracy of 80% and F1 score of 0.80.

An ML based algorithm was presented by Asad et al. [2] to assess the level of depression in social media users. A structural model with Naïve Bayes and SVM was created. After gathering data from Facebook and Twitter, it underwent preprocessing, which included tokenization, lowercase text conversion, and the removal of stop words, @mentions, and retweets. TF-IDF vectorization was used for feature extraction. Normal, mild, moderate, borderline, severe, and extreme are the six severity ranges into which the ML model was trained to classify depression levels. If the percentage is higher than the borderline of 55%, the user is considered depressed. A dataset of 50 Facebook users and 100 Twitter users was used to evaluate the model. Results showed that 65% of Twitter users and 38% of Facebook users had significant levels of depression. A questionnaire validated the results, resulting in a 74% accuracy and 100% precision for the Naïve Bayes classification.

A methodological approach for the analysis of mental health indicators from social media data was proposed by Arora and Arora [49]. The dataset was collected through Twitter's streaming Application Programming Interface (API). It underwent extensive preprocessing steps like converting text to lowercase, eliminating hashtags, URLs, unnecessary spaces, and punctuation, and replacing emoticons with sentiment keywords. Stop words were removed and sentiment-containing words were replaced by standard sentiment labels to decrease the size of the dataset. For feature extraction, the text was converted into base word forms and token sequences using stemming and tokenization, respectively. It further used sentiment extraction and Part of Speech (POS) tagging to support the feature set. The study used the Support Vector Regression (SVR) and Multinomial Naïve Bayes for classification. Multinomial Naïve Bayes and SVR models achieved classification accuracy of 78% and 79.7%, respectively.

An ML based approach to detect Postpartum Depression (PPD) using Reddit posts was proposed by Fatima et al. [3]. The model analyzed text-based posts from 21 subreddits to find linguistic patterns indicating PPD. 94 attributes were extracted by the LIWC tool, including word count, emotional tone, and a customized absolutist word dictionary. 20 key features were chosen with LASSO. Three categories were present in the dataset: PPD-specific content, general discussion, and depressed content. A two-layer machine learning model was used, with the first layer distinguishing depressive content from general discussion and the second identifying PPD-specific content. SVMs, MLP, and LR were used. In the first layer, the LR model provided the best performance, while in the second, the MLP model showed the best performance. The accuracy of the Depressive Content Classification (D-CC) was 91.7% and Postpartum Depression Content Classification (PPD-CC) obtained 86.91% accuracy.

Ding et al. [4] introduced a model to recognize depression in college students using Weibo data collected via third-party APIs and a custom web crawler. Preprocessing involved cleaning and sorting data, then extracting features like high-frequency words and emoji usage. High-frequency words were categorized into six levels. Key features included emoji frequency, word frequency, number of posts, followers, and followees, reducing an initial 1325 feature dimensions to 36 via deep neural networks. The core model used an integrated SVM with the AdaBoost algorithm to enhance classification. Training and testing were on datasets with 563 and 130 users, respectively. The DISVM classifier outperformed the Radial Basis Function Neural Network (accuracy: 82.31%, precision: 0.8366), standard SVM (accuracy: 80%, precision: 0.8277), and K-Nearest Neighbor (accuracy: 79.23%, precision: 0.8152) with an accuracy of 86.15% and precision of 0.8810. Results showed better recognition with 24-month data collection, highlighting depression's instability over time.

Skaik and Inkpen [5] used three datasets: D1 and D2 for depression estimation and P1 for Canadian population inference. D1 includes 292,564 tweets from 1,402 depressed users and 3,953,183 tweets from 5,610 non-depressed users. D2 consists of 327 self-reported depressed and 1,146 control users. P1 contains tweets from a 2015 Canadian representative sample, with demographics inferred from birth records. The study employed five classical and three deep learning models, utilizing features such as statistical measures, TF-IDF, LIWC categories, POS tags, topic modeling, and sentiment analysis. D1 dataset was utilized for training and 10-fold cross-validation, and D2 for testing. The Gradient Boosting Decision Tree (GBDT) model obtained the highest F1-score of 0.961 on D1 and performed best on D2 with an accuracy of 91.1% and precision of 0.875. In contrast, the deep learning model FastText-Crawl CNN obtained an F1-score of 0.898 and an accuracy of 91% on D2. Depression prevalence from P1 matched official statistics, validating the model for population-level analysis.

Rajaraman et al. [50] introduced a methodology to detect depressive tweets using various machine learning models. They evaluated TF-IDF Classifier, LSTM, Naïve Bayes, Linear Support Vector Machine (LSVM), and Logistic Regression on datasets: Sentiment 140, TWINT tweets, and Google Word2Vec embeddings. Preprocessing included tokenization, stemming, stop word removal, and vectorization. Performance metrics used were precision, recall, F1-score, support, and accuracy. LSTM had the highest accuracy of 99.52% in identifying depressive tweets, followed by TF-IDF and LSVM.

Alsagri and Ykhlef [6] developed a Depression Detection using Activity and Content Features (DDACF) model, combining user activity metrics and tweet content. Data preprocessing included tokenization, normalization, stemming, and a TF-IDF weighted document-term matrix. Integrated account features and activity measures served as input for DTs, Support Vector Machines (SVM) with linear and radial kernels, and Naïve Bayes (NB). A 10-fold cross-validation strategy was employed. The dataset included up to 3000 tweets per user from self-reported depressed and non-depressed individuals via Twitter's API. Feature engineering involved first-person pronouns, TF-IDF, information gain, sentiment analysis, and novel features like 'Dept-Sent' and 'Categoral.' Results showed that SVM had higher accuracy and stability compared to other models. The 'Dept-Sent' and 'Categoral' features improved all models except DT.

Govindasamy and Palanichamy [51] developed a ML framework to detect depression in social media users. Twitter data was collected and preprocessed, including tokenization, lemmatization, and stemming. Sentiment analysis was conducted using TextBlob to assess polarity and subjectivity. The dataset, with 1,000 and 3,000 tweets, was divided into 70-30 training and testing sets. Two classifiers, Naïve Bayes and NBTree, were employed to classify the tweets. The results indicated that both classifiers obtained an accuracy of 92.34% on the 1,000 tweet dataset and 97.31% on the 3,000 tweet dataset. Both models demonstrated high precision in detecting depressive tweets.

A supervised ML model was created by Hossain et al. [32] to predict depression from social media text. A total of 1,500 sentences were collected from Twitter, Facebook, and Instagram; 829 of these sentences were not depressed, and 669 of them were depressed. Text preprocessing included normalization, removal of null inputs, punctuation, and stop words, followed by tokenization and lemmatization. The dataset was split into 80% training and 20% testing sets. VADER's initial lack of success in sentiment analysis was resolved by experimenting with six machine learning classifiers (Multinomial Naïve Bayes, Linear Support Vector Classifier, kNN, RF, DT, and LR). With 95% accuracy, Multinomial Naïve Bayes and LR showed the best performance.

Chiong et al. [33] developed an ML framework for detecting depression in social media users, utilizing two distinct Twitter datasets: one explicitly labeled for depression created by Shen et al. [52] and another containing tweets with the term "depression" created by Eye [53]. A comprehensive feature set, derived from sentiment lexicons (SentiWordNet and SenticNet) and tweets characteristics was employed. The features were categorized into multiple groups to compare their effectiveness in depression detection. The models included standard classifiers like LR, SVM, DT, and MLP as well as ensemble methods such as Bagging Predictors (BP), RF, Adaptive Boosting (AB), and Gradient Boosting (GB). The models were evaluated using 10-fold cross-validation. Performance was evaluated using accuracy, precision, recall, and F1 score. SenticNet-based features outperformed SentiWordNet-based features, with further improvements observed through the integration of content-based and sentiment lexicon-derived features. Optimal performance was obtained when combining all sentiment lexicon and content-based features. The study

TABLE 1. Summary of depression detection using machine learning techniques.

Reference	Author	Year	Dataset	Technique
[43]	Islam et al.	2018	Facebook	SVM, DT, kNN, Ensemble (Boosted and Bagged trees, Subspace kNN)
[29]	Katchapakirin et al.	2018	Facebook	SVM, RF, DL
[30]	Aragón et al.	2019	eRisk 2017 and 2018	SVM
[3]	Fatima et al.	2019	Reddit	SVM, LR and MLP
[4]	Ding et al.	2020	Weibo	SVM with AdaBoost, SVM, RBF-NN, kNN
[5]	Skaik et al.	2021	Twitter	SVM, LR, RF, GBDT, XGBoost, DL
[33]	Chiong et al.	2021	Twitter [52], [53]	LR, SVM, DT, MLP, BP, RF, AB, GB
[56]	Aguilera et al.	2021	eRisk 2017 and 2018	OCC-kSS, gSC
[58]	Liaw et al.	2022	Twitter	LR, SVM, DT, RF, XGBoost
[62]	Adarsh at al.	2023	Reddit	Ensemble of kNN and SVM

found that while Eye's dataset exhibited higher accuracy, it struggled with recall due to class imbalance, which was effectively mitigated by ensemble models. Gradient Boosting emerged as the top-performing model, attaining over 98% accuracy on both datasets.

Chiong et al. [54] developed a methodology for detecting depression through social media text analysis using machine learning classifiers. Two publically available Twitter datasets, from Shen et al. [52] and Eye [53], were used for training and evaluation. Preprocessing included punctuation, number, and stop word removal, along with spelling, elongated words, and negative word correction. Lemmatization and POS tagging were used to improve text representation. The bag-of-Words model was applied to n-grams that were limited to trigrams to extract features. To address class imbalance, dynamic over- and under-sampling was applied. A set of machine learning classifiers LR, RF, SVM, MLP, DT, Adaptive Boosting, Bagging Predictors, and Gradient Boosting were evaluated using 10-fold cross-validation. Model robustness was assessed through external testing on Twitter, Reddit, and Facebook datasets. LR achieved the highest accuracy of 99.8% on Eye's dataset, while SVM with a linear kernel demonstrated superior performance on Shen's dataset, also attaining an accuracy of 99.8%. Interestingly, excluding the words "depression" and "diagnose" decreased accuracy on training datasets but improved performance on external datasets, suggesting enhanced model generalizability.

Chatterjee et al. [55] presented a comprehensive method to identify depression in social media users. Data was obtained from Twitter and Facebook using their respective APIs. Preprocessing included Tokenization, elimination of stop words, and constructing a Term Document Matrix (TDM) using TF-IDF vectorization. The dataset consisted of 7146 Facebook comments and an additional 8222 terms from a sentiment dictionary. Using this dictionary, each comment was categorized as positive, negative, or neutral. The NB model was used to classify the comments into depressed and non-depressed groups, with depression levels determined by the frequency of negative words. The Naïve Bayes classifier obtained an accuracy rate of 76.6%, a recall of 0.31, and a precision of 0.86.

Aguilera et al. [56] introduced a novel method for identifying mental illnesses like depression and anorexia in social media posts using One-Class Classification (OCC) such as One-Class Classification k-Strongest Strengths (OCC-kSS) and Global Strength Classifier (gSC). The methodology unifies every user post into a single document and determines the mass by calculating the frequency of domain-specific terms that are taken from a Data-based (DB) or Knowledge Based (KB) lexicon. Standard medical dictionaries are used as the KB lexicon, whereas training data is utilized to build the DB lexicon, which contains terms that people with the disorder frequently use. These masses influence document relevance, facilitating classification. Experiments carried out on datasets from the eRisk shared tasks show that gSC, with KB and DB lexicons, regularly outperforms OCC techniques and classic binary classifiers, obtaining noteworthy F1 scores on several datasets. Due to its resilience to less training data, gSC has the potential to be used in real-life scenarios with sparse labeled data.

In order to more accurately identify depression on social media, Titla-Tlatelpa et al. [57] created classifiers that are specific to user profiles. They hypothesized that users exhibiting comparable traits would express depression in a similar way, which led to the development of specialized classifiers based on attributes like age and gender. Using a single document representation of each user's post, they train classifiers on these representations. Instead of the more conventional BoW model, they used a dual word-sentiment representation known as Bag of Polarities (BoP), which captures both positive and negative contexts of words. Reddit and Twitter datasets were used to assess the BoP representation, and specialized lexicons were used to automatically infer user traits from the text. The classifiers, which were tuned

using the GridSearch method, demonstrated a considerable improvement over the single classifier approaches of the baseline when BoP was combined with trait-specific classifiers. Particularly, gender-based classifiers performed the best, obtaining F1 scores of 0.89 for Twitter and 0.71 for Reddit.

Liaw et al. [58] presented a novel method that builds extensive datasets and uses a variety of feature sets to identify depression in Twitter users. They generated two datasets, one for individuals with a diagnosis of depression and the other for a control group, using Twitter's API. Original tweets, retweets, quote tweets, replies, liked tweets, user details, and accounts followed were among the data. Text preprocessing included lowering text's case, removing HTML characters, punctuation, extra whitespaces, hyperlinks, mentions, newlines, and emoticons. Non-English tweets were excluded. Textual, user activity, network, and engagement features were included in feature engineering. Two feature sets were constructed: one replicating prior work, and another incorporating additional engagement and network features. A comparative analysis of LR, Linear SVM, DT, RF, and XGBoost models was conducted using accuracy, precision, recall, and F1-score. XGBoost was the top performer, achieving an F1-score of 0.8205. Notably, user engagement features, particularly depression keywords in liked tweets, significantly contributed to model performance.

A novel method for real-time depression diagnosis utilizing streaming data from Twitter's API was presented by Angskun et al. [59]. Data is extracted, transformed, and loaded into a Hadoop cluster by the system, which serves as the processing and storage infrastructure. Sentiment analysis of Tweets and demographic characteristics are used to categorize depression levels. The system has four user modes: individual, parent, advisor, and employer. After appropriate permissions are obtained, depression levels can be monitored using Twitter IDs. Data was acquired from three sources: PHQ-9 scores, personal information questionnaires, and Twitter API. This data is preprocessed, including extracting sentiment attributes and translating non-English words. Sentiment scores are then derived from the WordNet database. Several ML models such as SVM, NB, DT, RF, and DL techniques are evaluated using a test set of 192 Twitter users with various depression levels. Feature selection methods like RF, SVM-RFE, and ANOVA enhance the model's performance, which is benchmarked using metrics like accuracy, precision, and F-measure. The results indicate that RF achieves the highest accuracy for depression detection.

Vasha et al. [7] employed data mining techniques to predict depressive content within Bangla language social media data collected from Facebook. A dataset of 10,000 posts and comments was divided into 80% training and 20% testing sets. TF-IDF was used for feature extraction, converting text data into numeric values for machine learning models. Various classifiers such as RF, LR, DT, SVM, kNN, and Multinomial NB were applied to predict depressive content. SVM demonstrated superior performance, achieving a precision rate of 0.77 and an F1-score of 0.78.

An ML method to detect depression in Weibo users was presented by Guo et al. [60]. A dataset of posts from the depressed and normal groups was collected. Preprocessing was applied to the collected dataset, such as removing unnecessary text and using Jieba for Chinese word segmentation. The study added language related to depression to the DUT-SL sentiment lexicon. A number of ML models, such as LR, kNN, DTs, RF, SVMs, and ensemble approaches like XGBoost and LightGBM, were used. The WU3D dataset and a curated dataset were both used to assess the models. With the curated dataset, Soft Voting, an ensemble learning technique, achieved the highest accuracy of 94.3% and an F1 score of 0.94. LightGBM demonstrated remarkable performance on WU3D, with an accuracy of 96.1% and an F1 score of 0.93.

Aragón et al. [61] introduced a novel approach for detecting depression and anorexia among Reddit users by analyzing the fine-grained emotions embedded in their textual posts. The method relied on representing user-generated documents through a granular lens of sub-emotions, derived from the EmoLEX lexicon. Words associated with broad emotions were grouped into clusters using Affinity Propagation, and each cluster, or sub-emotion, was represented by a centroid vector. These sub-emotions were used to mask the original text, substituting each word with its closest sub-emotion label, allowing for a nuanced representation of the user's emotional state. The BoSE and Δ-BoSE representations were developed, with the latter incorporating temporal dynamics. The methodology was tested on the eRisk 2018 dataset, demonstrating the superior performance of BoSE and Δ-BoSE models over traditional BoW and deep learning models, achieving F1 scores of 0.63 and 0.53 respectively.

