Sustainable Futures 10 (2025) 100827

Contents lists available at ScienceDirect

Sustainable Futures

journal homepage: www.sciencedirect.com/journal/sustainable-futures

Detecting signs of depression on social media: A machine learning analysis and evaluation

Phi Ta [a], Nha Tran [a,b,c], Hung Nguyen [a,*], Hien D. Nguyen [b,c,d]

[a] *Ho Chi Minh City University of Education, Ho Chi Minh City, Viet Nam*
[b] *University of Information Technology, Ho Chi Minh City, Viet Nam*
[c] *Vietnam National University, Ho Chi Minh City, Viet Nam*
[d] *Computer Science Department, New Mexico State University, Las Cruces, USA*

ARTICLE INFO

Keywords:
Depression detection
Mental illness
Social media
Natural language processing
Survey

ABSTRACT

Depression has become a growing concern due to its detrimental effects on both personal functioning and interpersonal relationships. In contemporary society, it is of utmost urgency to research and develop systems capable of detecting symptoms of depression on social media. Our study is not merely a survey, but a comprehensive investigation aimed at uncovering valuable insights and trends in the detection of depression on social media platforms. Our findings present a consolidated map of current methodologies, highlight key trends, and, through experimental results, provide clear performance benchmarks for both post-level and user-level techniques. By integrating insights from both the extensive literature review and practical experiments, this work clarifies existing challenges, establishes performance baselines, and proposes empirically grounded future directions to advance the development of more effective and reliable depression detection systems on social networks. This work opens a promising future for addressing the challenge of detecting depression on social media and contributes to enhancing the effectiveness of depression detection systems, ultimately aiding individuals affected by the adverse effects of depression.

1. Introduction

Depression is a clinical mental health disorder characterized by a persistent feeling of sadness and loss of interest that interferes with daily functioning, negatively affecting the mood, thoughts, and behaviors of individuals. Depressive disorders can affect individuals of any gender or age [1]. Individuals with depression may experience a range of symptoms including disturbed sleep, changes in appetite, feelings of worthlessness, hopelessness, and thoughts of death, along with persistent fatigue and difficulty concentrating [2]. Depression is a major risk factor for suicide [3]. In particular, major depressive disorder impacts 15%–17% of the population, making it extremely common, and is linked to a substantial suicide risk of around 15% [4]. According to World Health Organization (WHO) statistics, an estimated 280 million people globally — approximately to 3.8% of the population, are affected by depression [1]. Of particular concern is the 28% increase in suicide rates in the Americas during this period, while Europe has recorded the highest suicide rate at 12.8 per 100,000 population [5]. Therefore, timely detection and intervention are crucial to mitigate the devastating consequences that depression causes.

The evolution of the internet has significantly fueled the growth and diversification of social media platforms such as Twitter, Facebook, Instagram, TikTok, and YouTube [6], offering users a diverse array of tools through which they can share their personal experiences and emotions. Indeed, users frequently post personal moments and feelings on these profiles [7]. However, these forms of online expression are not always positive. Social media can become an outlet for profound frustration, and in some cases, suicidal ideation. [8]. Notably, individuals suffering from depression tend to express distorted thoughts more frequently than others on social media [9]. Therefore, tapping into the rich resources available through social media platforms can present enormous potential for enhancing patient engagement, treatment quality, and healthcare outcomes [10], especially for examining mental health trends and issues [11].

Detecting depression on social media refers to the process of identifying signs and symptoms of depression or depressive trends in individuals based on their activities and shared content on social media platforms. This task can be approached at two distinct levels: post-level [12–14] and user-level [15–17]. At the post-level, the focus is

[*] Correspondence to: Ho Chi Minh City University of Education, 280 An Duong Vuong Street, Ward 4, District 5, Ho Chi Minh 70000, Vietnam.
E-mail addresses: 4701104156@student.hcmue.edu.vn (P. Ta), nhatt@hcmue.edu.vn (N. Tran), hungnv@hcmue.edu.com, hungnv@hcmue.edu.vn (H. Nguyen), hiennd@uit.edu.vn (H.D. Nguyen).

https://doi.org/10.1016/j.sftr.2025.100827
Received 30 December 2024; Received in revised form 20 May 2025; Accepted 7 June 2025
Available online 24 June 2025

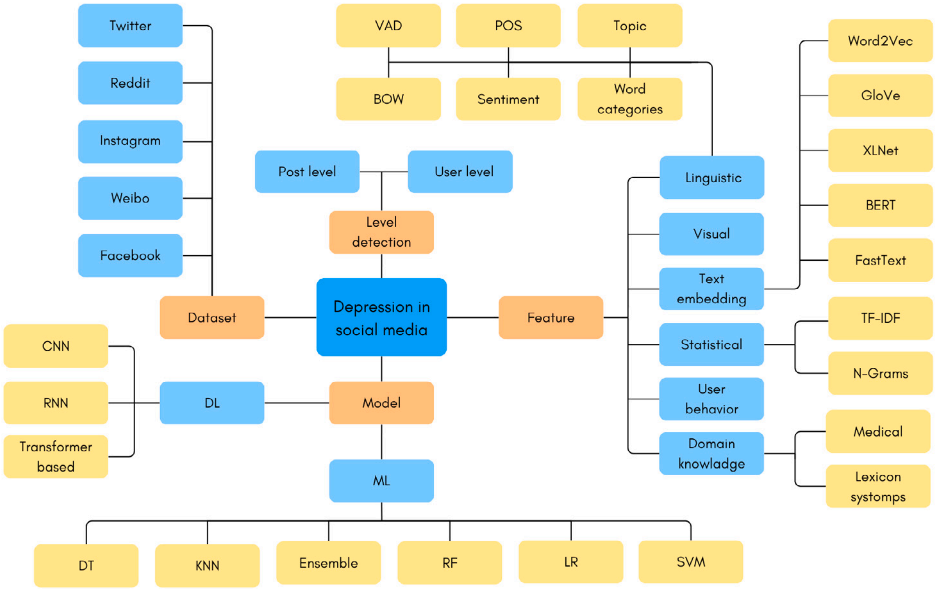

Fig. 1. Overview of depression detection on social media: Datasets, features, models, and detection levels.

on classifying depressive states based on individual social media posts. Each post is examined independently, with the classification process typically emphasizes the linguistic content, emotions, and other features of each post (which may include images) [14,18,19]. In contrast, at the user-level, the analysis involves examining depressive states based on a user's overall behavior and content on social media over a period of time. However, traditional methods of detecting depression have many limitations as they rely one standardized scales that require subjective inputs from patients or clinical diagnoses made by healthcare professionals [20]. Meanwhile, artificial intelligence (AI) is emerging as a promising technology in healthcare [21]. AI and data analysis methods have signaled a significant shift in the decision-making processes within the healthcare sector [10]. Therefore, analyzing users' linguistic patterns, online behaviors, and emotional cues on social media platforms, combined with advanced AI algorithms, can yield valuable insights into their mental health status. To this end, researchers have explored numerous techniques and features specific to each level, resulting in a diverse body of work. Consequently, a systematic synthesis is indispensable for consolidating this broad range of findings, mapping the state-of-the-art, identifying established practices, and outlining key challenges across the field. While efforts toward this systematic synthesis have been undertaken through existing reviews, they often demonstrate limitations in scope or focus. For example, study [22] systematically reviewed the literature to synthesize text-based methods for depression detection on social media, highlighting the potential of deep learning for early detection. However, being limited only to text data caused this study to overlook other multimodal factors or methods. In [23], researchers focused on surveying machine learning applications, pointing out important challenges such as sampling and ethical issues, thereby guiding future research. Nevertheless, the inclusion of only 17 studies over 30 years and the time limit to 2020 indicates a narrow survey scope that does not fully reflect recent advancements. Similarly, [24] surveyed 34 studies on large language models (LLMs) such as RoBERTa and BERT, demonstrating their effectiveness in classifying depression symptoms. However, this study was limited to English articles and focused solely on LLMs, thus omitting the evaluation of other important methods outside of LLMs. Additionally, the number of studies surveyed was also relatively small.

This study represents a significant advancement in the effort to explore the problem of detecting depressive states on social media. We comprehensively searched and synthesized studies from IEEE Explore, ACM Digital Library, Science Direct, Springer Nature, PubMed, and other health science journals. Results from over 70 studies published between 2017 and 2024 were meticulously selected. A thorough survey was undertaken to capture relevant information and identify notable trends. As shown in Fig. 1, the aspects considered included datasets, feature types, and modeling approaches integrated into the process of detecting depression from social media platforms. A key contribution of this work is its comprehensive synthesis of methods for detecting signs of depression on social media, covering both post-level and user-level analyses — a perspective that has been largely overlooked in previous surveys. This examination yields an overview of the diverse and rich methodologies currently employed, establishing a solid foundation for a deeper understanding of this field. Complementing this comprehensive literature review, our study further provides empirical validation through experimental evaluations. We experimentally evaluate a range of modern approaches, including deep learning and hybrid multi-feature analysis on established datasets, performing distinct analyses at both the post-level and user levels. At the post-level, it utilizes state-of-the-art Transformer-based language models to accurately classify individual posts by interpreting complex linguistic nuances and context, achieving high precision on platforms like Twitter and Reddit. Complementing this, the user-level analysis employs deep learning models (like LSTM and GRU) integrated with a diverse set of features, including emotional expression, posting times, and topics to build a comprehensive profile of user behavior indicative of depressive states. Experimental results provide clear performance benchmarks, revealing the relative strengths of each approach and establishing a baseline for designing more effective future detection methods. Significantly, the outcomes derived from our experimental evaluations closely mirror and substantiate the key trends and methodologies identified during our comprehensive literature survey. From these results, we also contribute to clarifying existing challenges as well as proposing future directions to improve the performance of depression detection systems on social networks.

The next section of this paper is an overview of the general structure of the depression detection system in social media. Section 3 presents and analyze the experimental process. Section 4 is experimental results and discussion. The last section concludes results of this study.

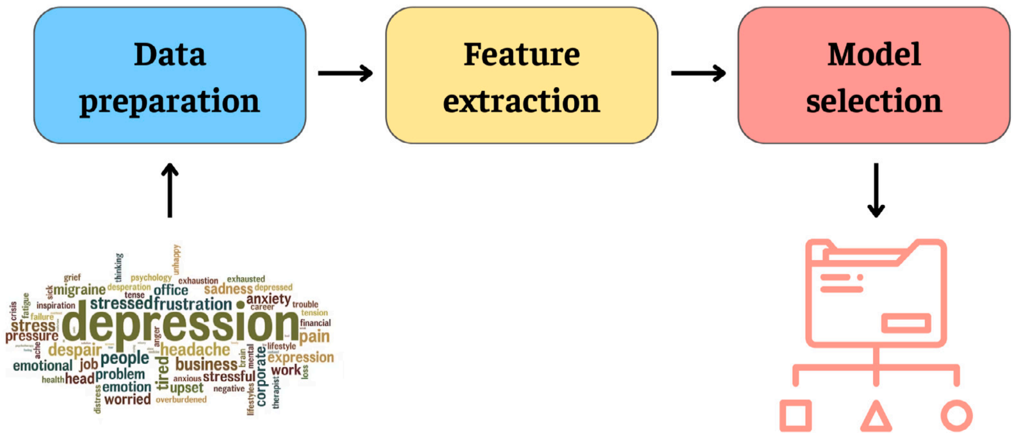

Fig. 2. General structure of depression detection systems on social networks.

