frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 11 September 2025
DOI 10.3389/frai.2025.1627078

OPEN ACCESS

EDITED BY
Kausik Basak,
JIS Institute of Advanced Studies and Research, India

REVIEWED BY
Arya Bhattacharya,
Mahindra University, India
Saurabh Pal,
University of Calcutta, India

*CORRESPONDENCE
Muhammad A. B. Fayyaz
m.fayyaz@mmu.ac.uk

RECEIVED 14 May 2025
ACCEPTED 18 August 2025
PUBLISHED 11 September 2025

CITATION
Hameed S, Nauman M, Akhtar N, Fayyaz MAB and Nawaz R (2025) Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models.
Front. Artif. Intell. 8:1627078.
doi: 10.3389/frai.2025.1627078

Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models

Sidra Hameed [1], Muhammad Nauman[1], Nadeem Akhtar[2], Muhammad A. B. Fayyaz[3]* and Raheel Nawaz[4]

[1]Faculty of Computing, The Islamia University of Bahawalpur, Punjab, Pakistan, [2]Department of Information Technology, FCIT, University of the Punjab, Lahore, Pakistan, [3]OTEHM, Manchester Metropolitan University, Manchester, United Kingdom, [4]Pro VC-Staffordshire University, Staffordshire, United Kingdom

Introduction: Mental disorders are highly prevalent in modern society, leading to substantial personal and societal burdens. Among these, depression is one of the most common, often exacerbated by socioeconomic, clinical, and individual risk factors. With the rise of social media, user-generated content offers valuable opportunities for the early detection of mental disorders through computational approaches.

Methods: This study explores the early detection of depression using black-box machine learning (ML) models, including Support Vector Machines (SVM), Random Forests (RF), Extreme Gradient Boosting (XGB), and Artificial Neural Networks (ANN). Advanced Natural Language Processing (NLP) techniques TF-IDF, Latent Dirichlet Allocation (LDA), N-grams, Bag of Words (BoW), and GloVe embeddings were employed to extract linguistic and semantic features. To address the interpretability limitations of black-box models, Explainable AI (XAI) methods were integrated, specifically the Local Interpretable Model-Agnostic Explanations (LIME).

Results: Experimental findings demonstrate that SVM achieved the highest accuracy in detecting depression from social media data, outperforming RF and other models. The application of LIME enabled granular insights into model predictions, highlighting linguistic markers strongly aligned with established psychological research.

Discussion: Unlike most prior studies that focus primarily on classification accuracy, this work emphasizes both predictive performance and interpretability. The integration of LIME not only enhanced transparency and interpretability but also improved the potential clinical trustworthiness of ML-based depression detection models.

KEYWORDS

mental illness detection, natural language processing, machine learning, explainable artificial intelligence, Local Interpretable Model-Agnostic Explanations (LIME)

1 Introduction

Mental disorders, also known as psychiatric disorders, encompass a wide range of conditions that disrupt thoughts, emotions, and behaviors, often impairing an individual's ability to function in daily life (Orabi et al., 2018; Zhang T. et al., 2022). These include depression, anxiety, schizophrenia, and bipolar disorder, with causes rooted in genetic, environmental, and biological factors. Among these, depression is particularly prevalent and debilitating, affecting health, relationships, and productivity. Despite the availability of treatments like psychotherapy and pharmacotherapy (Ive et al., 2020), depression remains a global public health concern. According to the World Health Organization, over 322 million people suffer from depression worldwide, yet many cases go undiagnosed due to stigma and limited access to mental health services, especially in low- and middle-income countries (Velupillai et al., 2018; World Health Organization, 2021).

To address these challenges, researchers are increasingly leveraging digital data sources—particularly social media—to detect mental health issues using Natural Language Processing (NLP) and Machine Learning (ML) techniques (World Health Organization, Regional Office for the Eastern Mediterranean). Social media platforms like Twitter (X), Reddit, and Facebook offer real-time insights into users' emotions, behavior, and potential mental distress. Recent studies have demonstrated the potential of analyzing linguistic and behavioral cues to predict mood disorders and related symptoms, such as stress, self-harm, and emotional deterioration, without requiring traditional clinical assessments (Yazdavar et al., 2020; Olusegun et al., 2023; AbuRaed et al., 2024; Chancellor and De Choudhury, 2020). These digital approaches offer scalable, non-intrusive alternatives to traditional mental health diagnostics, especially in underserved populations.

Natural Language Processing (NLP) is a subfield of Artificial Intelligence (AI) that has facilitated various tasks in recent years, including the management and analysis of large amounts of textual data, information extraction, sentiment analysis, emotion detection, and mental health surveillance, among others (Steinkamp and Cook, 2021a). Feature extraction techniques in NLP are essential for transforming unstructured textual data into structured numerical representations, thereby enabling machine learning models to perform tasks such as classification, sentiment analysis, and topic modeling. Traditional methods such as Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and N-grams convert text into sparse feature vectors by capturing lexical patterns and word co-occurrence frequencies (Salton and Buckley, 1988; Jurafsky and Martin, 2000). While effective for basic NLP tasks, these approaches cannot capture deeper semantic relationships and contextual nuances. To address these limitations, more advanced feature extraction techniques such as Word2Vec, GloVe (Pennington et al., 2014), and contextual embeddings from transformer-based models like BERT (Devlin et al., 2019) have been developed. These methods represent words in dense vector spaces and incorporate semantic and syntactic context, significantly improving performance across a wide range of downstream NLP applications.

In general, users can convey their emotions through various written formats, such as posts on social media platforms, interview transcripts, and professional notes that include patient descriptions of their mental states. X (formerly known as Twitter) is most commonly known as a platform for micro-blogging because it has a straightforward user interface that enables the publication of short stories of no more than 280 characters. Tweets posted by virtually every user are available to the general public and can be retrieved using the user's own X API (Govindasamy and Palanichamy, 2021). X enables users to analyse and understand current events and trends, regardless of their geographical location. More recently, the research work focused on determining whether a person is depressed by analyzing their tweets. More precisely, Sentiment analysis can determine whether a piece of writing has been produced in a positive, negative, or neutral tone. Comments and posts from other X users can reveal whether a user is happy or sad at any given moment. Each tweet is evaluated based on its positive, negative, or neutral sentiments. The NLP system may classify tweets as either depressive or non-depressive by identifying depressive symptoms.

More recently, black box ML algorithms have demonstrated exceptional performance in text classification and analysis, yielding accurate and efficient results across various applications. However, their intrinsic black box nature poses notable challenges to transparency, interpretability, and trustworthiness, especially in critical domains where comprehensible decision-making processes are essential (Cesarini et al., 2024; Jahromi et al., 2024). The lack of transparency in ML and its black box nature are significant issues in its implementation in critical domains, such as healthcare (Khan et al., 2024; Weerts et al., 2019; Chakraborty et al., 2021; Kawakura et al., 2022; Nauman et al., 2021). Explainable Artificial Intelligence (XAI) is a subdomain of AI that aims to improve transparency by explaining the internal decision-making processes of such models (Chen et al., 2021).

One the contrary, a few known efforts to explain the black box models in literature include SHAP (Chelgani et al., 2023) and LIME (Hakkoum et al., 2020). Recent research in explainability focuses on revealing the primary features that significantly impact a model's decision-making process (Zhang Y. et al., 2022). As AI-based systems only make predictions without explaining their rationale, there is a need for mechanisms to explain and interpret their decisions. Furthermore, Local Interpretable Model-agnostic Explanations (LIME) is an explainability technique used to interpret the predictions made by machine learning models. The LIME technque approximates a complex model with a local, interpretable one around the prediction to be explained, thereby offering insights into the model's behavior on individual predictions (Ribeiro et al., 2016b). This method is particularly valuable in domains where understanding the decision-making process is crucial, such as healthcare and mental health diagnosis, as it helps build trust and provides transparency in the model's decisions (Ribeiro et al., 2016a). By applying LIME to the ML models, we can identify which features contribute most to the predictions, thus making the model's decisions more understandable and actionable for stakeholders.

Recent research on depression detection from social media platforms like Twitter and Reddit has shown that ML and deep

learning models—such as CNNs, LSTMs, and BERT—are effective in identifying mental health indicators from user-generated content (Amanat et al., 2022b; Ji et al., 2022b). However, key gaps remain: many models operate as black boxes with limited interpretability, which is problematic in clinical contexts requiring transparency (Guo et al., 2023a; Ibrahimov and Ali, 2024). Most studies rely on a single feature representation method, overlooking the benefits of combining traditional and semantic features. Comparative evaluations across diverse models ranging from classical ML to deep neural networks are rare, limiting insight into their relative performance. Additionally, while XAI tools like LIME and SHAP are gaining traction, their integration into end-to-end depression detection systems remains limited. Finally, many datasets are weakly labeled, often based on heuristics or self-reports without clinical validation, undermining the reliability of resulting models (Chancellor and De Choudhury, 2020; Yazdavar et al., 2020).

This research distinguishes itself from existing literature by employing multiple feature extraction techniques and the LIME (Ribeiro et al., 2016b) method to elucidate the internal decision-making processes of machine learning models. This work focuses on interpreting the detection decisions made by ML models to enhance early mental disorder detection and support healthcare professionals. The research results will enable physicians to identify life-threatening diseases in their early stages, ultimately facilitating a healthier society. This work aims to leverage advanced ML techniques and XAI to facilitate the early detection of myocardial infarction through the analysis of X data. By addressing the gap in understanding how social media conversations can be leveraged for mental health insights, this research contributes novel methodologies for pre-processing data, feature extraction, and model interpretation.

The main contributions of this research work are as follows:

- We applied LIME explainability uniformly to 28 different feature-classifier combinations (7 feature extraction methods $\times$ 4 classifiers), rather than limiting interpretation to the single highest-accuracy model as in most prior studies. This research work provides a comprehensive view of how different models make predictions in the depression detection context.
- The research findings are structured to separately rank feature extraction methods and classifiers, eliminating cross-category confusion and allowing researchers to see the independent effect of each.
- Beyond standard tokenisation and normalization, the proposed pipeline includes slang expansion, emoticon-to-text conversion, and a mental health-specific stopword list, in particular, tailored to the noisy and abbreviated nature of depression-related social media posts.
- The research connected linguistic patterns highlighted by LIME to known depression-related cues in psychology and linguistics literature, providing actionable insights for mental health professionals and validating black box ML model outputs beyond raw accuracy.
- We systematically applied LIME across all model configurations, selecting it for its model-agnostic nature, suitability for short-text explanations, and computational efficiency.

2 Background

2.1 Mental disorder detection

Modern society is plagued by a high prevalence of mental disorders, a significant source of personal and societal suffering. It is a complex, multifactorial disease influenced by several socioeconomic and clinical factors, as well as individual risk factors (Thornicroft et al., 2022). Depression is a typical mental condition that can affect functioning and cause suicidal thoughts or attempts (Jain et al., 2022). Millions of User worldwide suffer from depression each year, which is recognized as a medical condition. Persistent unhappiness or even minor stressful life events can lead to depression, illustrating the intricate relationship between mental health, NLP, ML, and AI.

Various computing algorithms for the automatic analysis and representation of human language are referred to as NLP (Cambria and White, 2014). Within Artificial Intelligence and Computer Science, the study of NLP is of utmost significance. Research into NLP employs a wide range of theoretical frameworks and methodological approaches to enable human-computer communication using natural language. NLP is an interdisciplinary field combining elements of computer science, linguistics, and mathematics, with the fundamental objective of converting human language into executable computer instructions. Natural Language Understanding and Natural Language Generation are the two basic areas of investigation in the field of NLP (Kang et al., 2020).

2.2 Mental disorder detection and social media

Social media's extensive use may present opportunities to lower the prevalence of undetected mental disorders. An increasing number of research projects are investigating the connection between social media and mental health. These studies attempt to determine whether there is a causal link between social media use and negative behaviors such as stress, anxiety, depression, and suicidality (Guntuku et al., 2017).

Social media networks such as X, LinkedIn, Instagram, Snapchat, and Facebook have surged in popularity, making them one of the most important sources of readily available and easily accessible information on all facets of life (Islam et al., 2018). Users of these platforms can express themselves freely, share their thoughts and feelings, and discuss any topic. Users suffering from mental illnesses, such as depression, may isolate themselves and avoid social engagement (Hemmatirad et al., 2020). However, online platforms allow users to convey their thoughts, opinions, and sentiments regarding various topics through applications such as Facebook, X, and Instagram (Islam et al., 2018).