A methodology for identifying depression in Arabic social media texts was presented by Sabaneh et al. [8]. The preprocessing step consisted of removing any irrelevant information, normalizing user posts, and translating from Arabic to English using OpenAI GPT-3.5-turbo. This was followed by extracting medical concepts using the UMLS database. Term weighting techniques, such as BoW and TF-IDF were used to create feature vectors. Five ML classifiers RF, NB, LR, SVM, and Stochastic Gradient Descent (SGD) were used. The BoW based RF model achieved the best performance with an accuracy of 80.2%, precision at 0.82, recall at 0.79, and F1 score at 0.8.

Adarsh et al. [62] introduced an ensemble-based technique for detecting suicidal thoughts in Reddit users. A Reddit API was used to retrieve the dataset. It was then preprocessed using a Neural Machine Translation (NMT) approach to remove noise. A one-shot decision strategy was used to address the problem of class imbalance. Term frequency and cluster alignment were identified by semantic network analysis employing TF-IDF vectorization. An ensemble of

SVM and kNN was applied to the pre-filtered data for user classification. The proposed model obtained an accuracy of 98%, precision of 0.96, recall, and F1 score of 0.97.

Helmi et al. [9] introduced an approach for detecting depression in Arabic social media data. Data was collected via the Twitter API. To tackle the language difficulties posed by social media text, preprocessing measures were employed, which included handling slang, abbreviations, and emoticons. TF-IDF and BoW were used to extract features. Supervised classifiers, including SVM, RF, LR, and LightGBM were used for text classification. Language-specific adaptations were incorporated to enhance model performance across English and Arabic text. Experiments on the Arabic depression corpus demonstrated the minimal impact of feature selection, with RBF-SVM obtaining an F1 score of 0.966 using TF-IDF and LR achieving an F1 score of 0.964 using BoW.

B. DEEP LEARNING TECHNIQUES

A brief overview of select studies regarding depression detection through deep learning methods is provided in Table 2.

Orabi et al. [13] introduced an advanced Neural Network (NN) architecture for depression detection from social media posts, using optimized word embeddings. Data was obtained from the CLPsych 2015 shared task and Bell Let's Talk datasets. Preprocessing involved removing retweets, URLs, mentions, and non-alphanumeric characters, while keeping specific pronouns. Word encoding utilized sequence indices, one-hot encoding, and padding/truncation for normalization. Word embeddings were generated using pre-trained Word2Vec (Skip-gram and Common Bag of Words (CBOW)) and random embeddings. Using both labeled and unlabeled data, multi-task deep learning and averaging were used to optimize word embeddings. CNN-based models outperformed RNN-based model, with CNNWithMax achieving an accuracy of 87.9% on the CLPsych 2015 dataset. The model demonstrated strong generalization on the Bell Let's Talk dataset.

Song et al. [63] used the Reddit Self-reported Depression Diagnosis (RSDD) dataset to develop and test a depression detection model. The dataset was divided into training, validation, and testing sets, with about 35,000 control users and 3,070 diagnosed patients in each. The task involved classifying users as depressed or control based on their posts. A post-level attention mechanism and four feature networks in accordance with psychological theories were combined to create the Feature Attention Network (FAN). These networks focused on depressive symptoms, sentiments, ruminative thinking, and writing style. The performance of the model was evaluated against several baselines and achieved a balanced performance with an F1 score of 0.56, showing the effectiveness of the attention mechanism in prioritizing relevant posts.

Cong et al. [34] proposed a novel DL architecture, X-A-BiLSTM, to enhance depression detection in social media posts. The architecture integrates XGBoost with an Attention-based Bidirectional Long Short-Term Memory (BiLSTM) network for data processing and classification. XGBoost initially filters data, directly outputting negative samples. Positive samples undergo further processing by the Attention-BiLSTM network, which employs an attention mechanism to capture word-level contextual importance. The RSDD dataset was used for training and evaluation. Preprocessing, including word embedding and BiLSTM-based sequence modeling, preceded the application of the attention mechanism to assign weights to word features. The combined model effectively addressed class imbalance. Experimental results demonstrated X-A-BiLSTM's superior performance over traditional models, achieving significant improvements in precision by 16.9%, recall by 17.8%, and F1-score by 17.6% for diagnosed users.

Uddin et al. [10] addressed the challenge of detecting depression in Bangla tweets through a Long Short-Term Memory-Recurrent Neural Network (LSTM-RNN) model. A custom dataset of 5,000 Bangla tweets was collected and preprocessed to retain alphanumeric characters, punctuation, and spaces. Manual labeling categorized tweets into depressive, non-depressive, ambiguous, and incomplete. The dataset was divided into 80% training, 10% validation, and 10% testing sets. Hyperparameters were tuned in three steps: LSTM Size, Batch Size and Number of Epochs, and Number of LSTM Layers. Hyperparameter tuning identified an optimal LSTM size of 128, with a batch size of 25 and 10 epochs demonstrating superior performance. A five-layer LSTM architecture achieved the highest accuracy of 86.3%. The study highlights how important hyperparameter adjustment is, especially when working with small datasets.

A Gated Recurrent Unit (GRU) model was presented by Uddin et al. [64] for detecting depression in Bangla social media data. A dataset comprising 5,000 Bangla tweets was created. It was supplemented by 210 manually gathered phrases expressing depression. During preprocessing, non-compliant characters were removed and a access list of Bangla characters was created. Tweets were manually labeled into four categories: incomplete, ambiguous, depressive, and non-depressive. A balanced dataset was subsequently formed and stratified to mitigate the effects of its small size during training. For training the GRU model, the dataset was divided into 80% training, 10% validation, and 10% testing subsets. Hyper-parameter tuning was performed in three steps: adjusting the GRU size, batch size with the number of epochs, and the number of GRU layers. The optimal GRU size was found to be 512, batch size of 5, 3 GRU layers, and three epochs yielding an accuracy of 75.7%. The model demonstrated potential for detecting depression in Bangla text, despite the limited dataset.

Gamaarachchige et al. [14] used social media data from the CLPsych 2015 shared task to introduce a unique multi-task learning method for detecting mental diseases, specifically PTSD and depression. A multi-channel CNN architecture,

incorporating three kernel sizes, was employed, outperforming traditional RNNs. Data preprocessing involved removing URLs, @mentions, hashtags, emoticons, and stopwords while retaining key features like first-person pronouns due to their correlation with mental illness. Vocabulary generation was optimized using a TF-IDF based approach, resulting in a dictionary that enhanced model performance by capturing critical term relationships. The model was trained on user tweet sequences, incorporating emotion detection and demographic features within a multi-task, multi-channel, multi-input architecture. This method outperformed RNN and SVM models with remarkable accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) values of 87.59%, 86.61%, 82.05%, 83.81%, and 92.24%, respectively.

Sekulić et al. [65] explored depression detection in social media using the SMHD dataset and a Hierarchical Attention Network (HAN) model, originally designed for document classification. The HAN architecture, employing GRU-based sequence encoders and attention mechanisms, represented users as sequences of social media posts, facilitating a nuanced analysis of textual content. The model outperformed baseline logistic regression and linear SVM classifiers, achieving an F1-score of 68.28. The model matched established depression markers by accurately capturing pertinent linguistic elements such as affective language and personal pronouns. The model's ability to recognize important words and phrases associated with depression was further validated by attention weight analysis, underscoring its potential for precise mental health detection using social media.

Shah et al. [35] used the CLEF eRisk 2017 dataset to study early social media depression detection. Based on self-reported diagnoses of depression, the dataset was split into depressed and control groups. Handcrafted metadata and a variety of feature sets, such as Word2VecEmbed, FastTextEmbed, GloVeEmbed, and TrainableEmbed, were extracted. A BiLSTM model was used for the classification, and the architecture was adjusted according to initial test results. The model was trained in two stages: first, on a manually curated dataset of depressive and non-depressive posts, and then on the CLEF eRisk 2017 dataset, with an emphasis on detecting depression early. In order to identify potential depressed posts, the study developed the concept of a risk window. Early Risk Detection Error (ERDE), Latency, and Latency-weighted F1 were the evaluation metrics. At a risk window of 23, the Word2VecEmbed+Meta feature set obtained the highest F1-score of 0.81. Even with the encouraging outcomes, the study recognized the difficulty in reducing detection times and recommended additional research to deal with this problem.

Yadav et al. [36] presented a novel Figurative Language-enabled Multi-task Learning (FiLaMTL) framework for detecting depressive symptoms in social media posts. The framework leverages a BERT-based model, integrating multi-task learning with symptom identification as the primary task and figurative language detection as the auxiliary task. Soft-parameter sharing enhances primary task performance by exploiting information from figurative language usage. The D2S dataset, developed by the authors and comprising 3,738 symptom-categorized depressive tweets, was employed for experimentation. FiLaMTL outperformed state-of-the-art models, achieving F1-scores of 75.03% and 75.55% for symptom identification and figurative language detection, respectively. Fine-tuned for depression detection, FiLaMTL attained accuracies of 97.44% on the D2S dataset and 70.79% on the CLPsych dataset.

A Multimodal Topic-Enriched Auxiliary Learning (MTAL) approach was presented by An et al. [17] for the detection of depression. It integrates textual and visual inputs to improve prediction accuracy. The framework comprised a primary task of depression detection and two auxiliary tasks of textual and visual topic modeling. The primary task involves encoding text and image pairs using BERT and VGG models respectively to predict depression levels. Texts and images are encoded into feature vectors, which are then fused and processed through an LSTM network. Auxiliary tasks leverage a modality-agnostic topic model to generate topic distributions from both textual and visual data. The topic representations from these auxiliary tasks are combined to support the primary task. Evaluated on a balanced dataset, MTAL achieved an accuracy, F1-score, precision, and recall of 84.2%, significantly surpassing baseline models.

Lin et al. [15] introduced SenseMood, a system for detecting and analyzing depression in social media users through a deep visual-textual multimodal learning approach. Integrating visual and textual data, the system surpassed previous methodologies in depression detection accuracy. BERT was used for textual features and CNN for visual features for processing Twitter data, which included user tweets and related photographs. A binary classifier, trained on a fusion of these features, distinguished between depressed and non-depressed users. SenseMood outperformed previous models with accuracy of 88.4%, F1-score of 0.936, precision of 0.903, and recall of 0.870.

A method for early depression identification using social media data from the eRisk 2018 dataset was presented by Bucur et al. [66]. The user's posts were divided into ten consecutive segments to replicate real-time monitoring. Text preprocessing involved lowercasing, removal of punctuation, stopwords, numbers, and URLs, followed by stemming. The preprocessed text underwent Latent Semantic Indexing (LSI) to extract topic embeddings. These embeddings were then fed into a fully connected neural network with Leaky ReLU activation and Dropout regularization. An innovative confidence score mechanism was implemented to evaluate the reliability of predictions, allowing delayed classification choices. The model attained an ERDE5 of 10.24%, an ERDE50 of 7.74%, an F1-score of 0.30, a precision of 0.25, and a recall of 0.38. The study highlighted the challenges of early detection, particularly in balancing the need for timely identification against the risk of false positives.

Lei et al. [67] proposed an end-to-end method for detecting depression specifically among teenagers on social media platforms. A multi-modal, multi-lingual, multi-attribute classification (M3) model was employed to initially classify users into under-18 and over-18 age groups. The M3 model integrated a DenseNet-based classifier for visual data and a two-stack bidirectional character-level LSTM network for textual analysis. Depression was detected using an LSTM network that focused on the temporal sequence of recent posts. The eRisk2017 dataset, comprising 886 Reddit posts, was utilized for training. The model achieved a precision of 0.93 for ''not depressed'' and recall of 0.83 for ''depressed,'' but also exhibited a high false positive rate of 55.50%. While this oversensitivity may facilitate early interventions, the model's performance necessitates further refinement.

The difficulties in determining depression intensity from noisy and unstructured social media data were addressed by Ghosh et al. [68]. They used a novel method combining dense labeling and feature engineering to overcome the shortcomings of the available datasets and the sparsity of features related to depression. Emotion polarity and semantic analysis were used to relabel an existing dataset with different levels of depression intensity. They extracted various features including emotional, topical, online behavior, and user-level attributes. An LSTM network predicted depression intensity, outperforming baseline models with an MSE of 1.42. The model also exhibited strong performance in binary classification with an accuracy of 87.14%.

Chiu et al. [18] introduced a multimodal methodology to identify depression in Instagram users. A dataset was created by gathering data from social media posts containing hashtags related to depression and analyzing user profiles. The collected data was then subjected to preprocessing techniques to improve its quality. CNNs and Word2Vec were utilized to extract features from images and texts, respectively, and Random Forest was used to classify behavior. A final depression score was obtained by a multimodal system that combined these variables using Adaboost-based score fusion. The system underwent validation with a sample size of 520 Instagram users. It obtained an accuracy rate of 84.2% for image classification using AlexNet, 78.5% for text classification using BiLSTM, and 90.08% for behavior classification. The multimodal model achieved a precision of 0.895, a recall of 0.782, and an F1 score of 0.835.

An advanced framework for automatic depression detection called DepressionNet was proposed by Zogan et al. [37]. It integrates user behavior and social media post history. To address the challenges of processing vast and noisy user-generated content, an extractive-abstractive summarization technique was employed. The BERT-BART model, combining BERT for sentence embedding and BART for text condensation, generated concise summaries. Summarized content was then processed through a CNN and Bidirectional Gated Recurrent Unit (BiGRU) with an attention mechanism to extract relevant textual features. User behavior was modeled through a feature set encompassing social interactions, emotional indicators, and domain-specific terms, extracted using LDA. A fusion of summarized post content and behavioral features formed a comprehensive representation of the user's mental state. DepressionNet achieved precision, recall, F1-score, and accuracy of 0.909, 0.904, 0.912, and 0.901, respectively.

Uban et al. [69] introduced a framework for identifying depression by analyzing linguistic markers in social media data. The study aimed to find linguistic patterns associated with depression through the analysis of word sequences, stopwords, and pronouns. The LIWC and NRC lexicons were employed to capture emotional states and cognitive styles. The model underwent training using datasets of social media posts, with depression proving to be the most difficult condition to detect. The HAN effectively captured subtle verbal cues of depression, outperforming CNNs, BiLSTMs, and transformers. The HAN model achieved an F1-score of 0.45 and an AUC of 0.83 in the task of detecting depression.

Twitter data was used by Rizwan et al. [70] to propose a deep transfer learning method for depression intensity classification. A dataset of 73,355 tweets on depression was classified using four small transformer-based language models: Electra Small Generator (ESG), Electra Small Discriminator (ESD), XtremeDistil-L6 (XDL), and Albert Base V2 (ABV). The preprocessing stage involved the removal of non-ASCII characters, hashtags, and URLs. This was followed by sentiment annotation using VADER and TextBlob. Twitter posts were categorized into three groups based on ICD-10 criteria: mild, moderate, and severe depression. Each model was fine-tuned with a softmax classification layer, and hyperparameters were optimized. ESG achieved an F1-score of 0.89, precision of 0.89, recall of 0.89, accuracy of 92%, and an average epoch training period of 130 seconds, demonstrating an optimal trade-off between training time and performance. This research demonstrates that compact transformer models compare with larger models such as DistilBERT to have the same depression severity classification performance but significantly less computation resources.