2. The methods for depression detection system in social media

2.1. General structure of depression detection systems on social networks

As illustrated in Fig. 2, detecting depression on social media, whether at the post or user level, requires a meticulous classification system encompassing three crucial steps:

Step 1 - Data Preparation: This stage involves data collection and preprocessing. The data collection process ensures the system has sufficient information to carry out its task. The diversity and completeness of the data are crucial to build a robust and accurate classification model. Data preprocessing helps standardize formats, remove noise, and mitigate issues that could affect data quality. This enhances model performance on training data and strengthens the model's generalization and prediction capabilities for new data.

Step 2 - Feature Extraction: Different features will be selected and extracted depending on the chosen approaches and specific characteristics of the dataset. This process requires a profound understanding of the data. Extracting relevant and appropriate features improves the model's performance.

Step 3 - Model Selection: Suitable models will be chosen based on the collected data and extracted features. Model selection also involves effectively managing cost and resource efficiency integrated into the system. An appropriate model ensures performance and resource-saving in the system's utilization.

2.2. Review of datasets

This section systematically examines the foundational datasets and data collection methods used in social media-based depression detection research. We review the evolution of dataset characteristics, identify key limitations, and discuss innovative efforts aimed at addressing these challenges, thereby establishing the context and rationale for a more comprehensive synthesis and in-depth analysis in this field.

Recently, online platforms and forums are now generating vast data offering deep insights into users' social, psychological, and behavioral traits [10], often publicly shared. Therefore, constructing datasets for detecting depression on social media is relatively open. Table 1 provides information about the datasets used for this task. To harness this wealth of online data, the foundational data collection process for depression detection research typically relies on two main methods: utilizing official Application Programming Interfaces (APIs) provided by social media platforms [25–27], or applying scraping techniques to extract publicly available data [28]. Among these platforms, X (formerly Twitter) and Reddit emerge as particularly popular data sources. This preference largely stems from the fact that both platforms offer relatively open and well-documented APIs, allowing researchers to systematically access and collect large volumes of textual data (such as posts and comments) along with associated metadata. APIs offer the advantages of greater stability and adherence to platform terms of service, making them the preferred method for data access; however, scraping may be necessary when API limitations restrict access to data.

Observation of Table 1 shows a diverse landscape in terms of dataset size, however, a prominent challenge arises from the significant proportion of datasets remaining relatively small in scale, often comprising only a few thousand to tens of thousands of samples [14, 26,41]. This limited scale presents substantial obstacles to the effective application of deep learning. With such modest data volumes, models often struggle to capture the full richness and complexity of linguistic expressions associated with depressive states, making them susceptible to overfitting and significantly impairing their ability to generalize to unseen data. This represents a core area requiring further research, as models trained on limited data often fail when applied to more diverse real-world situations. To overcome these limitations, several recent studies have made significant efforts to construct and utilize large-scale datasets for depression detection. For instance, the dataset introduced in [29] comprises an extensive collection of over 11.8 million tweets, 1106 images, and 553 bio-descriptions from users exhibiting depressive behaviors, alongside a control group with more than 16.6 million tweets, 1140 images, and 580 bio-descriptions. This dataset offers a rich multimodal foundation for deep learning models to explore both textual and visual cues. Similarly, the work in [40] assembled a dataset consisting of 531,453 posts from 892 users, enabling more robust user-level modeling of depressive language patterns. Furthermore, the dataset presented in [43] includes over 500,000 posts and comments from 887 individuals, among whom 135 have been clinically diagnosed with depression, offering valuable ground truth for supervised learning approaches. These large collections of data are important steps forward, letting us build better models. However, even these big datasets can have problems. They might be biased based on how the information was gathered or labeled. Also, they might not include enough information about all different groups of people. Because of this, there is still ongoing work needed to figure out how to create datasets that are large, fair to everyone, and properly represent all groups.

Another notable trend concerns the language of the datasets. The majority of existing datasets are predominantly in English [29–31], significantly restricting the applicability and effectiveness of depression detection systems across diverse linguistic and cultural contexts. This language bias represents a critical area needing research because it hinders the development of equitable mental health tools globally. To address this limitation, considerable efforts have been made to create social media depression datasets in languages other than English. In the study conducted by Cha et al. [25], the authors collected data from Twitter users, with users having languages primarily in Korean,

Table 1
Social media and corresponding datasets used in depression detection research.

Platform	Study	Description
Twitter	[12]	5000 tweets. 4 categories: Depressive (984 tweets), Non-depressive (2930 tweets), Ambiguous (699 tweets) and Incomplete sentence (387 tweets)
	[13]	64,000 tweets
	[14]	3082 posts depressed and 3082 posts non-depressed
	[16]	1402 users depressed and 5160 users non-depressed
	[17]	5899 users labeled positive, 508786 tweets from positive users, 5160 users labeled negative, 2299106 tweets from negative users
	[19]	489 depressed tweets, 1865 non-depressed tweets
	[29]	Depression: 11,890,632 tweets, 1106 images, and 553 bio-descriptions. Control group: 16,623,164 tweets, 1140 images, and 580 bio-descriptions
	[30]	9300 tweets
	[31]	447,856 tweets from 2159 depressed users, 1349,447 tweets from 2049 non-depressed users
	[25]	Korean: 14 thousand users, 921 thousand posts. English: 210 thousand users, 10 million posts. Japanese: 216 thousand users, 15 million posts
	[32]	5 million tweets were collected from the 2132 users with depression and 4.2 million tweets from the 2036 users with no declared depression
	[33]	61,938 tweets from 1464 users
	[34]	2626 users depressed, 5373 users non-depressed
	[35]	222 tweets, 1522 retweets and 16 hashtags, collected from 192 Twitter users, 50 of whom are moderately depressed, 74 are slightly depressed, and 68 show no sign of being depressed
	[36]	3754 tweets
	[37]	Social media data from 946 participants
	[38]	3703,135 texts
Reddit	[14]	6500 depressed posts and 6500 non-depressed posts
	[39]	135 depressed users and 752 control users
	[40]	531,453 posts of 892 different users
	[41]	1293 depression-indicative posts and 549 standard posts
	[42]	Reddit posts of 9210 depressed users and 108,731 control users
	[34]	214 depressed users and 1493 non-depressed users
	[43]	887 subjects, of which 135 have been diagnosed with depression, and encompasses more than 500,000 different posts and comments
	[44]	1293 depression-indicative posts and 548 standard posts
	[45]	9210 depressed users and 107 274 control users
	[46]	44035 posts from 90 users
	[47]	2500 depressed posts and 2500 non-depressed posts
	[48]	544 depressed users and 3809 non-depressed users
Facebook	[19]	826 depressed posts and 1461 non-depressed posts
	[49]	1105 posts from 22 depressed and 13 non-depressed users
	[50]	5000 posts
	[51]	4144 depressed comments 3001 and non-depressed comments
Instagram	[52]	43,950 photographs from 166 Instagram users, 71 of whom had a history of depression
	[53]	9458 posts from 260 users have depressive tendencies. 22286 posts from 260 users have non-depressive tendencies
Weibo	[26]	15,879 Weibo posts from 10,130 distinct Weibo users
	[54]	135 user depressed and 252 user non-depressed
	[55]	3711 depressed users and 19,526 non-depressed users
	[56]	22,245 normal users and 10325 depressed users
	[57]	965 users with depression and 58,265 microblogs. 903 users without depression and 52,787 microblogs

English, or Japanese. A large dataset was constructed, including 14 thousand Korean user accounts, 210 thousand English user accounts, 216 thousand Japanese user accounts, 921 thousand Korean posts, 10 million English posts, and 15 million Japanese posts. In a study on depression detection in the Thai community on the Facebook social media platform, Katchapakirin et al. built a dataset from 35 Facebook users over 18 years old [49]. The results yielded 1105 posts from 22 users with depression and 13 users without depression.

Exploring features beyond the text is a promising innovative approach [29,58]. This approach can provide a more in-depth understanding of how depressive states influence social relationships and diversify the analysis process, shifting the focus from text-only to incorporating insights from images and interactions. However, effectively integrating and interpreting multimodal data, while also addressing related privacy and ethical concerns, remains a dynamic research area with significant gaps to be filled. Moreover, the challenge of imbalanced data, with fewer instances of depressive states compared to non-depressive states, is a common issue faced by most studies [12,16,19,48].

2.3. Review of features

Feature extraction is fundamental to the decision-making process in both the human mind and deep learning architectures [59], serving as the cornerstone for informed and effective outcomes. There are no standardized general methods for detecting mental disorders; instead, effective and early detection relies on hybrid methodologies that utilize data from physiological signals, behavioral patterns, and online social media platforms [60]. However, the inherent complexity and unstructured nature of social media data render feature extraction a particularly challenging task within this domain. To navigate these challenges, researchers explore a wide spectrum of features derived from this data, broadly categorized as linguistic, domain knowledge-based, word embeddings, statistical, user behavioral, and visual. The features used for predicting depression can be categorized into the following types: linguistic, domain knowledge, word embedding, statistical, user behavior, and visual. Table 2 provides a comprehensive overview of the different types of features commonly used in depression detection research, along with the corresponding studies that utilized them.

Linguistic features comprise Word categories, Valence-arousal-dominance (VAD), Topic Modeling, Part of speech (POS), Bag-of-words (BoW), and Sentiment analysis.

- Word categories: This is a robust tool for language analysis and collecting psychological information from text. It calculates the occurrences of words and phrases belonging to emotional categories, word types, themes, personality traits, and other features in the text. Some standard tools for extracting these features include Linguistic Inquiry and Word Count (LIWC) [40,51], Empath [61], and Chinese Affective Lexicon Ontology (CALO) [54].
- Valence, Arousal, and Dominance (VAD): Affective Norms for English Words is a database consisting of nearly 14,000 English words evaluated by numerous volunteers based on three emotional dimensions: valence, arousal, and dominance [62], scored on a scale from 1 to 9. Valence reflects positivity or negativity, arousal indicates excitement or fatigue, and dominance reflects control or submissiveness. In Ghosh and Anwar's study, the authors extracted VAD features to explore emotional features [16].
- Topic modeling: This is a method in natural language processing (NLP) for automatic classification, identification, and extraction of central topics from a text corpus. The goal of topic modeling is to find hidden topics within a text dataset without the direct intervention of humans. Topic models often use unsupervised machine learning techniques to classify individual words or phrases into different topics. The most common method in topic modeling is Latent Dirichlet Allocation (LDA). Zogan et al. [31] used LDA to extract latent topic distribution from user tweets.

Table 2
Summary of feature types, extraction methods, and related studies in social media-based depression detection.

Type of features	Feature extraction	Study
Linguistic	Word categories	[29,37,40,44,51,54,61,63]
	VAD	[16,17,31]
	Topic	[16,17,31,44,63]
	POS	[36,61,63]
	BoW	[11,18,34,45,50,64]
	Sentiment	[16,36,61,63]
Word embedding	Word2vec	[25,40,47,53,57,60,65]
	Fasttext	[14,15,40,46,50]
	GloVe	[15,40]
	BERT	[25,31,58,66]
	XLNet	[47,56,67]
Statistical	TF-IDF	[45,47,50,57,60,63–65]
	N-Grams	[11,16,29,36,44,68]
Domain knowledge		[17,31,68,69]
User behavior		[16,17,31,49,53,54]
Visual		[29,52,53,58]

- Part of Speech (POS): This is a grammatical concept used to categorize words in a sentence based on their function and role. It is an essential step in syntactic analysis, aiding in understanding sentence structure and meaning. In the study by [36], a POS vector was constructed by counting the occurrences of various Part of Speech tags.
- BoW: This is a widely used NLP technique representing text as a set of words. It disregards the order of words and focuses only on the frequency of each word in the text. Some studies have used BoW in the feature extraction process such as [18,45].
- Sentiment analysis: This is a concept in natural language processing (NLP) used to determine the predominant emotion of a text by classifying it into different emotional segments, such as positive, negative, or neutral. Popular tools used for sentiment analysis include Valence Aware Dictionary for sEntiment Reasoning (VADER) or TextBlob - a Python sentiment analysis library using Natural Language ToolKit (NLTK). In the study of Tlachac and Rundensteiner [61], the authors used Textblob to create a polarity score as well as a sentiment score based on words in the tweet. In addition Skaik and Inkpen [63] also use VADER to exploit features related to emotions.

