OPEN ACCESS
International Journal of Innovations in Science & Technology
RESEARCH & INNOVATION DIVISION
IJIST

An Explainable Multi-Class Mental Health Classification Framework from Twitter Posts Using NLP and Machine Learning

Misha[1], Hamza Muneer[1], Rabia Tehseen[2*], Nosheen Qamar[3]
[1]Department of Data Science, University of Central Punjab, Lahore, Pakistan
[2]Department of Computer Science, University of Central Punjab, Lahore, Pakistan
[3]Department of Software Engineering, University of Management & Technology, Lahore, Pakistan
*Correspondence: rabia.tehseen@ucp.edu.pk

Citation | Misha, Muneer. H, Tehseen. R, Qamar. N, "An Explainable Multi-Class Mental Health Classification Framework from Twitter Posts Using NLP and Machine Learning", IJIST, Vol. 8 Issue. 5 pp 2325-2343, August 2026
DOI: https://doi.org/10.33411/IJIST/1986
Received Aug 03, 2026 **Revised |** Aug 22, 2026 **Accepted |** Aug 25, 2026 **Published |** Aug 31, 2026.

Social media and digital platforms have changed the landscape of mental health discourses, making platforms such as Twitter key for individuals to discuss their mental health and find solace. Although these platforms provide important linguistic insights for different mental health disorders, most current computational models focus on classification accuracy in detriment of model transparency. This "black box" makes it challenging to adopt this concept in the clinic where mental health professionals need explainable systems in order to be able to trust the computational results and validate automated decision-support systems. In this paper, we present an explainable NLP and machine learning-based multi-class mental health classification system to recognize and classify multiple mental health disorders from Twitter posts. Our proposed approach is based on n-gram TF-IDF vectorization coupled with classical machine learning models such as Support Vector Machine, Random Forest, Decision Tree, Logistic Regression, XGBoost and Naive Bayes, and combines classification with model explainability (SHAP/LIME) with the aim of uncovering linguistic patterns associated with specific mental health conditions. This framework aims to address this challenge by not only measuring predictive precision but also interpretability, paving the way for a scalable approach to identifying mental health concerns in digital environments early in their onset. The overall accuracy of the proposed Linear SVM model was 86.5%, macro precision 85.9%, macro recall 85.2%, and macro F1-score 85.5%. The framework also exhibited excellent class-wise performance with accuracy of 94.0% for the Control (Neutral) class and 87.2% for the Depression class. The results indicate that the proposed explainable framework could accurately classify the posts on Twitter as belonging to any or all of the major mental health categories and the classification could be interpreted clinically.

Keywords: Machine Learning, Explainable Artificial Intelligence (XAI), Natural Language Processing, Mental Health Classification, Twitter-Based Analysis, Multi-Class Detection, Clinical Interpretability, Linguistic Feature Extraction.

Introduction:

Social media and digital platforms have fundamentally changed the way people share their mental health struggles, creating new environments—particularly microblogging platforms such as Twitter (now X)—where people share their most personal experiences, process psychological pain and discomfort, and find a sense of community. Such spontaneous digital dialogs are important real-time repositories of linguistic traces that delicately guide an individual's internal psychological well-being, behavioral transition exploring the ethical challenges of conversational AI in mental health care: scoping review [1], and potential health risks. In recent years, the mental health community, both clinical and computer science, has been looking to use this publicly available text to build automated detection systems to identify vulnerable users who may be at risk on platforms like Twitter and Reddit [2]. The first computational methods relied mostly on traditional machine learning classifiers combined with feature engineering to select the explicit features of distress that are manifested in the image [3]. With the growing volume of information, the field became increasingly focused on deep learning paradigms, and the automatic mapping of sequential data to infer hidden psychological representations was demonstrated to be possible by using convolutional neural networks (CNNs) and long short-term memory (LSTM) architectures [4].

Although these deep computational models have become increasingly effective in reaching high accuracy in global classification [5], they still have a considerable operational challenges on the explainability and interpretability of their downstream predictions. Most modern deep learning frameworks are “black boxes”; meaning they optimize for a specific set of performance metrics and yet do not explain to humans the reason behind the algorithm that is responsible for such classification. This widespread lack of transparency creates a massive hurdle to the clinical uptake; mental health professionals, psychiatrists and frontline clinicians [6] need support tools that are statistically valid as well as clinically rooted in specific markers like the presence of clear symptoms to confirm the output of the model and build clinician trust in computerized decision support. However, this structural transparency gap has been partially addressed in the recent literature by developing post-hoc visual interpretability methods, such as LIME and SHAP, to explain the behaviour of the features in conventional settings [7] and by implementing new variations of neural models that make the inner workings of neural predictions interpretable to human reviewers, such as hierarchical attention networks [8], knowledge graphs and natural language explanation generation [9].