Mann et al. [71] proposed a novel approach for identifying depression on social media by employing a multiple-instance learning (MIL) framework. In contrast to conventional supervised learning, Multiple Instance Learning (MIL) considers the input as a bag of instances, enabling fine-grained classification over a period of time. The Demil (Depression with Multiple-Instance Learning) framework considers users as bags of posts, where each post is represented as a vector of features. To address accurate labeling, Demil introduces a three-step method: transforming instances, pooling them, and final transformation for classification. The pooling function adapts to the time-dependent characteristics of posts. Both transformer-based and LSTM architectures are used in the study. Sequential bias is introduced via positional embeddings in the transformer. By avoiding label duplication

between posts, this method avoids false correlations. An F1 score of 0.8, precision of 0.76, and recall of 0.83 was attained by the LSTM model.

A HAN was used by Han et al. [72] to detect depression among Twitter users based on both tweet content and metaphorical concept mappings (MCMs). A collection of tweets and related MCMs, with embeddings produced by a pre-trained BERT model, represent each user. The architecture includes layers of attention-based encoders for tweets and MCMs, followed by Feed-forward Neural Networks (FNNs) activated by ReLU and Softmax functions for classification. The attention weights assigned to the tweet and MCM features are used to train the HAN model, which classifies individuals as either depressed or not. A publicly available Twitter dataset was used for evaluation, and the method performed better in detecting depression than previous baselines, with an average F1 score of 0.972 across five random testing sets. The model's ability to explain its predictions was enhanced by the use of attention mechanisms, highlighting the importance of specific traits in the prediction of depression.

Zogan et al. [73] proposed a hybrid DL model to detect depression in social media users. The model integrated an MLP with HAN to manage the complexity of user posts and a range of online behaviors. The HAN model was utilized to encode the user tweets at both the word and tweet levels and the MLP model handled the user behavior data. Based on the combined outputs, a sigmoid activation layer was employed to classify users as either depressed or not depressed. Various aspects of the user profile were extracted, such as metrics for social interaction, analysis of sentiment conveyed by emojis, distribution of topics, and identification of phrases related to depression. Data preprocessing involved removing stop words, non-ASCII characters, and users with sparse posting histories. The model attained an accuracy of 89.5%, a precision of 0.902, a recall of 0.892, and an F1-score of 0.893 on a Twitter dataset, exhibiting superior performance compared to baseline models.

Kour et al. [11] developed an architecture for depression detection in social media users through a hybrid CNN-biLSTM model. The framework comprised data extraction, preprocessing, feature extraction, classification, and evaluation. Preprocessing involved normalization, tokenization, and stop word removal. Word embeddings converted text into numerical vectors, allowing the CNN to capture spatial features while the biLSTM processed sequential information. The model underwent training and evaluation using a Twitter dataset consisting of both depressed and non-depressed users. It outperformed traditional CNN and RNN models with an accuracy of 94.28%, precision of 0.9699, F1-score of 0.9478, specificity of 0.9635, and AUC of 0.9543.

Yang et al. [74] introduced a novel approach for detecting depression and stress in social media posts by leveraging advanced natural language processing techniques. A Context Aware Post (CAP) encoder, based on MentalRoBERTa, was employed to process text, incorporating a knowledge infusion process using the COMET model to extract sentence-level mental state features. The model was trained to classify posts as indicative of depression or stress. The CAP encoder generated deep bidirectional word embeddings, capturing contextual information crucial for accurate classification. Model surpasses baselines in detecting depression and stress on social media, achieving Precision, Recall, and F1-Score of 0.95 on the Depression_Mixed dataset.

Amanat et al. [75] introduced an LSTM-RNN-based approach for detecting depression in social media text. A Twitter dataset was preprocessed to remove noise, followed by tokenization, stemming, and lemmatization. One-Hot encoding transformed text data into binary features for LSTM-RNN input. Principal Component Analysis (PCA) was employed for dimensionality reduction. The model obtained an accuracy of 99.66%, precision, recall, and F1-score of 0.98 in 10-fold cross-validation, outperforming SVM and NB. This study highlights the advantage of LSTM-RNN for identifying depressive symptoms in textual data, with potential applications in real-time mental healthcare.

Nadeem et al. [76] introduced a comprehensive methodological approach for detecting depressive content in social media data. To address the limitations of keyword-based labeling, a manual annotation process guided by a mental health practitioner was employed, improving dataset reliability by considering both explicit and implicit mentions of depression. Data preprocessing involved removing duplicates, hyperlinks, hashtags, and special characters while retaining stop words for contextual analysis. Feature extraction utilized TF-IDF, N-grams for machine learning, and pre-trained embeddings like Word2Vec, FastText, and GloVe for deep learning. A hybrid deep learning model integrating LSTM, CNN, GRU, and a self-attention mechanism was developed. The model achieved accuracy and F1-score of 97.4% for binary classification and 0.829 for ternary classification.

Narayanan et al. [12] proposed a hybrid CNN-LSTM model for detecting depression in social media posts. A dataset of 1.58 million tweets, equally representing depressive and non-depressive content, was collected from Kaggle. Tokenization, stop word removal, and noise reduction were all part of preprocessing. Word embeddings were produced by Word2Vec, and then a CNN layer was used to extract features. For sequential processing, the CNN output was fed into an LSTM layer. The model obtained an accuracy, precision, recall, and F1-score of 0.97. The results showed that the best results were achieved when using CNN for feature extraction and LSTM for sequence prediction to detect tweets related to depression.

Depression detection was explored by Poświata et al. [77] using competition data obtained from Reddit posts. It was annotated into three classes, ''not depression'', ''moderate'', and ''severe''. Data preprocessing included deduplication, which revealed a significant imbalance, with the severe class

underrepresented. To optimize model training, part of the dev set was merged with the train set. The authors fine-tuned pre-trained language models like BERT, RoBERTa, and XLNet. $\text{RoBERTa}_{\text{large}}$ achieved the best dev set performance with an F1-score of 0.605. They further developed a domain-specific model, DepRoBERTa, pre-trained on mental health-related subreddits, enhancing the F1-score to 0.616. The ensemble approach of $\text{RoBERTa}_{\text{large}}$ and DepRoBERTa outperformed each model with an F1-score of 0.637 on the dev set and 0.583 on the test set, securing first place in the competition.

The DEPTWEET dataset which is a collection of annotated tweets for diagnosing depression severity, was introduced by Kabir et al. [78]. The text was preprocessed by removing stopwords, unnecessary symbols, and lowercasing. Several classifiers such as SVM, BiLSTM, BERT, and DistilBERT were tested. In the BiLSTM model, a bidirectional layer was used with 64 units. A maximum token length of 128 was used to fine-tune BERT and DistilBERT. Categorical Cross Entropy was used as a loss function, and AdamW was used as an optimizer. DistilBERT outperformed traditional classifiers with a AUC of 0.86 for severe depression.

Wang et al. [79] introduced a multitask learning-based framework, FusionNet, for detecting depressed users on Weibo. The framework integrated text, social behavior, and image data to classify users' depression levels. Data collection and labeling were performed using web crawlers, followed by feature extraction using XLNet for text and statistical methods for social behavior and images. A Bi-GRU model with attention mechanisms was employed for classification. FusionNet's multitask learning strategy optimized both word vector and statistical feature classification tasks simultaneously. Evaluated on the newly created WU3D dataset, FusionNet achieved an F1-score of 0.97, precision of 0.99, recall of 0.96, and accuracy of 97.7%, outperforming SVM and Naïve Bayes.

Aragón et al. [80] introduced DisorBERT, a language model for detecting mental illnesses through social media analysis. They employed a two-stage domain adaptation process: fine-tuning BERT on Reddit to capture social media language, followed by adaptation to the mental health domain using relevant subreddit datasets. A depression lexicon guided the masking process, enhancing the model's focus on mental health-related terms. Evaluated on eRisk 2019-2020 datasets for detecting anorexia, depression, and self-harm, DisorBERT achieved an F1-score of 0.69, Precision of 0.56, and Recall of 0.89 for depression detection, outperforming baseline models.

A novel multimodal technique was introduced by Yadav et al. [81] to identify fine-grained depression symptoms from memes. The authors employed pre-trained RESNET-152 and BERT models to encode the visual and textual elements of memes, respectively. They implemented 2D adaptive average pooling to generate orthogonal features. Orthogonal regularization ensured non-redundant, de-correlated features, enhancing model expressiveness. A multimodal fusion strategy, combining textual and visual features through conditional adaptive gating, was employed for depression detection. The method was evaluated on the RESTORE dataset, a collection of 4,664 manually annotated depressive memes curated from Twitter and Reddit. After being fine-tuned with AdamW optimization, the model surpassed other methods and achieved impressive results, with an F1-score of 0.65, precision of 0.69, and recall of 0.607.

Zhang et al. [82] presented a model for detecting depression severity levels from social media posts. A post encoder and a sentiment-guided Transformer architecture were integrated, leveraging MentalRoBERTa for semantic encoding and SentiLARE for sentiment capture. Sentence-by-sentence encoding generated embeddings combining semantic and sentiment information. Multi-head self-attention and a Sentiment-guided Co-Attention module within the Transformer blocks were employed. Severity-aware contrastive loss differentiated closely-labeled samples, enhancing classification accuracy. Experiments on two public Reddit datasets demonstrated the model's superior performance, obtaining a Graded Precision of 0.93, a Graded Recall of 0.82, an FScore of 0.87 on DsD, and a Graded Precision of 0.88, a Graded Recall of 0.90, and FScore of 0.89 on DepSign.

A Hierarchical Convolutional Neural Network (HCN) model was presented by Zogan et al. [83] to analyze tweets for detecting depression in social media users. The dataset included tweets from users before and during the pandemic for comparative analysis. The HCN+ model used a two-channel CNN with an MLP to encode words, which improved feature extraction from user posts. User tweets were encoded as word embeddings and processed through CNN channels to capture context. The HCN+ model used attention mechanisms at both word and tweet levels to highlight relevant content. HCN+ outperformed BiGRU and CNN with an accuracy, F1-score, recall of 0.86, and precision of 0.87.

Zong et al. [84] presented an innovative approach for using social media analysis for early detection of depression. The Emotion Cause Detection (ECD) model, leveraging a pre-trained BERT language model, extracted contextualized representations from tweets. A recurrent feature encoder extracted fine-grained emotional, causal, and semantic features. These features were combined to predict a user's depression tendency. The ECD model handled depression detection as a binary classification problem. Evaluated on the eRisk2018 shared task one dataset, the model demonstrated good performance, achieving a precision of 0.78, recall of 0.47, and F1 score of 0.59, surpassing BiLSTM, BERT, and BioBERT baselines.

Ilias et al. [85] introduced a Multi Task Learning (MTL) approach for depression and stress detection in social media posts. Utilizing the Dreaddit dataset (3,553 posts) for stress detection and a dataset of 2,822 posts for depression detection. The posts were preprocessed using the BERT tokenizer.

TABLE 2. Summary of depression detection using deep learning techniques.

Reference	Author	Year	Dataset	Technique
[63]	Song et al.	2018	RSSD [113]	FAN
[14]	Gamaarachchig et al.	2019	CLPSych 2015 [112]	CNN
[65]	Sekulic et al.	2019	SMHD [114]	HAN
[17]	An et al.	2020	Twitter [110]	BERT, VGG, LSTM
[37]	Zogan et al.	2021	Twitter [52]	BERT, BART, CNN, BiGRU
[72]	Han et al.	2022	Twitter [52]	BERT, HAN
[79]	Wang et al.	2022	WU3D [79]	XLNet, BiGRU
[38]	Bucur et al.	2023	multiRedditDep [117], Twitter [110]	EmoBERTa, CLIP, Time2VecTransformer
[92]	Anshul et al.	2024	Twitter [52], COVID Tweets	LR, SVM, DT, RF, XGBoost
[19]	Zafar et al.	2024	multiRedditDep [117], Twitter [110]	EmoBERTa, CLIP, ViLBERT

Two MTL architectures were introduced: Double Encoders and Attention Fusion Network. The Double Encoders model utilized separate task-specific BERT layers, while the Attention Fusion Network integrated shared and task-specific layers. Experimental results demonstrated superior performance of both MTL models compared to single-task learning and transfer learning methods. The Double Encoders model achieved a Precision of 0.77, Recall of 0.86, F1 Score of 0.81, Accuracy of 0.79, and Specificity of 0.7. The Attention Fusion Network achieved a Precision of 0.82, Recall of 0.85, F1 Score of 0.83, Accuracy of 0.82, and Specificity of 0.8.

A novel methodology to assess the severity of depression based on social media posts was proposed by Liu et al. [86]. A Text Graph Convolutional Network (Text GCN) was employed to encode individual tweets into post embeddings, representing the corpus as heterogeneous graphs. To represent the emotional range of historical tweets, the Plutchik Transformer processed text and emotional features. A Time-aware Long Short-Term Memory (T-LSTM) accounted for temporal irregularity between consecutive posts. The model, trained using ordinal regression loss, effectively prioritized users based on depression severity. Experiments on two Reddit datasets demonstrated the model's superiority over baselines, with TIDE achieving Graded Precision, Graded Recall, Graded FScore of 0.97 on the first dataset and Graded Precision of 0.95, Graded Recall of 0.89, Graded FScore of 0.92 on the second dataset.

Cai et al. [87] presented a novel method for depression detection among Weibo users. The DSTS approach utilized a process of extracting features from a multivariate time series to accurately capture the changes in depression symptoms across time. The study identified key signs of depression by using the criteria outlined in the DSM-5 and analyzing language used on social media. Tweets were then given scores based on a pre-trained multilingual embedding model. The time series features were created by sequentially combining these scores. Text was preprocessed by removing special characters and stopwords, and normalizing the contents. Time series data was padded for consistent input lengths. Multiple ML and DL classifiers were utilized, and among them, InceptionTime (IT) demonstrated the best results, attaining an accuracy of 92% and precision, recall, and F1 scores of 0.92.

A Multimodal Hierarchical Attention (MHA) model was presented by Li et al. [88] to detect symptoms of depression in social media users. The MHA model utilized an attention mechanism to focus on the most relevant features for the task, both within and between the modalities, using text, images, and auxiliary data. For image feature extraction, they employed the ResNet-18 pre-trained model on ImageNet, and for text, they applied TextCNN. In addition, auxiliary data includes user interactions and activity patterns. A distribution normalization technique was used to rectify imbalances in the frequency of image posting. Evaluated using Weibo data, the MHA model outperformed baseline models with an F1-score of 92.78% and an accuracy of 92.84%.

Bucur et al. [38] proposed a transformer-based method to detect depression in multimodal social media posts. They analyzed sequences of text and images using pre-trained models Contrastive Language-Image Pretraining (CLIP) and EmoBERTa which was followed by a cross-modal encoder and transformer with time2vec positional embeddings. The Time2VecTransformer, utilizing time2vec, improved the model's ability to detect temporal patterns in posts. Mean pooling and attention masking were used to handle noisy data and uneven distribution of content. The model was validated using datasets from Reddit and Twitter, and it attained state-of-the-art performance in multimodal settings. The Time2VecTransformer achieved an F1 score, Precision, Recall, and Accuracy of 0.931 on the Twitter dataset, demonstrating the effectiveness of integrating time-aware

components for identifying depression from social media posts.

Wu et al. [89] proposed Mood2Content, a framework for early depression risk detection in COVID-19 patients using Twitter data. The methodology focused on extracting and analyzing tweets posted prior to depression onset. Users who self-reported COVID-19 infection and subsequent depression were identified. Tweets were collected and preprocessed. A BERT-based encoder, COVID-Twitter-BERT-v2, was used to extract content and mood features. User representations were produced by combining mood and content encoders, self-attention mechanisms, and positional embeddings to incorporate temporal information. Prediction accuracy and mood-content alignment were balanced by the total loss function. The framework showed exceptional performance in the early diagnosis of depression, attaining an AUROC of 0.9317 and an AUPRC of 0.8116 on the DepCOV dataset.