Word Embedding: Word embedding is a technique for representing words as numerical vectors. These vectors can represent words' meaning, relationships between words, context, and more. Commonly used word embedding models for this task include Word2vec [47,60], Fasttext [14,46], GloVe [15], BERT [25], and XLNet [56].

- Word2vec: Word2vec is a neural network model representing each input word as a numerical vector. With Word2vec, words that appear together in a sentence tend to be related in meaning, and semantically similar words are positioned close to each other in the vector space. Therefore, it captures the semantic relationships between words and the context of a sentence.
- Fasttext: This model provides word representations based on embeddings of sub-words (n-grams). Thanks to the ability to create embedding for n-grams, if an out-of-vocabulary appears, Fasttext can still create embedding for this word.
- GloVe: This unsupervised learning algorithm constructs word-word co-occurrence matrices by minimizing the difference between pairs of representation vectors for words and their dot product. GloVe captures semantic relationships between words across the entire text.
- BERT: A pre-trained model widely used for various NLP tasks. With its bidirectional text processing capability, BERT captures the meaning and context of each word, effectively handling polysemy and generating vector embeddings based on word context.
- XLNet: Another pre-trained model with capabilities similar to BERT (bidirectional text processing). Recognizing context well, XLNet produces vector embeddings based on the context of a word.

Statistical:

- TF-IDF: TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical method indicating the relevance of a word based on its frequency and importance. It is a simple and easily deployable method with good computational efficiency. Many studies have leveraged TF-IDF to construct feature sets for depression classification models on social media [50,64].
- N-Grams: N-Grams are a sequence of n consecutive words in a text. The frequency of occurrence of these n-grams is calculated to understand features related to words. Numerous studies have utilized N-grams for feature extraction [11,44,68].

Domain Knowledge Features: These features are extracted based on domain-specific knowledge about depression [70]. This process explores features related to medical symptoms and illnesses. In the study by Zogan et al. [17], the researchers counted the number of times depression symptoms were mentioned in users' tweets. The authors also concentrated on antidepressant medications, creating a vocabulary from the Wikipedia page "Antidepressant" and counting the number of listed names for antidepressant drugs. Besides, the study in [69] proposes a novel methodology for quantifying user amplification factors and content creation metrics within social networks. These factors have gained significant importance in the context of marketing strategies. Additionally, the study delves into the measurement of content creation scores, assessing the alignment between user-generated content and their expressed passions within the social media ecosystem.

User Behavior Features: Refer to the frequency and type of behavioral activities and user interactions. Features related to user behavior may include the number of comments, followers, and posting frequency. In the study [17], to extract features related to user behavior on social media, the authors calculated the distribution of posting times for each user by counting the number of tweets users posted every hour. In [54], the authors used a combination of 6 behavior features: frequency of forwardings, comments, praises, sending postings, mentions, and posting a time of microblogs.

Visual Features: These features are extracted from image data. These features are often combined with text-based data features to enhance the training of models. Many studies have exploited visual features to augment features for the depression detection process on social media [52,53].

The evolution of feature engineering for depression detection via social media analysis reveals a trajectory from rudimentary linguistic and statistical metrics towards complex, multimodal, and context-aware representations. Initial research predominantly relied on linguistic features, such as word categories derived from lexicons like LIWC [40,51], sentiment analysis scores [61,63], POS tags [36], and basic statistical methods like BoW [18] or TF-IDF [50,64]. While offering some interpretability and ease of implementation, these early approaches were significantly limited by their inability to capture deep semantic meaning, contextual nuances, and the inherent sparsity of textual data. Furthermore, classifying individuals based solely on lingual dialects presents considerable challenges [71], and established linguistic markers, such as certain LIWC category distributions, were found not to be equally informative across diverse demographic groups, raising concerns about generalizability and potential bias [72].

To address semantic limitations, word embedding techniques like Word2vec [47,60], GloVe [15], and FastText [14,46] emerged, providing dense vector representations that better captured lexical relationships. However, a key limitation of these static embeddings was their failure to resolve polysemy effectively. The subsequent advent of deep learning ushered in contextualized embeddings from pre-trained language models such as BERT [25,58] and XLNet [47,56]. These models significantly advanced the field by capturing context-dependent word meanings and long-range dependencies within user posts. Nonetheless, applying models pre-trained on generic text data faces limitations when confronted with the highly domain-specific language prevalent in mental health discussions online, necessitating effective domain adaptation strategies [73]. Concurrently, this phase of leveraging complex deep learning architectures also introduced significant challenges related to increased model complexity and reduced interpretability.

Realizing that just considering text content is not enough to fully understand a person's mental state, later studies tried to add different kinds of information. They started using information based on expert knowledge about depression, for example, looking for mentions of specific symptoms (like from the DSM-IV list) or names of antidepressant drugs [17]. At the same time, they also considered information about user behavior, such as user interactions, the times they post, and the structure of their friend networks and connections [17,54]. Adding this information helps find more specific signs about health and behavior that analyzing text alone might miss. This trend of combining multiple data types also led to multimodal approaches, meaning adding information from pictures users share [52,53] along with text and behavior data. The purpose is to get a more comprehensive view and assess the user more accurately. Despite these advancements, fundamental challenges persist. The inherent nature of the data source remains a significant hurdle, as users' expressions on social media platforms may not fully, objectively, accurately, or consistently reflect their actual mental state or depression-related symptoms [74]. Furthermore, effectively combining different data modalities and ensuring robust interpretability for complex models are issues that require more thorough research.

2.4. Review of techniques

To address the problem of depression detection on social media, various approaches have been proposed, including traditional machine learning models and deep learning models. Table 3 presents commonly used models in studies focused on depression detection on social media.

Initial studies often applied traditional machine learning models, including popular algorithms such as Support Vector Machine (SVM), Random Forest, Logistic Regression, Naive Bayes, Decision Tree, and K-Nearest Neighbors [11,18,26,51,57,64,75]. These models have the advantage of being relatively easy to implement and interpret, while proving effective in classification tasks when provided with high-quality, clearly defined input feature sets [83]. However, their inherent weakness is that they rely too much on manual feature extraction techniques [83]. Furthermore, with the increasing scale and complexity of social media data, these shallow machine learning models may lack the capability to effectively learn complex patterns and deep semantic representations hidden within the data [84]. They can also be sensitive to class imbalance, a common issue in mental health data [75].

To overcome these limitations and better exploit the richness of the data, researchers have increasingly turned to DL models. Although DL models typically require higher computational resources, they offer strong learning capabilities and can automatically extract complex features from raw data [84], thereby reducing the reliance on manual feature engineering. Initially, Convolutional Neural Networks (CNNs) were employed, primarily for their ability to capture local patterns in text, akin to n-gram features [15,65]. Subsequently, Recurrent Neural Networks (RNNs) and their variants such as Long Short-Term Memory and Gated Recurrent Units gained popularity due to their effectiveness in processing sequential data and modeling long-term dependencies [16,50]. Among these, Bi-directional LSTM (BiLSTM) models combined with word embeddings have shown particularly promising results [40]. More recently, the Transformer architecture has become dominant, thanks to its self-attention mechanisms that can model complex, long-range contextual relationships more effectively than RNN-based approaches [47,82]. This architecture now forms the foundation of many state-of-the-art large pre-trained language models.

Table 3
Summary of traditional machine learning and deep learning techniques applied in depression detection research.

Type of technique	Model	Study
Traditional machine learning	Logistic Regression	[11,26,57,61,64,75,76]
	Naive Bayes	[18,57,61,64,75,76]
	Decision Tree	[11,29,51,57,75]
	Random Forest	[11,26,29,50,57,61,64,75,76]
	K-Nearest Neighbors	[50,51,61,75]
	Support Vector Machine	[11,18,26,29,50,51,57,64,75,76]
	Ensemble	[11,29,51,61,77–79]
Deep learning	Convolutional Neural Networks Based	[15,65,71,80,81]
	Recurrent Neural Network Based	[16,50,65,71]
	Transformer Based	[47,82]
	Multilayer Perceptron	[11,26,29,44]

2.5. Post-level and user-level

As mentioned earlier, the problem of detecting depression on social media involves two levels: post-level and user-level. Both detecting depressive text and identifying depressed users are highly challenging tasks [71]. Table 4 details some representative studies in this task, including social media platforms, utilized features, approach methods, and best performance.

Post-level analysis focuses on evaluating the content of individual text units, such as a tweet or a Facebook status, to determine the immediate presence of depressive signs. Initially, this approach often relied on classic machine learning models like SVM or Multinomial Naive Bayes, combined with traditional text vectorization techniques such as BoW or TF-IDF, as seen in the works of [18,50]. Over time, there has been a clear shift towards applying deep learning architectures and more sophisticated word embedding methods. Models like BiGRU [50], or the combination of CNN and LSTM [14] using dense vector representations like Fasttext, have demonstrated superior effectiveness in capturing semantics. Other studies also integrate linguistic and psychological features such as LIWC, topic modeling (LDA), and N-Grams with neural networks like MLP [44], achieving high performance. However, the inherent limitation of the post-level approach is that it only provides a snapshot of the user's psychological state at a specific point in time. This can lead to inaccurate assessments by overlooking the persistent and fluctuating nature of depressive states over time [16,39,40], as well as difficulties with the short, unstructured, and sometimes ambiguous characteristics of social media text [81].

Table 4
Summary of feature types, detection methods, and performance metrics for post-level and user-level depression detection.

Type	Study	Platform	Feature	Method	Performance
Post-level	[50]	Facebook	BoW	Kernel SVM-radial basis function	74% (Acc), 0.73 (F1), 0.73 (Precision), 0.74 (Recall)
			TF-IDF	Kernel SVM-linear	78% (Acc), 0.76 (F1), 0.77 (Precision), 0.78 (Recall)
			Text embedding (Fasttext)	BiGRU	81% (Acc), 0.81 (F1), 0.81 (Precision), 0.81 (Recall)
	[18]	Twitter	BoW	SVM	79% (Acc), 0.7973 (F1), 0.804 (Precision), 0.79 (Recall)
				Multinomial Naive Bayes	83% (Acc), 0.8329 (F1), 0.836 (Precision), 0.83 (Recall)
	[14]	Reddit	Text embedding (Fasttext)	CNN + LSTM	0.87 (Acc)
		Twitter			0.88 (Acc)
	[44]	Reddit	Word categories (LIWC) + LDA + Bigram	MLP	91% (Acc), 0.93 (F1), 0.90 (Precision), 0.92 (Recall)
User-level	[54]	Weibo	Behavior + Word categories (CALO)	Multi-kernel SVM	83.46% (Acc), 0.76 (F1), 0.76 (Precision), 0.77 (Recall)
	[40]	Reddit	Text embedding (Fasttext) + Domain knowledge + Word categories (LIWC)	BiLSTM	0.81 (F1)
	[16]	Twitter	Word categories (LIWC) + Sentiment + VAD + Topic modeling (LDA) + Behavior + N-Gram	LSTM	87.14% (Acc), 1.42 (MSE)
	[58]	Twitter	Text embedding (BERT) + Visual	Neural network	88.4% (Acc), 0.936 (F1), 0.903 (Precision), 0.870 (Recall)
	[63]	Twitter	Text embedding (Fasttext)	CNN	91% (Acc), 0.898 (F1), 0.882 (Precision), 0.914 (Recall)

To overcome the temporal context limitations of post-level analysis, the user-level approach was developed. This method assesses an individual's depression risk based on the overall analysis of their activity and profile over a longer period. This requires processing and integrating a larger and more diverse volume of data, including not only posts but also interactions, profile information, and even multimodal data such as images. The core characteristic of this level is the combination of various feature types, ranging from behavioral features, psychological lexicon categories (LIWC, CALO), sentiment analysis (VADER), topic modeling (LDA), to advanced text embeddings and visual features [16,40,54,58]. Complex deep learning models such as LSTM, BiLSTM, CNN, and deep neural networks, often combined with powerful pre-trained embedding techniques like BERT [58], have become mainstream tools, enabling the modeling of long-term trends and user mood dynamics [63]. Despite the potential to provide more comprehensive and accurate assessments, user-level analysis faces significant challenges. These include the complexity of managing and processing large, heterogeneous datasets [49,85], difficulties in accurately determining the onset and duration of depressive episodes, especially when users experience alternating depressive and non-depressive periods [86]. Furthermore, analyzing a user's entire history is speculative [52], and different observation time frames are applied in studies [52,63]. Another major challenge is the lack of interpretability in complex deep learning models, making it difficult to understand the decision-making mechanism [87], along with ethical concerns and unintended consequences from automated diagnostic systems [88].

