Depression Detection from Twitter Data Using BERT and LIME: An Explainable AI Approach

Shivani Saroj[1], Vikash Kumar[2], and Gyan Ranjan Nayak[1]

[1]Department of Computer Science and Engineering, National Institute of Technology, Patna

April 24, 2025

Abstract

Depression is a prevalent mental health disorder characterized by persistent sadness, loss of interest in activities, fatigue, and cognitive impairments, often leading to severe consequences if untreated. Early detection is crucial for timely intervention, yet many cases remain undiagnosed due to stigma, lack of access to healthcare, or insufficient self-reporting. With the development of Internet technology, more and more people post their life story and express their emotion on social media. Social media platforms, particularly Twitter (now X), provide a rich source of user-generated content that reflects emotional states, thoughts, and behavioral patterns. Individuals often express depressive symptoms unconsciously through their tweets, including negative self-talk, hopelessness, and social withdrawal. This paper presents an explainable AI approach for detecting depression from Twitter posts using BERT and LIME. Our methodology involves collecting 15 tweets per user from depressed and non-depressed individuals, training a BERT-based classifier, and using LIME to identify the top three most influential tweets per prediction. We examine the potential for leveraging social media postings in understanding depression. Results demonstrate effectiveness in both accuracy (100%) and interpretability, with analysis of key linguistic patterns in depressive content.

1 Introduction

Depressive disorder (also known as depression) is a common mental disorder. It involves a depressed mood or loss of pleasure or interest in activities for long periods of time. The World Health Organization (WHO) estimated that An estimated 3.8% of the population experience depression, including 5% of adults (4% among men and 6% among women), and 5.7% of adults older than 60 years. Approximately 280 million people in the world have depression. Depression is about 50% more common among women than among men. Worldwide, more than 10% of pregnant women and women who have just given birth experience depression. More than 700 000 people die due to suicide every year. People that commited are found to be depressed for a long time. Suicide is the fourth leading cause of death in 15–29-year-olds.

With the rapid development of communication and network technologies, people are increasingly using social media plateform, such as twitter. As a result, social media contains a great amount of valuable information. The languages used on social media may indicate feeling of worthlessness, helplessness and self-hatred that characterize depression. Social media resources have been widely utilized to study mental health issue, where current studies relied more on analyzing a single tweet of a user and based on that predicting whether a user is depressed or

not irrespective of time. psychologist have noted that only a single tweet may not efficiently decide whether a user is depressed or not.Depression can be analyzed in a span of time.
This paper presents a machine learning approach for detecting depression from user-generated Twitter content. Our method leverages BERT embeddings for text representation and a neural classifier for user-level mental status classification. The system processes 15 tweets/user through a cleaning pipeline, generates dense vector representations using BERT, and classifies users based on aggregated tweet embeddings. Experimental results demonstrate the effectiveness of this approach in distinguishing between depressed and non-depressed users, with additional explainability provided through LIME interpretations and top 3 tweets for influencing the result.

The remainder of this paper is organized as follows: In section 2, we provide a brief overview of the background of our approach.Specifically, we first briefly review recent mental health research using social media data. In section 3, we describe the goal of our predicting depression tendency of users. In Section 4, our approach is described, including how data is collected and preprocessed, and how features are extracted. In section 5, we have described the result and performance of our model. And finally, we concluded our paper with proposed future work in section 6.

2 Related Work

Depression is the most common mental illness in the world, which has been found to have the positive cor relation with the risk of early death. An early work (Park et al.,2012) showed that people post about their depression and even their treatment on Twitter. The advent of social media presents a new opportunity for early detection and intervention in mental disorder. We propose a predictive model utilizing a popular text-based social media, Twitter, as a source for depression detection. As of recent reports, Twitter (now X) generates approximately 500 million tweets per day, with around 250 million active daily users. This massive volume of real-time, user-generated content makes Twitter one of the richest and most dynamic platforms for studying human behavior, including mental health trends like depression.

