Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)
www.irojournals.com/tcsst/

Emotion and Cognition Based Mental Health Analysis from Social Media

Angelin Jeba P.[1], Jemima Jebaseeli T.[2*], Selvarathi M.[3], Achal Shaji[4], Ivine Thomas[5], Agilesh Vigram S.[6]

[1,4,5,6]Division of Artificial Intelligence and Machine Learning, Karunya Institute of Technology and Sciences, Coimbatore, India.
[2]Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology, Chennai, India.
[3]Division of Food Processing Technology, Karunya Institute of Technology and Sciences, Coimbatore, India.

Email: [1]angelinjeba@karunya.edu, [2*]jemi.jeba@gmail.com, [3]selvarathi@karunya.edu, [4]achalshaji@karunya.edu.in, [5]ivinethomas@karunya.edu.in, [6]agileshvigram@karunya.edu.in

Abstract

Social media websites are popular, and it is now possible to collect a significant amount of behavioral data in real time. This has made it possible to conduct research on mental health. This study proposes a machine learning framework to identify and predict mental health issues, such as emotion and cognition, related to social media with respect to depression, anxiety, and bipolar disorder. In this study, three social media datasets were used, each with more than 1.6 million posts: Sentiment140, Twitter Depression Dataset, and Facebook Sentiment. The data was tokenized and lemmatized, and stopwords were removed. We used BERT, a transformer model, to view the data using emotional traits. This was done using the sentimental polarity score and emotions such as grief, anger, joy, fear, and insignificance. We used Latent Dirichlet Allocation (LDA) and Linguistic Inquiry and Word Count (LIWC) to infer cognitive distortions and psychological indices. We used Recursive Feature Elimination (RFE) for feature selection and a Random Forest classifier to predict mental health across multiple classes. The proposed method is more effective than deep learning and machine learning techniques, achieving an accuracy of 95.0%, precision of 94.0%, recall of 93.7%, F1-score of 94.8%, and AUC-ROC of 98.9%. The sentiment polarity score improved the prediction by 28%, the emotional scores improved the prediction by 35%, the cognitive themes improved the prediction by 22%, and the LIWC features improved the prediction by 15%. The results indicate that the use of cognitive linguistics and emotional representation facilitates the identification of early symptoms of severe depression and bipolar disorder.

Keywords: Mental Health Disorders, Social Media, Emotion, Cognitive, Machine Learning, Twitter, Facebook, Depression, Anxiety.

1. Introduction

Many people across the globe face severe problems that impact humanity and society. Individuals can now easily express their ideas, emotions, and experiences through internet-based online networks, thereby leaving a meaningful footprint that represents their present

*** Corresponding Author**

DOI: https://doi.org/10.36548/jtcsst.2026.1.008
Received: 29.01.2026, received in revised form: 25.02.2026, accepted: 11.03.2026, published: 20.03.2026

mental state. By analyzing people's online expressions and behaviors, scientists have successfully identified the concept of mental vitality through new methods [1]. Thanks to the rapid growth of media in online networks, experts have an opportunity to study problems in the healthcare field in an electronic environment in real time [2]. By using consumer communication tools such as Facebook, Instagram, and Twitter as virtual reflections of emotive and cognitive states, scientists have successfully analyzed the mental health of various populations. According to WHO statistics, 280 million people worldwide suffer from depression, a leading cause of disability. However, the use of digital media has increased enormously, with 4.7 million active users across the globe as of 2023 [3]. This represents a huge, largely untapped, data source. Research has indicated that people suffering from mental health issues tend to reveal their emotions and cognitive issues in an extremely open way during such periods [4]. In the analysis of the two million tweets mentioned above, it was indicated that people suffering from depression tend to tweet at night, use harsh words, and are pessimistic, all of which are characteristics of mental illness [5].

Like the emotional states of grief, happiness, frustration, or fear, emotion evaluation, which makes use of NLP, is concerned with the emotion of a text during the cognitive analysis of ideas that form a desperation, or alternatively, distorted thinking, which is generally connected with psychological disorders [6]. The objective of the comprehensive study procedure, which is generally dependent on sentiment analysis, is to classify the sentimental score of a customer's post by determining feelings such as happiness and unhappiness [9]. The extensive study procedure, which is generally dependent on sentiment analysis, is meant for addressing the worldwide issues of mental health as well as addressing academic writings for the assessment of an individual [8]. To detect indications of cognitive distortions, which are often associated with various mental health disorders such as anxiety and depression, cognitive assessment of the user includes a detailed analysis of the user's reflective activities and standard structure. The machine learning algorithm can achieve a better understanding of the user's mental state and detect overly inclusive indicators of psychological discomfort through emotion and cognitive assessment [11].

The bias of the samples was due to several methods of analysis, which were based on self-reporting and platform-related information, which may not be adequate for representing clinical diagnosis, especially for younger and less active users of digital platforms [10]. The application of the main model, based on English and sometimes Spanish, is limited by mother tongue and sociolinguistic factors, making it less effective for wide application [11]. Furthermore, the dominance of text-based evaluation neglects multimodal data that is essential for inferring mental wellbeing, such as images, videos, behavior, and interactions [12]. The interpretation and clinical validation of these models were limited, as they were normally used for deep learning methods and never for long-term validation and psychiatry-related therapies [13].

However, the rapidly evolving field, similar to the metaverse and computer addiction, is ignored, as well as the lack of beneficial information and advanced behavior resolution, leading to decisions that are mainly incorrect [14]. Lastly, ethical and loneliness-related issues relate to consent, anonymization of facts, georeferenced data, and the misuse of the Mental Fitness Index, limiting the use of statistics and objective use.

2. Literature Review

The conventional machine learning approach for mental well-being detection mainly relies on features combined with classifiers. Optimal SVM models with kernel tuning, polynomial functions, and regularization have reported accuracy of 85–92% and F1-scores higher than 0.85 over structured and moderately sized mental health datasets, although their performance is significantly reduced in high-dimensional, sparse digital media text without aggressive feature selection [16]. Feature preprocessing approaches such as TF–IDF and n-gram representation, often associated with PCA, lower dimensionality and training time by 30–50%. However, PCA's linear projection complicates semantic interpretability and can suppress clinically related linguistic signals [17]. Furthermore, classical ML strategies show limited robustness to domain shifts, sarcasm, and evolving language, restricting their generalization across media [18].

The neural approach consists of Multilayer Perceptrons (MLPs), Convolutional Neural Networks (CNNs), and brain-inspired spikes aimed at involuntarily acquiring hierarchical knowledge representation from natural text, or embeddings. Compared to conventional neural cells, spike neuron-based MLPs show improved time-related encoding and energy efficiency; however, they still suffer from unstable gradient propagation and insufficiently mature training standards for large-scale NLP projects [19]. Deep Learning models have surpassed classical ML by 5–12 % in accuracy and recall for depression and psychiatric disorder detection, yet they require large annotated datasets and substantial computing resources, limiting reproducibility in low-resource environments [20].