Overall, the field of depression detection on social media has witnessed significant progress, with an increasing trend towards using deep learning models and sophisticated feature representation techniques at both levels of analysis. The performance of systems has continuously improved; however, no single method proves optimal for all situations, emphasizing the importance of selecting methods appropriate for the specific context and data. The integration of diverse feature types, particularly at the user level, shows great potential. Nevertheless, challenges related to complex data processing, modeling temporal dynamics, ensuring interpretability, and addressing ethical issues remain key factors that require continued research and resolution in the future.

3. Challenges and future directions

3.1. Datasets

3.1.1. Data scale and representativeness

As extensively discussed in Section 2.2, much of the current research still relies on relatively small-scale datasets. This limits the capacity of deep learning models to fully capture the complexity of language expressions related to depression, leading to the risk of overfitting and poor generalization ability on real-world data. Therefore, there is a need to construct larger-scale datasets that are more diverse in terms of demographics, language, and culture. Additionally, techniques such as data augmentation and transfer learning should be leveraged to optimize the use of existing data. In particular, promoting collaborative and responsible data sharing initiatives is crucial. Some privacy-preserving measures, such as de-identification and requiring IRB approval for data access, have facilitated the sharing of some data among research groups [72].

3.1.2. Reliability of data collection and labeling

In the process of mining social media data for depression, the identification of depression status often relies on indirect methods such as keyword-based searches or user self-reporting [26,29,89]. These methods are prone to generating noisy labels, missing subtle or implicitly expressed cases, and lack clinical validation. The stigma surrounding depression often leads individuals to conceal their struggles and postpone seeking help [86]. Meanwhile, collecting control group data poses significant challenges, as the samples may include individuals with depression who do not openly disclose their mental health status on their profiles [32], as well as uncertainty regarding whether users mentioning their condition have received a formal diagnosis [42]. Developing labeling methods with the involvement of psychological experts is needed. Combining social media data with more reliable information sources (e.g., standardized clinical questionnaires) can create higher-quality ground truth. Illustrating this principle, one study generated more accurate labels than typical social media mining by screening participants for depression using the standardized CES-D clinical scale and confirming the absence of depression history in the control group before collecting their Instagram data [52]. Additionally, researching models robust to label noise is also a viable approach.

3.1.3. Data imbalance

A common challenge in depression detection datasets is the severe class imbalance, where the number of samples representing the depressive state is significantly fewer than those representing the non-depressive state. This situation directly results in machine learning models being biased towards the majority class, reducing their predictive effectiveness for the minority class (which is the primary target class of interest). It is necessary to apply advanced techniques for handling imbalanced data, including resampling methods such as oversampling the minority class or undersampling the majority class, or designing special loss functions aimed at minimizing the negative impact of data imbalance.

3.1.4. Unstructured and noisy nature of data

Analyzing social media data presents significant challenges due to its inherently noisy and unstructured nature [16]. This data typically contains abundant slang, emojis, typographical errors, and non-standard forms of expression, all of which pose difficulties for traditional natural language processing (NLP) techniques. Consequently, there is a pressing need to develop more robust NLP preprocessing and modeling techniques capable of adapting to the unique diversity and noise inherent in social media language. Specifically, leveraging advanced approaches like contextual embeddings and attention mechanisms is crucial for achieving a deeper understanding of the underlying meaning embedded within this complex data.

3.2. Feature engineering and utilization

Extracting meaningful and effective features from raw data is a foundational yet challenging step in social media depression detection research. Although selecting and combining multiple features can improve model performance [44], appropriately selecting features to remove redundant or irrelevant ones is also crucial, as this leads to reduced feature space dimensionality, shorter training times, and enhanced model accuracy and generalization [90]. But, many features are designed and selected in an ad-hoc manner, lacking a strong connection to psychological or psychiatric theories regarding how depression manifests through language and behavior. This leads to difficulty in explaining why some features yield predictive performance while others do not, reducing the model's reliability and potential for improvement.

Future development directions should focus on enhancing close interdisciplinary collaboration between computer science and psychology and psychiatry experts. This collaboration aims to identify online linguistic and behavioral markers that are genuinely clinically significant. Concurrently, there is a need to systematically integrate specialized knowledge repositories such as psychological lexicons or diagnostic standards (such as DSM — Diagnostic and Statistical Manual of Mental Illnesses) into the feature design and selection process itself. Prioritizing the development and use of highly interpretable features is also a key factor for enhancing transparency and trustworthiness.

3.3. Interpretability

Although many modern machine learning models, particularly deep neural networks, have achieved high performance in detecting depression from social media data, they often operate as black boxes. The lack of clear explanations for why a specific prediction is made not only reduces the model's trustworthiness — especially in sensitive contexts like mental healthcare — but also poses challenges in debugging, optimizing the model, and identifying potential biases. Recent studies have explored various approaches to build interpretable depression detection models from social media data. The research by [74] focuses on applying Large Language Models (LLMs) to automatically fill the Beck Depression Inventory (BDI) questionnaire based on Reddit user posts, while simultaneously generating textual explanations for each prediction via prompt engineering techniques. [17] propose MD-HAN, a hybrid architecture combining Hierarchical Attention Networks (HAN) to analyze text content and an MLP network to process multi-aspect features (such as behavior, emotion, topics), with explainability based on HAN's attention weights. Meanwhile, [87] introduce METN, a multimodal transformer network based on Temporal Convolutional Networks (TCN), emphasizing the integration of temporal factors with image and text data, and utilizing attention maps to enhance interpretability. All of these methods aim to improve the reliability and transparency of automatic depression detection systems on social media platforms.

Despite advancements in interpretable depression detection, current methodologies exhibit significant limitations. Models employing Large Language Models (LLMs) can generate natural language explanations, yet they face inherent risks regarding faithfulness due to the potential for hallucination, yielding interpretations that may be inaccurate or lack grounding in empirical data evidence. Consequently, the evaluation of their explainability often relies on indirect metrics insufficient for confirming content reliability in critical medical applications. Furthermore, a substantial and largely unaddressed challenge persists in elucidating the complex interactions among multi-aspect features or across different data modalities within sophisticated hybrid or multimodal architectures. These collective deficiencies underscore the urgent need for more robust interpretability frameworks and stringent evaluation standards. To this end, established explanation methods such as SHAP or LIME could be employed to provide a more granular analysis of individual feature contributions to prediction outcomes. Beyond correlational insights, a pivotal future direction involves shifting focus towards identifying clinically meaningful causal factors, a pursuit expected to enhance not only model reliability and trustworthiness but also its practical utility for informing effective diagnosis and intervention strategies.

3.4. Multimodal approach

Depression is a complex condition that can manifest through diverse information channels, not limited to just speech or text. From images, videos, and voice, to the way a person interacts on social media, all can contain important signs. One of the main limitations is that the exploration of these diverse data sources is still very restricted. The majority of current research still focuses primarily on text analysis (e.g., posts, comments, message) [13,47,51]. Although text provides valuable information, relying solely on it can overlook other subtle signs. Efforts to combine data from multiple modalities, such as images, video, audio, or network interaction data, are still quite rudimentary. Even when integration occurs, it often stops at combining two basic modalities, text and image, failing to truly harness the potential of the fuller information picture [29,58].

To enhance the effectiveness of depression detection systems, it is necessary to invest in researching and developing more robust and flexible multimodal fusion techniques. These techniques must be capable of intelligently integrating information from diverse, heterogeneous sources, capturing the interactions and complementary aspects among the modalities. Leveraging advanced models such as deep learning models based on attention mechanisms is also a potential direction. These models can automatically focus on the most important modalities or features in each specific case, thereby enhancing the overall performance of the depression detection system. In parallel, the development of high-quality, carefully annotated multimodal datasets is a crucial factor. These datasets will serve as the foundation for training and evaluating complex analytical models.

3.5. Ethical and privacy considerations

Leveraging data from social media for mental health research opens up many opportunities, but simultaneously raises serious concerns about ethics and privacy. The sensitive nature of personal information related to mental health demands utmost caution during collection, storage, and analysis. A major challenge is that users are often not fully aware or have never explicitly consented to their data being used for this research purpose. This situation carries a significant risk of personal information being misused, potentially leading to discrimination or inadvertently causing further harm, anxiety, or stigma for individuals already in vulnerable states. This risk is compounded by the fact that even when participants were informed during dataset construction, guaranteeing strict anonymity remained nearly impossible [52].

To address these complex issues responsibly, building and implementing a robust ethical and technical framework is essential. The focus should be on developing and strictly applying advanced technological techniques aimed at protecting privacy. The process for obtaining user consent must be designed and must ensure clarity, transparency, and be truly understandable, helping users make fully informed decisions about how their data will be used and protected. Furthermore, whenever the research goals permit, prioritizing the use of anonymized or aggregated data instead of potentially identifiable data is a crucial principle.

Table 5
Class label and corresponding number of tweets in Twitter dataset.

Category	Original	Obtain
Depression	3082	2449
Non-depression	4687	4318

Table 6
Class label and corresponding number of posts in Reddit dataset.

Category	Original	Obtain
Depression	1293	1293
Non-depression	548	546

4. Experimental analysis and comparison

We conducted experiments and analyzed the depression detection method on social media at both the post and user levels. Models were trained on widely used benchmark datasets corresponding to each level of depression detection.

4.1. Post-level detection

This study presents a method for detecting depression at the post-level by applying three transformer encoder-based language models to classify posts. The approach is illustrated in Fig. 3.

Data collection: To detect depression at the post-level, we conducted experiments on two datasets from two social media platforms, Twitter and Reddit. The Twitter dataset provided by Cho [91] consists of 7779 tweets, where 3082 tweets are labeled as depressive and 4687 tweets as non-depressive. The Reddit dataset provided by Pirina and Çöltekin [92] includes 1841 posts, with 1293 posts labeled as depressive and 548 posts labeled as non-depressive.

Data pre-processing: The data pre-processing was carried out meticulously, removing stopwords, URLs, retweets, and mentions from the data. Each dataset row was tokenized, breaking the text into tokens or words. The tokenized words then underwent stemming and lemmatization. To ensure adequate information in each tweet, we retained only those tweets over five words. For the Twitter dataset, the final result consisted of 6767 tweets, with 2449 labeled as depressive and 4318 labeled as non-depressive. For the Reddit dataset, 1839 posts were obtained, with 1293 posts labeled as depressive and 546 posts labeled as non-depressive. Tables 5 and 6 present information about the datasets used for the depression detection task on social media at the post-level.