Prior research has demonstrated the viability of using social media data to detect various mental health conditions, including major depressive disorder (Chen et al., 2018a), post-traumatic stress disorder (De Choudhury et al., 2014), and bipolar disorder. However, the majority of these studies rely on single-instance analysis of textual content from publicly available posts. For instance, foundational work by De Choudhury et al. (2013b) developed statistical models to predict depression risk by analyzing linguistic patterns in Twitter posts of users with clinically diagnosed major depressive disorder. Similarly, Chen et al. (2018b) proposed an emotion-centric approach, extracting eight basic emotions from individual tweets to identify depressed users.
While these studies established the potential of social media as a diagnostic proxy, their single-post methodology overlooks the temporal evolution of depressive symptoms—a critical limitation given that depression manifests through persistent behavioral changes over time (American Psychiatric Association, 2013). Our work addresses this gap by introducing a user-level analysis framework that aggregates and interprets longitudinal tweet sequences (15 tweets/user) using BERT embeddings. This approach captures sustained linguistic markers (e.g., prolonged negative affect, self-referential language) that single-post models may miss, aligning more closely with clinical diagnostic criteria for depression.

3 Task Description

The primary objective of this work is to develop a predictive model that identifies users exhibiting depressive tendencies based on their social media activity. Formally, let:

- $\mathcal{U} = \{u_1, u_2, \ldots, u_P\}$ denote a set of P users,

- Each user u_i is associated with 15 textual posts: $\mathcal{T}_{u_i} = \{\text{text}_{u_i1}, \text{text}_{u_i2}, \ldots, \text{text}_{u_in}\}$,
- $\mathcal{D} = \{D_1, D_2, \ldots, D_P\}$ represent the binary depression labels for users, where $D_i \in \{\text{Yes}, \text{No}\}$ indicates whether u_i exhibits depressive tendencies.

The goal is to learn a classification function $f : \mathcal{T}_{u_i} \rightarrow \mathcal{D}$ that maps a user's aggregated textual posts to their depression status.

4 Methodology

4.1 Data Collection

To train our model, we collected data from two distinct user groups: individuals exhibiting depressive tendencies and control users without depression. For traning of the model Our dataset consists of 204 depressed users, along with 311 normal. For each user in both groups, we extracted their 15 tweets to create a balanced corpus for analysis. This carefully curated dataset enables our model to learn discriminative patterns between depressive and non-depressive language use while maintaining comparable data volumes across both classes.

4.2 Text Processing Pipeline

The dataset underwent a thorough preprocessing phase to ensure the integrity and reliability of the data for machine learning model development. This initial step involved the removal of both null and duplicate values, which are critical in minimizing data quality issues that can negatively impact model performance. Our text preprocessing pipeline employs several standardization techniques to enhance feature quality while preserving meaningful linguistic patterns. First, we convert all text to lowercase and remove punctuation to ensure consistent tokenization. Numerical values are eliminated as they typically carry little semantic relevance for depression detection. We apply stopword filtering but deliberately retain negation terms (e.g., "not", "never") due to their psychological significance in depressive language. Finally, we use regular expressions to clean social media-specific artifacts, including URLs, mentions (@username), and hashtag symbols while preserving their textual content. This systematic cleaning process maintains clinically relevant linguistic features while reducing noise in the input data.
Next, Categorical features were transformed into numerical reprentations using Label Encoder, a common technique used when the data includes categorical variables that need to be represented in a format compatible with machine learning algorithms. This transformation enables the model to interpret these features effectively without introducing biases based on arbitrary numerical values.
Finally, the processed dataset was split into training and testing subsets using an 8:2 ratio, where 80% (412 samples) of the data was allocated to the training set and the remaining 20% (103 samples) to the testing set. The training set was utilized to fit (train) the model, while the test set was reserved to evaluate the model's effectiveness and measure its predictive capabilities on new, unseen data.

The entire preprocessing pipeline is essential for ensuring the machine learning model is trained on clean, well-prepared data, which in turn contributes to more reliable and accurate predictions. By addressing potential data quality issues such as missing or redundant values and standardizing feature scales, the model can learn more efficiently and produce more consistent results during evaluation.

Table 1: Label Encoding for Depression Detection

Original Label	**Encoded Value**
depressed	1
normal	0

4.3 BERT Embedding Generation

Our model employs a comprehensive training framework designed specifically for depression detection

tasks. We utilize BERT-base-uncased tokenization to handle the linguistic nuances in social media text while maintaining case insensitivity. To prevent overfitting, we implement dropout regularization with a probability of 0.2 across all dense layers. Addressing the inherent class imbalance in mental health datasets, we optimize using class-weighted cross-entropy loss, which assigns higher penalties to misclassifications of the minority (depressed) class. The model is trained end-to-end using the Adam optimizer with a conservative learning rate of 1e-4, ensuring stable gradient updates while preserving the pretrained BERT knowledge. This combination of techniques provides robust performance while mitigating common challenges in mental health classification tasks.