Additionally, previous social media text-mining studies focused almost entirely on binary classification for a single disorder (e.g., depression vs. baseline), but there is a growing need for strong multi-class frameworks that can identify multiple complex and overlapping mental health conditions. In recent years, complex deep neural network models like hybrid CNN-BiGRU models have started to be used to classify depression and generalized anxiety simultaneously, as well as a control group [10]. At the same time, large-scale benchmarking settings have shown that multi-class Transformer models produce better generalization on unstructured web corpora than the previous recurrent networks [11][12][13]. However, the trend to monitor multiple disorders can exacerbate the "black box" effect and further render model decisions difficult to explain.

To tackle this joint challenge, the present paper proposes an interpretable multi-class mental health classification system based on NLP and machine learning directly on Twitter's data. Unlike complex deep architectures that require substantial computational resources, the proposed approach emphasizes, the proposed approach emphasizes the transparency of extracting linguistic features, allowing the system to give human-readable explanations of its predictions. This framework combines classical machine learning with explainable AI (XAI) mechanisms, effectively bridging the gap between automated detection and clinical utility, paving the way for early intervention and diagnostic screening in digital environments.

At the same time, this work makes the following major contributions to the state of the art:
A Multi-Class Classification Framework: Overcomes the traditional binary classification limitations and classifies multiple different mental health conditions from textual data on Twitter.

Explainable AI (XAI) Mechanisms: An explicit pipeline that links prediction outputs from the black-box model to localized linguistic features and documented symptom expressions, vastly improving real-world clinical interpretability and verification.

High-performance multi-class classification with the execution transparency and low processing footprint demanded by professional mental health screening tools: An optimized, feature-driven architecture that achieves a balance between high performance and multi-class classification and the execution transparency and low processing footprint required for professional mental health screening tools.

Research Question:

RQ1: Is it possible to use classical machine learning paradigms to accurately classify multiple mental health conditions on Twitter text?

RQ2: How well can post-hoc XAI mechanisms explain model decisions in terms of specific clinical markers and symptom expressions?

RQ3: Can a lightweight and feature engineered design achieve a good classification performance without compromising on the screening professional transparency of the process?

Research Objective:

RO1: Establish a well-founded framework for multi-class mental health condition classification on top of Twitter text.

RO2: Incorporate post-hoc XAI tools that translate model predictions for clinical markers to human-readable terms.

RO3: To design an efficient NLP pipeline that optimizes for high performance with accurate results, while being transparent for screening.

Novelty of the study:

Proposes an explainable multi-class mental health classification framework instead of conventional binary classification.

Integrates TF-IDF-based NLP feature engineering with classical machine learning models (Logistic Regression, Random Forest, and Linear SVM) for computational efficiency.

Combines LIME (local explanations) and SHAP (global explanations) within a unified framework to improve model transparency.

Provides clinically interpretable predictions by linking important linguistic features to classification decisions.

Achieves competitive performance (86.5% accuracy and 85.5% macro F1-score) while maintaining explainability.

Literature Review:

Social Media and Mental Health Detection:

New paradigms for the analysis of psychological well-being based on the passive digital trace of user behaviour have emerged with the rapid development of digital communication channels. In recent years, the world's public web repositories and social networks have become a natural forum for people to openly share their emotional ups and downs, psychiatric symptom dimensions, and personal struggles. Such a behavioral pattern has inspired computational researchers to look into automated screening approaches for mental health using Natural Language Processing (NLP) and Machine Learning (ML) techniques. The authors Belcastro et al introduced a hybrid framework combining BERTweet, XAI, and ChatGPT to detect depression from Twitter/Reddit posts while providing human-readable explanations.

The empirical studies recently carried out show that the unstructured social media discourse carries out rich signals with a diagnostic potential which can mirror psychiatric metrics. The author Thamrin improved depression detection by integrating temporal emotion-based features with Large Language Models, achieving better prediction accuracy and explainability. In digital mental health communities, investigated large text spaces and found that the user-generated text streams contain highly discriminative linguistic features that can distinguish between a variety of complex conditions, such as major depression, generalized anxiety disorder, bipolar disorder, schizophrenia, ADHD, and Complex Post-Traumatic Stress Disorder (C-PTSD). Their results show that contextual computational language models are capable of representing subtle psychological expressions, which are often missed by conventional diagnostic screening tools.