Fig. 4 visually represents the distribution of the number of words in posts across the two datasets used. A noticeable difference in the length of posts between the two datasets is observed, with the Reddit dataset having longer posts. This serves as evidence of the instability of post data on social media, which may lead to reduced generalizability of models when tested on different data samples.

Method: BERT (Bidirectional Encoder Representations from Transformers), RoBERTa (Robustly optimized BERT approach), and DistilBERT are all transformer-based models developed for natural language

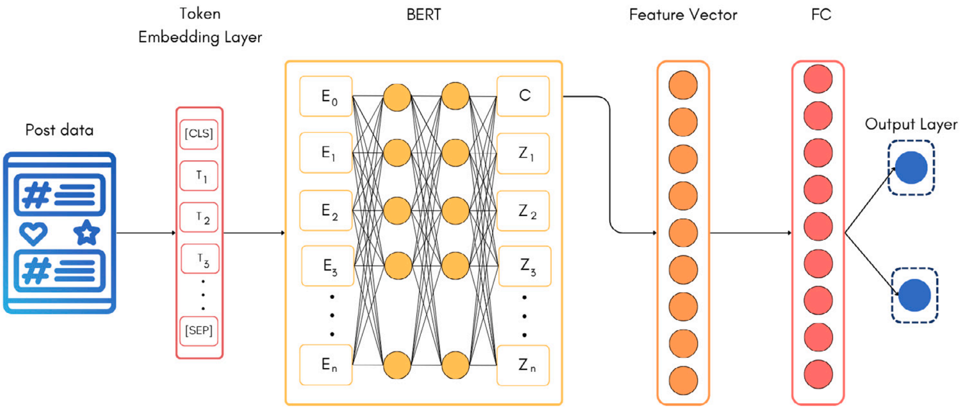

Fig. 3. Architecture to classify depression with pre-trained models.

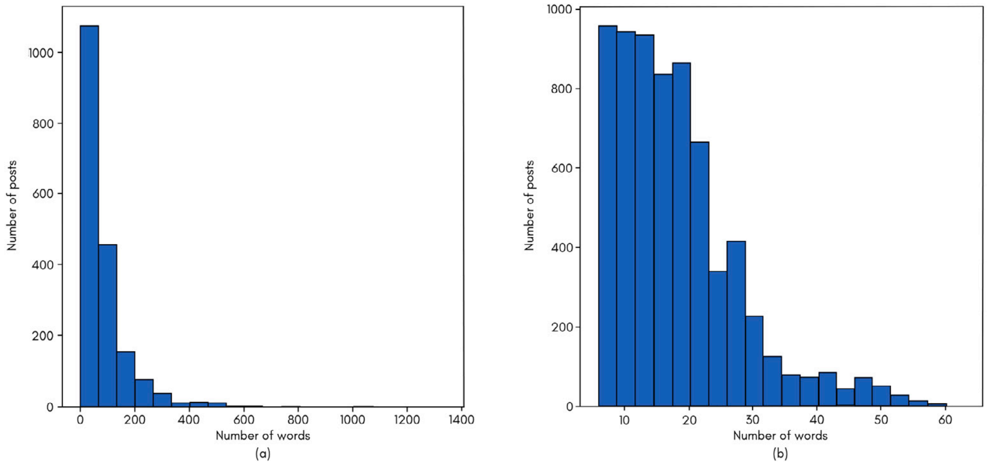

Fig. 4. Distribution of post lengths in the Reddit dataset (a) and Twitter dataset (b).

processing tasks. Unlike RNN and LSTM, which process data sequentially, Transformers excel in simultaneously processing long sequences such as sentences. The self-attention mechanism activates this processing power, enabling Transformers to capture rich contexts and leverage vast datasets for training. These models can generate rich and efficient language representations, facilitating deep insights into language context. The key feature of these transformer-based models is the pre-training phase with a large corpus of text data, followed by fine-tuning for specific tasks. During the pre-training phase, the model learns to predict missing words in sentences or understand relationships between words. This unsupervised learning method allows models to capture language states and build flexible representations.

BERT [93] introduced the concept of bidirectional attention, allowing the model to consider both directions of context for each word in the sentence. This bidirectional understanding enhances the model's ability to capture complex relationships and dependencies in language structure.

RoBERTa [94], an extension of BERT, improves pre-training by removing the next sentence prediction objective and training with smaller batches and a more significant learning rate.

DistilBERT [95] focuses on model compression. It is a distilled version of BERT designed to maintain similar performance while reducing the number of parameters, making DistilBERT more suitable for applications with limited computational resources.

Experimental setup: Table 7 lists the three selected models for this study: bert-base-uncased, roberta-base, and distilbert-base-cased. For each model, the corresponding tokenizer from the Huggingface library was used to tokenize the posts. In the Twitter dataset, the maximum token length for a tweet is 63, with a minimum of 5. All tweets were padded to a fixed length of 70. Meanwhile, the Reddit dataset has longer posts, so posts were padded to a fixed length of 256. The datasets were split into a 70% training set, a 15% validation, and a 15% testing set.

The hyperparameters used during fine-tuning were set as follows: the learning rate is 8e-5, the optimizer is Adam, and the batch size is 32. The entire experimental process was carried out on an NVIDIA Tesla T4 GPU provided by Kaggle.

4.2. User-level detection

Data collection: To detect depression at the user-level, we conducted experiments on a dataset provided by Shen et al. [89]. The dataset was constructed from the Twitter platform and divided into three parts: Depressed dataset D1, including 2558 users, users were considered depressed if their fixed-line tweets strictly adhered to the rule "(I'm/I was/I am/I've been) diagnosed depression."; Non-depressed dataset D2, consisting of 5304 users who had never posted any content containing the string "depress."; Depression Candidate dataset D3 is a large-scale dataset with 58,810 users selected if their sample posts loosely had the string "depress." In this study, we used only datasets D1 and D2.

Data pre-processing: We removed stopwords, punctuations, hashtags, mentions, emojis, and URLs from the tweets. Subsequently, stemming and lemmatization processes were applied to transform words into their base forms. After this process, all tweets with a length of 0 were removed to ensure analyzable user information. Next, users with at most ten tweets were excluded from the dataset. The result was 5459 users, with 2330 labeled as depressed and 3129 labeled as non-depressed. Details about the dataset are presented in Table 8.

Table 7
Language models used in the study.

Model	Parameter	Language	Huggingface link
bert-base-uncased	110M	English	https://huggingface.co/bert-base-uncased
roberta-base	125M	English	https://huggingface.co/roberta-base
distilbert-base-cased	66M	English	https://huggingface.co/distilbert-base-cased

Table 8
Class label and corresponding number of users.

Category	Users	Tweets
Depressed	2330	458,581
Non-depressed	3129	1884,944

Features extraction: After conducting a systematic survey, we selected and extracted five features for this task: Word Categories, VAD, Sentiment Polarity, Posting Time, and Topic.

- Word categories: Linguistic Inquiry and Word Count dictionary (LIWC) [96] are utilized, which is widely used in psychological and linguistic analysis. In this dictionary, words are categorized into one or more language features. Typically, non-depressed users do not usually use negative words, while users with depression often use personal pronouns and emotionally negative words in their posts [16]. Therefore, there are seven language features have been chosen [97,98]: personal pronouns (first-person singular, first-person plural, second-person, third-person singular, third-person plural), positive emotion, and negative emotion.
- VAD: For each word in a tweet, VAD values were extracted from the Affective Norms for English Words (ANEW) database [99]. Inspired by Hutto and Gilbert [100], if one of the three words preceding the current word indicates negation, the VAD value of that word is inverted. Subsequently, the VAD value of a tweet is calculated by averaging the values of its words, and the user's VAD feature is the average across all tweets.
- Sentiment polarity: The NLTK library was employed to examine emotional expressions in tweets, primarily relying on the Valence Aware Dictionary and Sentiment Reasoner (VADER) algorithm. This algorithm classifies emotions into four categories (Negative, Neutral, Positive, and Compound), with values ranging from −1 to 1. For each user's tweet, sentiment scores for the four emotion categories were calculated at both word and sentence levels. Then, we computed the average of each feature across all tweets to obtain the user's emotional features.
- Posting time: To exploit features related to users' posting times, we divided the 24-hour day into eight-time intervals, each consisting of three consecutive time slots. Specifically, the time within each interval is as follows: (1) 23 h, 0 h, 1 h; (2) 2 h, 3 h, 4 h; (3) 5 h, 6 h, 7 h; (4) 8 h, 9 h, 10 h; (5) 11 h, 12, 13; (6) 14 h, 15, 16 h; (7) 17 h, 18 h, 19 h; (8) 20 h, 21 h, 22 h. From this, we calculated the posting frequency of each user distributed across these time intervals to create a model input feature.
- Topic modeling: We employed TopSBM, a flexible and principled framework for topic modeling based on a stochastic block model (SBM) with non-parametric priors to extract features related to users' topics. Analyzing artificial and real-world data demonstrated that the SBM approach leads to better topic models than LDA regarding statistical model selection [101]. For each user, we extracted 14 topic-related features based on their tweets.

Method: LSTM [102] and Gate Recurrent Unit Networks (GRU) [103] are both advancements from the traditional RNN. The drawback of the conventional RNN model lies in its difficulty in capturing long-term dependencies, encountering issues such as gradient vanishing and gradient explosion. LSTM addresses these problems through a memory cell structure, each memory cell controlled by three gates: input gate, forget gate, and output gate. These gates determine which information is added, forgotten, and output from the memory cell. GRU, a simpler version of LSTM, employs two gates: the reset gate and the update gate. The reset gate decides how much the hidden state of forgetting is, while the update gate determines how much input to use in updating the hidden state. Compared to LSTM, GRU requires less computation but generally exhibits slightly lower performance. The architecture of the two models is illustrated in Figs. 5.

Experimental setup: Both LSTM and GRU models are constructed with a three-layer architecture, using ReLU as the activation function. The final layer is a fully connected layer with one hidden node, activated by the Sigmoid function for user classification. Both models utilize the Adam optimizer with a learning rate of 0.001, employ binary-cross-entropy as the loss function, and are trained for 100 epochs with a batch size of 32. The dataset is divided into a 70% training set (3815 users), a 15% validation set (818 users), and a 15% testing set (817 users). The experiment is conducted on Google Colaboratory using the NVIDIA Tesla T4 GPU.

5. Experimental results and discussions

5.1. Evaluation metrics

The evaluation metrics used to assess the performance of models include accuracy, recall, precision, and F1 score. These values are calculated using formulas (1), (2), (3), and (4), where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

5.2. Post-level detection

Each model undergoes a training process on a training dataset, with validation procedures conducted on a separate validation dataset throughout the training process. The weights of the best-performing model are saved and used to predict labels on the test dataset to calculate the performance of each model for each test. The results from the prediction process on the two test datasets are presented in Tables 9.