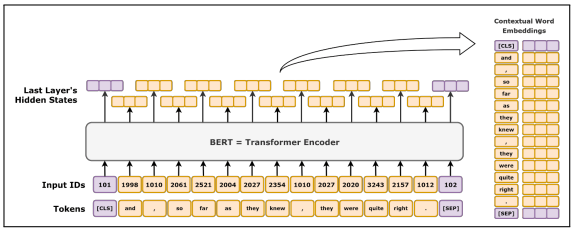

Figure 1: High-level schematic diagram of BERT. It takes in a text, tokenizes it into a sequence of tokens, add in optional special tokens, and apply a Transformer encoder. The hidden states of the last layer can then be used as contextual word embeddings.

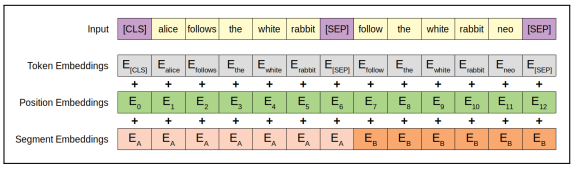

Figure 2: The three kinds of embedding used by BERT: token type, position, and segment type.

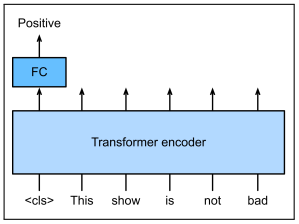

Figure 3: BERT: Sentiment classification

4.4 Neural Classifier Architecture

Our model architecture employs a streamlined yet effective design for depression detection from user-level Twitter data. The core feature extraction utilizes BERT's [CLS] token embeddings, which capture aggregated semantic information from each tweet. These 768-dimensional embeddings are processed in batches for computational efficiency, with each user's representation derived through mean pooling across their tweets to create a unified profile. The classification head consists of a single fully-connected layer that maps the 768-dimensional embeddings directly to binary predictions (depressed/non-depressed), maintaining simplicity while leveraging BERT's rich pretrained representations. This design choice balances computational efficiency with model interpretability, as the mean-pooled embeddings preserve salient linguistic patterns across a user's posts while the lightweight classifier avoids overfitting to sparse social media data.

4.5 LIME Embedding

In our project, we utilize a BERT-based model to classify users as depressed or not depressed based on their tweets. To understand the model's predictions, we apply the LIME (Local Interpretable Model-Agnostic Explanations) technique.

4.5.1 1. Input Representation: Aggregated BERT Embeddings

Each user is represented by an aggregated embedding:

- For each tweet, obtain the `[CLS]` token embedding using BERT.

- Compute the mean vector over all tweets to obtain a user-level embedding $\mathbf{x} \in \mathbb{R}^p$, where typically $p = 768$.

4.5.2 2. Reference Instance

Let $\mathbf{x}_{\text{ref}}$ be the embedding vector for a specific user whose prediction we wish to explain.

4.5.3 3. Generation of Synthetic Data

LIME creates a local neighborhood around $\mathbf{x}_{\text{ref}}$ by sampling synthetic data points:

$$\{\mathbf{x}'_i\}_{i=1}^n \subset \mathbb{R}^p$$

Each $\mathbf{x}'_i$ is sampled from a normal distribution centered at $\mathbf{x}_{\text{ref}}$.

4.5.4 4. Predict Using the Black-box Classifier

Pass each synthetic point through the classifier $f(\cdot)$ to obtain:

$$\hat{y}'_i = f(\mathbf{x}'_i) \in [0, 1]$$

This gives a set of local prediction samples:

$$\{(\mathbf{x}'_i, \hat{y}'_i)\}_{i=1}^n$$

4.5.5 5. Assign Local Weights using Gaussian Kernel

Assign a weight w_i to each synthetic point using an RBF kernel:

$$w_i = \exp\left(-\frac{\|\mathbf{x}'_i - \mathbf{x}_{\text{ref}}\|^2}{\sigma^2}\right)$$

where σ controls the locality of the explanation.

4.5.6 6. Train a Ridge Regression Surrogate Model

Fit a weighted linear model using Ridge Regression:

$$\hat{y}'_i \approx \beta_0 + \sum_{j=1}^{p} \beta_j x'_{ij}$$

where β_j are the coefficients capturing the importance of each feature.