Gopalakrishnan et al. [90] investigated the role of social media behavior in PPD detection using deep learning. Data from new mothers, collected through the Postpartum Depression Screening Scale (PDSS) and public Facebook profiles, underwent extensive preprocessing, including removing irrelevant content, tokenization, and limiting post length and frequency. Natural language processing and sentiment analysis were employed to identify PPD indicators. Key features were extracted using GloVe embeddings and psycholinguistic style analysis, followed by classification through neural networks, including an Attribute Selection Hybrid Network (ASHN). The ASHN achieved a Precision of 0.78, Recall of 0.72, Accuracy of 74.7%, and F1-Score of 0.75 in detecting depression-related posts.

A deep learning system called AIRCNN-LSTM was presented by Bhuvaneswari et al. [91] to Detect Depression (DD) in social media (SM) data. This model integrated an attention mechanism with CNNs and LSTM networks. Tokenization, stop word removal, stemming, and hashtag and URL removal were among the various steps in the data preprocessing procedure. The authors used the TFIDF-MIG scheme for feature weighting and the IEHA method for feature selection. The Dreaddit and Sentiment Tweets datasets were used to evaluate the model. In the Sentiment Tweets dataset, AIRCNN-LSTM outperformed C-LSTM, CNN, DT, and SVM with a classification accuracy (AC) of 98.3%. In the Dreaddit dataset, AIRCNN-LSTM classified depressive posts with 97.8% accuracy.

Anshul et al. [92] proposed a multimodal feature-based ensemble learning (MFEL) approach for detecting depression among social media users. The methodology integrated textual, visual, and user-specific features for a comprehensive user representation. Intrinsic features from tweet content and extrinsic features from external sources linked in the tweets were extracted. A novel CNN-based model, VNN, extracted visual features from images. An ensemble technique combined the outputs of various classifiers to improve detection accuracy. The proposed model was validated using the Tsinghua and COVID-19 datasets. The model achieved a precision of 0.93, recall of 0.9, F1 score of 0.91, and accuracy of 91.7% on the COVID-19 dataset.

Liu [93] proposed the Weighted Graph-enhanced RoBERTa Neural Network (WGRNN) for depression detection. Their framework integrates RoBERTa with graph convolutional layers to process mixed input, including text and social tags, for enhanced feature representation. Nodes comprising words from RoBERTa's vocabulary, document indices, and social tags are interconnected using pointwise mutual information to form a graph structure. The model utilizes a dual-input mechanism where retokenized graph inputs and RoBERTa-encoded vectors are passed through separate weighting modules to compute their importance, which are then concatenated and classified using a softmax layer. The model was evaluated on a fine-grained Chinese micro-blog dataset. The WGRNN using only text input achieved state-of-the-art results in both depression polarity prediction and cause detection tasks. When using mixed input, the model showed considerable performance gains, with F1-micro and F1-macro scores increasing by 7.7% and 6.6%, respectively.

Elmajali et al. [94] introduced an DL approach for classifying depression symptoms in Arabic tweets. Pre-trained models AraBERT and MARBERT were used to identify tweets corresponding to PHQ-9 depression symptoms and normal tweets. A dataset of 1,471 Arabic tweets underwent preprocessing, including URL, stop word, and irrelevant content removal. Text normalization and label conversion to numeric values were performed. The normal class was augmented using ChatGPT. Fine-tuned AraBERT and MARBERT models were employed for classification. AraBERT outperformed MARBERT, obtaining an accuracy of 99.3%, precision of 0.99, recall of 0.988, and F1 score of 0.989.

Abbas et al. [95] presented an innovative approach for detecting depression using a combination of BERT-based DL models and traditional ML classifiers. Data preprocessing involved tokenization, normalization, and stemming. Features were extracted by using BERT and BERT-RF models. Classifiers such as LR, RF, and kNN were evaluated. RF had the highest performance with BERT features, achieving 71% accuracy and 0.72 precision, recall, and F1 score. With BERT-RF features, all models improved significantly, with LR reaching a near-perfect score of 0.99 across all metrics, demonstrating BERT-RF's superiority in depression detection. This study highlights the potential of combining DL with traditional classifiers for improved depression detection from social media data.

Zafar et al. [19] introduced Multi-Explainable Temporal-Net (METN), a multimodal temporal network for depression detection using social media posts. The model combined a Temporal Convolutional Network (TCN) with a multimodal transformer framework. Pre-existing foundation models, such as CLIP and EmoBERTa were used for image processing and text encoding respectively. Cross-modal alignment was

done using ViLBERT. Attention maps made it easier to interpret the model. Reddit and Twitter datasets were used to test and train the METN model. It outperformed earlier methods, achieving F1 scores of 0.945 and 0.913 on these datasets, respectively.

TAM-SenticNet, a Neuro-Symbolic Artificial Intelligence (AI) framework for early depression identification, was introduced by Dou et al. [96]. To improve interpretability and accuracy, the system combined symbolic logic with a Time-Aware Affective Memories (TAM) network. Modules for semantic processing, integration, T-LSTM, and emotion processing were integrated into the TAM network. SenticNet added symbolic logic to the TAM network to enhance it. With a precision of 0.665, recall of 0.881, and F1 Score of 0.758, the framework outperformed other models in the evaluation conducted on the CLEF eRisk 2017 and 2018 datasets.

A novel CNN-BiLSTM Attention (CBA) model for depression detection in social media texts was proposed by Thekkekara et al. [97]. The model is a combination of CNN and BiLSTM networks. In order to concentrate on important language components, the model integrated an attention mechanism. Text normalization and irrelevant element removal were part of the preprocessing step. The model was trained and tested on the CLEF2017 eRisk dataset. It outperformed baseline models, including LSTM, BiLSTM, CNN, and CNN-LSTM. The model obtained an accuracy of 96.71%, F1 macro score of 0.89, AUC of 0.85, and precision of 0.89.

Liu et al. [98] introduced a hybrid model for detecting depression on social media platforms, combining a BERT-based model with an attention mechanism. The model architecture comprised a BERT-based layer for deep text comprehension and a fusion layer driven by attention scores for interpretability. The pre-trained ALBERT model extracted semantic representations from user-generated content. Textual features were passed through an embedding layer, transformed into vectors, and processed by the attention mechanism. The model was tested on the WU3D dataset, obtaining an F1-score, Precision, Recall of 0.94, and Accuracy of 94.7% outperforming other models.

A multi-channel CNN with individual attention (MCNN-IA) was proposed by Dalal et al. [99] to detect depression in social media data. The user documents were represented using GloVe vectors and a one-dimensional convolutional layer with four channels was used to extract local textual features. The most important features were extracted by global max pooling and then further tuned by specialized Bahdanau attention layers. The model was tested on the CLEF-eRisk 2017 dataset. It achieved an accuracy of 91%, precision of 0.65, recall of 0.76, and F1-score of 0.7. Despite the dataset imbalance, MCNN-IA showed competitive performance, with higher recall than a similar model without attention mechanisms. Although precision was lower, the attention layers identified important n-grams for depression classification.

A Twitter-based corpus called SetembroBR was presented by Santos et al. [100] to identify depression and anxiety disorders in the Portuguese Language. Data collection focused on pre-diagnosis/treatment tweets. Preprocessing involved removing URLs, hashtags, emoticons, and non-standard characters. Four models LR, LSTM, CNN, and BERT were evaluated. Features included TF-IDF counts, static word embeddings, and context-dependent BERT embeddings. The BERT model which was fine-tuned on Portuguese text, achieved the best results, with a Precision of 0.85, Recall of 0.77, and F1 score of 0.63 for depression detection. These experiments demonstrate the potential of the SetembroBR corpus for improving mental health prediction through social media analysis.

Farruque et al. [101] proposed a Deep Temporal model of User-level clinical Depression (TUD) for detecting depression using Twitter data. Using tweets annotated by clinicians, a mental health pre-trained BERT model was fine-tuned. The model's performance was enhanced through a semi-supervised learning framework, which involved harvesting additional tweets related to depression from users identified as depressed. Extracted features included Depression Symptoms Expression Vectors (DSEVs) and temporal patterns like Depression Recurrence Frequency (DRFS) and Inertia scores. The TUD model analyzed these features to identify depressive episodes and predict clinical depression. The model was tested across three datasets (CLPsych-2015-Users, IJCAI-2017-Ongoing-Users, and Mixed-Users), and achieved a precision of 0.76, recall of 0.66, and F1 score of 0.7 using all features on CLPsych-2015.

Zhang et al. [102] introduced a comprehensive approach for detecting depression in social media data, integrating medical domain knowledge into a deep learning model. A depression ontology model with prevalence data was developed to capture depression-related terminologies, symptoms, life events, and treatments. Entities linked to depression diagnoses in social media posts were identified. Temporal information and relevancy from medical knowledge and pre-trained language models enhanced these entities. A knowledge-aware DL model, incorporating an attention mechanism to refine the output was used for classification. The model was validated using the eRisk 2018 and 2020 datasets, demonstrating superior performance compared to traditional and state-of-the-art DL models, achieving an AUC of 0.824, F1 Score of 0.828, Precision of 0.836, and Recall of 0.824.

Singh et al. [103] proposed AP-CBWS, a hybrid meta-heuristic approach for detecting mental depression using social media data. The model integrated Black Widow Optimization (BWO) and Crow Search Algorithm (CSA) to optimize feature selection and improve the performance of various ML models, including CNN, LSTM, NNs, RF, and SVM. Population diversity and local optima issues in BWO were addressed by incorporating CSA. The model optimized parameters such as tree depth in RF, hidden neurons in

CNN, LSTM, and NN, and maximum epochs in SVM. The model was evaluated on four datasets and outperformed conventional models. The AP-CBWS-CNN-ENS ensemble achieved 96.6% accuracy, 0.98 precision, 0.96 recall, and 0.97 F1-score on the first dataset.

Wang et al. [104] introduced the Early Sensing Depression Model (ESDM) for early detection of depression in social media users. ESDM consists of a Classification with Partial Information (CPI) module and a Decision for Classification Moment (DCM) module. CPI employs BERT and LSTM models with an attention mechanism to highlight depression-related content from user posts. DCM determines whether to finalize predictions or continue processing more posts. The model is trained using an early detection loss function that balances timely predictions with accuracy. Evaluated on the eRisk-17 dataset, ESDM obtained an F1 score of 0.66 for early detection of depression and 0.72 for traditional depression detection, with ERDE50 and ERDE5 scores of 0.077 and 0.109, respectively.

Alhamed et al. [105] conducted a comprehensive study to detect linguistic changes in social media posts before and after a reported depression diagnosis. Data was collected from X (formerly known as Twitter). Classical ML models, such as SVM and RF, and advanced transformer-based models, including BERT, RoBERTa, and MentalBERT, were employed to classify posts. ML models were optimized through grid search and utilized Word2Vec embeddings. Results indicated that transformer-based models significantly outperformed classical approaches, with MentalBERT achieving accuracy, precision, recall, and F1 scores of 0.98. Large Language Models (LLMs), such as GPT-3, GPT-3.5, Google Bard, and Alpaca, were also experimented with but they exhibited inconsistent performance. The study concludes that transformer-based models effectively detect depression-related linguistic shifts, whereas LLMs need further refinement for reliable mental health tasks.

Malhotra et al. [39] proposed an explainable Transformer-based system for mental healthcare monitoring using social media data. Data preprocessing included cleaning noisy social media text, such as removing URLs, emoticons, and punctuation, followed by lemmatization and tokenization. Supervised training fine-tuned pre-trained Transformer models, including BERT, DistilBERT, and domain-specific models like MentalBERT and PsychBERT, using four mental health datasets. Explainability was achieved by interpreting model predictions using SHAP and LIME. Additionally, BERTopic was used to extract latent themes. MentalBERT achieved the highest F1-score of 0.885 on Dataset 1 and PHSBERT achieved the highest F1-score of 0.967 on Dataset 2.

Tejaswini et al. [106] proposed a hybrid deep learning approach for detecting depression from social media text. The methodology utilized a combination of CNN and LSTM architectures, leveraging the strengths of both global feature extraction and long-term dependency modeling. The datasets used included Reddit posts from prior work and Twitter data from a Kaggle repository. Preprocessing steps involved tokenization, lowercasing, removal of stop words, stemming, and lemmatization. FastText embeddings were used for text representation, which addressed vocabulary limitations by generating n-gram representations for out-of-vocabulary words. The processed text data was padded to ensure uniform input lengths. The hybrid model termed FCL, combined FastText embeddings with CNN and LSTM and used ReLU activation. Evaluations on two datasets demonstrated an accuracy of 87.0% for Dataset 1 and 88.0% for Dataset 2, surpassing baseline models utilizing Word2Vec and GloVe embeddings.

V. GENERIC ARCHITECTURE

This section outlines the generic architecture of a depression detection system, detailing each component involved in the process, from data acquisition to evaluation. Figure 4 provides a visual representation of the key stages in this architecture.

A. DATA ACQUISITION

Data acquisition is the primary step in the architecture, which involves collecting social media data, such as posts, comments, images, and metadata from platforms like Twitter, Reddit, Facebook, and Instagram. This data can be obtained using two primary methods:

- **Self-collection:** Data can be collected directly from social media websites using APIs like Twitter API or Reddit API. This approach allows data collection depending on certain criteria, such as time periods, user groups, or topics.
- **Public Datasets:** Several publicly available datasets are mentioned in Section VI. They contain labeled data for depression detection tasks. These datasets are already collected and annotated, making them a valuable resource for researchers.

B. DATA PREPROCESSING

Preprocessing is a key step after collecting data to ensure its quality and consistency for depression detection models. It helps make sure the data is clean, standardized, and ready for feature extraction and model input.

- **Text Cleaning:** Text cleaning involves removing irrelevant or noisy elements that do not provide useful information for depression detection models. This includes:
  - Irrelevant characters such as punctuation, special characters, or formatting symbols.
  - URLs, which are usually unrelated to the sentiment or meaning of the text.
  - Hashtags and emoticons, which may add noise unless they are specifically treated as features.

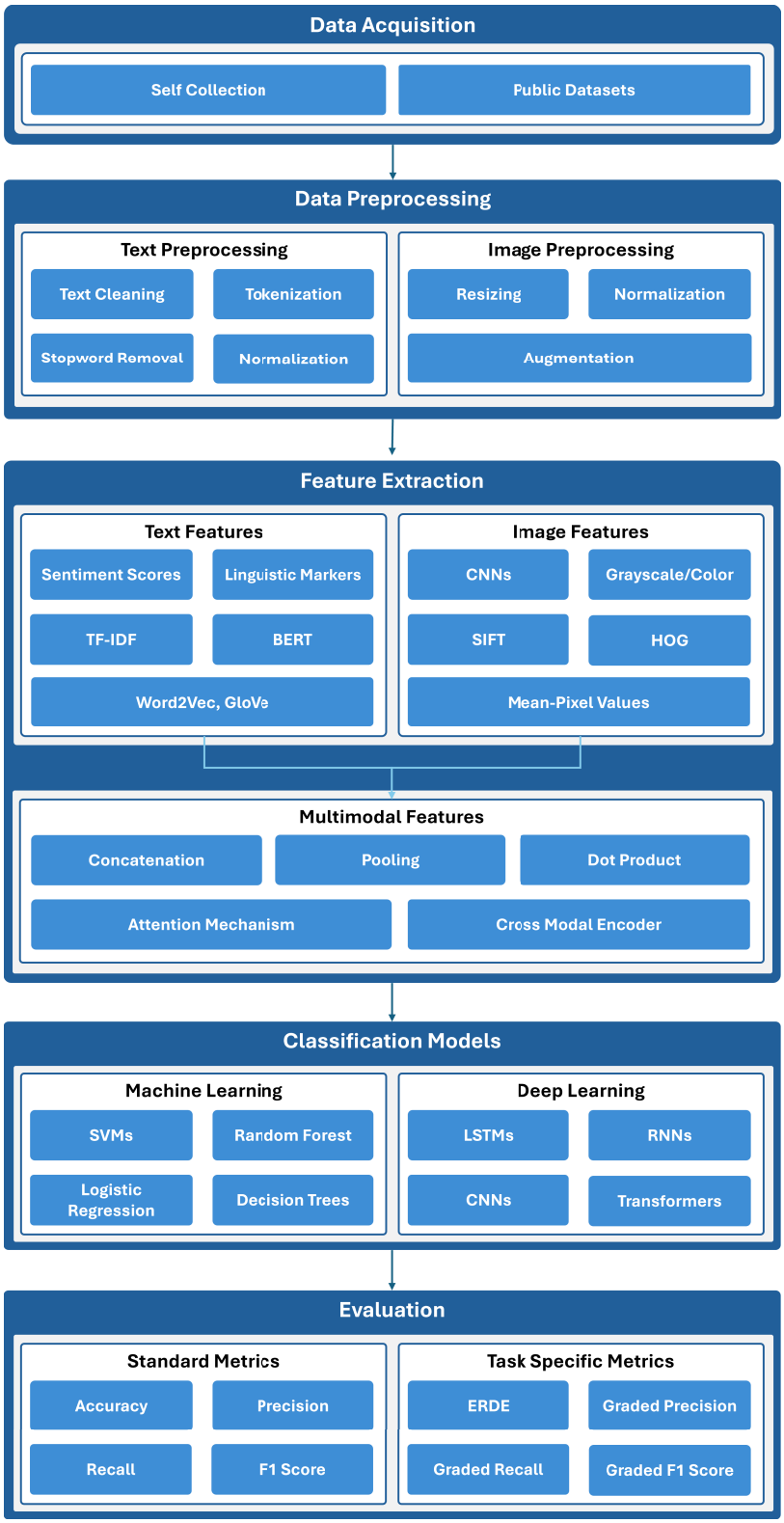

FIGURE 4. Generic architecture.