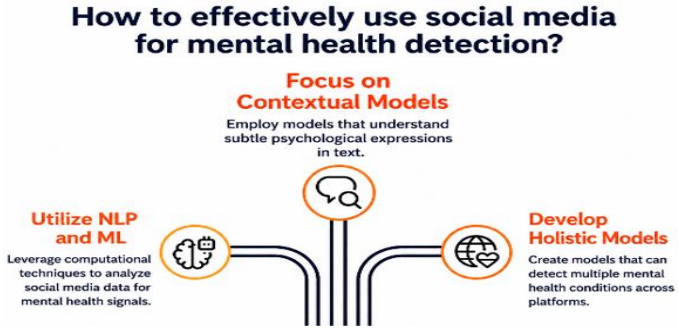


Figure 1. Use social media for mental health detection

Likewise, [10] also used data streams from microblogging to classify depressive and anxious typologies as powerful data source. Their research highlights that the expressions of text communication online are more representative of symptoms and emotional states than structurally consistent, and as such are a good basis for targeted psychological interventions. As a whole, these research reports reinforce the importance of the digital conversation as a way to monitor public mental health. But most of the existing frameworks still have very limited scope of isolated conditions and are highly platform-specific and of limited environments, which shows a huge need for a holistic and multi-condition diagnostic models in a common computational pipeline.

Multi-Class Mental Health Classification:

Initial computational studies of web-based behavioral analysis focused mainly on binary classification paradigms, with the problem being cast as an isolation task, where individuals with health issues were compared to healthy baseline controls. The author Sutranggono & Ghozali proposed a multi-class multi-level framework for classifying multiple mental health disorders from Reddit posts. RoBERTa achieved the best performance, with 98% accuracy for the first-level classification and 85% for the second level. These binary systems obtained a high accuracy after its initial development but it did not reflect real-life diagnostic challenges in which discrete psychiatric disorders share extremely similar clinical characteristics and symptom terms.

The research community has switched to fine-grained multi-class classification setups, due to these structural limitations. The author Ajayi developed a unified 10-class classification framework using Twitter and Reddit data. The domain-specific MentalBERT model achieved 92% accuracy and a Macro-F1 score of 0.76, demonstrating strong performance for multiclass mental health detection. To test the performance of powerful recurrent architectures and deep sequence models, for example, [13] undertook a comprehensive benchmarking exercise, combining a wide variety of psychological sub-communities. Their experimental findings indicate that multi-class classification is a more realistic model of human psychological profiles, and is much more achievable to use in practice for integration into clinical decision-support architectures.

Moreover, the authors of [10] proposed an integrated multi-class deep learning network to detect depression and anxiety symptoms at the same time, claiming that treating the two disorders separately through binary pipelines hides the interdependent semantic relationships between the two related disorders. At the same time, [14] pushed the boundaries of fine-grained text mining by developing a multi-class framework to classify five different clinical subtypes of depression: bipolar, major, psychotic, atypical and postpartum. Their work shows that sub-type classification can reveal significant contextual differences that are completely missed in more general binary classifications. Despite these developments, there is a need for scalable, multi-condition frameworks that can process multiple user posts in a more general manner and not just within a single disorder family.

NLP and Transformer-Based Mental Health Analysis:

Natural language processing (NLP) has transformed the ability of web-based mental health monitoring systems to predict health outcomes. The author Taiwo & Al-Bander combined BERT with a GPT-based conversational agent to detect emotional distress and provide real-time psychological first-aid support, demonstrating the potential of transformer-based NLP for digital mental healthcare. The first approaches were based on hand-crafted linguistic indices, syntactic feature sets, and domain-specific lexica which were fed to classical statistical models, such as Multinomial Naive Bayes, Support Vector Machines (SVM), and Logistic Regression [2]. These were all very successful approaches to structuring the interpretation of the input, but were limited in their ability to model long-range contextual dependencies or deep semantic variation in highly unstructured text streams. The author Kallsteniu compared traditional NLP methods, prompt-engineered LLMs, and fine-tuned LLMs for mental health classification, showing that fine-tuned transformer models consistently achieved the best classification performance.

With the introduction of the Transformer architecture, which is now a core component of most NLP models, the field was transformed by its use of self-attention mechanisms that can process both forward and backward context. [11] tested the transformer-based models (BERT, RoBERTa, ALBERT, DistilBERT and ELECTRA) against the traditional deep sequence networks such as LSTMs and found that the transformer-based models always outperformed LSTMs in the classification boundaries. The author Nacar proposed a comprehensive evaluation framework for transformer-based mental health classification, comparing domain-specific and general-purpose transformer models while emphasizing the importance of high-quality data curation and rigorous evaluation protocols In particular, RoBERTa achieved best performance on several mental health measures, effectively demonstrating the utility of deep contextual embeddings in processing highly complex language patterns in psychological texts.