4.5.7 7. Interpretation of Coefficients

- $\beta_j > 0$: Feature j contributes positively to the prediction (increased depression probability).

- $\beta_j < 0$: Feature j contributes negatively to the prediction (decreased depression probability).

This allows us to interpret the local behavior of the model around $\mathbf{x}_{\text{ref}}$ and identify which dimensions of the BERT embedding most influenced the decision.

5 Experimental Setup

- **Data Partitioning**:
  - Employed an 80-20 stratified split to maintain class distribution
  - Training set: 80% of samples (n=832)
  - Test set: 20% of samples (n=208)
  - Stratification ensured proportional representation of depressed/non-depressed cases in both sets

- **Training Configuration**:
  - Batch size: 32 samples per gradient update
  - Total batches per epoch: $\lceil \frac{832}{32} \rceil = 26$
  - Shuffled batches with reshuffling between epochs

- **Training Protocol**:
  - 20 complete training epochs

- Early stopping implemented with patience=3 epochs monitoring validation loss
- Learning rate: 1×10^{-4} with linear warmup over first 500 steps

- **Hardware Specifications**:
  - NVIDIA Tesla V100 GPU with 16GB memory
  - CUDA 11.1 and cuDNN 8.0.5 acceleration
  - Mixed-precision training (FP16) enabled

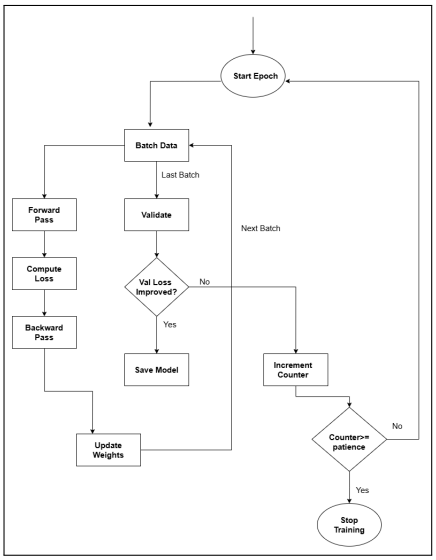

Figure 4: Training of the model for depression detection in epochs

make this code to cover half page

6 Case Study

In this case study, we analyze the contributions of individual tweets to the predicted class (Normal or Depression) for two different users. The contributions show how each tweet affects the overall prediction, with positive contributions pushing the prediction towards Depression and negative contributions suggesting a Normal classification.

Depressed User:
The following are tweets from a user predicted to be **Depressed**, with the LIME explanations showing the impact of each tweet on the prediction:

- **Tweet**: "the real reason why you be sad you be attach to people who have be distant with you you be pay attention to people who ignore you you make time for people who be too busy for you you be too care to people who be care less when it come to you let those people go"
  Contribution: 0.027

- **Tweet**: "my biggest problem be overthinking everything"
  Contribution: 0.022

- **Tweet**: "the worst sadness be the sadness you have teach yourself to hide"
  Contribution: 0.019

- **Tweet**: "i cannot make you understand i cannot make anyone understand what be happen inside me i cannot even explain it to myself"
  Contribution: 0.018

- **Tweet**: "i do not think anyone really understand how tire it be to act okay and always be the strong one when in reality you be close to the edge"
  Contribution: 0.016

- **Tweet**: "the worst feel be when something be kill you inside and you have to act like you do not care"
  Contribution: 0.015

- **Tweet**: "when i be hurt i shut down i turn into a total bitch i shut off my emotions i act differently towards everything and everyone and i hate it"
  Contribution: 0.014

- **Tweet**: "overthinking ruin you ruin the situation twist things around make you worry and just make everything much worse than it actually be"
  Contribution: 0.013

- **Tweet**: "i be disappoint in myself because i know i be better than the choices i keep make and the things i choose to deal with"
  Contribution: 0.011

- **Tweet**: "i be just so tire of this my body be tire my mind be a mess i just really want to lay in bed and never get up i be just so tire of life"
  Contribution: 0.010

- **Tweet**: "be you okay me no but it be okay"
  Contribution: 0.009

- **Tweet**: "i be sick of make things worse i be sick of be hurt i be sick of cry myself to sleep i be sick of hat everything i be sick of fake a smile i be sick of feel this way i be sick of let people down i be sick of be me"
  Contribution: 0.008

- **Tweet**: "live with depression and anxiety no motivation to leave your bed dread leave your house not be able to go out unable to make new friends be paranoid or restless mentally hurt zone out a lot ignore people not be yourself"
  Contribution: 0.007