– Special symbols, such as mathematical notations or system-generated metadata, which are not useful for natural language understanding.

The aim is to strip down the text to its most relevant components, making it easier for models to focus on meaningful content [14], [66], [69], [78], [87].

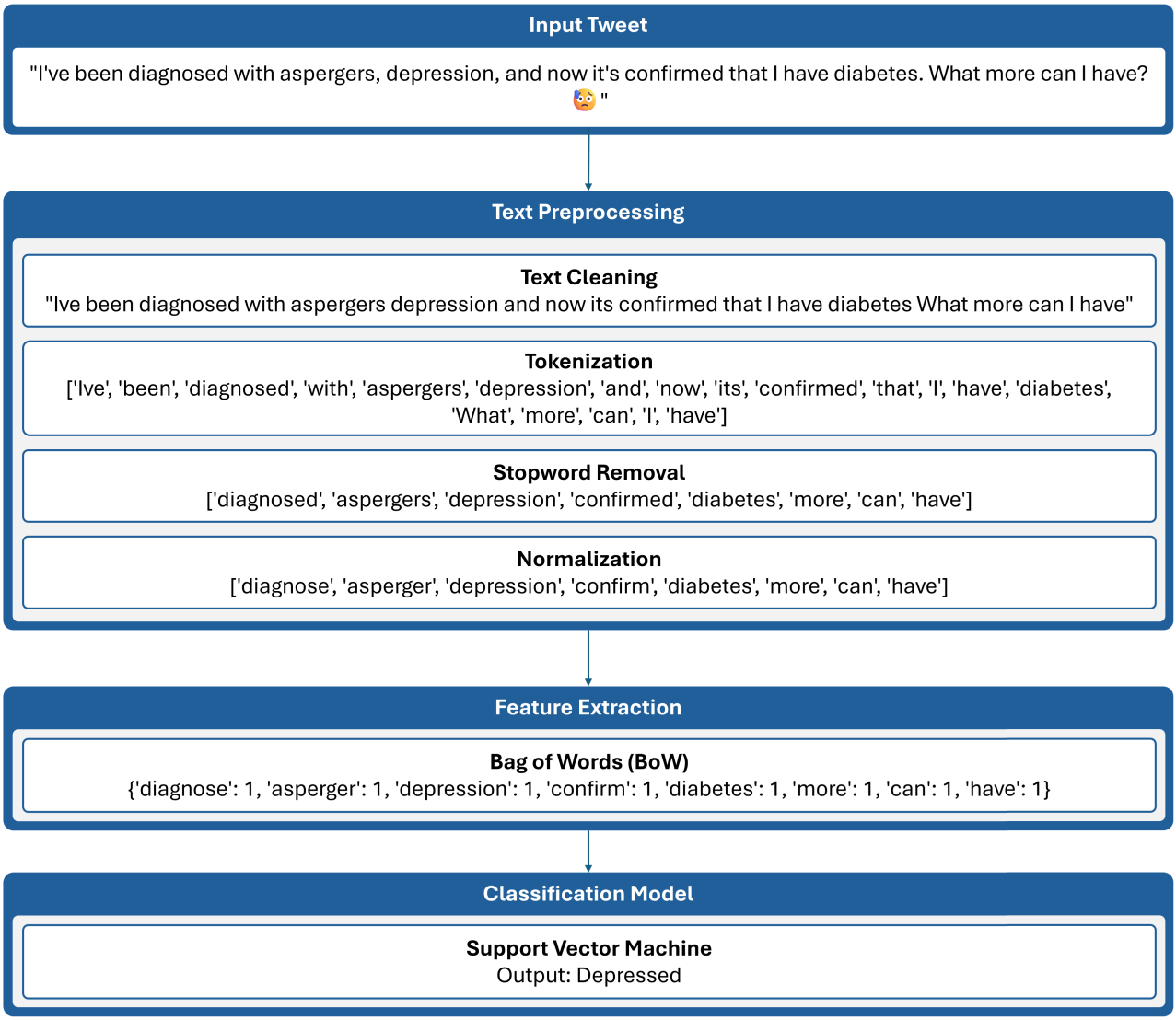

FIGURE 5. Illustration of tweet processing by a depression detection system.

- **Tokenization:** Tokenization refers to breaking down a sentence or document into smaller pieces, known as tokens. Each token typically represents a word, though in some cases it could be a subword or character. Tokenization helps in analyzing the structure and semantics of text:
  - Word-based tokenization splits the text into individual words.
  - Subword tokenization (like in Byte-Pair Encoding) is useful for rare or compound words.

  This process converts raw text into a structured format that can be analyzed by ML and DL models [2], [43], [50], [85], [95].
- **Stopword Removal:** Stopwords, such as articles, prepositions, or conjunctions (e.g., "the", "is", "in"), frequently occur in the text but offer little information for tasks like sentiment or depression detection. Removing them reduces dataset dimensionality and helps the model focus on more relevant terms. However, caution is needed to avoid removing emotionally significant words in certain contexts, especially in depression detection [14], [31], [49], [50], [69].
- **Normalization:** Normalization [6], [11], [13], [32], [94] involves making the text uniform to reduce the variability that can arise from different forms of the same word. It includes:
  - Converting text to lowercase so that words like "depression" and "Depression" are treated as the same.
  - Removing extra whitespace or irrelevant formatting to standardize input text.
  - Stemming/Lemmatization: Stemming converts words to their base form (e.g., "playing" to "play"), whereas lemmatization considers the

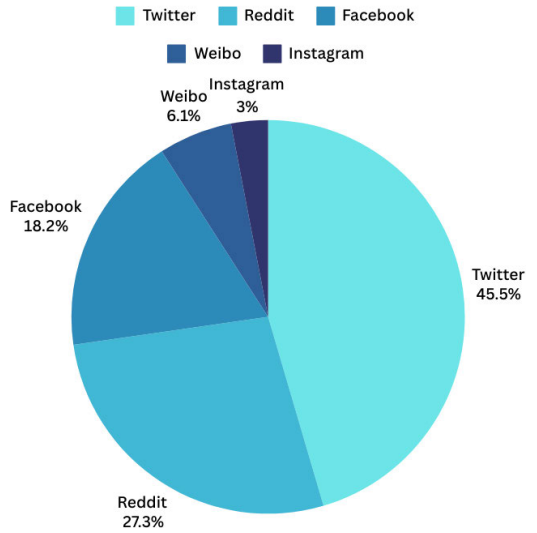

FIGURE 6. Data source for machine learning techniques.

word's morphological analysis (e.g., "better" to "good"). Both techniques reduce grammatical and derivational forms of words to improve model consistency.

- **Image Processing:** For multimodal approaches, where image data is involved alongside text, image preprocessing ensures that images are ready for input into deep learning models. Common steps include:
  - Resizing: Ensuring all images are of the same dimensions, as deep learning models often require fixed-size inputs.
  - Normalization: Standardizing pixel values (e.g., scaling pixel intensities to the range [0, 1] or normalizing them using a mean and standard deviation) to improve the stability of the learning process.
  - Augmentation: Applying transformations (e.g., rotation, flipping, cropping, color adjustment) to increase the variability of the dataset, which can improve the generalization ability of the model by simulating different real-world conditions.

  These steps help ensure that image inputs are consistent and feature-rich for neural networks [17], [18], [19], [38], [79], [88].

C. FEATURE EXTRACTION

Feature extraction involves converting raw data into meaningful representations:

- **Text-Based Features:** For machine and deep learning, features can include sentiment scores, linguistic markers (e.g., frequency of sad words), TF-IDF, or word embeddings like Word2Vec or GloVe. For deep learning features can be extracted using text encoders like BERT.
- **Image-Based Features:** For deep learning, features from images can be extracted using CNNs, grayscale features, mean pixel values of channels, Histogram of Oriented Gradients (HOG), Scale Invariant Feature Transform (SIFT), and color-based features.
- **Multimodal Features:** In multimodal approaches, features from both text and images are extracted separately. Later, these features can be fused using methods like concatenation, pooling, dot product, attention mechanisms, and cross-modal encoder to create a more comprehensive model.

D. CLASSIFICATION MODELS

After feature extraction, classification models are applied to detect depression:

- **Machine Learning Models:** Algorithms such as SVMs, RF, LR and DT are commonly used for text-based classification.
- **Deep Learning Models:** Text is processed using models like RNNs, LSTMs, and transformer-based models (e.g., BERT, RoBERTa), while CNNs and Visual Transformers (ViTs) are used for image data. These models are capable of detecting complex patterns in large datasets.

E. EVALUATION

Various evaluation metrics are employed to assess the performance of the classification models:

- **Standard Metrics:** Metrics like accuracy, precision, recall, F1-score, and AUC are widely used in both ML and DL models.
- **Task-Specific Metrics:** Depression detection requires specific metrics like ERDE, Graded Precision, Graded Recall, and Graded F1, which are tailored to the goal of early and accurate depression detection.

This architecture provides a robust framework for building depression detection systems using social media data, with the flexibility to integrate both text and image modalities and apply a range of ML and DL methods.

Figure 5 depicts a scenario demonstrating the processing of a tweet by a depression detection model.

VI. DATASETS

In recent years, a vast range of datasets have been created to facilitate the detection of depression from social media data. These datasets vary in size, language, source platform, and labeling methodology. Table 3 provides a summary of these datasets. Figures 6 and 7 illustrate the source of data used for ML and DL techniques, respectively.

- **eRisk 2018 task**: This dataset contains 214 users labeled as having depression and 1,493 control users. It was released as part of the CLEF eRisk challenge in 2018, aiming to develop methods for early risk detection of depression [107].
- **eRisk 2017 task**: The eRisk 2017 dataset includes 135 users with depression and 752 control users. Similar to the 2018 task, this dataset was part of the

TABLE 3. Summary of datasets for depression detection in social media.

Reference	Dataset	Source	Depressed Users	Control Users
[107]	eRisk 2018[1]	Reddit	214	1,493
[108]	eRisk 2017[2]	Reddit	135	752
[109]	eRisk 2020[3]	Reddit	145	618
[52]	TTDD[4]	Twitter	1,402	5,160
[110]	Gui et al.[5]	Twitter	1,402	5,160
[53]	EYE[6]	Twitter	2,314	8,000
[111]	Copper et al.	Twitter	441	5,728
[112]	CLPsych 2015[7]	Twitter	327	572
[113]	RSDD[8]	Reddit	9,210	107,274
[114]	SMHD[9]	Reddit	14,139	335,952
[36]	D2S	Twitter	3,738	8,417
[115]	DepSign	Reddit	9,586	3,081
[116]	Depression Severity[10]	Reddit	282 (Severe)	2,587 (Minimum)
[87]	SWDD[11]	Weibo	3,711	19,526
[117]	multiRedditDep	Reddit	6.6M posts	8.1M posts
[79]	WU3D[12]	Weibo	10,325	22,245
[100]	SetembroBR[13]	Twitter	1,684	11,788

CLEF eRisk challenge and focuses on early detection methodologies [108].

- **eRisk 2020 task**: The 2020 edition of the eRisk challenge introduced a new focus on self-harm, including 145 users identified with self-harm tendencies and 618 control users [109].
- **TTDD**: This dataset includes 1,402 users diagnosed with depression and 5,160 non-depressed users. It was developed by Shen et al. for training and evaluating depression detection models [52].
- Gui et al.: A multimodal Twitter dataset containing 1,402 depressed users and 5,160 controls, was curated by Gui et al. and is widely used for benchmarking in this domain [110].

[1]https://tec.citius.usc.es/ir/code/eRisk.html
[2]https://tec.citius.usc.es/ir/code/eRisk2017.html
[3]https://erisk.irlab.org/2020/eRisk2020.html
[4]https://github.com/sunlightsgy/MDDL
[5]https://drive.google.com/file/d/11ye00sHFY5re2NOBRKreg-tVbDNrc7Xd
[6]https://www.kaggle.com/datasets/bababullseye/depression-analysis
[7]https://www.cs.jhu.edu/ mdredze/clpsych-2015-shared-task-evaluation/
[8]https://ir.cs.georgetown.edu/resources/rsdd.html
[9]https://ir.cs.georgetown.edu/resources/smhd.html
[10]https://github.com/usmaann/Depression_Severity_Dataset
[11]https://github.com/cyc21csri/SWDD
[12]https://github.com/aidenwang9867/Weibo-User-Depression-Detection-Dataset
[13]https://drive.google.com/drive/folders/1MXFRs0u8iF1RNUWABTA0Oz8_Ix1skqZT

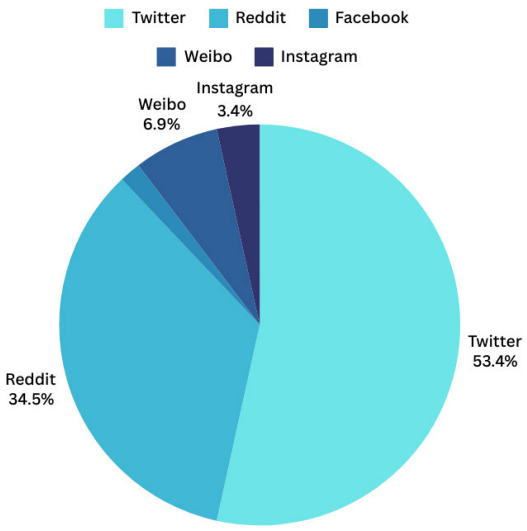

FIGURE 7. Data source for deep learning techniques.

- **EYE**: This dataset is composed of 2,314 users diagnosed with depression and 8,000 control users. It is utilized for tasks related to depression detection in social media [53].
- Copper et al.: A dataset curated by Copper et al. containing 441 depressed users and 5,728 controls. It is often used in studies examining the linguistic and behavioral patterns associated with depression [111].