From the results obtained when training the models on the Twitter dataset, it can be observed that the bert-base-uncased model achieves the highest accuracy in most metrics with 0.92 accuracy, 0.93 recall, and 0.90 F1 score, and 0.87 precision, demonstrating its ability to classify accurately on the test dataset. Distilbert-base-uncased and roberta-based also achieve relatively high accuracy, with 0.92 and 0.91, respectively. Both bert-base-uncased and distilbert-base-uncased models exhibit stable precision and recall, indicating good classification accuracy and sensitivity. The bert-base-uncased model performs well due to its large number of parameters and the ability to understand deep and wide language representations. Distilbert-base-uncased, a

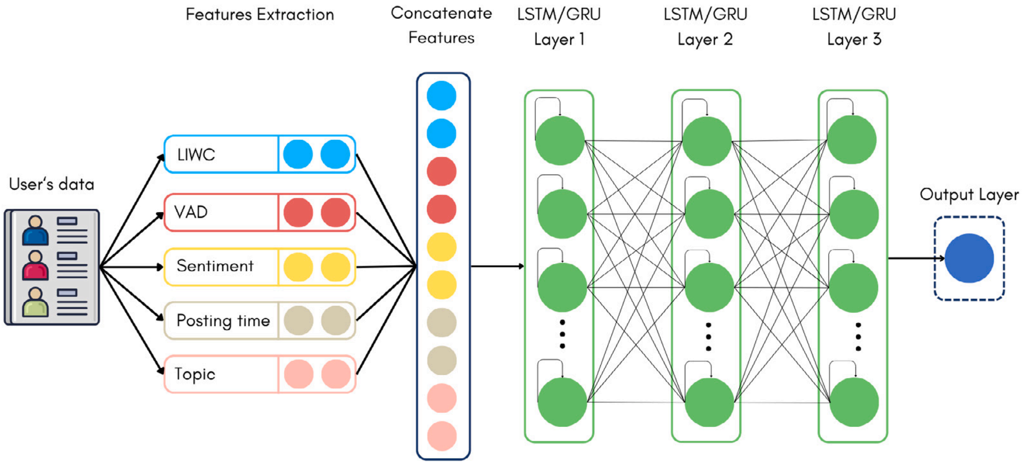

Fig. 5. Architecture of the LSTM and GRU models to detect depression on social networks.

Table 9
Evaluation scores of language models across Twitter and Reddit datasets.

Dataset	Model	Accuracy	F1-score	Recall	Precision
Twitter	bert-base-uncased	**0.92**	**0.90**	**0.93**	0.87
	roberta-based	0.91	0.88	0.88	0.88
	distilbert-base-uncased	**0.92**	0.89	0.89	**0.90**
Reddit	bert-base-uncased	**0.90**	0.92	0.91	0.93
	roberta-based	0.89	0.92	0.93	0.91
	distilbert-base-uncased	**0.90**	**0.93**	**0.96**	0.90

Table 10
Comparison of the proposed method with existing approaches on Twitter and Reddit datasets.

Dataset	Method	Accuracy	F1-score	Recall	Precision
Twitter	FastText & CNN + LSTM [14]	0.88	0.88	0.87	**0.89**
	Transformer based (Ours)	**0.92**	**0.90**	**0.93**	0.87
Reddit	LIWC + LDA + bigram & MLP [44]	**0.91**	0.93	0.92	0.90
	Emotion-based attention network [41]	**0.91**	**0.94**	**0.96**	**0.92**
	Transformer based (Ours)	0.90	0.93	**0.96**	0.90

Table 11
Performance of model with randomly combined features.

Feature	Model	Accuracy	F1-score	Recall	Precision
Topic + Posting time	LSTM	0.822	0.776	0.732	0.826
	GRU	0.817	0.783	0.744	0.825
VADER + VAD + LIWC	LSTM	0.822	0.792	0.755	0.832
	GRU	0.666	0.559	0.483	0.662
Posting time + LIWC	LSTM	0.811	0.723	0.631	0.859
	GRU	0.787	0.715	0.609	0.865
VADER + VAD + Posting time	LSTM	0.788	0.721	0.633	0.835
	GRU	0.817	0.775	0.740	0.815
VADER + VAD + Topic + Posting time	LSTM	0.814	0.774	0.741	0.809
	GRU	0.830	0.774	0.674	**0.908**
All Features	LSTM	**0.854**	**0.822**	**0.784**	0.865
	GRU	0.846	0.797	0.729	0.879

lighter version of BERT with significantly fewer parameters, still performs well and has the highest precision among the three models at 0.90.

The achieved results across the models for the Reddit dataset are also awe-inspiring. Among them, the distilbert-base-uncased model attains the highest accuracy with 0.90, a 0.93 F1 Score, 0.96 Recall, and 0.90 Precision. The bert-base-uncased model also demonstrates comparable performance with 0.90 accuracy, although F1 Score and Recall are lower than the distilbert-base-uncased model. Still, it achieves the highest precision among the three models at 0.93. Despite lower performance, the roberta-based model maintains a decent level of performance, approaching the other two models with 0.89 accuracy, 0.92 F1 Score, 0.93 Recall, and 0.91 Precision.

Comparing the results of the models on both datasets, it can be observed that the models perform better when trained on the Twitter dataset. As illustrated in Fig. 4, Twitter posts are relatively short and consistent, allowing the models to learn information more comprehensively. In contrast, the Reddit dataset contains posts with varying lengths, including very long posts, leading to padding and potentially missing information during model training, resulting in lower performance. Nevertheless, both two models still achieved high results on the test datasets despite being trained on imbalanced datasets.

We compare the performance of our best model with other approaches on the same dataset, as shown in Tables 10 and Fig. 6. The results indicate that our model achieves performance levels nearly equivalent to other methods on the Reddit dataset. Conversely, our model exhibits impressive performance for the Twitter dataset, outperforming other methods across most metrics. This underscores the suitability of our approach for various social media datasets. However, the high computational cost of Transformer models like BERT often makes them unsuitable for large-scale deployment without significant computational resources [73].

5.3. User-level detection

During the experimental process, to better understand the roles and contributions of various features in identifying users with depression, we trained the model by combining different features. The models predicted on the same test dataset to assess performance and the results are presented in Table 11.

Combining emotional features shows better performance compared to other combinations, indicating a high contribution of emotions in depression detection since emotions expressed through words are

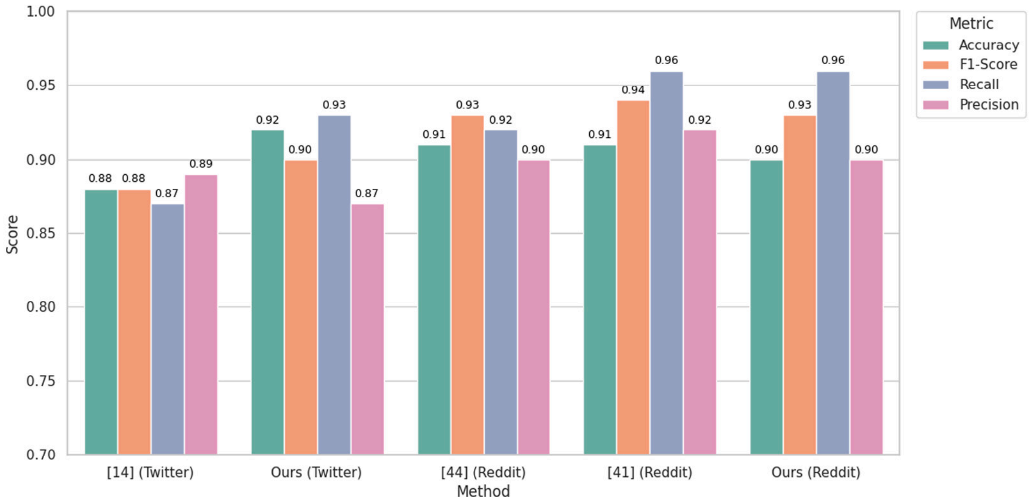

Fig. 6. Performance comparison of proposed and existing methods on Twitter and Reddit datasets.

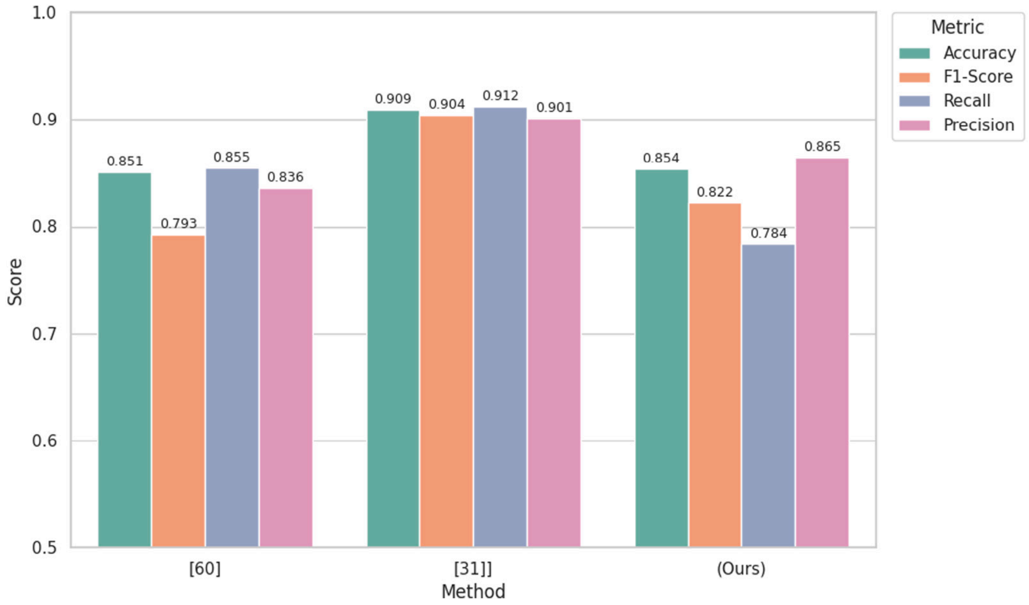

Fig. 7. Comparison of depression detection models on Shen et al. dataset [89].

Table 12
Results of the existing approaches on the dataset provided by Shen et al. [89].

Study	Features	Accuracy	F1-score	Recall	Precision
CNN [63]	Text embedding (Fasttext)	0.851	0.793	0.855	0.836
CNN + BiGRU [31]	Text embedding (BERT) + User behavior	**0.909**	**0.904**	**0.912**	**0.901**
LSTM (Ours)	LIWC + VADER + VAD + Topic + Posting time	0.854	0.822	0.784	0.865

directly related to the user's mood. Additionally, combining features related to posting time also yields good performance because depressed users often experience sleep disturbances at night [104].

Observing the results shows that the LSTM model outperforms GRU in most cases. Both models perform best when trained on the entire set of features. Specifically, LSTM excels with an accuracy of 0.854, precision of 0.865, recall of 0.784, and F1 score of 0.822. The results also reveal that the recall for LSTM and GRU is considerably lower than precision, indicating misclassification, particularly in class 1 (depression). The cause of this error is attributed to the imbalanced dataset, with a significantly larger number of samples in class 0 (non-depression).

As presented in Table 12 and Fig. 7, we compare the best performance achieved by our model with other approaches conducted on the dataset provided by Shen et al. [89]. While achieving lower performance than other methods, the difference in metrics is slight. On the other hand, our model utilizes relatively few features but remains influential. This helps our model maintain impressive performance without requiring excessive computational resources, making it easier to integrate into social media depression detection systems.

5.4. Discussions

Deep learning models, including Transformer-based architectures such as BERT and DistilBERT, along with recurrent neural networks (RNNs) like LSTM and GRU, have demonstrated significant potential in detecting depression at both the post and user levels. The performance differences between datasets highlight the critical role of data characteristics in shaping model capabilities. On the Twitter dataset, where posts are concise and uniform, Transformer models excel, effectively capturing language patterns related to depression. In contrast, on Reddit dataset, the diversity in post length and structure poses challenges, but models like DistilBERT adapt well, showcasing their ability to handle complex data with high computational efficiency. At the user level, integrating emotional and behavioral features, such as posting time, has significantly enhanced the performance of RNN models, particularly LSTM. This underscores the importance of combining behavioral data with text analysis, reflecting aspects such as sleep disturbances or emotional expressions, which are often associated with depressive states. These results open up potential practical applications in mental health monitoring, where multidimensional signals can improve accuracy in identifying at-risk individuals.