2.3 Feature extraction methods

2.3.1 Latent Dirichlet Allocation

Both NLP and ML make use of a probabilistic model known as Latent Dirichlet Allocation (LDA) for topic modeling. LDA is based

on the assumption that each text contained within a corpus can be modeled as a mixture of a limited number of underlying themes and that each word contained within a document is taken from one of those subjects. This assumption guides LDA's operation. By utilizing the well-known LDA approach, the limit focuses on three of the seventeen Sustainable Development Goals, while simultaneously summarizing and presenting linked subtopics (Al Qudah et al., 2022).

To construct and effectively employ an LDA model, one must first ascertain the composition of the target document's latent themes, such as θ and z. The following is the revised Equation 1, where γ and ϕ are the parameters of the posterior distribution of θ and z, respectively.

$$\varphi_{ni} \infty \beta_{iw_n} \exp\{\Psi(\gamma_i)\}, \gamma_i = \alpha_i + \sum_{n=1}^{N} \varphi_{ni} \tag{1}$$

The LDA model's parameters were able to be estimated after the distribution of the hidden variables had been discovered, which made the process much simpler. M represents the total number of documents, d stands for the document ID, and dni represents the greatest possible value for n that can be derived from the expectation step. Equation 2 displays these three variables in their respective spots.

$$\alpha = \alpha - H(\alpha)^{-1} g(\alpha), \beta_{ij} \infty \sum_{d=1}^{M} \sum_{n=1}^{N} \varphi(d)_{ni}^{*} w(d)_{n}^{j} \tag{2}$$

In addition, the results of making inferences about hidden variables could be utilized in calculating the target document's generation probability value, denoted by $p_{LDA}(x|\beta)$. According to LDA, a document cannot be comprehensive unless it draws from various themes because it requires drawing from a pool of ideas. In this study, these subjects were tagged and connected into thematic groups that helped distinguish between users of diabetic mobile apps who had good and negative sentiments toward them (Ossai and Wickramasinghe, 2023).

2.3.2 Term frequency-inverse document frequency

It is possible to quantify the significance or relevance of string representations (words, phrases, lemmas, and so on) by making use of the TF-IDF measure, which is utilized in the disciplines of Information Retrieval and ML that are included in a collection of documents. This can be done by comparing the document to another collection of documents. A k-best selection method and a modified version of the TF-IDF-based approach are developed as part of the feature vectorization process. Text vectorization based on modified TF-IDF, pre-trained embedding based on Google News Corpus, and a deep neural network are all components of this system (Dey and Das, 2023).

The Term Frequency (TF) of a term or word indicates the proportion of the document's total words that are comprised of instances of that term, as defined in Equation 3.

$$TF = \frac{\text{\# of times term appears in doc}}{\text{\# of terms in doc}} \tag{3}$$

A term's Inverse Document Frequency (IDF) reveals how frequently it appears in the total number of documents in the corpus. Equation 4 defines the equation for calculating the IDF. Words that do not appear in many papers (such as terms used in technical jargon, for example) are given more consideration than those used repeatedly throughout the entire work.

$$IDF = \log\left(\frac{\text{\# of doc in corpus}}{\text{\# of doc containing term in corpus}}\right) \tag{4}$$

Multiplying a term's TF and IDF scores yields its TF-IDF, defined in Equation 5.

$$TF - IDF = TF \times IDF \tag{5}$$

TF-IDF benefits many tasks involving natural language processing. For instance, search engines use it to determine how relevant a document is to a user's query. Text summarization, topic modeling, and categorization are some other applications of TF-IDF.

It's important to remember that there are several ways to determine an individual's IDF score. The logarithm to the base 10 is frequently used. However, a natural logarithm is used by some bookstores. To further prevent division by zero, a single can be added to the denominator in the following manner in Equation 6.

$$IDF = \log\left(\frac{\text{\# of doc in corpus}}{\text{\# of doc contain term in corpus} + 1}\right) \tag{6}$$

The TF-IDF technique is a widely utilized algorithm in the domain of text classification. The algorithm's formula is composed of two components: TF and IDF. The TF quantifies the occurrence of words within a specific class, effectively capturing their frequency in the text. In contrast, the IDF assesses the importance of a word by considering its rarity across a collection of documents, thus mitigating the influence of commonly occurring words that provide less informational value. In this study, the TF-IDF method capitalizes on the relationship between feature words and the number of texts in which appear. However, it does not account for the variation in feature word distribution across different categories, which can adversely affect classification accuracy. Despite this limitation, the TF-IDF algorithm remains a cornerstone in text classification due to its simplicity and effectiveness in various applications (Xiang, 2022).

2.3.3 N-grams

N-grams are contiguous sequences of n items extracted from a given sample of text or speech. These sequences are fundamental in various NLP applications, such as text prediction, language modeling, and information retrieval. The strength of N-grams lies in their ability to model local context within text sequences effectively, capturing the dependencies between words or characters, which is crucial for tasks like machine translation and speech recognition (Manning and Schütze, 1999).

2.3.4 Bag of words

Bag of Words (BoW) model is a fundamental method in NLP and IR, representing text data as a collection of words without

considering grammar or word order (Harris, 1954). The BoW model involves creating a vocabulary from all unique words in a corpus and then representing each document as a vector based on the frequency of each word within the document. This simple yet powerful technique has been widely used in tasks such as document classification, sentiment analysis, and IR, as it effectively captures the presence of words in documents, which can be indicative of their content (Manning and Schütze, 1999). However, one of the limitations of the BoW model is that it disregards the semantics and context of words, which can lead to a loss of important information (Jurafsky and Martin, 2000). Despite these limitations, BoW remains a popular choice due to its simplicity and effectiveness in various applications.

2.3.5 GloVe

Global Vectors for Word Representation (GloVe) is an unsupervised learning algorithm for obtaining vector representations for words (Pennington et al., 2014). Unlike traditional count-based methods such as the BoW or TF-IDF, GloVe leverages the global statistical information of a corpus. It constructs a co-occurrence matrix of words and captures the ratios of word co-occurrences to encode semantic relationships in a lower-dimensional space. This method allows GloVe to preserve linear substructures in the vector space, enabling analogical reasoning and capturing semantic similarities between words. GloVe has shown superior performance in various natural language processing tasks, including word analogy and word similarity benchmarks, and has become a popular choice for generating word embeddings that are used in downstream tasks such as text classification, machine translation, and sentiment analysis (Brochier et al., 2019).

The GloVe model effectively bridges the gap between count-based methods and predictive models like Word2Vec by combining the strengths of both approaches. While Word2Vec captures local context through sliding windows, GloVe integrates this with global statistical information, leading to more robust and meaningful word vectors (Levy and Goldberg, 2015). The ability of GloVe to capture both syntactic and semantic relationships between words is further enhanced by its ability to scale efficiently across large datasets, making it ideal for tasks that require high-quality word embeddings (Li et al., 2018). Additionally, GloVe's embeddings have been shown to perform well across different languages and domains, contributing to its widespread adoption in the NLP community for applications ranging from machine translation to question-answering systems (Bojanowski et al., 2017).

2.4 Prediction models

2.4.1 Artificial Neural Network

Artificial Neural Networks (ANNs) are computational models inspired by the structure and functioning of biological neural networks. An ANN consists of layers of interconnected nodes, or neurons, that process and transmit information. These networks are typically organized in an input layer, one or more hidden layers, and an output layer (LeCun et al., 2015). Each neuron applies a nonlinear activation function to the weighted sum of its inputs, enabling the network to capture complex patterns in data. ANNs have been widely applied in various domains, including computer vision, natural language processing, and speech recognition, due to their ability to learn from data and generalize to unseen examples (Schmidhuber, 2015). One of the key advantages of ANNs is their ability to perform hierarchical feature extraction, where higher-level representations are built from lower-level features (Goodfellow et al., 2016). This makes ANNs particularly effective in tasks that involve high-dimensional and unstructured data.

2.4.2 Random Forest

Random Forest (RF) is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees (Breiman, 2001). The core idea behind Random Forest is to combine the predictions of multiple decision trees, each trained on a random subset of the data, to improve accuracy and control overfitting. This approach reduces variance by averaging the results, making RF highly robust against noisy data and overfitting, especially in high-dimensional spaces (Liaw and Wiener, 2002). Moreover, Random Forest provides an intrinsic measure of feature importance, which can be valuable in interpreting the model's decisions (Ho, 1998). Due to its versatility and performance, Random Forest has been widely adopted in various fields, including bioinformatics, finance, and remote sensing.

2.4.3 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an advanced implementation of gradient boosting designed to enhance the performance and efficiency of machine learning models (Chen and Guestrin, 2016). XGBoost builds upon the principle of gradient boosting, where models are trained sequentially to correct the errors of previous models by optimizing a loss function. XGBoost introduces several innovations, including a regularization term to prevent overfitting, and efficient handling of sparse data and missing values (Chen et al., 2015). Furthermore, XGBoost is designed to be highly scalable, capable of running on distributed systems and handling large datasets with millions of examples (Zhang et al., 2017). Due to its ability to deliver high accuracy, speed, and scalability, XGBoost has become one of the most popular and widely used machine learning algorithms, particularly in competitive data science and applied machine learning.

2.4.4 Support Vector Machine

Support Vector Machine (SVM) is a powerful supervised learning algorithm used primarily for classification tasks, but it can also be applied to regression problems (Cortes and Vapnik, 1995). SVM works by finding the optimal hyperplane that maximally separates data points of different classes in a high-dimensional space. The main objective is to maximize the margin between the nearest points of different classes, known as support vectors, to the hyperplane (Vapnik, 1998). This approach makes SVM highly effective in high-dimensional spaces and well-suited for complex

datasets where the classes are not linearly separable. To handle such cases, SVM employs the kernel trick, which implicitly maps input features into higher-dimensional spaces, enabling the algorithm to find non-linear decision boundaries (Schölkopf and Smola, 2002). Due to its robustness and high accuracy, SVM has been widely used in various fields, including text classification, image recognition, and bioinformatics.

2.5 LIME

Local Interpretable Model-agnostic Explanations (LIME) is a popular XAI technique designed to interpret predictions made by complex, black box ML models (Ribeiro et al., 2016b). LIME has been applied to enhance the interpretability of models that predict mental disorders from social media data. By providing explanations for individual predictions, LIME helps in understanding which features (words, phrases, or patterns) in the text contribute most to the detection of conditions like depression or anxiety. This transparency is crucial for validating the model's decisions and ensuring that they align with clinical knowledge and intuition. LIME is also valuable in educational settings and research. It aids in demonstrating the internal workings of machine learning models to students and researchers. For example, in research focusing on detecting depression from X data, LIME can be used to show the significance of specific keywords or patterns, facilitating a better understanding of the model's behavior and improving its design and accuracy (Guo et al., 2023b).

By incorporating LIME into mental disorder detection models, researchers and practitioners can ensure that their models are not only accurate but also interpretable and trustworthy. This makes LIME a valuable tool in developing and deploying AI-based mental health diagnostics. The use of LIME enhances the transparency of machine learning models. In mental health applications, this transparency helps in gaining the trust of clinicians and patients, as they can see which features are influencing the model's predictions and assess whether these align with clinical expertise and evidence (Khoo et al., 2024; Akhtar et al., 2025).

3 Literature review

The detection of mental disorder through social media content is garnering significant attention from the research community (Santos et al., 2023; Amanat et al., 2022a; Chanda et al., 2022; Ji et al., 2022a; Haque et al., 2022; Wani et al., 2022; Rizwan et al., 2022; Ramírez-Cifuentes et al., 2021; Ghosh and Anwar, 2021; Mohammed et al., 2021). Table 1 provides a summary of related studies in the literature on mental disorder detection using machine learning and NLP techniques.

More recently, Ibrahimov et al. (2025) emphasized model transparency alongside performance, introduced Depression X, a knowledge-infused residual attention model achieving a 7% F1 improvement while providing interpretable insights. Similarly, Qasim et al. (2025) utilized transformer-based architectures (e.g., BERT/RoBERTa) to assess depression severity directly from social media text. In another study, Friedman et al. (2024) proposed

TABLE 1 Comparison table of some literature review.