- **CLPsych 2015**: The CLPsych 2015 dataset includes 327 depressed users and 572 control users. It was used in the Computational Linguistics and Clinical Psychology workshop to facilitate research on mental health detection from social media [112].
- **RSDD**: The Reddit Self-Reported Depression Diagnosis (RSDD) dataset is one of the largest, containing 9,210 users diagnosed with depression and 107,274 control users. It is a pivotal resource for research in this field due to its size and the richness of the user-generated content [113].
- **SMHD**: The Self-reported Mental Health Diagnoses (SMHD) contains 14,139 users identified as having depression and 335,952 control users. This dataset is particularly valuable for large-scale studies [114].
- **D2S**: The D2S dataset comprises 3,738 tweets labeled as depressive and 8,417 non-depressive users [36].
- **DepSign shared task**: This dataset, developed for the DepSign shared task, categorizes users based on the severity of their depression: 3,081 not depressed, 8,325 moderately depressed, and 1,261 severely depressed [115].
- **Depression Severity**: This dataset classifies users into four categories based on depression severity: Minimum (2,587), Mild (290), Moderate (394), and Severe (282) [116].
- **SWDD**: The Social Web Depression Detection (SWDD) dataset contains 3,711 users with depression and 19,526 non-depressed users. It is utilized for various social media-based depression detection tasks [87].
- **multiRedditDep**: This extensive multimodal dataset includes 6.6 million posts from depressed users and 8.1 million posts from non-depressed users, making it one of the largest and most comprehensive resources for studying depression in social media [117].
- **WU3D**: The WU3D dataset contains 10,325 depressed users and 22,245 non-depressed users, used primarily in research involving deep learning techniques for depression detection [79].
- **SetembroBR**: This dataset includes 1,684 depressed users and 11,788 control users, with a focus on Brazilian Portuguese content, providing insights into depression detection in non-English speaking populations [100].

A more extensive collection of datasets is available in this GitHub repository.[14]

Summary of Social Media Depression Detection Datasets: In summary, the development of datasets for depression detection in social media has advanced significantly, offering diverse resources on platforms like Twitter, Reddit, and Weibo. These datasets are crucial for improving depression detection in terms of both early diagnosis and severity estimation. However, there are still challenges, especially with diversity and dataset generalization. Most of the datasets focus on English-speaking populations which limits the global applicability of models. Future directions should focus on creating datasets that better represent different languages, demographics, and cultural contexts, as well as integrating multimodal data sources such as images and videos and expanding to underexplored platforms like Instagram and TikTok. This will improve the generalizability and robustness of depression detection models which would lead to the development of more precise and comprehensive mental health assessment tools. The performance of various techniques on some datasets is presented in Table 4.

[14]https://github.com/bucuram/depression-datasets-nlp

VII. APPLICATION AREAS

The detection of depression using social media data and ML and DL techniques has been applied in various contexts, spanning different languages, specific demographics, and particular life circumstances. Below are some key application areas:

A. CLINICAL SUPPORT AND EARLY INTERVENTION

Automated depression detection systems can act as auxiliary tools for healthcare professionals by identifying individuals who may require mental health support. Early detection through social media posts allows clinicians to intervene promptly, potentially preventing the escalation of depressive symptoms. For instance, studies have demonstrated that patterns in language usage [41], and engagement levels [118] on social media can effectively predict depressive states, providing critical insights to clinicians. Additionally, by tracking changes in social media behavior and sentiment over time, clinicians can assess whether a particular treatment is yielding positive results or if adjustments are necessary.

B. PREDICTING SUICIDAL IDEATION

Depression often leads to suicide, but early detection through social media monitoring may help save lives. Models can flag at-risk users based on analyses of patterns in language use and sentiment shifts and engagement trends [119]. This capability provides a critical opportunity for timely intervention by crisis hotlines and mental health professionals.

C. SEASONAL AND TEMPORAL TRENDS

Research has revealed that mental health conditions, including depression, exhibit distinct temporal patterns that can be tracked using social media data. Seasonal variations are frequently associated with depressive disorders [43]. Similarly, daily variations in mood have been observed, with depression showing measurable differences depending on the time of day [29], [43]. Understanding these trends can guide the scheduling of therapeutic content delivery or awareness campaigns during high-risk periods, ultimately contributing to better mental health outcomes.

TABLE 4. An overview of the performance of various techniques across some datasets.

Reference	Dataset	Technique	F1 Score
[34]	RSDD	X-A-BiLSTM	0.60
[14]	CLPSych 2015	CNN	0.83
[11]	TTDD	CNN-BiLSTM	0.947
[19]	Gui et al.	EmoBERTa, CLIP, ViLBERT	0.945
[19]	multiRedditDep	EmoBERTa, CLIP, ViLBERT	0.913

D. DEMOGRAPHIC AND GEOGRAPHIC CONTEXT

Depression manifests differently across various demographics and geographic locations due to cultural, social, and economic factors. Research has explored these variations by focusing on specific languages and countries. Studies focusing on language include investigations in Thai [29], Chinese [60], Portuguese [100], Bangla [7], [10], [64], and Arabic [8], [9], [94]. Similarly, research has targeted specific national populations, such as in Canada [5], providing insights into the prevalence and patterns of depression within those contexts. This highlights the importance of considering cultural nuances and linguistic variations when developing depression detection models.

E. VULNERABLE POPULATIONS

Several studies focus on populations particularly vulnerable to depression. Postpartum depression, affecting new mothers, has been a significant area of focus. Studies like [3] and [90] have employed machine learning and deep learning, respectively, to detect signs of postpartum depression in social media data, aiming to provide early intervention and support. The mental health of college students, another vulnerable group, has also received considerable attention. Reference [4] used ML approaches to identify signs of depression among college students. These studies underscore the potential of social media data to identify and support individuals at heightened risk of depression.

F. IMPACT OF GLOBAL EVENTS AND MEDIA CONTENT

External events and specific life circumstances can significantly impact mental health. The COVID-19 pandemic, a global crisis, has been a key focus. Studies such as [89] and [92] specifically investigated depression among Twitter users during the pandemic, revealing the psychological impact of such events on social media users. This research emphasizes the real-time monitoring capabilities of social media analysis for tracking mental health trends during crises. Additionally, unconventional media content like memes, often perceived as light-hearted or humorous, has been explored as indicators of depression. This innovative approach demonstrates the potential of analyzing non-textual data for mental health detection.

In summary, the application of ML and DL techniques for depression detection in social media encompasses diverse areas, including clinical support, predicting suicidal ideation, analyzing temporal trends, and addressing cultural and linguistic variations. By focusing on vulnerable populations and assessing the mental health impacts of global events and unconventional data, such as memes, these approaches provide valuable insights into depression's multifaceted online presence. This emphasizes the importance of adaptable models for effective mental health detection.

VIII. OPEN RESEARCH AREAS

Depression detection through social media analysis is a growing area of research. However, several critical challenges and open research areas remain:

A. PLATFORM FOCUS

According to Figures 6 and 7, the majority of work in depression detection has focused on Twitter and Reddit. While these platforms provide rich data sources, other social networking platforms such as Facebook, TikTok, and Instagram remain underexplored. Expanding research to these platforms could offer new insights and help create more comprehensive models that better reflect diverse online behaviors.

B. DATASET DIVERSITY AND GENERALIZATION

Current datasets predominantly focus on specific demographics, with an emphasis on the USA and Canada [19], [21], [38], often leading to models that struggle with generalization across different population groups. There's a need for more diverse datasets that encompass various age groups, ethnicities, and cultural backgrounds. Enhancing dataset diversity will enable models to perform better across different user segments, improving the generalizability and robustness of depression detection models.

C. MULTIMODAL APPROACHES

While most existing research focuses on textual data, there is a growing interest [15], [17], [18], [19], [38], [81], [88], [92] in multimodal approaches that incorporate images, videos, and other non-textual data alongside text. This area is still in its infancy, with significant potential for exploring how different data modalities can be combined to improve detection accuracy. Future research could look into the integration of various data types and the development of

multimodal models that better capture the complex nature of depression.

D. MULTILINGUAL MODELS

Depression detection models are often language-specific [120], limiting their applicability to users who communicate in different languages. Developing multilingual models capable of detecting depression across various languages and cultural contexts is a crucial research area. Such models would require robust cross-lingual transfer learning techniques and the incorporation of culturally specific expressions of mental health.

E. CROSS-PLATFORM USER ANALYSIS

Users often engage with multiple social media platforms, and their online behavior varies across these platforms. A viable research direction is the development of techniques that can aggregate and analyze data from multiple platforms for a single user [120]. By taking into account a user's activities on various online platforms, this method may offer a more thorough knowledge of their mental state.

F. DATA AUGMENTATION AND IMBALANCE

One of the key challenges while detecting depression is that the data is imbalanced. Mostly the samples of depressed individuals are significantly less as compared to normal individuals. It requires developing effective augmentation techniques to overcome this imbalance and enhance the model's ability to handle small and underrepresented samples. Studies on advanced augmentation methods and synthetic data generation [19] could have a big impact on the area, especially when it comes to multimodal data where images are less compared to text. To address the issue of class imbalance, methods such as distribution normalization [88], one-shot decision strategy [62], and oversampling and undersampling [54] have been used.

G. UNIFIED MODELS ACROSS PLATFORMS

Developing a single model that can detect depression across multiple social media platforms is a challenging and relatively unexplored area. To achieve consistent performance, it is important to account for the unique features and data structures of each platform. Research in this field could result in more scalable and versatile depression detection systems. These systems would be applicable in various online environments.

H. ENHANCING LARGE LANGUAGE MODELS (LLMS)

Recent research [105] has shown that LLMs frequently perform poorly on tasks related to mental health. This is because LLMs are likely to randomly guess or hallucinate. This highlights the need for more research on training and fine-tuning LLMs for depression detection. Enhancing the sensitivity, accuracy, and robustness of LLMs in this area is essential. These improvements could significantly benefit the field. They would also make these models more effective tools for mental health applications.

IX. ETHICAL CONSIDERATIONS

Social media-based mental health research offers valuable insights but raises critical ethical challenges. This section highlights concerns around privacy, consent, risks, and strategies to ensure responsible and ethical research practices.

A. DATA PRIVACY AND USER CONSENT

The collection of social media data for research often occurs without explicit consent from users, raising significant privacy concerns. Many users remain unaware that their public posts are being utilized for studies, which challenges the fundamental principle of informed consent. To mitigate these ethical issues, researchers should prioritize anonymizing data, obtaining necessary ethical clearances, and, where possible, actively seeking consent from users. These measures help balance the pursuit of valuable insights with the responsibility to respect users' privacy and autonomy.

B. POTENTIAL RISKS AND MISUSE

There are potential risks associated with using social media data for mental health predictions. For instance, incorrect or biased predictions may lead to stigmatization or discrimination against individuals identified as "at-risk." Furthermore, there is a risk of misuse by third parties, such as employers or insurance companies, who could exploit mental health predictions to make adverse decisions.

C. MITIGATION STRATEGIES

To address these ethical concerns, researchers should:

- Implement robust anonymization techniques to ensure that individuals cannot be identified from the data.
- Obtain ethical approval from Institutional Review Boards (IRBs) or equivalent bodies before conducting studies.
- Develop transparent models that allow for explainability and accountability in predictions.
- Collaborate with ethicists and mental health professionals to ensure that research aligns with ethical and clinical standards.

D. BALANCING BENEFITS AND RISKS

While social media-based mental health predictions have the potential to improve early detection and intervention strategies, researchers must balance these benefits with the risks. Adopting a human-centric approach that prioritizes user rights and welfare is essential for ensuring that advancements in this field are both ethical and impactful.

X. CONCLUSION

This comprehensive review looked into the use of ML and DL techniques to detect depression in social media. We have highlighted the key changes in classification methods by looking at how these techniques have evolved

from traditional ML models to advanced DL systems. The generic architecture for depression detection systems shows how data preprocessing, feature extraction, model selection, and evaluation are interconnected.

The several datasets utilized in this field have also been thoroughly discussed, highlighting the significance of carefully choosing the right resources in order to produce accurate and reliable detection results. The application areas of depression detection illustrate the real-life impact and potential of these techniques in identifying mental health trends and enabling early intervention.

Despite the progress made, several open research areas remain, including the need for more diverse and representative datasets, improved model interpretability, and the ethical considerations surrounding privacy and data use. Addressing these issues will be critical to the future development of more effective and responsible depression detection systems.

In conclusion, while significant progress has been made in the domain, collaboration and further research across disciplines are essential to refine these technologies and ensure they are used to their full potential in supporting mental health on social media platforms.

REFERENCES

[1] World Health Org. (2022). *Mental Disorders*. Accessed: Jun. 6, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[2] N. A. Asad, Md. A. Mahmud Pranto, S. Afreen, and Md. M. Islam, "Depression detection by analyzing social media posts of user," in *Proc. IEEE Int. Conf. Signal Process., Inf., Commun. Syst. (SPICSCON)*, Nov. 2019, pp. 13–17.

[3] I. Fatima, B. U. D. Abbasi, S. Khan, M. Al-Saeed, H. F. Ahmad, and R. Mumtaz, "Prediction of postpartum depression using machine learning techniques from social media text," *Expert Syst.*, vol. 36, no. 4, Aug. 2019, Art. no. e12409, 10.1111/exsy.12409.

[4] Y. Ding, X. Chen, Q. Fu, and S. Zhong, "A depression recognition method for college students using deep integrated support vector algorithm," *IEEE Access*, vol. 8, pp. 75616–75629, 2020.

[5] R. Skaik and D. Inkpen, "Using Twitter social media for depression detection in the Canadian population," in *Proc. 3rd Artif. Intell. Cloud Comput. Conf.*, New York, NY, USA, Dec. 2020, pp. 109–114, doi: 10.1145/3442536.3442553.

[6] H. S. Alsagri and M. Ykhlef, "Machine learning-based approach for depression detection in Twitter using content and activity features," *IEICE Trans. Inf. Syst.*, vol. E103.D, no. 8, pp. 1825–1832, Aug. 2020, doi: 10.1587/TRANSINF.2020EDP7023.

[7] Z. N. Vasha, B. Sharma, I. J. Esha, J. Al Nahian, and J. A. Polin, "Depression detection in social media comments data using machine learning algorithms," *Bull. Electr. Eng. Informat.*, vol. 12, no. 2, pp. 987–996, Apr. 2023. [Online]. Available: https://beei.org/index.php/EEI/article/view/4182/3162

[8] K. Sabaneh, M. A. Salameh, F. Khaleel, M. M. Herzallah, J. Y. Natsheh, and M. Maree, "Early risk prediction of depression based on social media posts in Arabic," in *Proc. IEEE 35th Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2023, pp. 595–602.

[9] A. Helmy, R. Nassar, and N. Ramdan, "Depression detection for Twitter users using sentiment analysis in English and Arabic tweets," *Artif. Intell. Med.*, vol. 147, Jan. 2024, Art. no. 102716. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0933365723002300

[10] A. H. Uddin, D. Bapery, and A. S. M. Arif, "Depression analysis from social media data in Bangla language using long short term memory (LSTM) recurrent neural network technique," in *Proc. Int. Conf. Comput., Commun., Chem., Mater. Electron. Eng. (IC4ME2)*, Jul. 2019, pp. 1–4.

[11] H. Kour and M. K. Gupta, "An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM," *Multimedia Tools Appl.*, vol. 81, no. 17, pp. 23649–23685, Jul. 2022, doi: 10.1007/s11042-022-12648-y.

[12] S. R. Narayanan, S. Babu, and A. Thandayantavida, "Detection of depression from social media using deep learning approach," *J. Positive School Psychol.*, vol. 6, no. 4, pp. 4909–4915, 2022.