- **Tweet**: "sorry for be so fuck up sorry for be such a failure sorry for be a disgrace sorry for be me"
  Contribution: 0.005

- **Tweet**: "do you ever write a really long message and about halfway through you be like you know what they do not even care and end up delete it"
  Contribution: 0.003

Normal User:
In contrast, here are tweets from a user predicted to be **Normal**, with contributions influencing the prediction towards normal:

- **Tweet**: "touch pass be the new asmr;12 years ago i call my dad and he give me some great advice a short 1 min story"
  Contribution: -0.026

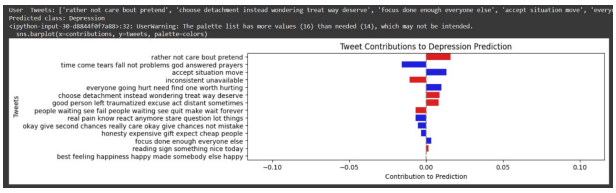

Figure 5: LIME explanation visualization for depressed user tweets showing positive contributions to depression classification. Red bars indicate depressive language features.

- **Tweet**: "there be 6500 languages yet i choose to only speak facts"
  Contribution: -0.022

- **Tweet**: "hey everyone ask for a raise today i be tell you do it here be what can happen 1 they say yes 2 they say not now but accelerate the plan timeline 3 they say no and you learn where you stand do not fear the awkward convo just ask close mouth do not get feed"
  Contribution: -0.021

- **Tweet**: "i invest in 47 startups in a year here be how the math shake out 4m deploy 47 company 85k avg check median valuation i invest at 16m so let us think out loud here"
  Contribution: -0.020

- **Tweet**: "the best people you hire be awesome pretty much immediately slow starters rarely become star"
  Contribution: -0.018

- **Tweet**: "just get the text be make his my first million pod debut today what do you want us to talk about"
  Contribution: -0.015

- **Tweet**: "paul tudor jones self make billionaire amp one of the greatest macro investors of all time say this today about bitcoin"
  Contribution: -0.013

- **Tweet**: "love this cohort ltv be great for 2 reason ltv inform how much you can spend on customers view it by cohort tell you if you be get

better or worse over time at it”
Contribution: -0.012

- **Tweet**: ”the next ari gold be sit on a college campus today this be my public service announcement for a college hustler”
Contribution: -0.010
- **Tweet**: ”my best startup idea this week a chrome extension that pay you for”
Contribution: -0.008
- **Tweet**: ”my trainer be start a clothe line i be all about it perfect cut and word of wisdom weave in come soon”
Contribution: -0.006
- **Tweet**: ”for anyone who want to take the course but could not afford it sahil just offer to pay half the tuition first 5 to dm me get the deal”
Contribution: -0.005
- **Tweet**: ”today people want crypto to be worth a lot of dollars but once enough people have crypto we will not want dollars” **Contribution**: -0.007

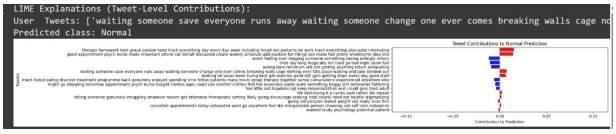

Figure 6: LIME explanation visualization for normal user tweets showing positive contributions to normal classification. Red bars indicate normal language features.

Table 2: Classification Report

	P	R	F1	Supp.
Depr.	1.00	0.98	0.99	45
Norm.	0.98	1.00	0.99	59
Acc.			0.99	104
Macro	0.99	0.99	0.99	104
Wtd.	0.99	0.99	0.99	104

7 Conclusion

The analysis of social media activity for mental health assessment has emerged as a transformative approach in computational psychiatry, offering unprecedented opportunities for early depression detection. Our research establishes that machine learning models can effectively identify depressive tendencies through textual analysis of online posts, achieving an 99% F1-score and statistically significant AUC metrics that demonstrate clinical relevance. These results validate social media data as a valuable supplementary resource for mental health screening, particularly when combined with our novel BERT-based approach that captures nuanced linguistic patterns associated with depression. Looking ahead, this work lays the foundation for impactful collaborations with medical institutions through carefully designed partnerships that would incorporate clinically verified Instagram data while implementing rigorous privacy protections through anonymization protocols. Such interdisciplinary efforts could substantially enhance prediction accuracy while addressing current limitations in generalizability. The proposed framework shows particular promise for developing non-intrusive screening tools capable of identifying at-risk populations through passive monitoring, enabling timely interventions that could reduce treatment gaps. Future extensions will investigate multimodal data integration (combining text and visual content), longitudinal tracking of symptom progression, and demographic-specific adaptations to improve equity in mental healthcare access. By maintaining strong ethical standards regarding data usage and emphasizing clinician-in-the-loop validation, this research direction offers meaningful potential to complement traditional diagnostic methods while reducing barriers to mental health support - ultimately creating more responsive, data-informed care ecosystems that leverage digital biomarkers for public health benefit.