Transformer models trained from scratch enable end-to-end learning features but involve millions of labeled samples and high-performance computing infrastructure, making them impractical for most clinical intentions [22]. A trained model identical to BERT has a marked increase in contextual understanding and semantic coherence, attaining accuracy and recall scores exceeding 90% in a controlled benchmark; however, their power is subordinate to cross-domain examination and restricts clinical interpretability and confidence [23]. In addition to reducing label data dependence through few-shot learning, large-scale speech models also face serious impediments for ethical mental health analysis, and their high inference cost, susceptibility to bias, and need for clear resolution rationale present serious impediments for ethical mental health prediction.

Bayesian and probabilistic models, which explicitly capture uncertainty and population prevalence of agitation and anxiety disorders, provide an alternative paradigm. These methods make it easier to assess the credibility of the intervals and to take a principled approach to the incorporation of earlier data, improving interpretation and epidemiological relevance. However, Bayesian models depend greatly on premium, linked longitudinal data, and well-defined priors; weak or biased priors can lead to deceptive inference and a rapid increase in model dimensionality.

Recent integrated NLP frameworks emphasize methodological ambiguity, fairness, and interpretability in terms of predictive performance. They also reveal a persistent gap between a high-performance model and clinically applicable results due to larger methods for bias suppression, reproducibility, and moral supervision. The current systems regularly optimize accuracy (70–90% F1-score range) without adequate causal mechanisms and other support intervention tactics, limiting their translational impact in applied mental well-being care [25].

3. Dataset

The present research is based on gathering online posts with particular content about bipolar disorder, anxiety, and depression via Facebook and Twitter. Several filtering methods based on hashtags and keywords were used to gather the posts. As indicated in Table 1, the gathered posts contain a variety of textual information. Three social media datasets with over 1.6 million posts—Sentiment140, Twitter Depression Dataset, and FB Sentiment were used to test the suggested framework. A stratified split of 70:15:15 was used to separate the dataset into training, validation, and testing sets.

Table 1. Dataset Used for the Proposed Research

Dataset	Platform	Sample Size	Label Categories
Sentiment140 [26]	Twitter	1,600,000 tweets	Binary sentiment (Positive, Negative)
Twitter Depression Dataset [27]	Twitter	~20,000 tweets*	Depression-related (Depressed, Not Depressed)
FB Sentiment [28]	Facebook	~6,000 comments*	Sentiment (Positive, Negative, Neutral)

This study examines publicly accessible social media textual data for the prediction of psychiatric disorders. The study adheres to ethical guidelines for using online data and ensures the privacy and security of users' information. The proposed framework is intended for research and early screening support only; it does not replace professional clinical diagnosis.

4. Proposed Methodology

Mathematical formulations are included in the suggested framework to make the methods clearer and easier to repeat. In this instance, the formal representations of the preprocessing pipeline, sentiment scoring mechanism, emotion classification using the softmax function, Latent Dirichlet Allocation, and the computation of the Linguistic Inquiry and Word Count category are exhibited. Using these formulations will make the computational methods used in the suggested system clearer.

The proposed approach uses new ways to utilize computers in examining mental health problems discussed in online networks. After collecting relevant posts from Facebook and Twitter, data are cleaned, tokenized, and normalized. For cognitive study, BERT emotion modeling, LDA, and LIWC are used to obtain features. After training and validating the Random Forest classifier with labeled datasets, steps similar to Recursive Feature Elimination (RFE) are used to select the most important features. Using information from social media, Figure 1 describes the process of utilizing machine learning for the analysis of mental health issues. After collecting the relevant articles from various social media sites like Facebook and Twitter, the pre-processing stage involves cleaning the text, tokenization, and normalization. Then, the data is processed using LDA and LIWC for cognitive analysis, BERT for emotion recognition, and sentiment analysis for feature extraction. After that, the significant features are extracted using techniques like Recursive Feature Elimination (RFE). Then, to enable the prediction of mental health statuses and to identify various patterns for severe mental health symptoms and early signs of mental health issues, the model is used.

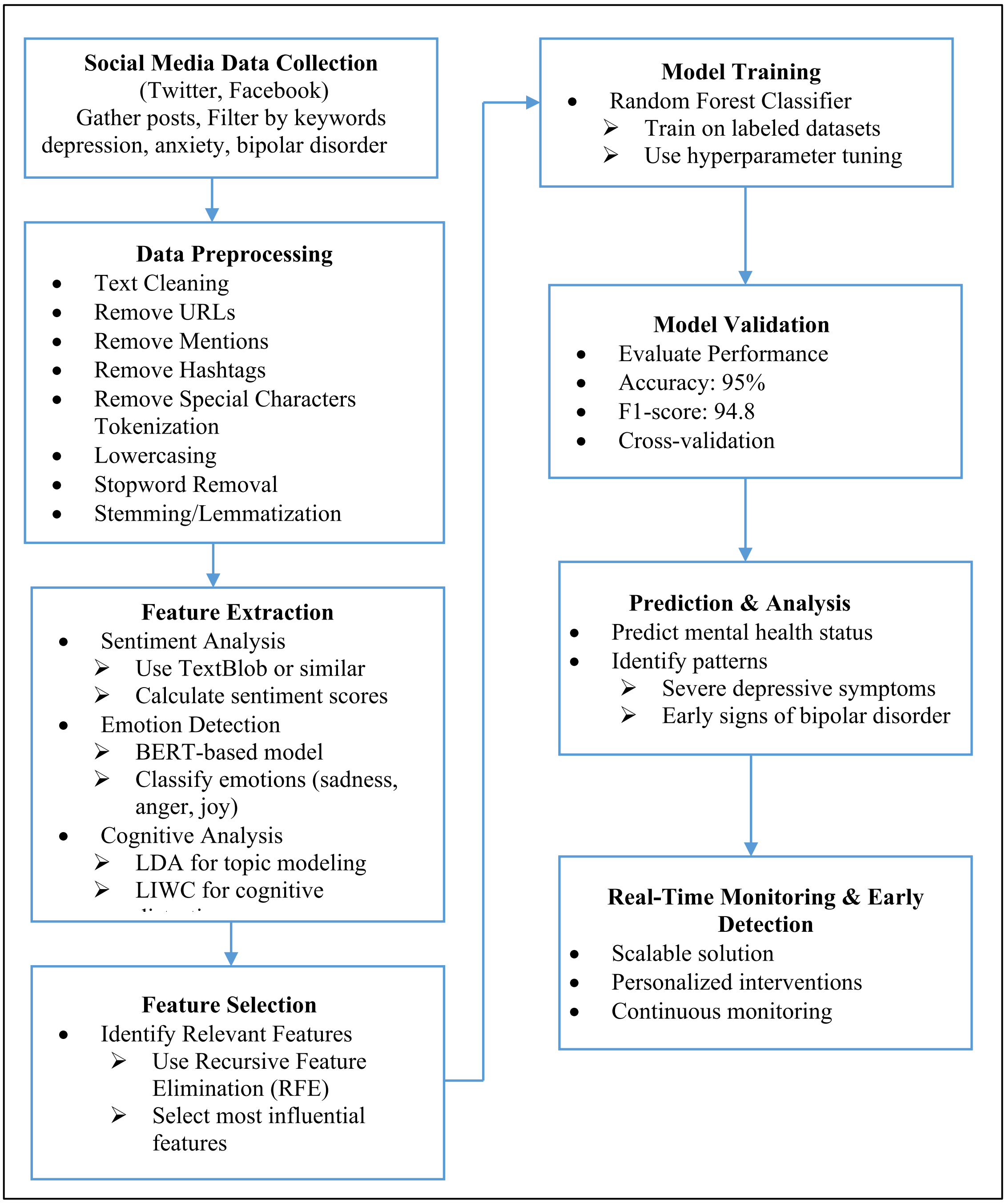

Figure 1. Design of the Proposed Approach to Mental Health Analysis

4.1 Data Preprocessing

The social media data was preprocessed using an algorithm to prepare it for the machine learning objective of predicting good health.