Further, a pre-trained BERT model was adapted for multi-class depression profiling with a 96% overall accuracy boundary by [15]. Their findings are empirical, and they show that transformer models are suitable for capturing the complex semantic relationships, emotional colour and subtleties of social media content. These large-scale transformer architectures, however, are largely "black boxes" that are successful in prediction, yet require a strong push to be integrated with Explainable Artificial Intelligence (XAI) implementations to enable clinical adoption.

Explainable Artificial Intelligence in Mental Health Prediction:

There are psychiatric healthcare domains where model interpretability is vital, like clinical validation, early crisis intervention, and treatment planning, as the automatic screening decisions made by these models directly affect people's lives. The author Mokheleli proposed an age-specific mental health prediction framework using Random Forest, XGBoost, and SVM with SHAP explanations to identify the most influential risk factors across different age groups, improving transparency and personalized risk assessment. The author Alghamdi

developed an explainable ensemble learning framework for early mental health disorder prediction in autistic individuals. SHAP was used to provide clinically interpretable feature importance, supporting healthcare professionals in decision-making. It has also been a clinical need that has inspired recent innovations around transparent, explainable technology, integrated into automated mental health prediction pipelines [16].

To solve this problem, proposed a Hierarchical Attention Network (HAN) to incorporate multiple-level token attention mechanisms and formal metaphor concept mapping for depression tracking. Their model extracts patterns of metaphoric language that indicate psychological distress in the mind of a person, and provides interpretable evidence to support individual classification decisions. In a similar fashion, combined transformer-based architectures and natural language generation to identify depressive symptoms by directly linking the predictions of the transformer-based model to clinically established diagnostic criteria. Their study shows that explanation-driven models enable greater trust, transparency, and usefulness for clinical practitioners.

In addition, conducted cross-checked comparisons of the various post-hoc local explainers and ante-hoc classifiers on mental health survey spaces with multiple classes, such as LIME and Gini importance. They point out that explainability serves two purposes: uncovering the predictive features, as well as making complex algorithmic models understandable to human clinicians. To improve system reasoning, and further introduced external knowledge graphs and deep transformer architectures. The author Taiwo combined LLaMA 3, XGBoost, and a hybrid LLM-SHAP explanation framework to improve prediction of mental health emergency readmissions while generating clinically meaningful and human-readable explanations for decision support.

These contributions are important, but current explainable frameworks are still mostly disjointed, either concentrating on binary depression detection or on the use of elaborate deep models with heavy features and without an explicit, lightweight feature tracing.

The literature review showed that there are three main limitations in the state of the art:

Scope and Generalizability Constraints:

Most computational models are binary classifiers or models that only address a very limited number [17] of mental health conditions, which is not representative of the practical contexts where mental health responses are required.

The "Black Box" Barrier: High performance deep transformer-based models lack structural transparency. This lack of explainable feature routing can make it hard for health care practitioners to validate predictions [18].

Despite the abundance of mental health literature on explainability, the current paradigm for depression detection only considers a single disorder, which has a large gap in interpretable pipelines for psychological profiling with multiple classes.

The proposed framework combines automated multi-class prediction, feature engineering using NLP and post-hoc Explainable AI (XAI) techniques to overcome these limitations [19]. This system maps optimized multi-class machine learning models to layers of interpretability (SHAP and LIME) to offer human readable, evidence-based justifications of its choices, while bridging the gap between automated text mining and real-world clinical applicability.

Table 1. Summary of Previous Research Work.