[13] A. Husseini Orabi, P. Buddhitha, M. Husseini Orabi, and D. Inkpen, "Deep learning for depression detection of Twitter users," in *Proc. 5th Workshop Comput. Linguistics Clin. Psychol., Keyboard Clinic*, K. Loveys, K. Niederhoffer, E. Prud'hommeaux, R. Resnik, and P. Resnik, Eds., Jun. 2018, pp. 88–97. [Online]. Available: https://aclanthology.org/W18-0609

[14] P. K. Gamaarachchige and D. Inkpen, "Multi-task, multi-channel, multi-input learning for mental illness detection using social media text," in *Proc. 10th Int. Workshop Health Text Mining Inf. Anal. (LOUHI)*, E. Holderness, A. Jimeno Yepes, A. Lavelli, A.-L. Minard, J. Pustejovsky, and F. Rinaldi, Eds., Nov. 2019, pp. 54–64. [Online]. Available: https://aclanthology.org/D19-6208

[15] C. Lin, P. Hu, H. Su, S. Li, J. Mei, J. Zhou, and H. Leung, "SenseMood: Depression detection on social media," in *Proc. Int. Conf. Multimedia Retr.*, New York, NY, USA, Jun. 2020, p. 407, doi: 10.1145/3372278.3391932.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, J. Burstein, C. Doran, and T. Solorio, Eds., Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[17] M. An, J. Wang, S. Li, and G. Zhou, "Multimodal topic-enriched auxiliary learning for depression detection," in *Proc. 28th Int. Conf. Comput. Linguistics*, Barcelona, Spain, D. Scott, N. Bel, and C. Zong, Eds., 2020, pp. 1078–1089. [Online]. Available: https://aclanthology.org/2020.coling-main.94

[18] C. Y. Chiu, H. Y. Lane, J. L. Koh, and A. L. P. Chen, "Multimodal depression detection on Instagram considering time interval of posts," *J. Intell. Inf. Syst.*, vol. 56, no. 1, pp. 25–47, Feb. 2021, doi: 10.1007/s10844-020-00599-5.

[19] A. Zafar, D. Aftab, R. Qureshi, Y. Wang, and H. Yan, "Multi-explainable temporalNet: An interpretable multimodal approach using temporal convolutional network for user-level depression detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 2258–2265.

[20] W. A. Gadzama, D. Gabi, M. S. Argungu, and H. U. Suru, "The use of machine learning and deep learning models in detecting depression on social media: A systematic literature review," *Personalized Med. Psychiatry*, vols. 45–46, Jul. 2024, Art. no. 100125. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468171724000115

[21] A. Aldkheel and L. Zhou, "Depression detection on social media: A classification framework and research challenges and opportunities," *J. Healthcare Informat. Res.*, vol. 8, no. 1, pp. 88–120, Mar. 2024, doi: 10.1007/s41666-023-00152-3.

[22] A. Helmy, R. Nassar, and N. Ramadan, "Depression detection from social media platforms: A systematic literature review," in *Proc. Intell. Methods, Syst., Appl. (IMSA)*, 2023, pp. 387–393.

[23] K. M. Hasib, M. R. Islam, S. Sakib, Md. A. Akbar, I. Razzak, and M. S. Alam, "Depression detection from social networks data based on machine learning and deep learning techniques: An interrogative survey," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 4, pp. 1568–1586, Aug. 2023.

[24] M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in *Proc. 5th Annu. ACM Web Sci. Conf.*, New York, NY, USA, May 2013, pp. 47–56, doi: 10.1145/2464464.2464480.

[25] X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, and Z. Bao, "A depression detection model based on sentiment analysis in micro-blog social network," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*, vol. 7867, Berlin, Germany. Springer, 2013, pp. 201–213, doi: 10.1007/978-3-642-40319-4_18.

[26] P. Nambisan, Z. Luo, A. Kapoor, T. B. Patrick, and R. A. Cisler, "Social media, big data, and public health informatics: Ruminating behavior of depression revealed through Twitter," in *Proc. 48th Hawaii Int. Conf. Syst. Sci.*, Jan. 2015, pp. 2906–2913.

[27] X. Wang, C. Zhang, and L. Sun, "An improved model for depression detection in micro-blog social network," in *Proc. IEEE 13th Int. Conf. Data Mining Workshops*, Dec. 2013, pp. 80–87.

[28] P. Resnik, W. Armstrong, L. Claudino, and T. Nguyen, "The university of Maryland CLPsych 2015 shared task system," in *Proc. 2nd Workshop Comput. Linguistics Clin. Psychol., Linguistic Signal Clin. Reality*, 2015, pp. 54–60. [Online]. Available: https://aclanthology.org/W15-1207

[29] K. Katchapakirin, K. Wongpatikaseree, P. Yomaboot, and Y. Kaewpitakkun, "Facebook social media for depression detection in the Thai community," in *Proc. 15th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, Jul. 2018, pp. 1–6.

[30] M. E. Aragón, A. P. López-Monroy, L. C. González-Gurrola, and M. Montes-y Gómez, "Detecting depression in social media using fine-grained emotions," in *Proc. Conf. North American Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Minneapolis, Minnesota, J. Burstein, C. Doran, and T. Solorio, Eds., Jun. 2019, pp. 1481–1486. [Online]. Available: https://aclanthology.org/N19-1151

[31] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in Reddit social media forum," *IEEE Access*, vol. 7, pp. 44883–44893, 2019.

[32] M. T. Hossain, M. A. R. Talukder, and N. Jahan, "Social networking sites data analysis using NLP and ML to predict depression," in *Proc. 12th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Jul. 2021, pp. 1–5.

[33] R. Chiong, G. S. Budhi, and S. Dhakal, "Combining sentiment lexicons and content-based features for depression detection," *IEEE Intell. Syst.*, vol. 36, no. 6, pp. 99–105, Nov. 2021.

[34] Q. Cong, Z. Feng, F. Li, Y. Xiang, G. Rao, and C. Tao, "X-A-BiLSTM: A deep learning approach for depression detection in imbalanced data," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 1624–1627.

[35] F. M. Shah, F. Ahmed, S. K. S. Joy, S. Ahmed, S. Sadek, R. Shil, and M. H. Kabir, "Early depression detection from social network using deep learning techniques," in *Proc. IEEE Region 10 Symp. (TENSYMP)*, Jun. 2020, pp. 823–826.

[36] S. Yadav, J. Chauhan, J. P. Sain, K. Thirunarayan, A. Sheth, and J. Schumm, "Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework," in *Proc. 28th Int. Conf. Comput. Linguistics*, Barcelona, Spain, D. Scott, N. Bel, and C. Zong, Eds., 2020, pp. 696–709. [Online]. Available: https://aclanthology.org/2020.coling-main.61

[37] H. Zogan, I. Razzak, S. Jameel, and G. Xu, "DepressionNet: Learning multi-modalities with user post summarization for depression detection on social media," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 133–142, doi: 10.1145/3404835.3462938.

[38] A.-M. Bucur, A. Cosma, P. Rosso, and L. P. Dinu, "It's just a matter of time: detecting depression with time-enriched multimodal transformers," in *Proc. Eur. Conf. Inf. Retr.*, Dublin, Ireland. Berlin, Germany: Springer, Apr. 2023, pp. 200–215, doi: 10.1007/978-3-031-28244-7_13.

[39] A. Malhotra and R. Jindal, "XAI transformer based approach for interpreting depressed and suicidal user behavior on online social networks," *Cognit. Syst. Res.*, vol. 84, Mar. 2024, Art. no. 101186. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389041723001201

[40] I. Pirina and Ç. Çöltekin, "Identifying depression on Reddit: The effect of training data," in *Proc. EMNLP Workshop SMM4H, 3rd Social Media Mining Health Appl. Workshop Shared Task*, Brussels, Belgium, G. Gonzalez-Hernandez, D. Weissenbacher, A. Sarker, and M. Paul, Eds., Oct. 2018, pp. 9–12. [Online]. Available: https://aclanthology.org/W18-5903

[41] N. Ramirez-Esparza, C. Chung, E. Kacewic, and J. Pennebaker, "The psychology of word use in depression forums in English and in spanish: Testing two text analytic approaches," in *Proc. Int. AAAI Conf. Web Social Media*, Sep. 2021, vol. 2, no. 1, pp. 102–108. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/18623

[42] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, Apr. 2009, doi: 10.1561/1500000019.

[43] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq, "Depression detection from social network data using machine learning techniques," *Health Inf. Sci. Syst.*, vol. 6, no. 1, p. 8, Dec. 2018.

[44] D. E. Losada, F. Crestani, and J. Parapar, "ERISK 2017: CLEF lab on early risk prediction on the Internet: Experimental foundations," in *Proc. Int. Conf. Cross-Lang. Eval. Forum Eur. Lang.*, Dublin, Ireland, Jan. 2017, pp. 346–360.

[45] D. E. Losada, F. Crestani, and J. Parapar, "Overview of erisk: Early risk prediction on the internet," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, P. Bellot, C. Trabelsi, J. Mothe, F. Murtagh, J. Y. Nie, L. Soulier, E. SanJuan, L. Cappellato, and N. Ferro, Eds., Cham, Switzerland: Springer, 2018, pp. 343–361.

[46] P. Ekman and R. J. Davidson, *The Nature of Emotion: Fundamental Questions*. New York, NY, USA: Oxford Univ. Press, 1994.

[47] P. Thavikulwat, "Affinity propagation: A clustering algorithm for computer-assisted business simulations and experiential exercises," in *Proc. Develop. Bus. Simulation Experiential Learn.*, vol. 35, 2014, pp. 1–5. [Online]. Available: https://api.semanticscholar.org/CorpusID

[48] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," in *Proc. Trans. Assoc. Comput. Linguistics*, vol. 5, 2017, pp. 135–146, doi: 10.1162/tacl_a_00051.

[49] P. Arora and P. Arora, "Mining Twitter data for depression detection," in *Proc. Int. Conf. Signal Process. Commun. (ICSC)*, Mar. 2019, pp. 186–189.

[50] P. V. Rajaraman, A. Nath, P. R. Akshaya, and G. C. Bhuja, "Depression detection of tweets and a comparative test," *Int. J. Eng. Res.*, vol. 9, no. 3, pp. 422–425, Mar. 2020, doi: 10.17577/ijertv9is030270.

[51] K. A. Govindasamy and N. Palanichamy, "Depression detection using machine learning techniques on Twitter data," in *Proc. 5th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, May 2021, pp. 960–966.

[52] 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 *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 3838–3844, doi: 10.24963/IJCAI.2017/536.

[53] Kaggle. (2020). *Depression Analysis*. Accessed: Jul. 31, 2020. [Online]. Available: https://www.kaggle.com/datasets/bababullseye/depression-analysis

[54] R. Chiong, G. S. Budhi, S. Dhakal, and F. Chiong, "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts," *Comput. Biol. Med.*, vol. 135, Aug. 2021, Art. no. 104499. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521002936

[55] R. Chatterjee, R. K. Gupta, and B. Gupta, "Depression detection from social media posts using multinomial naive theorem," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 1022, no. 1, Jan. 2021, Art. no. 012095, doi: 10.1088/1757-899X/1022/1/012095.

[56] J. Aguilera, D. I. H. Farías, R. M. Ortega-Mendoza, and M. Montes-Y-Gómez, "Depression and anorexia detection in social media as a one-class classification problem," *Appl. Intell.*, vol. 51, no. 8, pp. 6088–6103, 2021, doi: 10.1007/s10489-020-02131-2.

[57] J. De Jesús Titla-Tlatelpa, R. M. Ortega-Mendoza, M. Montes-Y-Gómez, and L. Villaseñor-Pineda, "A profile-based sentiment-aware approach for depression detection in social media," *EPJ Data Sci.*, vol. 10, no. 1, p. 54, Dec. 2021, doi: 10.1140/epjds/s13688-021-00309-3.

[58] A. S. Liaw and H. N. Chua, "Depression detection on social media with user network and engagement features using machine learning methods," in *Proc. IEEE Int. Conf. Artif. Intell. Eng. Technol. (IICAIET)*, Sep. 2022, pp. 1–6.

[59] J. Angskun, S. Tipprasert, and T. Angskun, "Big data analytics on social networks for real-time depression detection," *J. Big Data*, vol. 9, no. 1, p. 69, Dec. 2022, doi: 10.1186/s40537-022-00622-2.

[60] Z. Guo, N. Ding, M. Zhai, Z. Zhang, and Z. Li, "Leveraging domain knowledge to improve depression detection on Chinese social media," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 4, pp. 1528–1536, Aug. 2023.

[61] 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. Affect. Comput.*, vol. 14, no. 1, pp. 211–222, Jan. 2023.

[62] V. Adarsh, P. Arun Kumar, V. Lavanya, and G. R. Gangadharan, "Fair and explainable depression detection in social media," *Inf. Process. Manage.*, vol. 60, no. 1, Jan. 2022, Art. no. 103168. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0306457322002692

[63] H. Song, J. You, J.-W. Chung, and J. C. Park, "Feature attention network: Interpretable depression detection from social media," in *Proc. 32nd Pacific Asia Conf. Lang., Inf. Comput.*, S. Politzer-Ahles, Y.-Y. Hsu, C.-R. Huang, and Y. Yao, Eds., Dec. 2018, pp. 1–3. [Online]. Available: https://aclanthology.org/Y18-1070

[64] A. H. Uddin, D. Bapery, and A. S. Mohammad Arif, "Depression analysis of Bangla social media data using gated recurrent neural network," in *Proc. 1st Int. Conf. Adv. Sci., Eng. Robot. Technol. (ICASERT)*, May 2019, pp. 1–6.

[65] I. Sekulic and M. Strube, "Adapting deep learning methods for mental health prediction on social media," in *Proc. 5th Workshop Noisy User-Generated Text (W-NUT)*, Hong Kong, 2019, pp. 322–327. [Online]. Available: https://aclanthology.org/D19-5542

[66] A.-M. Bucur and L. P. Dinu, "Detecting early onset of depression from social media text using learned confidence scores," in *Proc. 7th Italian Conf. Comput. Linguistics*, 2020, pp. 73–78, doi: 10.4000/books.aaccademia.8305.

[67] S. Lei and W. Su, "An end-to-end method for teenagers potential depression detection on social media," in *Proc. IEEE Int. Perform., Comput., Commun. Conf. (IPCCC)*, Oct. 2021, pp. 1–2.

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

[69] A.-S. Uban, B. Chulvi, and P. Rosso, "An emotion and cognitive based analysis of mental health disorders from social media data," *Future Gener. Comput. Syst.*, vol. 124, pp. 480–494, Nov. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X21001825

[70] M. Rizwan, M. F. Mushtaq, U. Akram, A. Mehmood, I. Ashraf, and B. Sahelices, "Depression classification from tweets using small deep transfer learning language models," *IEEE Access*, vol. 10, pp. 129176–129189, 2022.

[71] P. Mann, E. H. Matsushima, and A. Paes, "Detecting depression from social media data as a multiple-instance learning task," in *Proc. 10th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2022, pp. 1–8.

[72] S. Han, R. Mao, and E. Cambria, "Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings," in *Proc. 29th Int. Conf. Comput. Linguistics*, Gyeongju, South Korea, Oct. 2022, pp. 94–104. [Online]. Available: https://aclanthology.org/2022.coling-1.9

[73] 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, Jan. 2022, doi: 10.1007/s11280-021-00992-2.

[74] K. Yang, T. Zhang, and S. Ananiadou, "A mental state knowledge-aware and contrastive network for early stress and depression detection on social media," *Inf. Process. Manage.*, vol. 59, no. 4, Art. no. 102961. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457322000796

[75] A. Amanat, M. Rizwan, A. R. Javed, M. Abdelhaq, R. Alsaqour, S. Pandya, and M. Uddin, "Deep learning for depression detection from textual data," *Electronics*, vol. 11, no. 5, p. 676, Feb. 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/5/676

[76] A. Nadeem, M. Naveed, M. Islam Satti, H. Afzal, T. Ahmad, and K.-I. Kim, "Depression detection based on hybrid deep learning SSCL framework using self-attention mechanism: An application to social networking data," *Sensors*, vol. 22, no. 24, p. 9775, Dec. 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/24/9775

[77] R. Poświata and M. Perełkiewicz, "OPI@LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pre-trained language models," in *Proc. 2nd Workshop Lang. Technol. Equality, Diversity Inclusion*, Dublin, Ireland, B. R. Chakravarthi, B. Bharathi, J. P. McCrae, M. Zarrouk, K. Bali, and P. Buitelaar, Eds., 2022, pp. 276–282. [Online]. Available: https://aclanthology.org/2022.ltedi-1.40

[78] M. Kabir, T. Ahmed, M. B. Hasan, M. T. R. Laskar, T. K. Joarder, H. Mahmud, and K. Hasan, "DEPTWEET: A typology for social media texts to detect depression severities," *Comput. Hum. Behav.*, vol. 139, Feb. 2023, Art. no. 107503, doi: 10.1016/j.chb.2022.107503.