Study	Data source	Methods and accuracy
Helmy et al. (2024)	English & Arabic Tweets	TF-IDF, BOW Lgbm 96.3%, RF 95.7% L-svm 95.9%, Rbf-svm 20%, LR 96.4%
Santos et al. (2023)	Twitter/X	LIWC 58%, 67%, 56%
Adarsh et al. (2023)	Reddit	SVM + KNN 98.05%, SVM 84.92%, DT 86.16%, RF 86.64%, XGBoost 88.48%, CNN 89.42%
Kabir et al. (2023)	Twitter/X	SVM 51%, 51%, 51%, 54% BiLSTM 62%, 56%, 79%
Amanat et al. (2022a)	Text Tweets	RNN 99%
Chanda et al. (2022)	Twitter/X	SVM 71%, KNN 62%, RF 54%, DT 52%
Ji et al. (2022a)	Twitter/X & Reddit	CNN 78%, LSTM 80%, BiLSTM 82%, RCNN 80%, SSA 81%, RN 83%
Haque et al. (2022)	Twitter/X	ML 93%, TN 94.0%, TN 92.5%, BiLSTM 93%
Saha et al. (2022)	Twitter/X	CNN 41%, RU44%, LSTM 45% Bi-GRU 41%, Bi-LSTM 41%
Wani et al. (2022)	Twitter/X, FB Youtube	CNN 98.15%, Word2Vec LSTM 92.19% CNN + LSTM 91.48%
de Souza et al. (2022)	Reddit	LSTM 65%, CNN 79%, Hybrid 72%
Tong et al. (2022)	TTDD, CLPsych 2015 LSVT, Statlog, Glass	86%, 85%, 87%, 86%, 87%, 87%
Santhosh Baboo and Amirthapriya (2022)	Twitter/X	RF 73%, LR 77%, SGB 72%
Zhang T. et al. (2022)	Tweets based	CNN 17%, RNN 36%, Transformer based methods 17%, hybrid-based methods 30%
Jain et al. (2022)	Reddit	NB 74.35%, SVM 77.12% LR 77.29%, RF 77.29%
Ramírez-Cifuentes et al. (2021)	Reddit	88%
Ghosh and Anwar (2021)	Twitter/X	Depression score 91%
Mohammed et al. (2021)	Bangala Data	DT 81.56%, RF 91.64%, AB 85.12% XGB 92.80%, GNB 91.06%, MLP 87.29%
Hemmatirad et al. (2020)	Twitter/X & Reddit	Twitter/X 95%, Reddit 73%
Tlachac and Rundensteiner (2020)	Twitter	CNN 86%, LSTM 90%, Naive Bayes 82% NN-BiLSTM with Attention model 97%
Fatima et al. (2020)	eRisk 2018	LR 76%, NB 67%, SVC 67%
Tadesse et al. (2019)	Reddit & Twitter/X	91%

EAC-Net, an emotion-aware encoder leveraging contrastive learning and self-attention, demonstrating superior recall on depression and stress detection across multiple datasets. Expanding into multimodal input, Cha et al. (2024) presented MOGAM, which integrates video, text, and metadata via graph-attention mechanisms, achieving 0.87 accuracy on clinically labeled users. Additionally, Al Asad et al. (2024) developed a BERT+Bi-LSTM pipeline for both English and Arabic, highlighting the importance of explainability in achieving top F1 scores across languages.

In another work, Ibrahimov et al. (2024) highlighted the critical role of XAI frameworks in making mental health AI models transparent and trustworthy. Moving beyond surveys, Chen and Lin (2025) developed LLM-MTD, a large-language-model based multi-task system that simultaneously classifies depression and generates medically informed explanations, achieving state-of-the-art results on the RSDD benchmark. Empirical studies, such as those by Hoque et al. (2025), demonstrate the effective application of explainability tools like SHAP and LIME in real-world educational datasets. Their model attained over 91% accuracy in detecting depression in Bangladeshi university student posts, reinforcing the value of interpretability in high-risk settings.

Tadesse et al. (2019) proposed an approach to identify depression-related posts on Reddit using NLP and ML techniques. Their approach highlighted the significant improvement in detection accuracy by using a combination of linguistic features and classifiers, achieving up to 91% accuracy with a Multilayer Perceptron (MLP) classifier. Their work underscores the importance of feature selection and combination in enhancing the performance of depression detection systems. Furthermore, Guntuku et al. (2017) explored the detection of depression through social media data, highlighting significant advances in NLP and ML that facilitate large-scale mental health screening. Despite these technological advances, the generalizability of such studies to diverse populations and alignment with established clinical criteria remains uncertain. Furthermore, recent work highlighted the pressing need to address ethical, legal, and clinical considerations, particularly concerning data ownership, privacy protection, and the integration of these methods into existing healthcare systems.

Cho et al. (2019) proposed an approach to a comprehensive evaluation of the research that has been conducted on the use of ML algorithms for the diagnosis of depression, as well as recommendations for the practical uses of ML and predicted clinical remission following treatment with citalopram for twelve weeks. The dataset consisted of 1949 sad individuals who were participating in level 1 of the Sequenced Therapy Options to Relieve Depression study. In analyzing mental health using ML techniques, the primary focus is on providing a supervised learning environment for classification.

Another work by Helmy et al. (2024) explored the application of machine learning techniques to identify signs of depression in X data. It introduces manually labeled Arabic and automatically labeled English depression corpora, evaluates various pre-processing, feature extraction, and supervised classification techniques, and demonstrates the viability of machine learning for early depression detection despite recent trends favoring deep learning. This work underscores the importance of diverse, language-specific corpora and provides valuable insights into effective combinations of methodologies for predicting depression severity. The experiments demonstrated the significant impact of feature representation and resampling techniques on classifier performance, with Random Forest (Breiman, 2001) and RBF-SVM (Schölkopf and Smola, 2002; Cortes and Vapnik, 1995) models showing high effectiveness across different scenarios.

Santos et al. (2023) proposed the use of Mixture of Experts models combined with BERT-based approaches for predicting depression and anxiety from self-reports on social media in Portuguese was proposed. Their findings indicate that models outperform traditional feature engineering methods while also allowing for more interpretable models. The finding suggests potential improvements through modifications such as attention mechanisms, hierarchical mixtures, and multi-task learning. More recently, Amanat et al. (2022a) investigated the application of deep learning models for detecting signs of depression in textual data sourced from social media. The proposed framework leveraged LSTM networks and RNNs to analyse and classify text, achieving an impressive accuracy of 99% in early depression detection. The findings underscore the potential of advanced machine learning techniques to enable timely and precise identification of depressive tendencies, offering valuable support for mental health interventions and early assistance strategies.

Guo et al. (2023b) investigated mental health detection using text data from online forums, employing advanced machine learning techniques, including CNNs and LSTM networks. These models captured complex patterns and contextual nuances in textual data. To enhance the interpretability of these inherently black box models, the authors used the LIME technique. The LIME provided insights into the specific language patterns and features that influenced the model's predictions, enabling researchers to link certain textual expressions to mental health conditions. The interpretability increased the model's trustworthiness and supports its integration into clinical settings, where understanding decision rationale is critical for adoption and application.

Furthermore, Joyce et al. (2023) introduced the TIFU framework to enhance the trustworthiness of AI in psychiatry by focusing on transparency and interpretability. The author emphasized the importance of explainable AI, particularly through methods like LIME, to make complex models more understandable for healthcare professionals and patients, enhancing their reliability and acceptance in mental health applications. Abd Yusof et al. (2017) developed a computational model to identify potential causes of depression by analyzing user-generated content. This work identified prominent causes of depression and how they evolved, highlighting differences between individuals with varying levels of neuroticism. Another study, Sabaneh et al. (2023) integrated several advanced methodologies, including the use of ChatGPT-3 for translating Arabic text to English, QuickUMLS (Soldaini and Goharian, 2016) for extracting medical concepts from the translated text, and machine learning algorithms for classification. The researchers utilized a variety of classification algorithms, such as RF, SVM, and LR, with RF achieving the highest accuracy of 80.24%.

Saxena et al. (2022) explored the challenge of multi-class causal categorization of mental health issues on social media, focusing on the problem of incorrect predictions due to overlapping causal explanations. Their work identified inconsistencies in causal explanations as a key reason for varying accuracy by fine-tuning

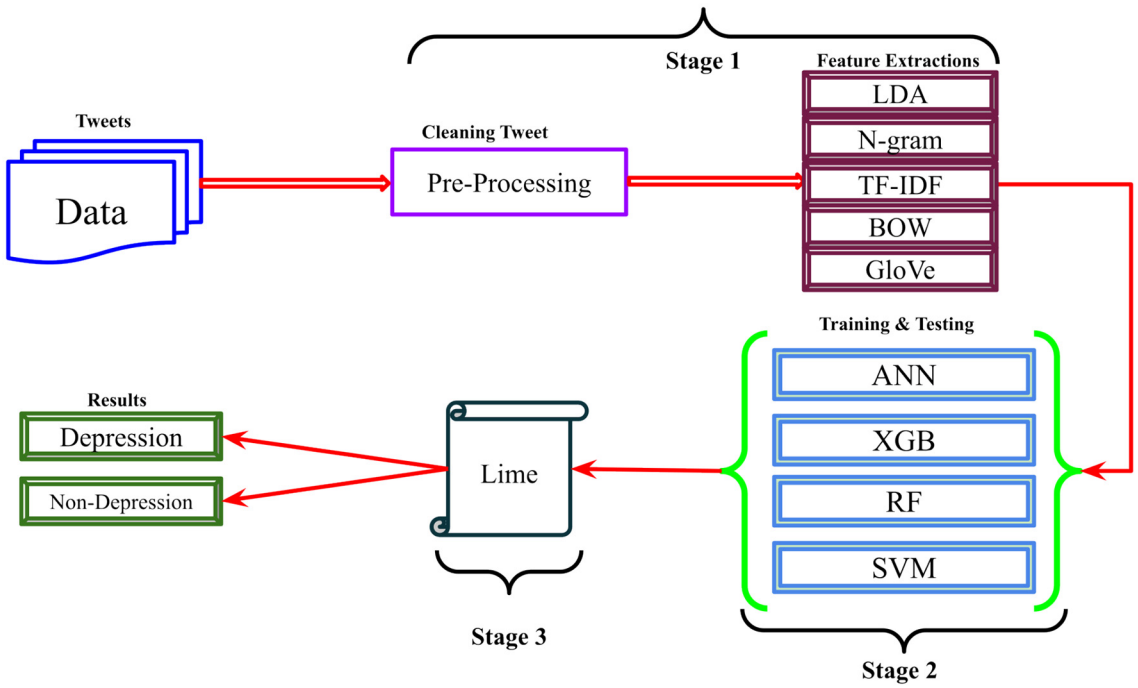

FIGURE 1
The proposed research methodology for depression detection using NLP and XAI techniques.

classifiers and applying LIME and Integrated Gradient methods (Sundararajan et al., 2017; Kokhlikyan et al., 2020; Ancona et al., 2018). The proposed approach was validated on the CAMS dataset, achieving category-wise average scores of 81.29% and 0.906 using cosine similarity and word mover's distance, respectively. Furthermore, Adarsh et al. (2023) utilized LIME to enhance the explainability of their classification model. The LIME was employed to identify and highlight specific words within social media posts that significantly contribute to the classification of posts as either containing suicidal ideations or not. The LIME was used in their approach to enhance the transparency and interpretability of the depression detection model, making it easier to understand and trust the model's decisions, particularly in identifying critical language markers of suicidal ideation.

4 Methods

This research work presents a robust approach for detecting depression from X posts. The proposed approach consists of three steps: First, preprocessing techniques and feature extraction methods; second, machine learning classifiers; and third, interpretability analysis using LIME. The comprehensive methodology is depicted in Figure 1.

4.1 Data collection

The reliability and accuracy of any proposed system are intrinsically linked to the quality and representativeness of the data collected. As such, data serves as the cornerstone of the system's overall effectiveness and performance (Jain et al., 2020). Therefore, the collection and preparation of an appropriate dataset are essential to achieve the desired objectives. The experiments conducted in this work utilized publicly available datasets hosted on Kaggle[1]. In this work, we utilize a dataset derived from posts and comments on X. The analysis focuses on key factors, such as indications of mental illnesses like depression, as reflected in user posts and interactions. Representative instances of the data set are presented in Table 2 for illustrative purposes. Table 3 demonstrates the words for the posts in both categories, like "depression" and "non-depression" which are topically specific.