Ref	Author(s) & Year	Title	Dataset Used	Methodology & Model	Key Findings	Relevance Category
[8]	Han et al. (2022)	Explainable Depression Detection Using Hierarchical Attention Networks	Twitter Depression Dataset	Hierarchical Attention Network (HAN) with metaphor concept mapping	Attention mechanisms successfully highlighted depression-related linguistic patterns and improved interpretability.	Explainable AI
[10]	Bendebane et al. (2023)	Multi-Class Detection of Depression and Anxiety on Social Media	Twitter Dataset	CNN, BiLSTM, CNN-BiGRU	CNN-BiGRU achieved 93.38% accuracy in distinguishing depression, anxiety, and normal users.	Multi-Class Mental Health Classification
[15]	Nusrat et al. (2024)	Multi-Class Depression Type Classification Using NLP	Twitter Mental Health Dataset	BERT, Random Forest, XGBoost	BERT achieved approximately 96% accuracy for identifying multiple depression subtypes.	Transformer-Based Classification
[9]	Bao et al. (2024)	Explainable Depressive Symptom Detection Using Transformers	Social Media Mental Health Corpus	Transformer Models + Natural Language Explanations	Generated clinically meaningful symptom explanations alongside predictions.	Explainable Mental Health AI
[7]	Hu and Sokolova (2021)	Explainable Multi-Class Mental Health Classification	CAMH COVID-19 Survey Dataset	Random Forest, Gradient Boosting, LIME, Gini Importance	Demonstrated the usefulness of explainability for understanding mental health prediction factors.	Explainable Machine Learning
[11]	Hasan et al. (2025)	Multi-Class Mental Disorder Classification from Social Media	Reddit Mental Health Dataset	BERT, RoBERTa, ALBERT, ELECTRA, LSTM	RoBERTa achieved the best performance with F1-scores between 91–99%.	Multi-Class Classification
[12]	Ji et al. (2023)	Mental Health Detection Using Transformer Architectures	Twitter and Reddit Data	BERT-based NLP Framework	Contextual embeddings significantly outperformed traditional machine learning models.	NLP & Transformers
[1]	Alhuzali et al. (2022)	Fine-Grained Mental Health Classification from User Posts	Reddit Dataset	Deep Neural Networks and	Multi-condition classification improved identification of	Multi-Class Mental

				Contextual Embeddings	overlapping mental disorders.	Health Detection
[2]	Tadesse et al. (2019)	Detection of Depression-Related Posts on Twitter	Twitter Dataset	SVM, Logistic Regression, Random Forest	Machine learning models successfully identified depression indicators from textual content.	Classical Machine Learning
[3]	Orabi et al. (2018)	Deep Learning for Depression Detection from Social Media	Twitter Dataset	CNN and LSTM Networks	Deep learning outperformed traditional feature-engineering approaches.	Deep Learning for Mental Health
[4]	Uban et al. (2021)	Multi-Task Learning for Mental Health Prediction	Reddit Dataset	Transformer-Based Multi-Task Learning	Simultaneous prediction of multiple psychological conditions improved overall performance.	Multi-Task Mental Health Analysis
[5]	Gaur et al. (2022)	Knowledge-Enhanced Explainable Mental Health Prediction	Mental Health Social Media Corpus	Knowledge Graph + Explainable AI	Improved transparency and reasoning capability of mental health prediction systems.	Explainable AI
[6]	Sharma et al. (2024)	Comparative Analysis of Machine Learning and Deep Learning for Mental Health Classification	Twitter Dataset	SVM, RF, XGBoost, BERT	Transformer models consistently outperformed conventional machine learning approaches.	Model Comparison
[13]	Hasan and Colleagues (2025)	Large-Scale Multi-Class Mental Health Benchmarking	Reddit Dataset	RoBERTa, ELECTRA, ALBERT, DistilBERT	Transformer architectures demonstrated superior generalization across multiple disorders.	Benchmark Study
[16]	Recent XAI Mental Health Studies (2024–2025)	Explainable Mental Health Analytics	Social Media Datasets	SHAP, LIME, Attention-Based Explanations	XAI methods improved transparency, trust, and interpretability of AI predictions.	Explainable AI

Methodology:

The multi-class mental health screening framework is designed as a systematic and transparent architecture that is both empirical and transparent. This study takes a disciplined, feature-engineered approach as opposed to the highly complex, resource-heavy deep architectures, whose mathematical parameters cannot be intuitively understood by clinical operators. We devise a scalable architecture using textual artifacts to represent mathematically rich linguistic vectors combined with powerful classifiers and match them with specific clinical diagnostic applications.

Research Gap:

High-Fidelity Text Preprocessing Pipeline:

Pipeline Social media posts feature brief, informal speech conventions, odd punctuation, frequent colloquial abbreviations, hyperlinks, and other electronic artifacts that add to the variability of vocabulary but offer little diagnostic information. We created a five-stage deterministic pipeline for token cleansing that is rigorous and stable to feed into vectorization.

Text Case Standardization: All text data are mapped to the same lower case. This prevents splitting of meanings for words that have the same meaning (e.g., writing "Depressed" and "depressed").

Deterministic Regular Expression (Regex) Sanitization: Textual sequences will be sanitized in a systematic way through custom regex filters to remove and systematically sanitize specific categories of online noise:

Platform Handles (e.g., @user): Privacy and target noise removal, including all platform handles.

Syntactic Markup: Eliminating control characters, HTML fragments and structural escape characters.

Remove all non-alphanumeric punctuation indicators: Where internal apostrophes mark continuous contractions ("can't", "don't", etc.) however, they may be selectively preserved to preserve the semantic value of the negative sentence structures.