[79] Y. Wang, Z. Wang, C. Li, Y. Zhang, and H. Wang, "Online social network individual depression detection using a multitask heterogenous modality fusion approach," *Inf. Sci.*, vol. 609, pp. 727–749, Sep. 2022, doi: 10.1016/j.ins.2022.07.109.

[80] M. Aragon, A. P. Lopez Monroy, L. Gonzalez, D. E. Losada, and M. Montes, "DisorBERT: A double domain adaptation model for detecting signs of mental disorders in social media," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, Toronto, ON, Canada, 2023, pp. 15305–15318. [Online]. Available: https://aclanthology.org/2023.acl-long.853

[81] S. Yadav, C. Caragea, C. Zhao, N. Kumari, M. Solberg, and T. Sharma, "Towards identifying fine-grained depression symptoms from memes," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, Toronto, ON, Canada, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Jul. 2023, pp. 8890–8905. [Online]. Available: https://aclanthology.org/2023.acl-long.495

[82] T. Zhang, K. Yang, and S. Ananiadou, "Sentiment-guided transformer with severity-aware contrastive learning for depression detection on social media," in *Proc. 22nd Workshop Biomed. Natural Lang. Process. BioNLP Shared Tasks*, Toronto, ON, Canada, D. Demner-fushman, S. Ananiadou, and K. Cohen, Eds., Jul. 2023, pp. 114–126. [Online]. Available: https://aclanthology.org/2023.bionlp-1.9

[83] 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. Biomed. Health Informat.*, vol. 28, no. 4, pp. 1815–1823, Apr. 2024.

[84] L. Zong, J. Zheng, X. Zhang, X. Liu, W. Liang, and B. Xu, "An early depression detection model on social media using emotional and causal features," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2023, pp. 4856–4858.

[85] L. Ilias and D. Askounis, "Multitask learning for recognizing stress and depression in social media," *Online Social Netw. Media*, vols. 37–38, Sep. 2023, Art. no. 100270, doi: 10.1016/j.osnem.2023.100270.

[86] Z. Liu, X. Ma, P. Zhang, C. Hao, S. Zhang, and L. Wang, "TIDE: Affective time-aware representations for fine-grained depression identification on social media," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2023, pp. 01–08.

[87] Y. Cai, H. Wang, H. Ye, Y. Jin, and W. Gao, "Depression detection on online social network with multivariate time series feature of user depressive symptoms," *Expert Syst. Appl.*, vol. 217, May 2023, Art. no. 119538, doi: 10.1016/j.eswa.2023.119538.

[88] Z. Li, Z. An, W. Cheng, J. Zhou, F. Zheng, and B. Hu, "MHA: A multimodal hierarchical attention model for depression detection in social media," *Health Inf. Sci. Syst.*, vol. 11, no. 1, p. 6, Jan. 2023, doi: 10.1007/s13755-022-00197-5.

[89] 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 *Proc. ACM Web Conf.*, New York, NY, USA, Apr. 2023, pp. 3968–3977, doi: 10.1145/3543507.3583867.

[90] A. Gopalakrishnan, R. Gururajan, R. Venkataraman, X. Zhou, K. C. Ching, A. Saravanan, and M. Sen, "Attribute selection hybrid network model for risk factors analysis of postpartum depression using social media," *Brain Informat.*, vol. 10, no. 1, p. 28, Dec. 2023, doi: 10.1186/s40708-023-00206-7.

[91] M. Bhuvaneswari and V. L. Prabha, "A deep learning approach for the depression detection of social media data with hybrid feature selection and attention mechanism," *Expert Syst.*, vol. 40, no. 9, p. 13371, Nov. 2023, doi: 10.1111/exsy.13371.

[92] A. Anshul, G. S. Pranav, M. Z. U. Rehman, and N. Kumar, "A multimodal framework for depression detection during COVID-19 via harvesting social media," *IEEE Trans. Computat. Social Syst.*, vol. 11, no. 2, pp. 2872–2888, Apr. 2024.

[93] Y. Liu, "Depression detection via a Chinese social media platform: A novel causal relation-aware deep learning approach," *J. Supercomput.*, vol. 80, no. 8, pp. 10327–10356, May 2024, doi: 10.1007/s11227-023-05830-y.

[94] S. Elmajali and I. Ahmad, "Toward early detection of depression: Detecting depression symptoms in Arabic tweets using pretrained transformers," *IEEE Access*, vol. 12, pp. 88134–88145, 2024.

[95] M. A. Abbas, K. Munir, A. Raza, N. A. Samee, M. M. Jamjoom, and Z. Ullah, "Novel transformer based contextualized embedding and probabilistic features for depression detection from social media," *IEEE Access*, vol. 12, pp. 54087–54100, 2024.

[96] R. Dou and X. Kang, "TAM-SenticNet: A neuro-symbolic AI approach for early depression detection via social media analysis," *Comput. Electr. Eng.*, vol. 114, May 2024, Art. no. 109071, doi: 10.1016/j.compeleceng.2023.109071.

[97] J. P. Thekkekara, S. Yongchareon, and V. Liesaputra, "An attention-based CNN-BiLSTM model for depression detection on social media text," *Expert Syst. Appl.*, vol. 249, Sep. 2024, Art. no. 123834, doi: 10.1016/j.eswa.2024.123834.

[98] J. Liu, W. Chen, L. Wang, and F. Ding, "A hybrid depression detection model and correlation analysis for social media based on attention mechanism," *Int. J. Mach. Learn. Cybern.*, vol. 15, no. 7, pp. 2631–2642, Jul. 2024, doi: 10.1007/s13042-023-02053-8.

[99] S. Dalal, S. Jain, and M. Dave, "Convolution neural network having multiple channels with own attention layer for depression detection from social data," *New Gener. Comput.*, vol. 42, no. 1, pp. 135–155, Nov. 2023, doi: 10.1007/s00354-023-00237-y.

[100] W. R. D. Santos, R. L. de Oliveira, and I. Paraboni, "SetembroBR: A social media corpus for depression and anxiety disorder prediction," *Lang. Resour. Eval.*, vol. 58, no. 1, pp. 273–300, Jan. 2023, doi: 10.1007/s10579-022-09633-0.

[101] N. Farruque, R. Goebel, S. Sivapalan, and O. Zaïane, "Deep temporal modelling of clinical depression through social media text," *Natural Lang. Process. J.*, vol. 6, Mar. 2024, Art. no. 100052, doi: 10.1016/j.nlp.2023.100052.

[102] W. Zhang, J. Xie, Z. (. Zhang, and X. Liu, "Depression detection using digital traces on social media: A knowledge-aware deep learning approach," *J. Manage. Inf. Syst.*, vol. 41, no. 2, pp. 546–580, Apr. 2024, doi: 10.1080/07421222.2024.2340822.

[103] P. Singh, G. Singh, A. Singh, and J. Singh, "Intelligent mental depression recognition model with ensemble learning through social media tweet resources," *Cybern. Syst.*, vol. 55, no. 2, pp. 471–510, Feb. 2024, doi: 10.1080/01969722.2022.2122009.

[104] B. Wang, Y. Zi, Y. Zhao, P. Deng, and B. Qin, "ESDM: Early sensing depression model in social media streams," in *Proc. Joint Int. Conf. Comput. Linguistics, Lang. Resour. Eval. (LREC-COLING)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., May 2024, pp. 6288–6298. [Online]. Available: https://aclanthology.org/2024.lrec-main.556

[105] F. Alhamed, J. Ive, and L. Specia, "Classifying social media users before and after depression diagnosis via their language usage: A dataset and study," in *Proc. Joint Int. Conf. Comput. Linguistics, Lang. Resour. Eval. (LREC-COLING)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., May 2024, pp. 3250–3260. [Online]. Available: https://aclanthology.org/2024.lrec-main.289

[106] V. Tejaswini, K. Sathya Babu, and B. Sahoo, "Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 23, no. 1, pp. 1–20, Jan. 2024, doi: 10.1145/3569580.

[107] D. E. Losada and F. Crestani, "A test collection for research on depression and language use," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, N. Fuhr, P. Quaresma, T. Gonçalves, B. Larsen, K. Balog, C. Macdonald, L. Cappellato, and N. Ferro, Eds., Cham, Switzerland: Springer, 2016, pp. 28–39.

[108] D. E. Losada, F. Crestani, and J. Parapar, "Erisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, G. J. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, L. Cappellato, and N. Ferro, Eds., Cham, Switzerland: Springer, 2017, pp. 346–360.

[109] D. E. Losada, F. Crestani, and J. Parapar, "erisk 2020: Self-harm and depression challenges," in *Proc. Eur. Conf. Inf. Retr.*, Berlin, Germany. Springer, Apr. 2020, pp. 557–563, doi: 10.1007/978-3-030-45442-5_72.

[110] T. Gui, L. Zhu, Q. Zhang, M. Peng, X. Zhou, K. Ding, and Z. Chen, "Cooperative multimodal approach to depression detection in Twitter," in *Proc. 33rd AAAI Conf. Artif. Intell., 31st Innov. Appl. Artif. Intell. Conf., 9th AAAI Symp. Educ. Adv. Artif. Intell.*, vol. 33, Jul. 2019, pp. 110–117, doi: 10.1609/aaai.v33i01.3301110.

[111] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in *Proc. Workshop Comput. Linguistics Clin. Psychol., Linguistic Signal Clin. Reality*, P. Resnik, R. Resnik, and M. Mitchell, Eds., Jun. 2014, pp. 51–60. [Online]. Available: https://aclanthology.org/W14-3207

[112] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell, "CLPsych 2015 shared task: Depression and PTSD on Twitter," in *Proc. 2nd Workshop Comput. Linguistics Clin. Psychol., Linguistic Signal Clin. Reality*, 2015, pp. 31–39. [Online]. Available: https://aclanthology.org/W15-1204

[113] A. Yates, A. Cohan, and N. Goharian, "Depression and self-harm risk assessment in online forums," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, Sep. 2017, pp. 2968–2978. [Online]. Available: https://aclanthology.org/D17-1322

[114] 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 *Proc. 27th Int. Conf. Comput. Linguistics*, Santa Fe, NM, USA, E. M. Bender, L. Derczynski, and P. Isabelle, Eds., Aug. 2018, pp. 1485–1497. [Online]. Available: https://aclanthology.org/C18-1126

[115] X. Lin, Y. Fu, Z. Yang, N. Lin, and S. Jiang, "BERT 4EVER@LT-EDI-ACL2022-detecting signs of depression from social media: Detecting depression in social media using prompt-learning and word-emotion cluster," in *Proc. 2nd Workshop Lang. Technol. Equality, Diversity Inclusion*, Dublin, Ireland, B. R. Chakravarthi, B. Bharathi, J. P. McCrae, M. Zarrouk, K. Bali, and P. Buitelaar, Eds., 2022, pp. 200–205. [Online]. Available: https://aclanthology.org/2022.ltedi-1.27

[116] U. Naseem, A. G. Dunn, J. Kim, and M. Khushi, "Early identification of depression severity levels on Reddit using ordinal classification," in *Proc. ACM Web Conf.*, New York, NY, USA, Apr. 2022, pp. 2563–2572, doi: 10.1145/3485447.3512128.

[117] A.-S. Uban, B. Chulvi, and P. Rosso, "Explainability of depression detection on social media: From deep learning models to psychological interpretations and multimodality," in *Early Detection of Mental Health Disorders By Social Media Monitoring*. Cham, Switzerland: Springer, 2022, pp. 289–320, 10.1007/978-3-031-04431-1_13.

[118] L. Y. Lin, J. E. Sidani, A. Shensa, A. Radovic, E. Miller, J. B. Colditz, B. L. Hoffman, L. M. Giles, and B. A. Primack, "Association between social media use and depression among U.S. Young adults," *Depression Anxiety*, vol. 33, no. 4, pp. 323–331, Apr. 2016, 10.1002/da.22466.

[119] M. Chatterjee, P. Kumar, P. Samanta, and D. Sarkar, "Suicide ideation detection from online social media: A multi-modal feature based technique," *Int. J. Inf. Manage. Data Insights*, vol. 2, no. 2, Nov. 2022, Art. no. 100103. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667096822000465

[120] U. Yadav, A. K. Sharma, and D. Patil, "Review of automated depression detection: Social posts, audio and video, open challenges and future direction," *Concurrency Comput., Pract. Exper.*, vol. 35, no. 1, Jan. 2023, Art. no. e7407. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.7407

WALEED BIN TAHIR received the bachelor's degree in electrical engineering from the National University of Sciences and Technology (NUST), Islamabad, Pakistan, where he is currently pursuing the master's degree in artificial intelligence with the School of Electrical Engineering and Computer Science (SEECS). With experience as an Artificial Intelligence Research Engineer, he has contributed to various projects in the field of AI for medicine. His research interests include artificial intelligence, computer vision, natural language processing, and deep learning. For more information, please visit his LinkedIn profile at https://linkedin.com/in/waleedbintahir.

SHAH KHALID received the M.S. degree from the University of Peshawar, Pakistan, and the Ph.D. degree from Jiangsu University, China. He is currently an Assistant Professor with the School of Electrical Engineering and Computer Science, National University of Sciences and Technology (NUST SEECS), Islamabad, Pakistan. He has been involved in several research projects in Pakistan and abroad. His research interests include information retrieval, web search engines, scholarly retrieval systems, recommender systems, knowledge graphs, social web, real-time sentiment analysis, web engineering, text summarization, federated search, and digital libraries. For more information, please visit his website at https://sites.google.com/view/shahkhalid.

SULAIMAN ALMUTAIRI received the bachelor's degree in computer science from Bluefield State University, USA, the master's degree in information technology from RMU, USA, and the Ph.D. degree in biomedical informatics from Rutgers University, NJ, USA. He has been a Tenure with Qassim University, since 2005. He is currently a Distinguished Data Governance Consultant and the Director of quality scholarship services with the Ministry of Education. He also holds the position of an Assistant Professor of health informatics with Qassim University. He has held various pivotal roles, including the Vice Dean of planning, development, and quality. He has showcased his expertise as a Data Governance Consultant with prominent tech firms in Saudi Arabia, such as Elm and SITE. Additionally, he served as a Cultural Mission Diplomat representing Saudi Arabia in Canada and contributed globally as a Data Consultant with Development Gateway and UNICEF. His research interests include data governance, bio-informatics, information and policy, and software development.

MOHAMMED ABOHASHRH received the Ph.D. degree in biomedical informatics from Rutgers, The State University of New Jersey–Newark. He is an experienced Associate Professor with a demonstrated history of working in the education management industry. He has skilled in statistics, research, clinical research, data analysis, and healthcare. Currently, he is working on various research grants under the Deanship of Scientific Research, King Khalid University, Saudi Arabia. He has published various research articles in the fields of artificial intelligence, bioinformatics, machine learning, big data, and IOMT in various domains, such as healthcare, disease detection, and diagnosis.

SUFYAN ALI MEMON received the Ph.D. degree in electronic systems engineering from Hanyang University, Republic of Korea, in 2016. He has been an Assistant Professor with the Department of Defense Systems Engineering, Sejong University, Seoul, Republic of Korea, since March 2021. His research interests include tracking, estimation, guidance, navigation, control, machine learning, and artificial intelligence.

JAWAD KHAN received the master's degree in computer science from the Kohat University of Science and Technology, Pakistan, and the Ph.D. degree in computer engineering from Kyung Hee University, South Korea. He was a Postdoctoral Researcher with the Department of Robotics, Hanyang University, South Korea, for three and half years. He is currently an Assistant Professor with the School of Computing, Gachon University, South Korea. His research interests include natural language processing, information retrieval, sentiment analysis/opinion mining, text processing, social media mining, artificial intelligence, machine learning, deep learning, the Internet of Things, and computer vision.