Posts were included in the dataset if they met specific inclusion criteria: they had to be written in English, contain at least five words to ensure sufficient linguistic context for natural language processing (NLP), and include depression-related keywords such as "depressed," "sad," or "alone," as identified in prior research as indicators of mental distress on social media (Chancellor and De Choudhury, 2020). Conversely, exclusion criteria were applied to remove posts that could compromise the reliability of the model. These excluded posts containing only emojis, links, or hashtags due to their lack of semantic and syntactic depth, advertisements or spam-like content that do not represent authentic user emotions and introduce noise, and explicitly sarcastic or humorous content, which may distort model predictions due to linguistic ambiguity.

For this work, a total of 1,600,000 tweets were collected, representing a diverse spectrum of user experiences related to critical factors such as depression and other mental health

1 https://www.kaggle.com/datasets/adedamolaajewole/training1600000processednoemoticon

TABLE 2 Labeled instances from the X dataset used for classification.

User	Tweet	Class
Scotthamilton	Is upset that he can't update his Facebook by texting it... and might cry as a result School today also. Blah!	Non Depression
Mattycus	@Kenichan I dived many times for the ball. Managed to save 50% The rest go out of bounds	Non Depression
BaptisteTheFool	Meh... Almost Lover is the exception...this track gets me depressed every time.	Depression
erika_strange	@infidelsarecool ugh how depressing. i want to punch something.	Depression
ACTinglikeamama	@gigdiary I know - was a little depressed that we ate so much last night there were no leftovers today	Depression

TABLE 3 Words frequently used in depressive text.

Depression	Non depression
Alone, break, blame, depressed,	Go, days, aww, almost, holiday
Unhappy, worry, exam, rubbish	UK, son, YouTube, liked, Chicago
Danny, office, upset, past, reason	Happy, dreamy, love, Faith, Games
Needs, dead, hmmm, random, sd	Awsome, money, Movie, Frnds, hills
Waiting, hurt, blocked, cry, lost	Really, half, mad, episode, loved
Headache, summer, death, sucks	Lucky, cute, girls, town, visit
Miley Cyrus, job, Painfull, Massive	Needs, rest, excited, joy, happy
Upset, kick, dumb, Unsuccessful	Haha, listening, high, puppy, oooh
Disappointed, kill, Sadly, end	Went, ago, finished, drink, milk
STILL, feeling, busy, dark, migraine	DAMN, please, play, song, dance

conditions. The data collection process involved filtering tweets using keywords indicative of mental health issues, including terms such as "depressed," "anxiety," and other relevant expressions.

4.2 Data pre-processing

Before feature selection and model training, NLP techniques were applied to pre-process the collected dataset. The initial step involved cleaning X posts from the data, resulting in a substantial dataset ready for feature extraction. The data pre-processing steps are illustrated in Figure 2.

The following pre-processing steps were implemented:

Tokenisation: the process of tokenization involves splitting the textual data into individual units, typically words or tokens, to facilitate further linguistic analysis. This step enables the transformation of unstructured text into a structured format suitable for feature extraction and modeling.

Noise removal: to enhance the quality of the dataset, noise removal techniques were applied. This included eliminating irrelevant elements such as URLs, punctuation marks, numerical values, and common stop words that do not contribute a significant semantic meaning. By refining the data set in this way, the subsequent analysis becomes more focused and meaningful.

Stemming: stemming techniques were employed to reduce words to their root or base forms, thereby minimizing variations of the same word (e.g., "running" and "ran" both reduced to "run"). This step helps to normalize inflected words and consolidate similar terms, leading to a more compact and informative feature space.

Normalization: normalization was carried out by converting all text into lowercase, ensuring uniformity across the dataset. This step prevents the algorithm from treating words with different cases (e.g., "Text" vs. "text") as distinct entities, thereby improving the consistency and reliability of the text representation.

4.3 Feature extraction

After pre-processing, we employed several standard feature extraction techniques to capture the linguistic and semantic characteristics of user-generated posts. To reduce feature dimensionality while preserving document-level semantic structure, LDA was applied, modeling 70 latent topics. The TF-IDF vectors were generated to weight word importance across the corpus, facilitating the identification of salient terms. Both unigrams and bigrams were extracted using the Scikit-learn library, limiting to the top 3000 most frequent n-grams to improve contextual understanding in short texts. Additionally, the BoW representation was used as a baseline, relying on sparse word counts without considering syntactic relationships. Finally, pre-trained GloVe embeddings were incorporated to capture semantic similarity and contextual relationships between words in dense vector form. These diverse feature representations were used as inputs for training multiple machine learning models for classification.

The selection of feature extraction methods in this study was guided by two key considerations. First, we aimed to include a mix of traditional lexical representations (TF-IDF, N-gram, BOW), topic modeling approaches (LDA), and dense vector embeddings (GloVe) to capture both surface-level and semantic aspects of text. This diversity allows us to evaluate how LIME explanations differ when models are trained on features with fundamentally different representational properties. Second, we selected methods that are widely used in prior depression detection and sentiment analysis research, enabling meaningful comparison with existing literature and ensuring reproducibility.

While alternative embedding methods such as FastText, Word2Vec, or contextual embeddings like BERT are available, GloVe was chosen because it offers strong semantic representation with relatively low computational cost, making it suitable for large-scale experiments across multiple classifiers. Additionally, GloVe embeddings are static, which ensures that any interpretability differences observed using LIME are attributable to the model and feature-classifier interaction, rather than dynamic embedding variability. This controlled setting aligns with our goal of producing a consistent, explainability-focused benchmark rather than exhaustively comparing all possible embedding types.

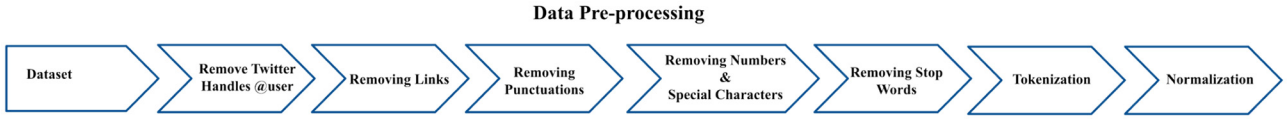

FIGURE 2
Data pre-processing.

GloVe was selected as the representative dense vector embedding method in our study for several reasons. First, GloVe captures global co-occurrence statistics, allowing it to encode semantic relationships between words effectively, an important factor for short-text, depression-related posts, where subtle semantic cues may indicate emotional state. Second, its extensive prior use in depression detection and sentiment analysis literature ensures comparability with existing work. Third, preliminary trials on our dataset indicated that GloVe produced slightly higher accuracy and more consistent performance across classifiers compared to FastText, whose subword-level advantages were less pronounced in our data due to the prevalence of short, informal tokens. By including GloVe alongside statistical (TF-IDF, N-gram, BOW) and probabilistic topic-modeling (LDA) methods, we aimed to evaluate LIME explainability across a diverse spectrum of feature representations.

In literature, TF-IDF is widely used to identify term importance in documents; it suffers from known drawbacks: high-dimensional, sparse vector representations, and an inability to capture semantic relationships such as synonymy or context, especially in short texts like tweets (Xu et al., 2013; Joshi et al., 2020). To address these issues, we limited the TF-IDF vocabulary to the top-N frequent terms (e.g., top 5,000) to reduce dimensionality and noise, following best practices in short-text feature selection. We also incorporated semantic embeddings such as GloVe to enrich the representations with contextual meaning beyond term frequency. Finally, we employed LIME to provide *post-hoc* interpretability, helping validate and visualize influential terms that drive classification decisions in our depression detection models.

4.4 Model training and validation

We divided the dataset into three distinct subsets: training, validation, and testing. The training set comprised 70% of the total data, amounting to 1,120,000 tweets, while the validation and testing sets included 15% each, consisting of 240,000 tweets for validation and 240,000 tweets for testing. This division allows us to effectively develop and fine-tune our models while ensuring an unbiased evaluation.

The class labels "depression" or "non-depression" were intuitively assigned based on the presence of specific depression-related keywords in the tweets, such as "depressed," "sad," "cry," and "alone." These labels were generated using a perception-based weak labeling approach, which is common in social media mental health detection studies. While this approach facilitates large-scale data collection, it may introduce noisy labels, which we mitigated through extensive pre-processing and validation using multiple classifiers. We acknowledge this limitation and highlight it in our Discussion section, suggesting the integration of expert-driven annotation in future work.

TABLE 4 Classifier hyperparameters used for depression detection, selected through empirical tuning and standard practices.

Model	Parameter	Value/setting
ANN	Input layer	TF-IDF or GloVe embeddings
	Hidden layers	Two: 64 and 32 neurons
	Activation	ReLU (hidden), Softmax (output)
	Optimizer	Adam (learning rate = 0.001)
	Loss function	Categorical Cross-entropy
	Epochs/batch size	20/32 with early Stopping
SVM	Kernel	Linear
	Regularization (C)	1.0
	Decision function	One-vs.-rest
Random forest	Number of trees	100
	Max depth	None (expand until pure)
	Criterion	Gini Index
	Bootstrap	True
XGBoost	Number of estimators	100
	Max depth	6
	Learning rate	0.1
	Objective	Binary:logistic

Table 4 outlines the key hyperparameters and settings used for each classifier. These values were chosen based on iterative testing on the validation set, informed by prior literature and practical experimentation.

To check how well our models worked, we used a common method called hold-out validation. We split the dataset into 80% for training and 20% for testing. The test data was not used during training or tuning, so we could see how well the model performs on new, unseen data. We also used five-fold cross-validation while training models like SVM, Random Forest, and ANN. This method helps us fine-tune model settings and reduce the chance of overfitting. To measure performance, we used standard metrics like accuracy, precision, recall, and F1-score. We also created confusion matrices and ROC curves to visualize how the models performed. These results were all based on the test data to make sure they were reliable. For the ANN, we used early stopping to avoid overfitting. This means the training stopped automatically when the model stopped improving on the validation data.

To classify tweets as indicative of depression or not, we selected and implemented several well-established machine learning models that have demonstrated strong performance in text classification tasks. We utilized the ANN, specifically a Multi-Layer Perceptron (MLP) architecture with two hidden layers of 4 and 16 neurons, respectively, as this provided a manageable level of complexity for evaluating various feature representations. The RF algorithm was selected for its robustness against noisy data and its ability to mitigate overfitting through the aggregation of multiple decision trees and random feature selection. We also employed XGBoost, chosen for its superior accuracy and regularization capabilities, which are achieved by sequentially optimizing weak learners to minimize classification error. Additionally, we incorporated the SVM because of its effectiveness in handling high-dimensional feature spaces, utilizing kernel methods to identify optimal separating hyperplanes. Each of these classifiers was trained using the same preprocessed and vectorized data, ensuring a fair and consistent comparison of their respective performance in detecting depression from tweets.

4.5 Performance evaluations

We assessed the performance of the model by applying well-known performance metrics, including accuracy and precision, and recall. The formulas of these evaluation metrics are shown below:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

5 Results

The robustnsess of the proposed approach lies in our preprocessing pipeline, which refers to its ability to effectively clean and normalize noisy, informal social media text, including irregular spellings, emoticons, repeated characters, and abbreviations which are common on platforms like X (Chancellor and De Choudhury, 2020; Steinkamp and Cook, 2021b). This pipeline improved the quality of feature extraction by reducing vocabulary sparsity and enhancing model generalizability. Our approach also ensured consistent performance across all tested classifiers (ANN, SVM, XGBoost, RF), demonstrating resilience against data variation, which is crucial when working with user-generated, unstructured data. Compared to existing methods (e.g., Ji et al., 2022a; Amanat et al., 2022a), our framework achieves higher or comparable accuracy using simpler architectures (e.g., GloVe+RF: 88%, SVM+TFIDF: 79%), while maintaining interpretability through LIME, which is rarely integrated in similar works (Ji et al., 2022b; Amanat et al., 2022b). This hybridization of multiple NLP features with black box models, accompanied by transparent explanations, offers a practical and explainable solution for early depression detection. Furthermore, our system does not require deep learning or transformer-based models, making it more computationally efficient and suitable for real-world deployment.