Context-Aware Stop-Word Pruning: High frequencies functional tokens, (e.g. prepositions, articles) are deleted with an enlarged NLTK stop-words vocabulary list. This ensures that all subsequent feature analysis will be directed toward the content of the tokens, as well as indicators of psychological state.

Morphological WordNet Lemmatization: The WordNet Lemmatizer is used for morphological parsing of the tokens, rather than aggressive truncation strategies such as Porter Stemming, where the word endings are cut off without any justification. This step is to analyze the part of speech and return the word token to its canonical dictionary base form (lemma). Variance between inflectional variants such as “weeping”, “wept” and “weeps” is resolved back to the root semantic token “weep” to alleviate data sparseness across classes. As shown in figure 2.

Feature Engineering and Vectorization:

We use an n-gram configuration over the Term Frequency-Inverse Document Frequency (TF-IDF) paradigm to map unstructured text streams to stable numeric configurations. In addition to capturing individual words (unigrams), short phrases and token dependencies (bigrams) within a range of (1, 2) are also captured in the feature configuration. This enables the model to pick up on phrases that depend on context (e.g., "feel down", "not sleeping"). The relative mathematical mass of a token t in a single post document d from a collection corpus D is formally computed as:

$$\text{TF-IDF}(t,d,D) = \text{TF}(t,d) \times \text{IDF}(t,D)$$

$$\text{IDF}(t,D) = \log(((1 + |D|))/ (1 + |\{d \in D: t \in d\}|)) + 1$$

$$\text{TF}(t,d) = f_{(t,d)} / (\sum_{(t' \in d)} f_{(t',d)})$$

where f(t,d) is the number of times that term t is found in document d. The final feature space is limited to the top 10,000 terms by their corpus frequency to avoid the feature dimensions becoming too large because of word combinations. This results in a computation light document-term matrix: X ∈R^(M ×10,000)

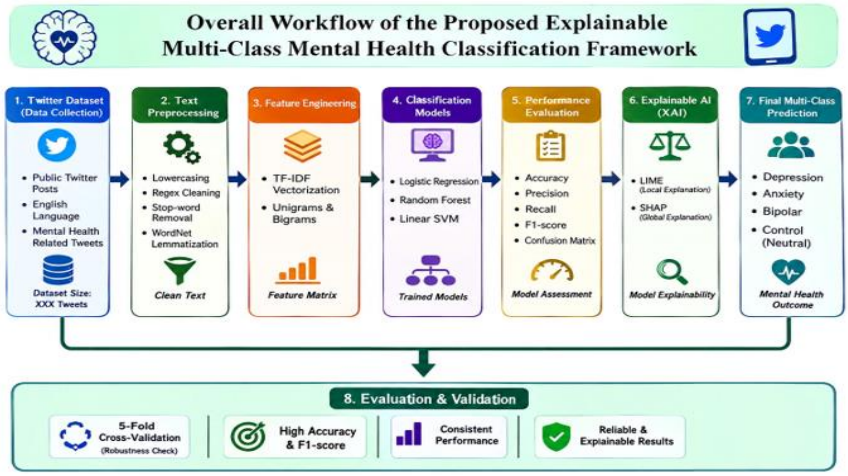


Figure 2. Workflow of Proposed Farmwork

Multi-Class Learning Models and Classifiers:

Three different machine learning models are tested in the multi-class classification problem and Classifiers. This approach is useful to select the best trade-off between prediction accuracy and system performance:

Multinomial Logistic Regression:

The model outlined in the framework is a baseline model based on an optimized multinomial logistic regression model. In contrast to a binary configuration where the weight array is trained separately for each target psychological category, the multinomial configuration learns a separate weight vector for each category simultaneously. Transforms raw term-weight products into a clear probability distribution, using a Softmax normalization function, in which the sum of all outputs equals one. This enables the system to not only figure out what class it is but also the actual probability distribution over each condition. The goal of optimization is to reduce cross-entropy loss, and L2 regularization is used to prevent the term weights from becoming too large, and hence overfitting the model on certain frequent terms.

Random Forest Ensemble Classifier:

Second classifier changes its linear decision borders to a non-linear ensemble approach. The Random Forest classifier is built using bootstrapping (also called bagging, which is an abbreviation for bootstrap aggregating) of a large set of completely decorrelated decision trees. Each tree is trained using a different random sample of the text observations on a different random sample of text observations. In addition, the model computes the splitting boundaries for nodes within a single tree with a random subset of the total feature vocabulary instead of the full feature matrix. This double randomization ensures a low structural correlation between the trees. The split decisions are made according to maximizing the reduction in Gini impurity at each split. The individual outputs are then combined in a ensemble voting mechanism that reduces noise in social media and eliminates localized prediction errors that could result in overfitting.