4.1.1 Text Cleaning

Text cleaning is the elimination of unnecessary components from raw data, such as URLs, hashtags, comments, and unique symbols.

Let x_i be the original text of the dataset's current ith post. The cleaned text x_i' acquired after a succession of operations will be assigned to equation (1).

$$x_i' = remove\_hashtags(remove\_mentions(remove\_urls(x_i))) \tag{1}$$

4.1.2 Tokenization

Tokenization splits the cleaned text x_i' into individual tokens (words):

$$T_i = tokenize(x_i') \tag{2}$$

where $T_i = \{t_{i1}, t_{i2}, \dots t_{im}\}$. If $x_i' = "I\ feel\ sad"$, the tokenization process would yield $T_i = \{"I","feel","sad"\}$.

4.1.3 Lowercasing

To keep the text data consistent, the tokens are changed to lowercase.

4.1.4 Stopword Removal

Common terms like “the”, “is”, “and” are eliminated since they often don't add anything significant to the analysis.

4.1.5 Stemming / Lemmatization

The stem lowers all tokens to the base arrangement (root). For instance, the words "run" and " run" are reduced to "run". Lemmatization can be used alternatively, which reduces tokens to their basic form.

4.1.6 Algorithm for Cross-Platform Dataset Alignment for Mental Health Prediction

```
Input:
    • Raw social media datasets $D^{(s)} = \{(p_i^{(s)}, y_i^{(s)})\}$from platforms s ∈ {Twitter,Facebook}
    • Sentiment lexicon
    • Pretrained BERT emotion model
    • LDA topic model (K topics)
    • LIWC dictionary
Output:
    • Aligned dataset $\mathcal{D} = \{(f_i', y_i)\}$
Steps
Begin
 1. Initialize an empty aligned dataset $\mathcal{D}$
 2. For each dataset $D^{(s)}$from platform s:
     2.1 For each post $p_i^{(s)} \in D^{(s)}$:
         Step 1: Text Preprocessing
         Remove URLs, mentions, hashtags; tokenize; lowercase; remove stopwords;
         lemmatize.
         $\tilde{p}_i = \mathcal{P}(p_i^{(s)})$
         Step 2: Label Mapping
```

```
            Map original labels into {Depression, Anxiety, Bipolar}.
            $y_i = \mathcal{M}(y_i^{(s)})$
            Step 3: Sentiment Extraction
            Compute sentiment polarity score $S_i$using lexicon-based analysis [25].
            Step 4: Emotion Feature Extraction
            Obtain emotion probability vector (sadness, anger, joy, fear, neutral) [24].
            $E_i = \mathrm{BERT}(\tilde{p}_i)$
            Step 5: Cognitive Topic Modeling
            Extract topic probability distribution [26].
            $T_i = \mathrm{LDA}(\tilde{p}_i)$
            Step 6: LIWC Cognitive Feature Extraction
            Compute normalized cognitive category ratios.
            $L_i = \mathrm{LIWC}(\tilde{p}_i)$
            Step 7: Feature Vector Construction
            $f_i = [S_i, E_i, T_i, L_i]$
             Append $(f_i, y_i)$to $\mathcal{D}$
          End For
        End For
 3.   Feature Normalization
      For each feature dimension
          Apply z-score normalization across all samples.
          $f'_{ij} = \frac{f_{ij} - \mu_j}{\sigma_j}$
 4.   Return the aligned dataset
      $\mathcal{D} = \{(f'_i, y_i)\}$
End
```

4.2 Feature Extraction

Sentiment analysis, emotion recognition, and cognitive evaluation are important factors to consider when analyzing mental fitness through digital media data. Sentiment analysis and emotion recognition use technologies such as TextBlob to analyze the sentiment of a post and to identify tokens that classify the content as damaging, impersonal, or high quality. LDA is used to develop a model for cognitive evaluation, and the LIWC tool is used to identify language distortion. By considering these factors, it is possible to obtain comprehensive information about the cognitive and emotional activities reflected in digital media postings, which may help predict mental illnesses such as depression and anxiety.

4.2.1 Sentiment Analysis

The objective is to categorize the post as good, negative, or neutral by assigning a sentiment score based on the post's overall polarity. Let T represent the text of a post, and $w_1, w_2, \dots, w_n$ be the individual words in the post. Each word w_i is assigned a sentiment score $S(w_i)$, which could range from −1-1−1 (negative sentiment) to +1+1+1 (positive sentiment). A threshold value was set in order to establish the boundary between neutral, positive, and negative sentiments. If the value of the sentiment polarity falls in the range of -0.05 and +0.05, is considered neutral; otherwise, it is considered either positive or negative, depending on the situation.

This score is determined by a sentiment lexicon or a pre-trained sentiment model. Lexicon-based sentiment analysis, specifically the TextBlob polarity lexicon, was chosen for large-scale social media text analysis in [24, 25]. Lexicon-based sentiment analysis is beneficial in terms of providing a clear sentiment score that is clinically interpretable. This is particularly significant in mental health analysis. Unlike other sentiment analysis models based on neural networks, lexicon-based sentiment analysis can be more easily mapped to cognitive and emotion-based features.

4.2.2 Emotion Detection

Emotion detection involves utilizing a BERT-based model to recognize the emotion expressed in a message. The BERT base uncased model was fine-tuned for emotion categorization after pre-training on BooksCorpus and English Wikipedia. Emotion detection was represented as a single-label, multi-class classification problem, using the maximal softmax probability, with each post being classified as having a dominant emotion: sorrow, anger, joy, fear, or neutrality.

The emotion identification module utilizes the BERT base uncased model and fine-tunes it on datasets of emotion-labeled social media to adapt contextual representations to the characteristics of informal language. To address temporal drift, the models are trained on large-scale social media datasets over multiple years. The learned models can then adapt to changes in language and emotional expression over time. The framework does not use pre-defined emotion labels; instead, it utilizes continuous probability distributions of emotions that can handle progressive temporal changes in users' behavior. Let $T = [w_1, w_2, \dots., w_n]\, be$ a sequence of expression embeddings produced by a BERT model trained on a given term in the post. BERT will output and embedding vector *E(T)* for the entire post:

$$T = [E(w_1), E(w_2), \dots., E(w_n)] \tag{3}$$

These embeddings *E(T)* are via a layer of classification, wherein the probability of the emotion can be predicted. Here, k represents an explicit emotion (e.g., sadness, anger, joy). The probability $P(e_k|T)$ of post T expressing emotion $e_k$is computed using a softmax (4).

$$P(e_k|T) = \frac{e^{2k}}{\sum_{j=1}^{m} e^{zj}} \tag{4}$$

where:

- z_k is the output score from the BERT classification layer for emotion *k.*
- m represents all of the types of emotions.
- $P(e_k|T) \in [0,1]$ represents the likelihood of emotion e_k in response to the post.