Once the performance of the ML model has been evaluated, it becomes essential to analyse and interpret the findings to gain deeper insights into the model's performance. This involves discerning the crucial features influencing the model's predictions, comprehending the relationships between these features and the target variable, and identifying any pertinent patterns or trends within the dataset. This work employed a comprehensive set of experiments to detect depression from X posts using various combinations of feature extraction methods and ML classifiers. To visualize the most prominent terms that express emotions, a word cloud representation is utilized. Figure 3a demonstrates the depressive user's feelings, experiences, and stories. This method summarizes the ideas and phrases most commonly linked to mental disorders in the analysis of social media discussions. Figure 3b illustrates the word cloud representing the positive core words of the dataset. Word clouds are visual representations that draw attention to the most prevalent terms in a collection of text. In addition, the prominence of a word in the cloud reflects its frequency of use in the corresponding tweets. This method summarizes the ideas and phrases most commonly linked to mental disorders in the analysis of social media discussions. On the contrary, Table 5 presents the words correlated with the specific topics generated from the posts. These topics comprise a lexicon of words commonly used among accounts associated with depression.

The results of research experiments are summarized in Table 6, which shows the accuracy and prediction probabilities for each combination of feature extraction method and classifier. Additionally, Table 7 provides a comparative analysis of several ML classifiers, namely ANN, XGBoost, RF, and SVM, using different feature extraction techniques. These features include LDA, TF-IDF, N-gram, BOW, GloVe, and various combinations of these techniques. The classifiers are evaluated based on their performance metrics: Precision, Recall, and F1-score.

It should be noted from Table 6 that the GloVe feature extraction technique combined with RF achieved the highest accuracy of 88%, demonstrating its ability to capture rich semantic information effectively. SVM also performed well with GloVe, achieving an accuracy of 85%. TF-IDF and N-gram modeling showed competitive performance, with XGB achieving an accuracy of 77% and 78%, respectively, demonstrating their effectiveness in capturing text features. BOW proved to be a reliable feature extraction method, with XGB and ANN yielding accuracies of 78% and 73%, respectively. The combination of LDA with TF-IDF and N-gram yielded an accuracy of 78% for ANN, showcasing the potential of combining multiple feature extraction techniques. Our experiments demonstrate the varied strengths of feature extraction methods and classifiers, with GloVe providing the most insightful semantic information. At the same time, N-gram and BOW maintained a consistent balance of performance across models.

Figure 4 presents a comparative analysis of four ML classifiers ANNs, XGB, RF, and SVM across various feature extraction methods: LDA, TF-IDF, N-gram, BOW, GloVe, LDA+TF-IDF+Ngram, and LDA+BOW+TFIDF. The performance of each

(a) Negative Words (b) Positive Words

FIGURE 3
Word cloud for training data to visualize important. **(a)** Negative and **(b)** positive words.

combination is evaluated and depicted in terms of accuracy percentages. The highest performance is achieved by the GloVe feature extraction method, followed closely by the SVM algorithm at 86%. Overall, the GloVe method consistently outperforms other feature extraction techniques, indicating its effectiveness in capturing word semantics and improving classification accuracy. Other notable performances include the LDA+TF-IDF+Ngram method, which demonstrates balanced accuracy across different classifiers, particularly with SVM classifiers. This comprehensive comparison underscores the importance of selecting suitable feature extraction methods to boost the predictive power of machine learning models in text classification tasks.

5.1 Overview of evaluation

The experiments were performed on the same dataset for all feature extraction methods and classifiers to ensure consistent comparison. We evaluated seven feature extraction methods (LDA, TF-IDF, N-gram, BOW, GloVe, LDA+TFIDF+N-gram, and LDA+BOW+TFIDF) and four classifiers (ANN, XGBoost, RF, SVM). To improve analytical clarity, results are presented in two separate views: (1) ranking of feature extraction methods averaged across all classifiers, and (2) ranking of classifiers averaged across all feature extraction methods. Detailed per-configuration precision, recall and F1-score values are provided in the Tables 8, 9.

5.2 Feature extraction method rankings

Table 8 reports the average accuracy of each feature extraction method calculated over all four classifiers. This ranking shows which feature representations perform best on average in our study.

Interpretation: GloVe embeddings provide the highest average accuracy (82.75%) across classifiers, indicating that dense semantic representations capture context useful for depressive-linguistic signals in our dataset. Traditional lexical representations (TF-IDF, N-gram, BOW) remain competitive and may be preferable in resource-constrained settings.

5.3 Classifier rankings

Table 9 shows the average accuracy of each ML classifier across all feature extraction methods. This ranking isolates classifier performance independent of any single feature choice.

6 LIME analysis

Despite their excellent performance, several ML models are often characterized as black boxes that produce outputs without offering explicit insights into the underlying reasoning behind their decisions. Understanding and interpreting the decision-making processes of such models is critical, particularly in applications where trust, transparency, and accountability are paramount. Consequently, it is imperative to examine the outputs of these models thoroughly and, more importantly, to develop methodologies that enable the generation of interpretable explanations for their decisions (Søgaard, 2021; Gohel et al., 2021). Providing explanations for a model's output enhances our ability to evaluate its predictions critically, thereby fostering greater confidence in determining whether to trust or question its outcomes.

To investigate the model's explainability, we employed LIME, a popular XAI technique that facilitates the interpretation of outputs without requiring direct inspection of the model's internal structure. LIME achieves this by perturbing the local features surrounding a specific target prediction and analyzing the corresponding changes in the model's output. In our experiments, the words surrounding a target entity were modified systematically, and the effects on the model's predictions were subsequently assessed to gain insights into the decision-making process.

Each subplot in Figure 5 illustrates the LIME analysis visualizations, providing an interpretable explanation of the predictions made by different classifiers for specific instances. Each

TABLE 5 Topics extracted with LDA.

Sr	Topics	Words
1	Daily activities	Haha, play, friend, year, stop haha play friend year stop yeah tell think today chang want left know word
2	Planning:	Hous, damn, love, plan, like hous damn love plan like trip time dear usual watch lost today tomorrow cook list want
3	SocialMedia usage	Gonna, enjoy, thing, fun, welcome gonna enjoy thing fun welcom weather X love come cool sure
4	Home life	Home, week, away, head, update home week away head updat night guess today stuck bought outside
5	Depression	Good, feel, morn, better, hope good feel morn better hope morning right like hate night realli make today coffe bed think sleep class
6	Affection nostalgia	Readi, pretti, miss, yay, love readi pretti miss yay love song tomorrow goodnight saturday amp sleep hang alreadi night
7	Planning anticipation	Day, ll, someth, think, later day ll someth think later start make beauti tomorrow happen days amp let money
8	Work productivity	Work, glad, time, snow, home work glad time snow home hard today earli okay tonight easter night
9	Celebrations greetings	Happi, birthday, wonder, peopl, mani happi birthday wonder peopl mani love best sorri repli realli sooo
10	School life	Today, school, life, break, lunch today school life break lunch hour watch hear spring funni till wanna room
11	Quotations humor	Quot, listen, cold, like, hahaha quot listen cold like hahaha music babi love problem hey th great haha brother smile song
12	Love relationships	Love, tweet, nice, tire, amp love tweet nice tire amp awesome twitter summer train join kinda ddlovato
13	Happy	Awesom, Watch, like, send, wear awesom watch like send wear place love asot400 way help house man make amp fail
14	Technology	Know, need, want, twitter, hello know need want twitter hello dont think phone like love realli mileycyru fm cute
15	Reading learning	Read, alway, book, final, food read alway book final food time gone dinner believ think iphon famili tonight pick
16	Weekend activities	Good, weekend, girl, sick, luck good weekend girl sick luck night rain dream wish fuck shower tuesday great time today
17	Online engagement	Http, com, thank, follow, twitpic http com thank follow twitpic www tinyurl thanks check twitter link
18	Future plans	Time, watch, soon, movi, long time watch soon movi long suck come today love want realli heard movie anoth real
19	Sleep relaxation	Great, sleep, time, post, hope great sleep time post hope bore bit late everyth ly free breakfast http day
20	Visuals photos	Look, like, wait, forward, picture look like wait forward pictur sound realli think ll gt welcome twitter awww

subplot highlights the contribution of individual words (features) toward the prediction of either "Depression" or "Non-Depression" labels. The importance of the words is represented as bars, where positive contributions toward "Depression" are shown in blue, and contributions toward "Non-Depression" are shown in orange.

Figure 5a demonstrates LIME analysis for ANN with LDA features extraction on the 10th instance. It can be noted that the word "snow" contributed significantly to the "Depression" label, with a probability score of 0.54, compared to 0.46 for "Non-Depression." This shows how specific context-sensitive words can influence predictions. Similarly, Figure 5b demonstrates LIME analysis for XGB and TF-IDF on the 960th instance, the words "wish," "sleep," and "tooo" strongly contributed to the "Depression" label, achieving a high probability score of 0.87 for "Depression". This highlights the model's ability to capture sentiment-relevant features using the TF-IDF approach.

In addition, Figure 5c demonstrates a dominant contribution of the words "miss" and "sorry" toward the "Depression" label, achieving a probability score of 0.93. This emphasizes the BOW model's capability to detect sentiment-related words. Similarly, Figure 5d visualizes that, RF with N-Gram on the 2000th instance, the words "work," "desk," and "10am" contribute strongly toward the "Depression" label, with a probability score of 0.91. This demonstrates how the N-Gram technique effectively captures contextual word co-occurrences. On the other hand, Figure 5e indicates that ANN with GloVe embeddings on the 3000th instance, provided a contrasting prediction, with the word "islandiva147" contributing more toward "Non-Depression" than "Depression." This suggests that GloVe embeddings capture semantic nuances but may misinterpret words out of context.

Overall, these visualizations illustrate the interpretability of the classifiers and their reliance on feature-specific contributions. While ANN and SVM demonstrated strong performance on LDA and BOW features, respectively, XGB and RF highlighted the importance of TF-IDF and N-Gram features. GloVe embeddings, while semantically rich, occasionally misinterpret specific instances, underscoring the importance of feature selection.

TABLE 6 Results of classification performance ML models and feature selection methods based on accuracy.

Sr	Features	ANN	XGBoost	RF	SVM
1	LDA	72	68	72	72
2	TF-IDF	72	77	78	79
3	N-gram	73	78	76	77
4	BOW	73	78	75	78
5	GloVe	72	86	88	85
6	LDA+TFIDF+Ngram	78	77	72	72
7	LDA+BOW+TFIDF	76	77	77	78

7 Discussion

Our study examines various feature sets and classifiers, achieving notable accuracies, particularly when combining GloVe with classifiers such as RF and SVM. GloVe+RF, achieved 88% accuracy obtained by Ji et al. (2022a). In comparison, Amanat et al. (2022a) reported accuracies up to 96.4% using a combination of TF-IDF, BOW, SVM, and RF. While our results for GloVe+RF (88%) and GloVe+SVM (85%) are slightly lower, they are still competitive given the different data sources and methodologies used. which is within the range reported by comparable studies

TABLE 7 Comparison of ML models and feature selection methods based on additional metrics such as precision, recall, and F-score.

Sr	Features	ANN			XGBoost			RF			SVM		
		Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
1	LDA	73	94	82	74	87	80	74	94	81	72	98	83
2	TF-IDF	82	80	81	80	90	85	80	92	86	81	92	86
3	N-gram	83	79	81	81	91	86	81	86	84	81	94	86
4	BOW	83	81	85	81	91	85	82	84	83	79	94	86
5	GloVe	72	99	83	86	88	87	97	92	94	83	94	90
6	LDA+TFIDF+Ngram	80	91	85	82	87	84	79	92	84	79	92	84
7	LDA+BOW+TFIDF	82	85	84	78	94	86	77	95	85	79	94	86

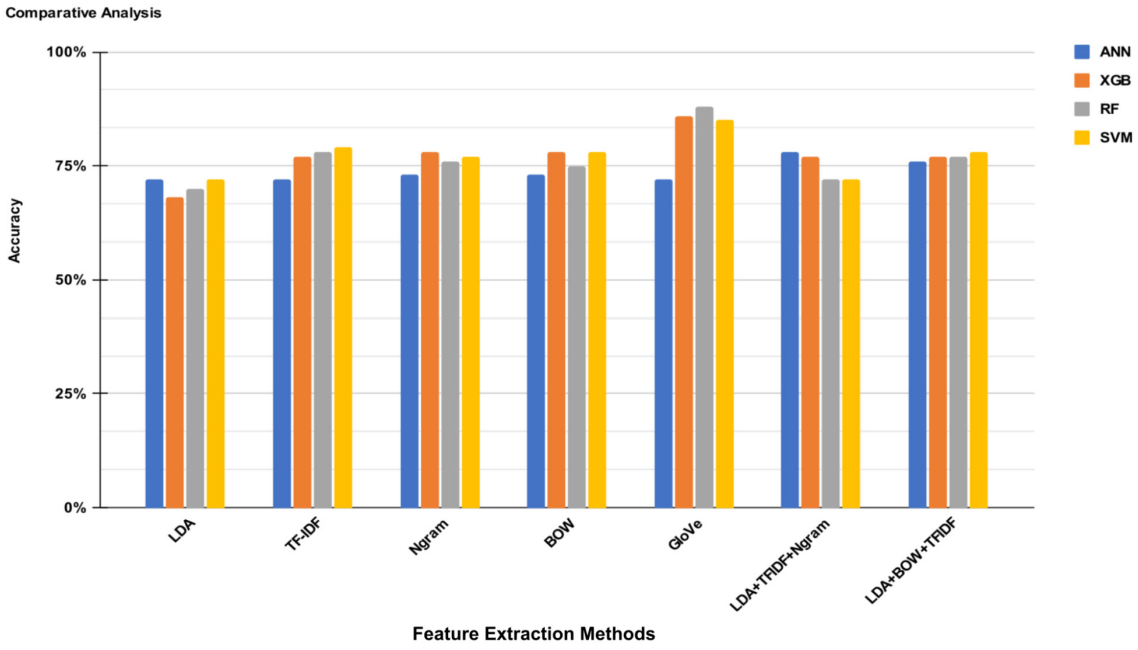

FIGURE 4
Comparison of ML models and feature selection methods based on accuracy.