When placed in the context of existing methods, our approach demonstrates significant competitiveness. At the post level, Transformer models outperform some combined methods like FastText & CNN + LSTM on the Twitter dataset, while maintaining high performance on Reddit with fewer feature requirements compared to emotion-based attention networks. At the user level, although LSTM's performance is slightly lower than advanced methods like CNN + BiGRU in some previous studies, our approach still achieves a balance between effectiveness and computational complexity. Using fewer features while maintaining competitiveness suggests optimization potential, especially in large-scale monitoring systems. A notable challenge is the issue of imbalanced data, particularly evident at the user level, where models tend to favor the majority class (non-depressed). This leads to missing some depression cases, reducing the comprehensiveness of the detection system. Additionally, the structural differences in data between platforms like Twitter and Reddit indicate that the model's generalization ability still needs improvement, especially when handling long or heterogeneous text inputs.

To address these limitations, future research could consider applying techniques such as resampling, class balancing, or data augmentation to enhance the detection of the minority class. Exploring lighter models, such as optimized variants of Transformers, could also reduce computational requirements while maintaining high performance. Furthermore, integrating multimodal data—such as images, user interactions, or other signals—promises to provide a more comprehensive picture of psychological states, thereby improving the accuracy and practicality of the system. Additionally, evaluating the robustness of models across various social media platforms can shed light on their adaptability in diverse contexts. This is particularly important considering the differences in user behavior and emotional expression across online communities.

6. Conclusion

This study discusses the detection of depression on social media and the related analytical techniques. Our research covered various aspects pertinent to this area, including databases, feature extraction techniques, and models from machine learning to deep learning. We also provided a analysis of methodologies for detecting depression on social media at both the post-level and user-level. In our experiments at both post and user levels, we compared several modern models, offering a broad perspective on the strengths and limitations of current techniques. Our findings particularly highlighted the effectiveness of deep learning models, especially BERT and its variants in post-level analysis. For user-level analysis, we investigated how different features influence model outcomes, finding that combining multiple features improves prediction accuracy. This research is expected to assist others in understanding critical information and trends in detecting depression on social media, contributing to the support and identification of individuals with depression. We aim to explore more significant features for training various classification models in the future. Additionally, we plan to extend our research to detect a broader range of mental illnesses, incorporating visual, audio, and physiological features, which we believe will be a significant step forward in this field.

CRediT authorship contribution statement

Phi Ta: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Nha Tran:** Writing – original draft, Visualization, Validation, Supervision, Resources, Methodology, Formal analysis, Data curation. **Hung Nguyen:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Hien D. Nguyen:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is funded by Ministry of Education and Training of Vietnam under Grant B2025-SPS-02, Project title: "Building a learner response model based on behavioral and emotional analysis by deep learning approach"

Data availability

The data that has been used is confidential.

References

[1] World Health Organization, Depressive disorder (depression), 2024, https://www.who.int/news-room/fact-sheets/detail/depression. Last (Accessed 14 February 2024).

[2] World Health Organization, Depression, 2025, https://www.who.int/health-topics/depression. .n.d Last (Accessed 04 April 2025).

[3] World Health Organization, WHO Guide for Integration of Perinatal Mental Health in Maternal and Child Health Services, World Health Organization, Geneva, 2022, p. 66, Mental Health, Brain Health and Substance Use (MSD). URL: https://www.who.int/publications/i/item/9789240057142.

[4] L. Orsolini, R. Latini, M. Pompili, G. Serafini, U. Volpe, F. Vellante, M. Fornaro, A. Valchera, C. Tomasetti, S. Fraticelli, M. Alessandrini, R. La Rovere, S. Trotta, G. Martinotti, M. Di Giannantonio, D. De Berardis, Understanding the complex of suicide in depression: from research to clinics, Psychiatry Investig. 17 (3) (2020) 207–221, http://dx.doi.org/10.30773/pi.2019.0171.

[5] World Health Organization, World Health Statistics 2021: Monitoring Health for the SDGs, Sustainable Development Goals, World Health Organization, 2021.

[6] C. Wang, Social media platform-oriented topic mining and information security analysis by big data and deep convolutional neural network, Technol. Forecast. Soc. Change 199 (2024) 123070.

[7] A.H. Yazdavar, M.S. Mahdavinejad, G. Bajaj, K. Thirunarayan, J. Pathak, A. Sheth, Mental health analysis via social media data, in: 2018 IEEE International Conference on Healthcare Informatics, ICHI, IEEE, 2018, pp. 459–460.

[8] R. Sharma, M. Sharma, S. Joshi, Social media mirage-the two actual selves of an individual: Conceptualization and scale development, Technol. Forecast. Soc. Change 206 (2024) 123502.

[9] K.C. Bathina, M. Ten Thij, L. Lorenzo-Luaces, L.A. Rutter, J. Bollen, Individuals with depression express more distorted thinking on social media, Nat. Hum. Behav. 5 (4) (2021) 458–466.

[10] S. Singha, H. Arha, A.K. Kar, Healthcare analytics: A techno-functional perspective, Technol. Forecast. Soc. Change 197 (2023) 122908.

[11] R. Chiong, G.S. Budhi, S. Dhakal, F. Chiong, A textual-based featuring approach for depression detection using machine learning classifiers and social media texts, Comput. Biol. Med. 135 (2021) 104499.

[12] A.H. Uddin, D. Bapery, A.S.M. Arif, Depression analysis from social media data in bangla language using long short term memory (LSTM) recurrent neural network technique, in: 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering, IC4ME2, IEEE, 2019, pp. 1–4.

[13] L. Ma, Y. Wang, Constructing a semantic graph with depression symptoms extraction from twitter, in: 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB, IEEE, 2019, pp. 1–5.

[14] V. Tejaswini, K. Sathya Babu, 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. 23 (1) (2024) 1–20.

[15] M. Trotzek, S. Koitka, C.M. Friedrich, Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences, IEEE Trans. Knowl. Data Eng. 32 (3) (2018) 588–601.

[16] S. Ghosh, T. Anwar, Depression intensity estimation via social media: A deep learning approach, IEEE Trans. Comput. Soc. Syst. 8 (6) (2021) 1465–1474.

[17] H. Zogan, I. Razzak, X. Wang, S. Jameel, G. Xu, Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media, World Wide Web 25 (1) (2022) 281–304.

[18] M. Deshpande, V. Rao, Depression detection using emotion artificial intelligence, in: 2017 International Conference on Intelligent Sustainable Systems, Iciss, IEEE, 2017, pp. 858–862.

[19] M.M. Aldarwish, H.F. Ahmad, Predicting depression levels using social media posts, in: 2017 IEEE 13th International Symposium on Autonomous Decentralized System, ISADS, IEEE, 2017, pp. 277–280.

[20] D. Liu, X.L. Feng, F. Ahmed, M. Shahid, J. Guo, Detecting and measuring depression on social media using a machine learning approach: Systematic review, J. Med. Internet Res. Ment Heal. 9 (3) (2022) e27244, http://dx.doi.org/10.2196/27244, URL: https://mental.jmir.org/2022/3/e27244.

[21] L.M. Meyer, S. Stead, T.O. Salge, D. Antons, Artificial intelligence in acute care: A systematic review, conceptual synthesis, and research agenda, Technol. Forecast. Soc. Change 206 (2024) 123568.

[22] D. William, D. Suhartono, Text-based depression detection on social media posts: A systematic literature review, Procedia Comput. Sci. 179 (2021) 582–589.

[23] D. Liu, X.L. Feng, F. Ahmed, M. Shahid, J. Guo, et al., Detecting and measuring depression on social media using a machine learning approach: systematic review, J. Med. Internet Res. Ment. Heal. 9 (3) (2022) e27244.

[24] M. Omar, I. Levkovich, Exploring the efficacy and potential of large language models for depression: A systematic review, J. Affect. Disord. (2024).

[25] J. Cha, S. Kim, E. Park, A lexicon-based approach to examine depression detection in social media: the case of Twitter and university community, Humanit. Soc. Sci. Commun. 9 (1) (2022) 1–10.

[26] A. Li, D. Jiao, T. Zhu, Detecting depression stigma on social media: A linguistic analysis, J. Affect. Disord. 232 (2018) 358–362.

[27] D.E. Losada, F. Crestani, A test collection for research on depression and language use, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2016, pp. 28–39.

[28] M.T.A. Hridoy, S.R. Saha, M.M. Islam, M.A. Uddin, M.Z. Mahmud, Leveraging web scraping and stacking ensemble machine learning techniques to enhance detection of major depressive disorder from social media posts, Soc. Netw. Anal. Min. 14 (1) (2024) 239.

[29] R. Safa, P. Bayat, L. Moghtader, Automatic detection of depression symptoms in twitter using multimodal analysis, J. Supercomput. 78 (4) (2022) 4709–4744.

[30] D. Mowery, H. Smith, T. Cheney, G. Stoddard, G. Coppersmith, C. Bryan, M. Conway, et al., Understanding depressive symptoms and psychosocial stressors on Twitter: a corpus-based study, J. Med. Internet Res. 19 (2) (2017) e6895.

[31] H. Zogan, I. Razzak, S. Jameel, G. Xu, Depressionnet: learning multi-modalities with user post summarization for depression detection on social media, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 133–142.

[32] A. Wongkoblap, M.A. Vadillo, V. Curcin, et al., Deep learning with anaphora resolution for the detection of tweeters with depression: Algorithm development and validation study, J. Med. Internet Res. Ment. Heal. 8 (8) (2021) e19824.

[33] H. Karamti, A.M. Mahmoud, A pre-protective objective in mining females social contents for identification of early signs of depression using soft computing deep framework, Sci. Rep. 13 (1) (2023) 14899.

[34] J. de Jesús Titla-Tlatelpa, R.M. Ortega-Mendoza, M. Montes-y Gómez, L. Villaseñor-Pineda, A profile-based sentiment-aware approach for depression detection in social media, Eur. Polym. J. Data Sci. 10 (1) (2021) 54.

[35] J. Angskun, S. Tipprasert, T. Angskun, Big data analytics on social networks for real-time depression detection, J. Big Data 9 (1) (2022) 69.

[36] P. Arora, P. Arora, Mining twitter data for depression detection, in: 2019 International Conference on Signal Processing and Communication, ICSC, IEEE, 2019, pp. 186–189.

[37] S.W. Kelley, C.M. Gillan, Using language in social media posts to study the network dynamics of depression longitudinally, Nat. Commun. 13 (1) (2022) 870.

[38] H. Jung, H.-A. Park, T.-M. Song, Ontology-based approach to social data sentiment analysis: detection of adolescent depression signals, J. Med. Internet Res. 19 (7) (2017) e259.

[39] S.G. Burdisso, M. Errecalde, M. Montes-y Gómez, A text classification framework for simple and effective early depression detection over social media streams, Expert Syst. Appl. 133 (2019) 182–197.

[40] F.M. Shah, F. Ahmed, S.K.S. Joy, S. Ahmed, S. Sadek, R. Shil, M.H. Kabir, Early depression detection from social network using deep learning techniques, in: 2020 IEEE Region 10 Symposium, TENSYMP, IEEE, 2020, pp. 823–826.

[41] L. Ren, H. Lin, B. Xu, S. Zhang, L. Yang, S. Sun, Depression detection on reddit with an emotion-based attention network: algorithm development and validation, J. Med. Internet Res. Med. Inform. 9 (7) (2021) e28754.

[42] D. Owen, D. Antypas, A. Hassoulas, A.F. Pardiñas, L. Espinosa-Anke, J.C. Collados, et al., Enabling early health care intervention by detecting depression in users of web-based forums using language models: Longitudinal analysis and evaluation, J. Med. Internet Res. AI 2 (1) (2023) e41205.

[43] F. Cacheda, D. Fernandez, F.J. Novoa, V. Carneiro, Early detection of depression: social network analysis and random forest techniques, J. Med. Internet Res. 21 (6) (2019) e12554.