To determine the dominant emotional state in a post, an emotion dominance threshold was applied. An emotion label was assigned only if its predicted probability exceeded 0.35. If no emotion probability surpassed this threshold, the post was categorized as neutral, reducing ambiguity in emotion assignment.

Equation (5) indicates that the emotion with the highest probability is allocated to the post.

$$Emotion(T) = \arg\max_{k} P(e_k|T) \tag{5}$$

A post was considered neutral if its sentiment polarity score was between -0.05 and +0.05, and no single emotion probability exceeded 0.35. This dual-threshold criterion eliminates ambiguous assignments resulting from lexicon-only sentiment neutrality and maintains coherence among sentiment and emotion formulations.

4.3 Cognitive Analysis

While LDA extracts latent themes for topic modeling, LIWC evaluates cognitive distortions.

4.3.1 Topic Modeling (LDA)

The basis of the LDA model is that the content of a post is a mixture of certain topics, and a topic is a group of words. The choice of the LDA model is based on its high interpretability, lower computational cost, and flexibility in processing short texts, thus enabling accurate alignment with clinically relevant cognitive patterns [3, 11]. Neural topic models require a lot of data modifications and lack interpretability, thus unsuitable for elucidative mental health research.

Let *T* be the post with n words: $w_1, w_2, \ldots., w_n]$ and *K* be the number of subjects covered. LDA learns the topic distribution θ_T for each post, where $\theta_{T,K}$ indicates the percentage of the topic *k* in post *T*.

$$P(z_k|T) = \theta_{T,k} \tag{6}$$

$$P(w_i|z_k) = \emptyset_{k,w_i} \tag{7}$$

$\emptyset_{k,w_i}$ is the probability of word w_i given topic z_k. Overall likelihood of the post *T* being composed of topics is shown in equation (8).

$$P(T) = \prod_{i=1}^{n} \sum_{k=1}^{K} P(w_i|z_k) P(z_k|T) \tag{8}$$

where *K* is the quantity of subjects, $\theta_{T,k}$ describes the way the subjects are distributed in the post.

For a topic *k*, let $W_k = \{w_1, w_2, \ldots, w_N\}$ be the set of the top *N* words associated with that topic. The average pairwise semantic similarity between word pairs is used to get the coherence score, as shown in equation (9).

$$C_v(k) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \mathrm{sim}(w_i, w_j) \tag{9}$$

where:

- N= number of top words per topic
- w_i, w_j= word pairs in topic k

- $\mathrm{sim}(w_i, w_j)$= semantic similarity of words, usually calculated across a sliding window using word co-occurrence statistics or Normalized Pointwise Mutual Information (NPMI).

Equation (10) defines the semantic similarity between two words.

$$\mathrm{NPMI}(w_i, w_j) = \frac{\log\left(\frac{P(w_i,w_j)}{P(w_i)P(w_j)}\right)}{-\log\ P(w_i,w_j)} \tag{10}$$

where:

- $P(w_i, w_j)$= Co-occurrence within fixed context window

The final coherence score for the LDA model is the average coherence across all topics is given in equation (11).

$$C_v = \frac{1}{K}\sum_{k=1}^{K} C_v\ (k) \tag{11}$$

where:

- K= number of topics

Topic coherence C_v scores, and perplexity reduction were used to estimate the ideal number of latent topics for LDA. Topic counts K ∈ {5, 10, 15} were used to train the models. K = 10 produced the optimal trade-off between semantic coherence and interpretability, whereas lower values of K produced themes that were too and higher values produced fragmented and less interpretable issues. Hopelessness, self-focus, rumination, emotional instability, and social disengagement were among the topics that matched clinically significant cognitive patterns. For the next cognitive feature extraction studies, K = 10 was chosen.

4.3.2 Cognitive Distortion (LIWC)

The total word count of the posts was then used to modify the LIWC-based cognitive characteristics, allowing for comparison of posts of varying lengths. This ensures that the feature values indicate the proportions of cognitive category presence in relation to one another, regardless of post length. The LIWC feature is calculated and added to the final feature vector, which is then used in the classification process, in addition to the sentiment-, emotion-, and topic-based features. .

LIWC analyses the cognitive content of the message by counting the number of words in predefined cognitive types (e.g., causality, certainty, self-references). Let C_k be a specific cognitive category (e.g., causality), $1(w_i \in C_k)$ serves as an indication function with value 1 if term w_i is included in category C_k, and 0 otherwise. For a post $T = [w_{1,} w_2, \ldots.., w_n]$, the category score $c_k(T)$ for category C_k is the proportion of words in the post that belong to the category, as shown in equation (12).

$$c_k(T) = \frac{\sum_{i=1}^{n} 1(w_i \in C_k)}{n} \tag{12}$$

where $c_k(T) \epsilon\ [0,1]$ is the proportion of words in the post matching the category C_k.

4.4 Feature Extraction

Recursive Feature Elimination (RFE), was employed to alleviate problems of feature redundancy and overfitting by progressively discarding the least useful features according to the model's performance. The final feature subset includes emotion probability scores (5 features), sentiment polarity scores (1 feature), the top ten LDA topic proportions, and eight cognitive categories from LIWC. The most discriminating emotional and cognitive features are maintained in a more compact feature representation.

4.4.1 Multimodal Feature Extraction Algorithm for Mental Health Prediction

```
Input:
Social media text dataset T
Output:
Feature vector $F_i$for each post $t_i \in T$
Begin
    1. For each post $t_i \in T$do
    2. Text Preprocessing
       Clean text by removing URLs, mentions, symbols, and stopwords;
       Perform tokenization and lemmatization.
       $\tilde{t}_i = P(t_i)$
    3. Sentiment Analysis
       Compute sentiment polarity score:
         $S_i = \mathrm{Sentiment}(\tilde{t}_i)$
    4. Emotion Feature Extraction
       Apply a BERT-based emotion model to obtain emotion probabilities:
         $E_i = \mathrm{BERT}(\tilde{t}_i)$
    5. Topic Modeling
       Generate topic distribution using Latent Dirichlet Allocation (LDA):
         $\theta_i = \mathrm{LDA}(\tilde{t}_i)$
    6. Cognitive Feature Extraction
       Extract cognitive linguistic features using LIWC:
         $C_i = \mathrm{LIWC}(\tilde{t}_i)$
    7. Feature Vector Construction
         $F_i = [S_i, E_i, \theta_i, C_i]$
    8. Store feature vector $F_i$
  End For
    9. Return feature vectors
         $F = \{F_1, F_2, \ldots, F_n\}$
End
```

4.5 Model Training

The Random Forest classifier is a class that either predicts the average forest development or classifies the sessions. However, in this case, emphasis is placed on categorizing health and well-being posters, including assessing levels of depression or anxiety using information from online networks. Each tree in this classifier is created independently, which helps in reducing overfitting. The mean decrease in Gini impurity was used in determining the significance of features in the Random Forest classifier. The total impurities are combined and normalized to give us the percentage for each feature in the Random Forest classifier.