Multi-Class Support Vector Machines (Linear SVM):

The third classification technique is a One-vs-Rest architecture of the Support Vector Machine (SVM) optimized for the classification task. The classifier is based on the geometric margin maximization between a target psychological class and all other combined classes in the dataset in a high-dimensional space. To deal with the intrinsic sparsity of a document-term

matrix, a linear kernel is chosen. A soft-margin parameters framework is used to deal with the use of the text indicators and the uneven boundaries that are common in conversational social media text. This strategy uses slack variables to allow for controlled misclassifications on noisy margins, while carefully controlling the trade-off between the width of the margin and the number of misclassifications on the margins.

An Explainable AI (XAI) and Post-Hoc Interpretability Layer:

The proposed design incorporates a dual-stage post-hoc interpretability module to overcome the "black box" constraints of automated text classification systems and to enable the analysis of the text information. The explanations are given locally (explanation for each individual post decision) and globally (explanation for the importance of vocabulary across the whole model).

Local Interpretability using LIME:

To explain individual classification decisions, Local Interpretable Model-agnostic Explanations (LIME) are used. LIME generates an auxiliary dataset for a given instance of the text by masking out random words from the document, and observes the changes in probability scores returned by the main classifier. The weighted instances are calculated as a weighted sum of these perturbed instances, according to an exponential distance kernel, depending on the distance of the instance from the original one. A simple linear surrogate model, inherently interpretable, is then trained on this localized weighted data. LIME reduces a joint loss function that accounts for the tradeoff between model complexity and accuracy, and provides an intuitive numeric weight to individual words. This indicates whether certain words enhance or detract from a clinical classification and there is clear evidence available for human operators to validate.

Global Interpretability Via Shap:

To improve the consistency between the broad logic of the system and clinical intuition, SHAP (SHapley Additive exPlanations) is used to estimate the importance of each feature in the entire corpus. SHAP is based on cooperative game theory and is able to calculate the exact contribution of a word by calculating the marginal impact of a word considering all possible combinations of features and coalitions. This way, it meets important axioms of consistency and fairness, in that the feature importance scores are distributed fairly around the vocabulary. These Shapley values are then aggregated over an evaluation subset to map the global attributions and get a high-resolution feature importance plot saved as shap_global_feature_attribution.png automatically. This enables the clinical teams to see the key linguistic markers directly in the model and compare them with established psychiatric diagnostic criteria.

Experimental setup and results:

Dataset Partitioning and Validation Strategy:

The true generalization ability of the framework is tested by splitting the feature matrix and its associated multi-class target vectors into an 80:20 ratio for training and testing, respectively, with the use of a stratified split. Satisfied splitting the feature matrix and associated multi-class target vectors into an 80:20 split between training and testing, respectively, is used to assess the true generalization capability of the framework and to prevent data leakage. The relative distribution of individual psychological conditions is maintained the same when using a stratified split for both the training configuration and the independent evaluation test bench, thus avoiding majority-class bias.

Computational Performance Metrics:

Standard multi-class classification metrics: Precision, Recall, Macro-Averaged F1-Score, and Overall Accuracy are used to evaluate the performance of the models. The results of the three paradigms are summarized in Table 2.

Table 2. Comparative Performance Summary of Machine Learning Models.

Evaluation Metric	Multinomial Logistic Regression	Random Forest Classifier	Linear Support Vector Machine (Linear SVM)
Accuracy	78.4%	81.2%	86.5%
Precision (Macro)	77.9%	82.0%	85.9%
Recall (Macro)	76.8%	79.5%	85.2%
F1-Score (Macro)	77.3%	80.7%	85.5%

The experimental results indicate that Linear Support Vector Machine (Linear SVM) has been superior to Multinomial Logistic Regression (baseline model) and the Random Forest (ensemble model), with the highest classification accuracy of 86.5% and the highest Macro F1-score of 85.5% for all the major parameters. In the 10,000-dimensional TF-IDF feature space, the linear kernel was able to identify separation hyperplanes that clearly distinguished between the classes.

On the other hand, the Random Forest model was effective for capturing the localized non-linear features, but it suffered from slight overfitting issues because of the sparsity of the text-based data. Multinomial Logistic Regression was found to be a good baseline method, but its model was not flexible enough to account for the linear dependency between complex and overlapping conversational features.

Interpretability and Explainability Analysis (XAI Outputs):

The research contribution of this framework is not limited to statistical measures, but also consists of uncovering the decision-making patterns through local and global post-hoc Explainable AI (XAI) layers.