TABLE 8 Average accuracy of feature extraction methods across ML classifiers.

Feature extraction	Accuracy (%)	Rank
GloVe	82.75	1
LDA + BOW + TF-IDF	77.00	2
TF-IDF	76.50	3
N-gram	76.00	4
BOW	76.00	5
LDA + TF-IDF + N-gram	74.75	6
LDA	71.00	7

TABLE 9 Average accuracy of ML classifiers across feature extraction methods.

Classifier	Avg accuracy (%)	Rank
XGBoost	77.29	1
SVM	77.29	2
Random Forest (RF)	76.86	3
ANN	73.71	4

on depression detection using traditional machine learning approaches. Variations in reported accuracy across the literature are largely attributable to differences in datasets, preprocessing steps, and feature-classifier combinations, making direct score comparisons less meaningful. Instead, the emphasis here is on showing that competitive performance can be achieved alongside enhanced interpretability.

Among feature extraction methods, GloVe embeddings achieved the highest average accuracy across classifiers (82.75%), followed by LDA+BOW+TF-IDF (77.00%) and TF-IDF (76.50%).

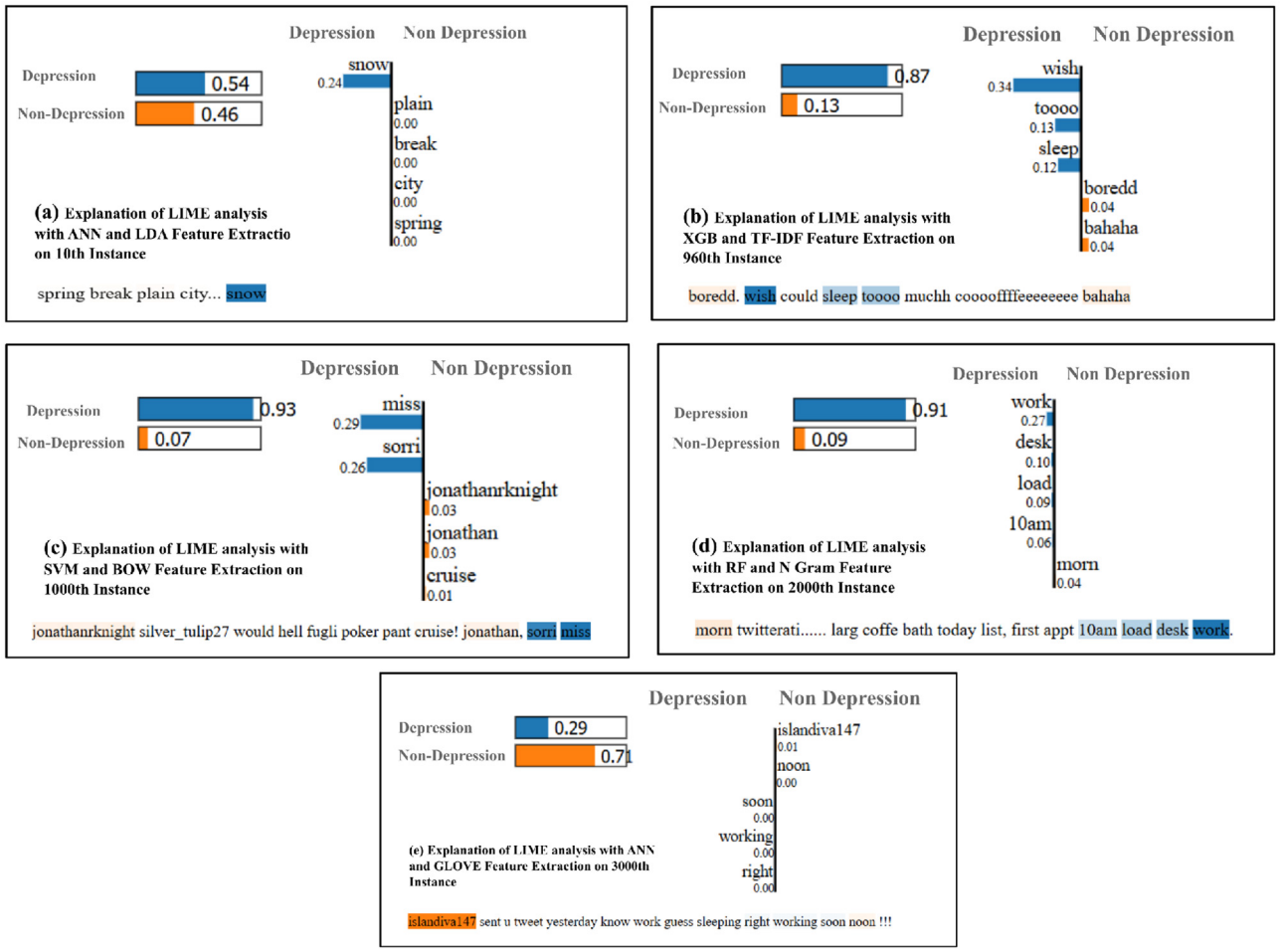

FIGURE 5
Example of explainable AI visualizations using LIME. **(A)** ANN and LDA feature extraction on 10th instance. **(B)** XGB and TF-IDF feature extraction on 960th instance. **(C)** SVM and BOW feature extraction on 1000th instance **(D)** RF and N gram feature extraction on 2000th instance. **(E)** ANN and GLOVE feature extraction on 3000th instance.

This indicates that semantic embeddings like GloVe capture richer contextual information than purely lexical representations. These results are consistent with prior work—for instance, Tong et al. (2022) reported 86% accuracy using GloVe-based features in a multi-classifier ensemble, whereas our GloVe+RF model reached 88% on this dataset. Although Amanat et al. (2022a) achieved higher performance (96.4%) with TF-IDF+BOW+SVM/RF, differences in datasets and preprocessing make direct comparisons indicative rather than conclusive. Notably, N-gram and BOW features, while ranking lower in average performance (76.00%), still matched or exceeded the accuracy of some deep learning models in the literature, such as the CNN (78%) and LSTM (80%) reported by de Souza et al. (2022), demonstrating that simpler representations can be competitive for certain datasets.

Among ML classifiers, XGBoost and SVM obtained the highest average accuracy across feature sets (77.29%), closely followed by Random Forest (76.86%), with ANN ranking lowest (73.71%). This suggests that tree-based ensembles and margin-based classifiers are generally better suited to the depression detection task when trained on short, noisy social media text. These trends align with the findings of Wani et al. (2022), where SVM achieved 71% and KNN 62% using N-gram features, and with Ji et al. (2022a), who reported competitive performance with SVM on short-text datasets. In our experiments, ANN performed best with the combined LDA+TF-IDF+N-gram feature set (78%), which slightly exceeds the BiLSTM performance (79%) reported by Haque et al. (2022), showing that under certain feature configurations, neural networks can still achieve strong results.

A key finding is that performance depends on the interaction between the feature set and the classifier. For example, while GloVe consistently ranks highest among features, its combination with RF (88%) and SVM (85%) outperformed its pairing with ANN (72%). Similarly, the LDA+TF-IDF+N-gram feature set worked particularly well with ANN (78%), but less so with RF (72%). These variations underscore the importance of evaluating both dimensions independently before selecting an optimal configuration.

Another distinctive aspect of this study is the systematic application of LIME across all feature-classifier combinations, rather than to a single model. LIME provided interpretable, instance-level explanations, identifying key linguistic cues such as first-person pronouns, negative emotion words, and self-referential

phrases that heavily influenced depressive content predictions. This transparency is essential for building trust with mental health professionals and differentiates our work from most prior studies, where explainability is rarely addressed at this scale.

It should be noted that the accuracy of depression detection models heavily relies on the authenticity and consistency of users' social media content. Factors such as bias for social desirability, stigma, and personal tendencies can influence the way users express themselves online, potentially leading to inaccuracies in detection (Shah et al., 2025). In general, detecting depression from social media posts inherently depends on the assumption that users share relevant depressive symptoms or emotional cues in their online interactions. However, not all individuals with depression disclose their condition or express depressive symptoms publicly on social media, which presents a notable limitation of computational models relying solely on social media text. This limitation leads to potential false negatives, where affected individuals who do not manifest depressive behavior online may be missed by such systems.

Similarly, another recent study (Aldkheel and Zhou, 2024) highlighted that social media detection methods that focus on observable linguistic, visual, and behavioral signals indicative of depression cannot account for users who do not publicly share or mask their symptoms due to privacy concerns, social stigma, or personal choice. Moreover, reliance on textual content alone restricts detection to expressed emotions and behaviors, which may not comprehensively represent every user's mental health status.

The necessity for combining social media analysis with clinical validation and offline data is emphasized to address these limitations and improve detection reliability. Verification of ground truth through clinical evaluations or integration of medical records along with social media data can help overcome false negatives caused by the absence of explicit online depressive expression (Aldkheel and Zhou, 2024). Thus, while social media-based depression detection offers valuable early screening potential, it cannot substitute for comprehensive clinical diagnosis and does not capture all cases, especially among users who do not disclose symptoms online.

GloVe word embeddings worked better with models like Random Forest and ANN because they capture the meaning and relationships between words. Unlike methods like TF-IDF or Bag of Words that just count word frequency, GloVe places similar words (like "sad" and "unhappy") close together in a way that shows their meaning. This helps the models better tell the difference between depressive and non-depressive posts, especially since social media posts are usually short. Traditional models like SVM and Random Forest did well with TF-IDF because they can handle large sets of sparse features. However, the ANN model didn't perform as well with TF-IDF or BoW since those features don't carry the deeper meaning that neural networks are designed to learn from. When we used LIME to explain the model's decisions, we found that GloVe helped the models focus on important words like "lonely," "worthless," and "help"— strong signs of depression. In contrast, TF-IDF often picked up common but less meaningful words, which made the learning less effective.

We used social media data from platform X (formerly Twitter) because it is public, real-time, and short in format making it ideal for spotting signs of mental health issues. Compared to sites like Reddit or medical records, Twitter has more variety in language, which helps our model work better in real-world situations (De Choudhury et al., 2013).

To get useful features from the text, we used both basic methods (like TF-IDF, LDA, N-gram, and Bag of Words) and word embeddings (like GloVe). While advanced models like BERT understand context better, they are slower and harder to explain. GloVe gave us good results without needing too much computing power, which is important if we want to use this in real-time systems (Pennington et al., 2014).

We picked simple models like SVM and Random Forest because they work well with smaller datasets, are faster to train, and are easier to understand—especially when used with tools like LIME. These models also don't need powerful hardware and can be used in apps or cloud platforms. We used LIME to explain how our models make decisions. It works with many kinds of models and helps make the results clearer for doctors and other users. This is important for mental health tools, where we need to be careful and ethical in how results are used (Ribeiro et al., 2016b).

Overall, our findings demonstrate the effectiveness of combining traditional and advanced NLP techniques with robust classifiers to achieve competitive performance in text classification tasks. Our results indicate that, while advanced deep learning models are powerful, conventional methods and hybrid approaches can also achieve competitive accuracy, offering more interpretable and computationally efficient alternatives.