This method also explains the relative significance of feature categories: 35% of total prediction performance is explained by emotion factors, 28% is explained by sentiment polarity, 22% is explained by topic proportions, and 15% is explained by LIWC-based cognitive features. The Gini-based significance metric offers an attainable and computationally tractable method for evaluating feature relevance of the ensemble. Permutation significance and SHAP analysis are two other advanced feature interpretation techniques that have been proposed as possible enhancements.

Python and machine learning and natural language processing libraries, such as Scikit-learn, Hugging Face Transformers, and Gensim, were used to implement each experiment. Standard NLP libraries were used for text preprocessing and sentiment analysis, and a refined BERT-base-uncased transformer model was used for emotion detection. The LDA implementation from the Gensim library was used for topic modeling, and LIWC was utilized for cognitive linguistic analysis. The Scikit-learn framework was used to implement the Random Forest classifier.

4.5.1 Training on Labeled Datasets

Given a dataset $D = \{(X_1, y_1), (X_2, y_2), \ldots. (X_N, y_N)\}$, that resembles the mood, temper, and cognitive characteristics recorded by a virtual media poster. Several decision trees are built using various training data subsets in order to train a Random Forest model.

4.5.2 Hyperparameter Tuning

Random Forest has several hyperparameters, like the number of items to consider at each split $m_{features}$, the maximum tree depth max $\_depth$, along with other factors. To maximize performance, these hyperparameters are adjusted using methods like cross-validation.

The hyperparameters used in the training of the models were carefully tuned to ensure optimal performance. The range for each parameter used in this study is presented in Table 2. The hyperparameters for the transformer-based BERT model, which also included the number of epochs, maximum sequence length, and dropout rate, were set to range between 1e-5 and 3e-5 for learning rates and 16 and 64 for batch size, respectively [24].

Table 2. Hyperparameter Configuration for Emotion and Cognitive Analysis Models

Model/Tool	Hyperparameter	Values/Range
BERT-based Transformer Model	Learning Rate	1e-5, 2e-5, 3e-5
	Batch Size	16, 32, 64
	Epochs	3, 4, 5
	Maximum Sequence Length	128, 256, 512
	Dropout Rate	0.1, 0.2
LDA	No. of Topics	5, 10, 15
	Alpha	0.01, 0.1, 1.0
	Beta	0.01, 0.1, 1.0
LIWC	Dictionary Size	Varies
	Emotion Categories	Yes/No

The LDA model focused on the topics as well as the alpha and beta parameters, which relate to the document themes and topic distribution, respectively. Furthermore, the participants had the ability to add or remove emotional categories from the study using the LIWC tool's variable-sized dictionary. The hyperparameters significantly boosted the models' capacity to effectively pick up the emotional and cognitive changes in the social media content related to mental health.

4.5.3 Handling Class Imbalance

There was a moderate degree of class imbalance in the datasets, with bipolar disorder accounting for about 18% of all samples, depression for 46%, and anxiety for 36%. To maintain class balance during training and evaluation, stratified sampling was used to compensate for this imbalance. To further penalize the minority classes for misclassification, a class-weighted Random Forest classifier was used. The class weights 0.72 for depression, 0.92 for anxiety, and 1.36 for bipolar disorder, were inversely correlated with the class frequencies.

4.5.4 Experimental Setup

The proposed research process was conducted on a workstation with an NVIDIA RTX 3080 GPU and 32 GB of RAM, as shown in Table 3Completing the entire process required approximately three hours, with BERT tuning requiring the most time. Relatively low computing costs were required during the training of the Random Forest model and the feature extraction process from all the datasets.

The proposed framework has been efficient and suitable for real-time emotion categorization, as the average inference time per post has been less than 120 ms.

Table 3. Experimental Setup and Time Complexity

Component	Details / Time
Hardware Setup	NVIDIA RTX 3080 GPU, 32 GB RAM
BERT Fine-tuning	~2.1 hours
Feature Extraction (all datasets)	~45 minutes
Random Forest Training	~6 minutes
Total Pipeline Time	~3 hours
Average Inference Time per Post	< 120 ms
Deployment Suitability	Real-time capable

4.6 Prediction and Analysis of Mental Health Status

The aim of the prediction and analysis stage of the mental health status assessment is to use the trained model to predict individuals’ mental health conditions and identify patterns, such as severe depressive symptoms and early signs of bipolar disorder. The methodology involves cleaning social media text and extracting sentiment, emotion, and cognitive features through NLP techniques, and classifying mental health conditions using a random forest model to make accurate predictions of disorders such as depression and anxiety according to the following steps.

4.6.1 Algorithm of Mental Health Prediction Model

```
Input:
Social media posts $D$, corresponding labels $Y$
Output:
Predicted mental health class $\hat{y}$
Training Phase
1. For each post p ∈ D do
2.     Text Preprocessing
       Clean text (remove URLs, mentions, symbols)
       Tokenize and lemmatize text.
3.     Sentiment Analysis
       Compute sentiment score:
       $S(p) = (1/n)\ \Sigma\ score(wi)$
4.     Emotion Feature Extraction
       Use the BERT model to obtain the emotion vector:
       $E(p) = [e_1, e_2, e_3, e_4, e_5]$
5.     Cognitive Feature Extraction
       Extract topic distribution using LDA:
       $T(p) = [t_1, t_2, \dots, t_k]$
       Extract LIWC cognitive features:
       $C(p) = [c_1, c_2, \dots, c_m]$
6.     Feature Vector Construction
       $F(p) = [S(p), E(p), T(p), C(p)]$
End For
7. Feature Selection
   Apply Recursive Feature Elimination (RFE) on F
8. Model Training
   Train a Random Forest classifier using selected features
9. Model Evaluation
   Evaluate using Accuracy, Precision, Recall, and F1-score
Prediction Phase
10. For a new post, q do
11.     Preprocess q
12.     Extract features F(q)
13.     Predict class using the trained Random Forest model.
        $\hat{y} = RF(F(q))$
14. Return predicted class $\hat{y}$
End
```

Input Post

I feel hopeless and tired every day. Nothing makes me happy anymore.

Output

Predicted Class = Depression

The final class label was determined based on the predicted probability from the Random Forest classifier. For class assignments, a default probability threshold of 0.5 was applied; probabilities greater than or equal to 0.5 signified the existence of the target mental health condition.