[44] M.M. Tadesse, H. Lin, B. Xu, L. Yang, Detection of depression-related posts in reddit social media forum, IEEE Access 7 (2019) 44883–44893.

[45] A. Trifan, R. Antunes, S. Matos, J.L. Oliveira, Understanding depression from psycholinguistic patterns in social media texts, in: European Conference on Information Retrieval, Springer, 2020, pp. 402–409.

[46] A. Pérez, J. Parapar, Á. Barreiro, Automatic depression score estimation with word embedding models, Artif. Intell. Med. 132 (2022) 102380.

[47] K. Malviya, B. Roy, S. Saritha, A transformers approach to detect depression in social media, in: 2021 International Conference on Artificial Intelligence and Smart Systems, ICAIS, IEEE, 2021, pp. 718–723.

[48] J. Aguilera, D.I.H. Farías, R.M. Ortega-Mendoza, M. Montes-y Gómez, Depression and anorexia detection in social media as a one-class classification problem, Appl. Intell. 51 (8) (2021) 6088–6103.

[49] K. Katchapakirin, K. Wongpatikaseree, P. Yomaboot, Y. Kaewpitakkun, Facebook social media for depression detection in the thai community, in: 2018 15th International Joint Conference on Computer Science and Software Engineering, Jcsse, IEEE, 2018, pp. 1–6.

[50] M.K. Kabir, M. Islam, A.N.B. Kabir, A. Haque, M.K. Rhaman, Detection of depression severity using bengali social media posts on mental health: study using natural language processing techniques, J. Med. Internet Res. Form. Res. 6 (9) (2022) e36118.

[51] M.R. Islam, M.A. Kabir, A. Ahmed, A.R.M. Kamal, H. Wang, A. Ulhaq, Depression detection from social network data using machine learning techniques, Heal. Inf. Sci. Syst. 6 (2018) 1–12.

[52] A.G. Reece, C.M. Danforth, Instagram photos reveal predictive markers of depression, Eur. Polym. J. Data Sci. 6 (1) (2017) 15.

[53] C.Y. Chiu, H.Y. Lane, J.L. Koh, A.L. Chen, Multimodal depression detection on instagram considering time interval of posts, J. Intell. Inf. Syst. 56 (1) (2021) 25–47.

[54] Z. Peng, Q. Hu, J. Dang, Multi-kernel SVM based depression recognition using social media data, Int. J. Mach. Learn. Cybern. 10 (2019) 43–57.

[55] Y. Cai, H. Wang, H. Ye, Y. Jin, W. Gao, Depression detection on online social network with multivariate time series feature of user depressive symptoms, Expert Syst. Appl. 217 (2023) 119538.

[56] Y. Wang, Z. Wang, C. Li, Y. Zhang, H. Wang, A multitask deep learning approach for user depression detection on sina weibo, 2020, arXiv preprint arXiv:2008.11708.

[57] G. Li, B. Li, L. Huang, S. Hou, et al., Automatic construction of a depression-domain lexicon based on microblogs: text mining study, J. Med. Internet Res. Med. Inform. 8 (6) (2020) e17650.

[58] C. Lin, P. Hu, H. Su, S. Li, J. Mei, J. Zhou, H. Leung, Sensemood: depression detection on social media, in: Proceedings of the 2020 International Conference on Multimedia Retrieval, 2020, pp. 407–411.

[59] J.P. Thekkekara, S. Yongchareon, V. Liesaputra, An attention-based CNN-BiLSTM model for depression detection on social media text, Expert Syst. Appl. 249 (2024) 123834.

[60] M.A. Wani, M.A. ELAffendi, K.A. Shakil, A.S. Imran, A.A. Abd El-Latif, Depression screening in humans with AI and deep learning techniques, IEEE Trans. Comput. Soc. Syst. 10 (4) (2022) 2074–2089.

[61] M. Tlachac, E. Rundensteiner, Screening for depression with retrospectively harvested private versus public text, IEEE J. Biomed. Heal. Inform. 24 (11) (2020) 3326–3332.

[62] A.B. Warriner, V. Kuperman, M. Brysbaert, Norms of valence, arousal, and dominance for 13,915 english lemmas, Behav. Res. Methods 45 (2013) 1191–1207.

[63] R. Skaik, D. Inkpen, Using twitter social media for depression detection in the canadian population, in: Proceedings of the 2020 3rd Artificial Intelligence and Cloud Computing Conference, 2020, pp. 109–114.

[64] G. Geetha, G. Saranya, K. Chakrapani, J.G. Ponsam, M. Safa, S. Karpagaselvi, Early detection of depression from social media data using machine learning algorithms, in: 2020 International Conference on Power, Energy, Control and Transmission Systems, ICPECTS, IEEE, 2020, pp. 1–6.

[65] H. Zhang, H. Wang, S. Han, W. Li, L. Zhuang, Detecting depression tendency with multimodal features, Comput. Methods Programs Biomed. 240 (2023) 107702.

[66] K. Yang, T. Zhang, S. Ananiadou, A mental state Knowledge–aware and Contrastive Network for early stress and depression detection on social media, Inf. Process. Manage. 59 (4) (2022) 102961.

[67] Y. Wang, Z. Wang, C. Li, Y. Zhang, H. Wang, Online social network individual depression detection using a multitask heterogenous modality fusion approach, Inform. Sci. 609 (2022) 727–749.

[68] F. Sadeque, D. Xu, S. Bethard, Measuring the latency of depression detection in social media, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 2018, pp. 495–503.

[69] T. Huynh, H.D. Nguyen, I. Zelinka, et al., A method to detect influencers in social networks based on the combination of amplification factors and content creation, PloS One 17 (10) (2022) e0274596.

[70] S.N. Hoang, B. Nguyen, N.P. Nguyen, et al., Enhanced task-based knowledge for lexicon-based approach in vietnamese hate speech detection, in: 14th International Conference on Knowledge and Systems Engineering, KSE 2022, IEEE, 2022.

[71] H. Kour, 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. 81 (17) (2022) 23649–23685.

[72] K. Harrigian, C. Aguirre, M. Dredze, On the state of social media data for mental health research, 2020, arXiv preprint arXiv:2011.05233.

[73] H. Bekamiri, D.S. Hain, R. Jurowetzki, Patentsberta: A deep nlp based hybrid model for patent distance and classification using augmented sbert, Technol. Forecast. Soc. Change 206 (2024) 123536.

[74] Y. Wang, D. Inkpen, P.K. Gamaarachchige, Explainable depression detection using large language models on social media data, in: Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology, CLPsych 2024, 2024, pp. 108–126.

[75] D.B. Victor, J. Kawsher, M.S. Labib, S. Latif, Machine learning techniques for depression analysis on social media-case study on bengali community, in: 2020 4th International Conference on Electronics, Communication and Aerospace Technology, ICECA, IEEE, 2020, pp. 1118–1126.

[76] S. Samanvitha, A. Bindiya, S. Sudhanva, B. Mahanand, Naïve Bayes classifier for depression detection using text data, in: 2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques, ICEECCOT, IEEE, 2021, pp. 418–421.

[77] N. Jagtap, H. Shukla, V. Shinde, S. Desai, V. Kulkarni, Use of ensemble machine learning to detect depression in social media posts, in: 2021 Second International Conference on Electronics and Sustainable Communication Systems, ICESC, IEEE, 2021, pp. 1396–1400.

[78] Z. Guo, N. Ding, M. Zhai, Z. Zhang, Z. Li, Leveraging domain knowledge to improve depression detection on Chinese social media, IEEE Trans. Comput. Soc. Syst. 10 (4) (2023) 1528–1536.

[79] L. Tong, Z. Liu, Z. Jiang, F. Zhou, L. Chen, J. Lyu, X. Zhang, Q. Zhang, A. Sadka, Y. Wang, et al., Cost-sensitive boosting pruning trees for depression detection on Twitter, IEEE Trans. Affect. Comput. 14 (3) (2022) 1898–1911.

[80] J.S.L. Figuerêdo, A.L.L. Maia, R.T. Calumby, Early depression detection in social media based on deep learning and underlying emotions, Online Soc. Netw. Media 31 (2022) 100225.

[81] H. Zogan, I. Razzak, S. Jameel, G. Xu, Hierarchical convolutional attention network for depression detection on social media and its impact during pandemic, IEEE J. Biomed. Heal. Inform. 28 (4) (2023) 1815–1823.

[82] M. Rizwan, M.F. Mushtaq, U. Akram, A. Mehmood, I. Ashraf, B. Sahelices, Depression classification from tweets using small deep transfer learning language models, IEEE Access 10 (2022) 129176–129189.

[83] C. Janiesch, P. Zschech, K. Heinrich, Machine learning and deep learning, Electron. Mark. 31 (3) (2021) 685–695.

[84] I.H. Sarker, Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions, Social Networks Comput. Sci. 2 (6) (2021) 420.

[85] H. Nguyen, N. Tran, H.D. Nguyen, L. Nguyen, K. Kotani, KTFEv2: multimodal facial emotion database and its analysis, IEEE Access 11 (2023) 17811–17822.

[86] W. Zhang, J. Xie, Z. Zhang, X. Liu, Depression detection using digital traces on social media: A knowledge-aware deep learning approach, J. Manage. Inf. Syst. 41 (2) (2024) 546–580.

[87] A. Zafar, D. Aftab, R. Qureshi, Y. Wang, H. Yan, Multi-explainable TemporalNet: An interpretable multimodal approach using temporal convolutional network for user-level depression detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 2258–2265.

[88] G. Cecere, C. Jean, F. Le Guel, M. Manant, Artificial intelligence and algorithmic bias? Field tests on social network with teens, Technol. Forecast. Soc. Change 201 (2024) 123204.

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

[90] J. Liu, M. Shi, A hybrid feature selection and ensemble approach to identify depressed users in online social media, Front. Psychol. 12 (2022) 802821.

[91] H. Chung, Twitter depression dataset, 2024, https://www.kaggle.com/datasets/hyunkic/twitter-depression-dataset. Last (Accessed 14 February 2024).

[92] I. Pirina, Ç. Çöltekin, Identifying depression on reddit: The effect of training data, in: Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task, 2018, pp. 9–12.

[93] J. Devlin, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[94] Y. Liu, Roberta: A robustly optimized bert pretraining approach, 364, 2019, arXiv preprint arXiv:1907.11692.

[95] V. Sanh, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2019, arXiv preprint arXiv:1910.01108.

[96] J. Pennebaker, C. Chung, M. Ireland, A. Gonzales, R. Booth, The development and psychometric properties of LIWC2007. Austin, TX, LIWC. Net, 2007, 2015.

[97] H.D. Nguyen, T. Huynh, S.N. Hoang, et al., Language-oriented sentiment analysis based on the grammar structure and improved self-attention network, in: 15th International Conference on Evaluation of Novel Approaches To Software Engineering, ENASE 2020, Prague, Czech Public, 2020, pp. 339–346.

[98] H.D. Nguyen, T. Huynh, S.N. Hoang, et al., Multi-level sentiment analysis of product reviews based on grammar rules of language, in: 20th International Conference on Intelligent Software Methodologies, Tools, and Techniques, SOMET 2021, Cancun, Mexico, 2021, pp. 444–456.

[99] M. Bradley, Affective norms for english words (ANEW): Instruction manual and affective ratings, 1999.

[100] C. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, 2014, pp. 216–225.

[101] M. Gerlach, T.P. Peixoto, E.G. Altmann, A network approach to topic models, Sci. Adv. 4 (7) (2018) eaaq1360.

[102] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[103] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, 2014, arXiv preprint arXiv:1406.1078.

[104] L. Lustberg, C.F. Reynolds III, Depression and insomnia: questions of cause and effect, Sleep Med. Rev. 4 (3) (2000) 253–262.