To address the identified research gaps in the literature, our study proposes an interpretable and comparative framework for depression detection using Twitter data, comprising four key components. First, multiple NLP feature representations are employed, including TF-IDF, BoW, LDA, N-grams, and GloVe embeddings, to effectively capture both statistical and semantic characteristics of textual content. Second, a range of ML classifiers, including SVM, RF, ANN, and XGBoost, were evaluated under identical experimental conditions to assess their predictive performance. Third, explainability was integrated into the framework using the LIME method, which highlights the contribution of individual features to classification outcomes, thereby enhancing transparency and trust in the model. Finally, a comparative evaluation is performed in which all models and feature extraction techniques are applied to the same dataset and evaluated using consistent metrics, accuracy, precision, recall, and F1 score, to ensure fairness, reproducibility, and applicability in the real world.

8 Conclusion

Mental illness is a prevalent social issue driven by socioeconomic, clinical, and individual risk factors, and the rise of social media has allowed the analysis of user-generated content for early detection of depression. In this work, we evaluated the effectiveness of various feature extraction methods and machine learning classifiers in detecting depression from X posts. Our findings demonstrate that social media data

can be effectively utilized for mental health monitoring, with methods such as N-gram, BOW, and TF-IDF providing significant insights. In particular, the combination of TF-IDF with XGB achieved the highest precision of 87%, while the GloVe embeddings with RF reached an accuracy of 88% with lower interpretability. The use of LIME highlighted the importance of balancing accuracy and interpretability in model outcomes. Future research should focus on incorporating advanced deep learning models such as Transformers and BERT, developing real-time detection systems, integrating multimodal data, expanding analyses to additional social media platforms, and addressing ethical and privacy concerns in mental health monitoring.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

SH: Conceptualization, Formal analysis, Methodology, Validation, Writing – original draft, Writing – review & editing. MN: Conceptualization, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. NA: Conceptualization, Supervision, Writing – original draft, Writing – review & editing. MF: Funding acquisition, Writing – original draft, Writing – review & editing. RN: Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted without any commercial or financial relationships that could potentially create a conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Abd Yusof, N. F., Lin, C., and Guerin, F. (2017). "Analysing the causes of depressed mood from depression vulnerable individuals," in *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)* (Cham: Springer), 9–17.

AbuRaed, A. G. T., Prikryl, E. A., Carenini, G., and Janjua, N. Z. (2024). Long COVID discourse in Canada, the United States, and Europe: topic modeling and sentiment analysis of twitter data. *J. Med. Internet Res.* 26:e59425. doi: 10.2196/59425

Adarsh, V., Kumar, P. A., Lavanya, V., and Gangadharan, G. (2023). Fair and explainable depression detection in social media. *Inf. Process. Manag.* 60:103168. doi: 10.1016/j.ipm.2022.103168

Akhtar, H. M. U., Nauman, M., Akhtar, N., Hameed, M., Hameed, S., Tareen, M. Z., et al. (2025). Mitigating cyber threats: machine learning and explainable AI for phishing detection. *VFAST Trans. Softw. Eng.* 13, 170–195. doi: 10.21015/vtse.v13i2.2129

Al Asad, N., Pranto, M. A. M., and Islam, M. M. (2024). "Explainable deep learning for mental health detection from English and Arabic social media posts," in *ACM Transactions on Asian and Low-Resource Language Information Processing Volume 23* (New York, NY: ACM).

Al Qudah, I., Hashem, I., Soufyane, A., Chen, W., and Merabtene, T. (2022). "Applying latent Dirichlet allocation technique to classify topics on sustainability using Arabic text," in *Intelligent Computing: Proceedings of the 2022 Computing Conference, Volume 1* (Cham: Springer), 630–638. doi: 10.1007/978-3-031-10461-9_43

Aldkheel, A., and Zhou, L. (2024). Depression detection on social media: a classification framework and research challenges and opportunities. *J. Healthc. Inform. Res.* 8, 88–120. doi: 10.1007/s41666-023-00152-3

Amanat, A., Rizwan, M., Javed, A. R., Abdelhaq, M., Alsaqour, R., Pandya, S., et al. (2022a). Deep learning for depression detection from textual data. *Electronics* 11:676. doi: 10.3390/electronics11050676

Amanat, A., Shah, S. A. A., Javaid, N., and Iqbal, F. (2022b). Detection of depression using convolutional neural networks and word2vec on reddit posts. *IEEE Access* 10, 14638–14648. doi: 10.1080/23311975.2022.2039087

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2018). Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv [Preprint]* arXiv:1711.06104. doi: 10.48550/arXiv.1711.06104

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* 5, 135–146. doi: 10.1162/tacl_a_00051

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Brochier, R., Guille, A., and Velcin, J. (2019). "Global vectors for node representations," in *The World Wide Web Conference* (New York, NY: ACM), 2587–2593. doi: 10.1145/3308558.3313595

Cambria, E., and White, B. (2014). Jumping NLP curves: a review of natural language processing research. *IEEE Comput. Intell. Mag.* 9, 48–57. doi: 10.1109/MCI.2014.2307227

Cesarini, M., Malandri, L., Pallucchini, F., Seveso, A., and Xing, F. (2024). Explainable AI for text classification: lessons from a comprehensive evaluation of *post hoc* methods. *Cogn. Comput.* 16, 3077–3095. doi: 10.1007/s12559-024-10325-w

Cha, J., Kim, S., Kim, D., and Park, E. (2024). Mogam: a multimodal object-oriented graph attention model for depression detection. *arXiv [Preprint]*. arXiv:2403.15485. doi: 10.48550/arXiv.2403.15485

Chakraborty, D., Ivan, C., Amero, P., Khan, M., Rodriguez-Aguayo, C., Başağaoğlu, H., et al. (2021). Explainable artificial intelligence reveals novel insight into tumor microenvironment conditions linked with better prognosis in patients with breast cancer. *Cancers* 13:3450. doi: 10.3390/cancers13143450

Chancellor, S., and De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. *NPJ Digit. Med.* 3, 1–11. doi: 10.1038/s41746-020-0233-7

Chanda, K., Roy, S., Mondal, H., and Bose, R. (2022). To judge depression and mental illness on social media using twitter. *Univers. J. Public Health* 10, 116–129. doi: 10.13189/ujph.2022.100113

Chelgani, S. C., Nasiri, H., Tohry, A., and Heidari, H. (2023). Modeling industrial hydrocyclone operational variables by shap-catboost-a "conscious lab" approach. *Powder Technol.* 420:118416. doi: 10.1016/j.powtec.2023.118416

Chen, D., Zhao, H., He, J., Pan, Q., and Zhao, W. (2021). "An causal XAI diagnostic model for breast cancer based on mammography reports," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Houston, TX: IEEE), 3341–3349. doi: 10.1109/BIBM52615.2021.9669648

Chen, T., and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY: ACM), 785–794. doi: 10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al. (2015). "Xgboost: a scalable tree boosting system," in *Proceedings of the 25th International Conference on Big Data* (Cham: Springer).

Chen, X., and Lin, X. (2025). Generating medically-informed explanations for depression detection using LLMS. *arXiv [Preprint]*. arXiv:2503.14671. doi: 10.48500/arXiv.2503.14671

Cho, G., Yim, J., Choi, Y., Ko, J., and Lee, S.-H. (2019). Review of machine learning algorithms for diagnosing mental illness. *Psychiatry Investig.* 16:262. doi: 10.30773/pi.2018.12.21.2

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1023/A:1022627411411

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). "Predicting depression via social media," in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*.

de Souza, V. B., Nobre, J. C., and Becker, K. (2022). DAC stacking: a deep learning ensemble to classify anxiety, depression, and their comorbidity from reddit texts. *IEEE J. Biomed. Health Inform.* 26, 3303–3311. doi: 10.1109/JBHI.2022.3151589

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "Bert: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, MN), 4171–4186.

Dey, R. K., and Das, A. K. (2023). Modified term frequency-inverse document frequency based deep hybrid framework for sentiment analysis. *Multimed. Tools Appl.* 12, 1–24. doi: 10.1007/s11042-023-14653-1

Fatima, B., Amina, M., Nachida, R., and Hamza, H. (2020). A mixed deep learning based model to early detection of depression. *J. Web Eng.* 19, 429–455. doi: 10.13052/jwe1540-9589.19344

Friedman, S., Allin, L., and Gray, W. (2024). "The world is your oyster": mothers' perspectives on the value and purpose of an independent Forest School provision. *Child. Geogr.* 22, 597–611. doi: 10.1080/14733285.2024.2321387

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

Gohel, P., Singh, P., and Mohanty, M. (2021). Explainable AI: current status and future directions. *arXiv [Preprint]*. arXiv:2107.07045. doi: 10.48550/arXiv.2107.07045

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Govindasamy, K. A., and Palanichamy, N. (2021). "Depression detection using machine learning techniques on twitter data," in *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)* (Madurai: IEEE), 960–966. doi: 10.1109/ICICCS51141.2021.9432203

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., and Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: an integrative review. *Curr. Opin. Behav. Sci.* 18, 43–49. doi: 10.1016/j.cobeha.2017.07.005

Guo, Y., Li, M., and Zhang, H. (2023a). Lime-based explainability for mental health predictions on social media. *J. Med. Internet Res.* 25:e43915. doi: 10.1016/j.scitotenv.2023.166118

Guo, Y., Zhang, Z., and Xu, X. (2023b). Research on the detection model of mental illness of online forum users based on convolutional network. *BMC Psychol.* 11:424. doi: 10.1186/s40359-023-01460-4

Hakkoum, H., Idri, A., and Abnane, I. (2020). "Artificial neural networks interpretation using lime for breast cancer diagnosis," in *Trends and Innovations in Information Systems and Technologies: Volume 38* (Cham: Springer), 15–24. doi: 10.1007/978-3-030-45697-9_2

Haque, R., Islam, N., Islam, M., and Ahsan, M. M. (2022). A comparative analysis on suicidal ideation detection using nlp, machine, and deep learning. *Technologies* 10:57. doi: 10.3390/technologies10030057

Harris, Z. S. (1954). Distributional structure. *Word* 10, 146–162. doi: 10.1080/00437956.1954.11659520

Helmy, A., Nassar, R., and Ramdan, N. (2024). Depression detection for twitter users using sentiment analysis in English and Arabic tweets. *Artif. Intell. Med.* 147:102716. doi: 10.1016/j.artmed.2023.102716

Hemmatirad, K., Bagherzadeh, H., Fazl-Ersi, E., and Vahedian, A. (2020). "Detection of mental illness risk on social media through multi-level SVMS," in *2020 8th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)* (Mashhad: IEEE), 116–120. doi: 10.1109/CFIS49607.2020.9238692

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 832–844. doi: 10.1109/34.709601

Hoque, A. M., Rahman, A., Hossain, M. E., Hossain, M. S., and Muhammad, G. (2025). Effective depression detection and interpretation: integrating machine learning, deep learning, language models, and explainable AI. *ARRAY* 25:100375. doi: 10.1016/j.array.2025.100375

Ibrahimov, K., and Ali, S. (2024). Importance of explainable artificial intelligence in mental health applications. *Front. Artif. Intell.* 7, 1–12.

Ibrahimov, Y., Anwar, T., and Yuan, T. (2024). Explainable AI for mental disorder detection via social media: a survey and outlook. *arXiv [Preprint]*. arXiv:2406.05984. doi: 10.4820/arXiv.2406.05984

Ibrahimov, Y., Anwar, T., and Yuan, T. (2025). Depressionx: knowledge infused residual attention for explainable depression severity assessment. *arXiv [Preprint]*. arXiv:2501.14985. doi: 10.48550/arXiv.2501.14985

Islam, M. R., Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H., Ulhaq, A., et al. (2018). Depression detection from social network data using machine learning techniques. *Health Inform. Sci. Syst.* 6, 1–12. doi: 10.1007/s13755-018-0046-0

Ive, J., Viani, N., Kam, J., Yin, L., Verma, S., Puntis, S., et al. (2020). Generation and evaluation of artificial mental health records for natural language processing. *NPJ Digit. Med.* 3, 1–9. doi: 10.1038/s41746-020-0267-x

Jahromi, M. N., Muddamsetty, S. M., Jarlner, A. S. S., Høgenhaug, A. M., Gammeltoft-Hansen, T., and Moeslund, T. B. (2024). Sidu-txt: an xai algorithm for nlp with a holistic assessment approach. *Nat. Lang. Process. J.* 7:100078. doi: 10.1016/j.nlp.2024.100078

Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., et al. (2020). "Overview and importance of data quality for machine learning tasks," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (New York, NY: ACM), 3561–3562. doi: 10.1145/3394486.3406477

Jain, P., Srinivas, K. R., and Vichare, A. (2022). Depression and suicide analysis using machine learning and nlp. *J. Phys. Conf. Seri.* 2161:012034. doi: 10.1088/1742-6596/2161/1/012034

Ji, S., Li, X., Huang, Z., and Cambria, E. (2022a). Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Comput. Appl.* 34, 10309–10319. doi: 10.1007/s00521-021-06208-y

Ji, S., Yu, C., and Fung, S. F. (2022b). Detecting depression on social media using Bert-based models: a case study on Twitter. *J. Affect. Disord. Rep.* 8:100313.