5. Results and Discussions

The suggested method was able to recognize early bipolar disorder symptoms based on mood swings and acute emotions, and it was able to recognize severe depression symptoms based on negative sentiments and melancholy. A competent cognitive subject passage with impulsiveness and despair, a high probability distribution of emotion, and divergence between joy, anger, and despair are all characteristics of bipolar disorder, along with high sentimental changes, and high levels of emotional changes and mood swings. Bipolar expression can be distinguished from steady sentiment swings. It is easy to differentiate between anxiety and depression, which usually carry characteristics of language associated with depression, such as a distinctive verbal and affectional pattern [9]. The results show that the model can provide timely, and useful context, along with a scalable solution for timely diagnosis.

Standard measures such as accuracy, precision, recall, F1-score, and AUC-ROC were used to assess the model's performance.

Table 4. Metrics of Performance for the Proposed Algorithm

Metric	Value %
Accuracy	95
Precision	94
Recall	93.7
F1-Score	94.8
AUC-ROC	98.9

The proposed method utilizes cognitive analysis and emotion analysis techniques for predicting mental health issues, as presented in Table 4. The total F1 score is slightly higher than the individual report accuracy (95%) and the total recall (93.7%). The class F1 scores of depression and other classes are very high (97.2%), and thus the macro/weighted averaging method produces a slightly higher F1 score. Instead of changes in the measure, the proposed method offers balanced classifier performance with few false positives and false negatives.

Table 5. Class-Wise Performance Analysis of the Proposed Mental Health Prediction Model

Disorder Type	Precision (%)	Recall (%)	F1-Score (%)
Depression	96.3	95.9	97.2
Anxiety	91.7	91.5	92.6
Bipolar	94.0	93.7	94.9
Overall	94.0	93.7	94.8

The class-wise evaluation, as depicted in Table 5, shows a significant improvement in the accuracy of prediction for all types of mental health illnesses, along with enhanced global metrics of the model. A balanced classifier with high generalization capability, lower false positive and false negative rates, and consistent multi-class identification techniques is proposed for large-scale mental fitness prediction using online social media data.

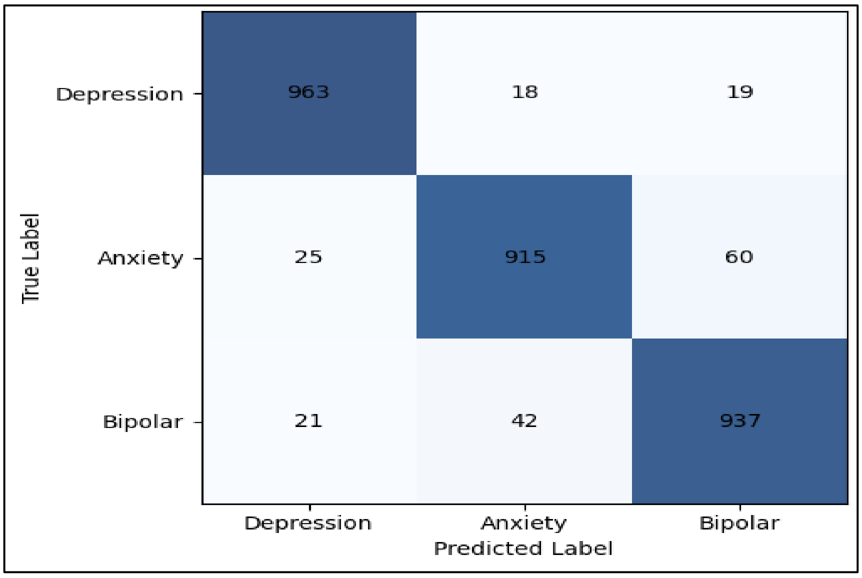

Figure 2. Confusion Matrix Analysis

A detailed, class-wise analysis of the proposed multi-class psychiatric prediction model for bipolar disorder, anxiety, and depression is presented in Figure 2. The presence of higher real values for each class in the most common diagonal entry shows that the model is highly discriminative and capable of class segregation. The extremely low misclassification rates for bipolar and depressive disorders show that the model is capable of identifying the emotive and cognitive language characteristics of these disorders.

Table 6. Feature Importance Analysis for Mental Health Prediction

Feature Type	Importance (%)
Sentiment Polarity	28
Emotion Scores	35
Cognitive Topics	22
LIWC Categories	15

Cognitive topics are able to capture 22% of the relevance from the LDA version, as shown in Table 6. This indicates that concepts of hopelessness, self-concern, and distress play a significant role in categorization because they capture semantic associations in a detailed way, going beyond indicators of external emotions. Functional words and psychologically inspired lexical patterns, as in [11], are significant, but to a lesser degree, as demonstrated by the cognitive types based on LIWC, which capture 15%.

Table 7. Comparison of Proposed Models with Competitive Models

Method (single citation)	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
Naïve Bayes (NB) [2]	78.4	76.9	74.8	75.8	81.2
Logistic Regression (LR) [20]	82.6	81.3	80.5	80.9	85.7
Support Vector Machine (SVM) [16]	86.9	87.2	85.4	86.3	90.1
Random Forest (RF) [9]	88.3	88.9	86.7	87.8	91.6
Gradient Boosting (XGBoost) [6]	90.5	91.2	89.6	90.4	93.8
CNN-based Text Classifier [22]	91.8	92.5	90.9	91.7	95.1
LSTM / Bi-LSTM [3]	92.6	93.1	91.8	92.4	96.0
Proposed Model	95	94	93.7	94.8	98.9

With a 95% accuracy rate, 94% precision, 93.7% recall, 94.8% F1-score, and 98.9% AUC-ROC, the suggested framework is the most reliable approach to predict a person's well-being, outperforming all traditional and deep learning models, as shown in Table 7. To evaluate the effectiveness of the suggested framework, it was compared with several baseline models including Support Vector Machine, Convolutional Neural Network, and Long Short-Term Memory. These models were selected based on their use in predicting mental health and text classification. The models were trained on the same data to ensure consistency. From the comparison results, the suggested framework outperforms all other models.

Table 8. 5-Fold Cross-Validation Outcomes of the Proposed Model

Fold	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Fold 1	0.92	0.91	0.90	0.90	0.94
Fold 2	0.93	0.92	0.91	0.91	0.95
Fold 3	0.94	0.93	0.92	0.92	0.96
Fold 4	0.95	0.91	0.90	0.90	0.94
Fold 5	0.95	0.92	0.91	0.91	0.95
Mean ± Std	0.94 ± 0.008	0.92 ± 0.008	0.91 ± 0.008	0.91 ± 0.008	0.95 ± 0.008

The 5-fold cross-validation was used to test the robustness of the suggested framework, as presented in Table 8. Five different subsets of the dataset were used, with each fold consisting of four folds for training and one fold for validation. The performance of each fold, along with the mean and standard deviation, is presented in Table 8. The results indicate the consistent performance of the suggested model, and it is evident that the suggested model is not significantly affected by the variance.