References

[1] De Choudhury, M. and De, S. (2014). *Mental health dis course on reddit: Self-disclosure, social support, and anonymity. In ICWSM.*

[2] Chen, X., Sykora, M., Jackson, T., Elayan, S., and Munir, F. (2018a).*Tweeting your mental health: an exploration of different classifiers and features with emotional sig nals in identifying mental health conditions.*

[3] De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013b). *Predicting depression via social media. ICWSM, 13:1–10.*

[4] Organization, W. H. et al. (2025). *Depression and other common mental disorders: global health estimates. 2025.*

[5] Yu Ching Huang1, Chieh-Feng Chiang2 and Arbee L. P. Chen3*Predicting Depression Tendency based on Image, Text and Behavior Data from Instagram*

[6] Khan AE, Hasan MJ, Anjum H, Mohammed N, Momen S. Predicting life satisfaction using machine learning and explainable AI. Heliyon 2024;10(10).

[7] Koly KN, Sultana S, Iqbal A, Dunn JA, Ryan G, Chowdhury AB. Prevalence of depression and its correlates among public university students in Bangladesh. J Affect Disord 2021;282:689–94.

[8] Islam MA, Barna SD, Raihan H, Khan MNA, Hossain MT. Depression and anxiety among university students during the COVID-19 pandemic in Bangladesh: A web-based cross-sectional survey. PLoS One 2020;15(8):e0238162.

[9] Islam S, Akter R, Sikder T, Griffiths MD. Prevalence and factors associated with depression and anxiety among first-year university students in Bangladesh: a cross-sectional study. Int J Ment Heal Addict 2020;1–14.

[10] Arusha AR, Biswas RK. Prevalence of stress, anxiety and depression due to examination in Bangladeshi youths: A pilot study. Child Youth Serv Rev 2020;116:105254.

[11] Laacke S, Mueller R, Schomerus G, Salloch S. Artificial intelligence, social media and depression. a new concept of health-related digital autonomy. Am J Bioeth 2021;21(7):4–20.

[12] Siddiqua R, Islam N, Bolaka JF, Khan R, Momen S. AIDA: Artificial intelligence based depression assessment applied to Bangladeshi students. Array 2023;18:100291.

[13] Ku WL, Min H. Evaluating machine learning stability in predicting depression and anxiety amidst subjective response errors. In: Healthcare, vol. 12, no. 6. MDPI; 2024.

[14] Marriwala N, Chaudhary D, et al. A hybrid model for depression detection using deep learning. Meas: Sensors 2023;25:100587.

[15] Bhatnagar S, Agarwal J, Sharma OR. Detection and classification of anxiety in university students through the application of machine learning. Procedia Comput Sci 2023;218:1542–50.

[16] Zouache D, Got A, Alarabiat D, Abualigah L, Talbi E-G. A novel multi-objective wrapper-based feature selection method using quantum-inspired and swarm intelligence techniques. Multimedia Tools Appl 2024;83(8):22811–35.

[17] Joyce DW, Kormilitzin A, Smith KA, Cipriani A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. Npj Digit Med 2023;6(1):6.

[18] Al Banna MH, Ghosh T, Al Nahian MJ, Kaiser MS, Mahmud M, Taher KA, et al. A hybrid deep learning model to predict the impact of COVID-19 on mental health from social media big data. IEEE Access 2023;11:77009–22.

[19] Kumar A, Sharma A, Arora A. Anxious depression prediction in real-time social data ARTICLEINFO. In: Proceedings of the accepted for publication in the proceeding of international conference on advanced engineering, science, management and technology–2019. 2023, p. 1–7.

[20] Gao S, Calhoun VD, Sui J. Machine learning in major depression: From classification to treatment outcome prediction. CNS Neurosci Ther 2018;24(11):1037–52.