Joshi, B., Shah, N., Barbieri, F., and Neves, L. (2020). The devil is in the details: evaluating limitations of transformer-based methods for granular tasks. *arXiv [Preprint]*. arXiv:2011.01196. doi: 10.48550/arXiv.2011.01196

Joyce, D. W., Kormilitzin, A., Smith, K. A., and Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digit. Med.* 6:6. doi: 10.1038/s41746-023-00751-9

Jurafsky, D., and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall.

Kabir, M., Ahmed, T., Hasan, M. B., Laskar, M. T. R., Joarder, T. K., Mahmud, H., et al. (2023). Deptweet: a typology for social media texts to detect depression severities. *Comput. Human Behav.* 139:107503. doi: 10.1016/j.chb.2022.107503

Kang, Y., Cai, Z., Tan, C.-W., Huang, Q., and Liu, H. (2020). Natural language processing (nlp) in management research: a literature review. *J. Manag. Anal.* 7, 139–172. doi: 10.1080/23270012.2020.1756939

Kawakura, S., Hirafuji, M., Ninomiya, S., and Shibasaki, R. (2022). Analyses of diverse agricultural worker data with explainable artificial intelligence: XAI based on shap, lime, and lightgbm. *Eur. J. Agric. Food Sci.* 4, 11–19. doi: 10.24018/ejfood.2022.4.6.348

Khan, N., Nauman, M., Almadhor, A. S., Akhtar, N., Alghuried, A., Alhudhaif, A., et al. (2024). Guaranteeing correctness in black-box machine learning: a fusion of explainable AI and formal methods for healthcare decision-making. *IEEE Access* 12, 90299–90316. doi: 10.1109/ACCESS.2024.3420415

Khoo, L. S., Lim, M. K., Chong, C. Y., and McNaney, R. (2024). Machine learning for multimodal mental health detection: a systematic review of passive sensing approaches. *Sensors* 24:348. doi: 10.3390/s24020348

Kokhlikyan, N., Miglani, V., Martin, E., Wang, W., Alsallakh, B., Reynolds, J., et al. (2020). Captum: a unified and generic model interpretability library for pytorch. *arXiv [Preprint]*. arXiv:2009:07896. doi: 10.48550/arXiv.2009:07896

LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Levy, O., and Goldberg, Y. (2015). "Improving distributional similarity with lessons learned from word embeddings," in *Transactions of the Association for Computational Linguistics* (Cambridge, MA: MIT Press), 211–225. doi: 10.1162/tacl_a_00134

Li, X., Jurafsky, D., and Liang, P. (2018). "Analogical reasoning with word vectors: a comparative study of algorithms," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Santa Fe, NM), 2150–2163.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomforest. *R News* 2, 18–22.

Manning, C. D., and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Mohammed, M. B., Abir, A. S. M., Salsabil, L., Shahriar, M., and Fahmin, A. (2021). "Depression analysis from social media data in Bangla language: an ensemble approach," in *2021 Emerging Technology in Computing, Communication and Electronics (ETCCE)* (Dhaka: IEEE), 1–6. doi: 10.1109/ETCCE54784.2021.9689887

Nauman, M., Akhtar, N., Alhazmi, O. H., Hameed, M., Ullah, H., Khan, N., et al. (2021). Improving the correctness of medical diagnostics based on machine learning with coloured petri nets. *IEEE Access* 9, 143434–143447. doi: 10.1109/ACCESS.2021.3121092

Olusegun, R., Oladunni, T., Audu, H., Houkpati, Y., and Bengesi, S. (2023). Text mining and emotion classification on monkeypox twitter dataset: a deep learning-natural language processing (NLP) approach. *IEEE Access* 11, 49882–49894. doi: 10.1109/ACCESS.2023.3277868

Orabi, A. H., Buddhitha, P., Orabi, M. H., and Inkpen, D. (2018). "Deep learning for depression detection of twitter users," in *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 88–97.

Ossai, C. I., and Wickramasinghe, N. (2023). Sentiments prediction and thematic analysis for diabetes mobile apps using embedded deep neural networks and latent Dirichlet allocation. *Artif. Intell. Med.*, 138, 102509. doi: 10.1016/j.artmed.2023.102509

Pennington, J., Socher, R., and Manning, C. D. (2014). "Glove: global vectors for word representation" in *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha). doi: 10.3115/v1/D14-1162

Qasim, A., Mehak, G., Hussain, N., Gelbukh, A., and Sidorov, G. (2025). Detection of depression severity in social media text using transformer-based models. *Information* 16:114. doi: 10.3390/info16020114

Ramírez-Cifuentes, D., Largeron, C., Tissier, J., Baeza-Yates, R., and Freire, A. (2021). Enhanced word embedding variations for the detection of substance abuse and mental health issues on social media writings. *IEEE Access* 9, 130449–130471. doi: 10.1109/ACCESS.2021.3112102

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016a). "Model-agnostic interpretability of machine learning," in *ICML Workshop on Human Interpretability in Machine Learning (WHI), Volume 31* (New York, NY), 91.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016b). "Why should i trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY: ACM), 1135–1144. doi: 10.1145/2939672.2939778

Rizwan, M., Mushtaq, M. F., Akram, U., Mehmood, A., Ashraf, I., Sahelices, B., et al. (2022). Depression classification from tweets using small deep transfer learning language models. *IEEE Access* 10, 129176–129189. doi: 10.1109/ACCESS.2022.3223049

Sabaneh, K., Salameh, M. A., Khaleel, F., Herzallah, M. M., Natsheh, J. Y., Maree, M., et al. (2023). "Early risk prediction of depression based on social media posts in Arabic," in *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)* (Atlanta, GA: IEEE), 595–602. doi: 10.1109/ICTAI59109.2023.00094

Saha, T., Reddy, S. M., Saha, S., and Bhattacharyya, P. (2022). Mental health disorder identification from motivational conversations. *IEEE Trans. Comput. Soc. Syst.* D10, 1130–1139. doi: 10.1109/TCSS.2022.3143763

Salton, G., and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inform. Process. Manag.* 24, 513–523. doi: 10.1016/0306-4573(88)90021-0

Santhosh Baboo, S., and Amirthapriya, M. (2022). Comparison of machine learning techniques on twitter emotions classification. *SN Comput. Sci.* 3, 1–8. doi: 10.1007/s42979-021-00889-x

Santos, W., Yoon, S., and Paraboni, I. (2023). Mental health prediction from social media text using mixture of experts. *IEEE Latin Am. Trans.* 21, 723–729. doi: 10.1109/TLA.2023.10172137

Saxena, C., Garg, M., and Ansari, G. (2022). "Explainable causal analysis of mental health on social media data," *International Conference on Neural Information Processing* (Cham: Springer), 172–183. doi: 10.1007/978-3-031-30108-7_15

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Schölkopf, B., and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT press. doi: 10.7551/mitpress/4175.001.0001

Shah, S. M., Gillani, S. A., Baig, M. S. A., Saleem, M. A., and Siddiqui, M. H. (2025). Advancing depression detection on social media platforms through fine-tuned large language models. *Online Soc. Netw. Media* 46:100311. doi: 10.1016/j.osnem.2025.100311

Søgaard, A. (2021). *Explainable Natural Language Processing*. San Rafael, CA: Morgan & Claypool Publishers.

Soldaini, L., and Goharian, N. (2016). "Quickumls: a fast, unsupervised approach for medical concept extraction," in *Proceedings of the AMIA Annual Symposium* (Washington, DC: American Medical Informatics Association), 1216–1225.

Steinkamp, J., and Cook, T. S. (2021a). Basic artificial intelligence techniques: natural language processing of radiology reports. *Radiol. Clin.* 59, 919–931. doi: 10.1016/j.rcl.2021.06.003

Steinkamp, J. M., and Cook, T. B. (2021b). Applications of natural language processing for mental health research on social media. *Int. J. Med. Inform.* 148:104399.

Sundararajan, M., Taly, A., and Yan, Q. (2017). "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)* (Sydney, NSW: PMLR), 3319–3328.

Tadesse, M. M., Lin, H., Xu, B., and Yang, L. (2019). Detection of depression-related posts in reddit social media forum. *IEEE Access* 7, 44883–44893. doi: 10.1109/ACCESS.2019.2909180

Thornicroft, G., Sunkel, C., Aliev, A. A., Baker, S., Brohan, E., El Chammay, R., et al. (2022). The lancet commission on ending stigma and discrimination in mental health. *Lancet* 400, 1438–1480. doi: 10.1016/S0140-6736(22)01470-2

Tlachac, M., and Rundensteiner, E. (2020). Screening for depression with retrospectively harvested private versus public text. *IEEE J. Biomed. Health Inform.* 24, 3326–3332. doi: 10.1109/JBHI.2020.2983035

Tong, L., Liu, Z., Jiang, Z., Zhou, F., Chen, L., Lyu, J., et al. (2022). Cost-sensitive boosting pruning trees for depression detection on Twitter. *IEEE Trans. Affect. Comput.* 14, 1898–1911. doi: 10.1109/TAFFC.2022.3145634

Vapnik, V. N. (1998). *Statistical Learning Theory*. Hoboken, NJ: Wiley.

Velupillai, S., Suominen, H., Liakata, M., Roberts, A., Shah, A. D., Morley, K., et al. (2018). Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances. *J. Biomed. Inform.* 88, 11–19. doi: 10.1016/j.jbi.2018.10.005

Wani, M. A., ELAffendi, M. A., Shakil, K. A., Imran, A. S., and Abd El-Latif, A. A. (2022). Depression screening in humans with AI and deep learning techniques. *IEEE Trans. Comput. Soc. Syst.* 10, 2074–2089. doi: 10.1109/TCSS.2022.3200213

Weerts, H. J., van Ipenburg, W., and Pechenizkiy, M. (2019). A human-grounded evaluation of shap for alert processing. *arXiv [Preprint]*. arXiv:1907.03324. doi: 10.48550/arXiv.1907.03324

World Health Organization (2021). Guidance on Community Mental Health Services: Promoting Person-Centred and Rights-Based Approaches. Geneva: World Health Organization.

World Health Organization, Regional Office for the Eastern Mediterranean (2019). *Mental Health Atlas 2017: Resources for Mental Health in the Eastern Mediterranean Region*.

Xiang, L. (2022). Application of an improved tf-idf method in literary text classification. *Adv. Multimed.* 2022:9285324. doi: 10.1155/2022/9285324

Xu, Z., Chen, M., Weinberger, K. Q., and Sha, F. (2013). "An alternative text representation to TF-IDF and bag-of-words," in *Proceedings of ICML* (*arXiv* [Preprint]. arXiv:1301.6770). doi: 10.48550/arXiv.1301.6770

Yazdavar, A. H., Mahdavinejad, M. S., Bajaj, G., Romine, W., Sheth, A., Monadjemi, A. H., et al. (2020). Multimodal mental health analysis in social media. *PLoS ONE* 15:e0226248. doi: 10.1371/journal.pone.0226248

Zhang, T., Schoene, A. M., Ji, S., and Ananiadou, S. (2022). Natural language processing applied to mental illness detection: a narrative review. *NPJ Digit. Med.* 5, 1–13. doi: 10.1038/s41746-022-00589-7

Zhang, W., Zhang, S., and Zhou, J. (2017). Distributed xgboost with column block splitting. *arXiv [Preprint]*. arXiv:1708.05721. doi: 10.48550/arXiv.1708.05721

Zhang, Y., Weng, Y., and Lund, J. (2022). Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics* 12:237. doi: 10.3390/diagnostics12020237