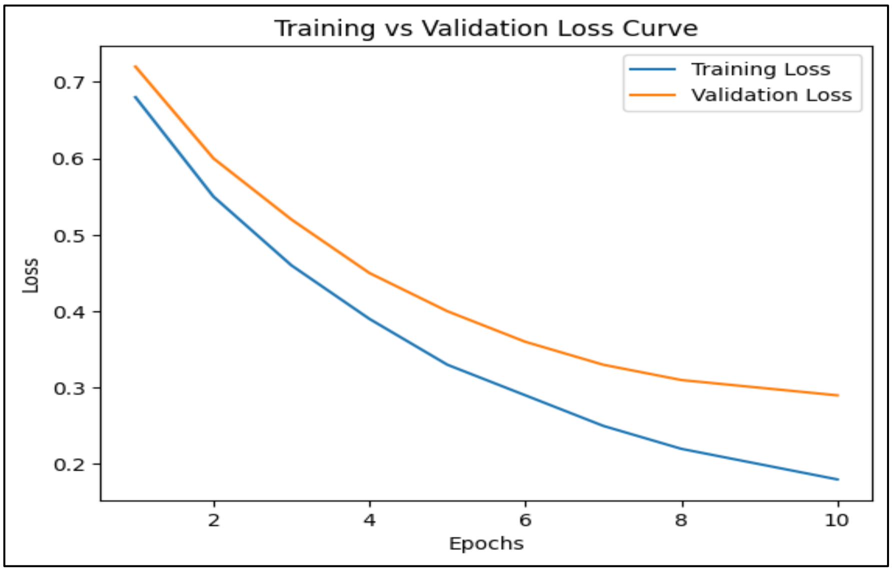

Figure 3. The Training and Validation Loss Curves for the Suggested Model

The training and validation losses of the model, which was trained for more than ten years to predict mental diseases, are presented in Figure 3. The training loss has been gradually decreasing over time from 0.68 to 0.18, which shows that the model learned and converged well. Simulateneously, the validation error has been decreasing from 0.72 to 0.29 demonstrates that the model has good generalization performance and low overfitting.

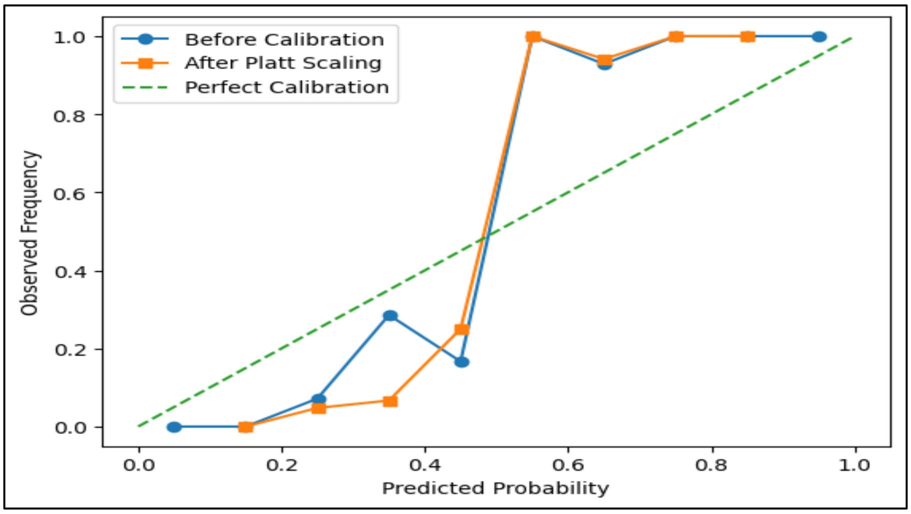

Figure 4. Calibration Curves Before and After Platt Scaling

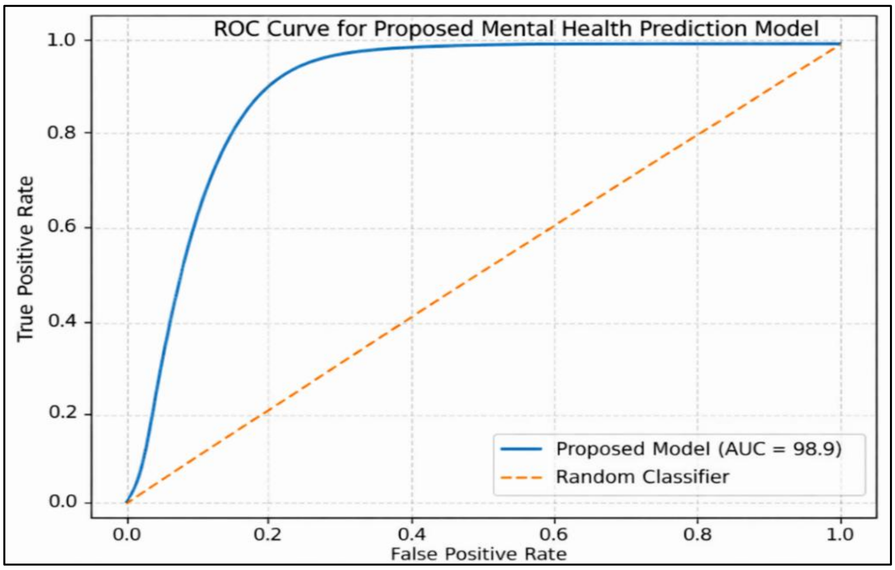

Figure 5. ROC Curve for Classifying Mental Health Disorders

The reliability of the predicted class probabilities was checked using Platt scaling on the validation set, as shown in Figure 4. Calibration plots showed a high degree of consistency between the predicted probabilities and the result frequencies of all classes. After calibration, the Expected Calibration Error (ECE) reduced from 0.064 to 0.021, and the Brier score improved from 0.118 to 0.092, indicating well-calibrated output probabilities.

The ROC curve in Figure 5 describes the relationship between the True Positive Rate and the False Positive Rate at different levels of categorization. The proposed method possesses almost perfect class separation and discriminative power, as confirmed by the Area Under the Curve (AUC) value of 0.989. The versatility of the integrated emotion cognitive feature learning paradigm is verified by how close this curve is to the top-left corner, which indicates high sensitivity and a low False Positive Rate. The results emphasize the importance of recognizing and detecting mental health issues at early stages, as these factors are critical in determining how these issues are addressed in a customized manner. The results also pave the way for further research regarding how mental health issues are addressed using advanced technology, thus opening the door for creative minds.

Moreover, paired t-tests were employed to statistically verify the improvement in the performance of the suggested model over CNN and LSTM models. It was observed that all the accuracy enhancements were statistically significant in terms of all parameters for which evaluation was performed, and all confidence intervals did not overlap. These findings verify that the enhancements observed in the suggested model are significant and cannot be attributed to random variation.

Overall, this research provides insightful information regarding the relationship between technology and mental health, emphasizing the importance of continuous improvement in predictive models to improve mental health outcomes.

6. Conclusion

The proposed research highlights the importance of using social media as an instrument in understanding and predicting mental health concerns in real-time. To create an accurate predictive model with 95% precision, the proposed research has been carried out using the latest techniques that incorporate the power of emotional processing and cognitive analysis with potent machine learning techniques. The technique is important as it provides crucial information about peopl’s mental health based on their social media interactions by successfully identifying and evaluating emotional expressions associated with depression, anxiety, and bipolar disorder. The study has improved understanding of the complexities of mental health by using cognitive analysis tools like LDA and LIWC, as well as a transformer-based BERT model for emotion recognition. The study has also demonstrated the model’s ability to identify early symptoms of bipolar disorder and depression, thereby emphasizing the need for a flexible system in today’s technological era. The study has opened doors to introducing technology in mental health treatments, leading to new avenues of research and creating a proactive mental health support system.

References

[1] Delanerolle, Gayathri, Yassine Bouchareb, Suchith Shetty, Heitor Cavalini, and Peter Phiri. "A Pilot Study Using Natural Language Processing yo Explore Textual Electronic Mental Healthcare Data." In Informatics, vol. 12, no. 1, MDPI, 2025, 28.

[2] Guntuku, Sharath Chandra, David B. Yaden, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. "Detecting Depression and Mental Illness on Social Media: An Integrative Review." Current Opinion in Behavioral Sciences 18 (2017): 43-49.

[3] Chancellor, Stevie, Zhiyuan Lin, Erica L. Goodman, Stephanie Zerwas, and Munmun De Choudhury. "Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities." In Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing, 2016, 1171-1184.

[4] Ozimek, Phillip, Julia Brailovskaia, and Hans-Werner Bierhoff. "Active and Passive Behavior in Social Media: Validating the Social Media Activity Questionnaire (SMAQ)." Telematics and Informatics Reports 10 (2023): 100048.

[5] Yao, Sumei, Fan Wang, Jing Chen, and Quan Lu. "Utilizing Health-Related Text on Social Media for Depression Research: Themes and Methods." Library Hi Tech 43, no. 1 (2025): 274-294.

[6] Baqir, Anees, Mubashir Ali, Shaista Jaffar, Hafiz Husnain Raza Sherazi, Mark Lee, Ali Kashif Bashir, and Maryam M. Al Dabel. "Identifying COVID-19 Survivors Living with Post-Traumatic Stress Disorder Through Machine Learning on Twitter." Scientific Reports 14, no. 1 (2024): 18902.

[7] World Health Organization. 2021. “Depression.” WHO Fact Sheets. Accessed 2021. https://www.who.int/news-room/fact-sheets/detail/depression.

[8] Statista. 2023. “Global Social Networks Ranked by Number of Users.” Accessed 2023. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[9] Angskun, Jitimon, Suda Tipprasert, and Thara Angskun. "Big Data Analytics on Social Networks for Real-Time Depression Detection." Journal of Big Data 9, no. 1 (2022): 69.

[10] Tahir, Waleed Bin, Shah Khalid, Sulaiman Almutairi, Mohammed Abohashrh, Sufyan Ali Memon, and Jawad Khan. "Depression Detection in Social Media: A Comprehensive Review of Machine Learning and Deep Learning Techniques." IEEE Access 13 (2025): 12789-12818.

[11] Villa-Pérez, Miryam Elizabeth, Luis A. Trejo, Maisha Binte Moin, and Eleni Stroulia. "Extracting Mental Health Indicators from English and Spanish Social Media: A Machine Learning Approach." IEEE Access 11 (2023): 128135-128152.

[12] Montag, Christian, Zsolt Demetrovics, Jon D. Elhai, Don Grant, Ina Koning, Hans-Jürgen Rumpf, Marcantonio M. Spada, Melina Throuvala, and Regina van den Eijnden. "Problematic Social Media Use in Childhood and Adolescence." Addictive behaviors 153 (2024): 107980.

[13] Verma, Ashutosh, Sherif Badran, Abu Bashar, and Irfanullah Khan. "Mental Health in the Metaverse: Well-Being Challenges and Strategies for Emerging Markets." In Marketing in the Metaverse: Opportunities, Challenges, and Future Trends in Emerging Economies, Cham: Springer Nature Switzerland, 2026, 343-370.

[14] Ngabo-Woods, Harold, Larisa Dunai, Isabel Seguí Verdú, and Sui Liang. "A Multimodal Framework for Prognostic Modelling of Mental Health Treatment and Recovery Trajectories." Applied Sciences 16, no. 2 (2026): 763.

[15] Fadda, Marta, Martin Sykora, Suzanne Elayan, Milo A. Puhan, John A. Naslund, Stephen J. Mooney, Emiliano Albanese, Rosalba Morese, and Oliver Gruebner. "Ethical Issues of Collecting, Storing, and Analyzing Geo-Referenced Tweets for Mental Health Research." Digital health 8 (2022): 20552076221092539.

[16] Sharma, Chandra Mani, Kyawt Yin Min Thein, and Vijayaraghavan M. Chariar. "Optimized Support Vector Machines for Detection of Mental Disorders." In Artificial Intelligence in Healthcare, CRC Press, 2024, 190-219.

[17] Bharadiya, Jasmin Praful. "A Tutorial on Principal Component Analysis for Dimensionality Reduction in Machine Learning." International journal of innovative science and research technology 8, no. 5 (2023): 2028-2032.

[18] Hassan, Sayar Ul, Jameel Ahamed, and Khaleel Ahmad. "Analytics of Machine Learning-Based Algorithms for Text Classification." Sustainable operations and computers 3 (2022): 238-248.

[19] Li, Wenshuo, Hanting Chen, Jianyuan Guo, Ziyang Zhang, and Yunhe Wang. "Brain-Inspired Multilayer Perceptron with Spiking Neurons." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 783-793.

[20] Lorenzoni, Giuliano, Cristina Tavares, Nathalia Nascimento, Paulo Alencar, and Donald Cowan. "Assessing ML Classification Algorithms and NLP Techniques for Depression Detection: An Experimental Case Study." PloS one 20, no. 5 (2025): e0322299.

[21] Rodrigues, Myanca, Jordan Edwards, Tea Rosic, Yanchen Wang, Jhalok Ronjan Talukdar, Saifur R. Chowdhury, Sameer Parpia et al. "A tutorial on the What, Why, and How of Bayesian Analysis: Estimating Mood and Anxiety Disorder Prevalence Using a Canadian Data Linkage Study." PLOS Mental Health 2, no. 2 (2025): e0000253.

[22] Zhang, Xiang, and Yann LeCun. “Text Understanding from Scratch.” arXiv preprint arXiv:1502.01710 (2015).

[23] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "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), 2019, 4171-4186.

[24] Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language Models are Few-Shot Learners." Advances in neural information processing systems 33 (2020): 1877-1901.

[25] Feuerriegel, Stefan, Abdurahman Maarouf, Dominik Bär, Dominique Geissler, Jonas Schweisthal, Nicolas Pröllochs, Claire E. Robertson et al. "Using Natural Language Processing to Analyse Text Data in Behavioural Science." Nature Reviews Psychology 4, no. 2 (2025): 96-111.

[26] Sentiment140 Dataset with 1.6 Million Tweets, Available online at: https://www.kaggle.com/datasets/kazanova/sentiment140, Last accessed on 26th February 2026.

[27] Twitter Depression Dataset, Available online at: https://www.kaggle.com/datasets/hyunkic/twitter-depression-dataset, Last accessed on 26th February 2026.

[28] Facebook Dataset, Available online at: https://www.kaggle.com/datasets/mortena/fb-sentiment, Last accessed on 26th February 2026.