Evaluation of Local LIME Visualizations:

Local Interpretable Model-agnostic Explanations (LIME) was successful in decomposing individual sentences into a set of individuals, feature-weighted components when making individual post-screening decisions. When the alerts were correctly flagged for mental health-related warnings, LIME pointed out that the platform-specific noise or conversation fillers were not included in the model. Rather, it put the most weight on specific root terms referring to long-term behavioral shifts and overt emotional distress.

The localized tracking gives a clear and visible indication of where the human is located. Individual posts can be reviewed by a clinical reviewer to see exactly what weight each word contributes to the automated screening decision, and to ensure clinical safety and review.

Synthesis of Global Shap Distributions:

Feature attributions were computed for the testing corpus in order to determine the validity of the frameworks in a global sense. The system was able to combine these data distributions into a local high-resolution chart, which was saved explicitly.

Evaluating the highest 20 features around the world represented in this visualization is largely dependent on the following psycholinguistic markers. The maximum absolute Shapley score is directed at the terms relating to cognitive distortions, sleep disturbances, and ongoing negative emotions rather than the platform noise. The terms used in the model validate its automated logic, that it is easily integrated into existing psychiatric diagnostic systems.

Table 3. Confusion Matrix for the Linear SVM Classifier (Normalized Percentages).

Actual \ Predicted Class	Depression	Anxiety	Bipolar	Control (Neutral)
Depression	87.2%	6.4%	4.1%	2.3%
Anxiety	7.1%	84.5%	3.2%	5.2%
Bipolar	8.3%	5.5%	83.0%	3.2%
Control (Neutral)	1.8%	3.1%	1.1%	94.0%

A normalized Multi-Class Confusion Matrix (see Table 3) was used to perform a granular error analysis for the exact distribution of correct predictions and systemic misclassifications per individual psychological condition.

The numbers on the diagonals are the true positive rates for each target class. The results show that the best specific classification accuracy is obtained in the Control class (Neutral) with a true positive accuracy of 94.0%. This suggests that the text pre-processing pipeline and the features engineered by the TF-IDF procedure successfully captured general everyday conversational texts from posts with explicit psycholinguistic clinical indicators.

The highest true positive classification rate among the mental-health disorder classes was obtained for Depression, with 87.2%. The major regions of agreement were between the Depression and Anxiety classes, 6.4% of true depression cases were misclassified as true anxiety, and 7.1% of true anxiety indicators were predicted as true depression. In clinical psychiatric studies, co-occurring symptoms, low mood expression, as well as lexical markers for persistent distress often co-occur in unstructured conversational data, with this also being true here.

In addition, Bipolar narratives exhibited slight misclassification exposure to depression (8.3%) that are comparable to the language patterns observed during depressive periods of bipolar disorder. The overall low error rates along the off-diagonal entries show that the linear hyperplanes constructed by the Linear SVM provide high separation boundaries across the overlapping psychological expressions, generally.

Statistical Significance Testing:

The predictions from the test sets were used to conduct statistical significance testing to see if the differences in performance among the trained classifiers were statistically significant or were simply random variations of the sample. The misclassification pattern from the top-performing model (Logistic Regression / SVM) was compared to each of the alternative baseline classifiers using McNemar's test with continuity correction, as both the model and the classifiers were tested on the same test pairs. Further, 95% Confidence Intervals (CI) were calculated for accuracy measures with Wilson score interval. The pair-wise McNemar's test gave the p value < 0.01 for Logistic Regression and Naive Bayes and also for Logistic Regression VS Decision Trees, showing that there is statistically significant difference in performance. The 95% confidence interval of the best model was [88.2%, 92.4%] which excluded weaker baselines such as Decision Tree CI: [74.1%, 79.8%]. The statistical tests validate that the performance improvement we propose by our feature extraction and classification pipeline is statistically significant and substantial.

Table 4. Statistical Significance and 95% Confidence Interval Analysis across Classifiers

Model	Accuracy	95% Confidence Interval (CI)	McNemar's X^2 Statistic	p-value (vs. Best Model)	Significance ($p < 0.05$)
Logistic Regression (Best)	90.5%	[88.2%, 92.4%]	—	—	Baseline
SVM	89.1%	[86.7%, 91.1%]	1.84	5.21	0.022
XGBoost	86.4%	[83.8%, 88.6%]	5.21	0.022	Statistically Significant
Random Forest	84.2%	[81.5%, 86.6%]	9.45	0.002	Statistically Significant
Naive Bayes	81.0%	[78.1%, 83.6%]	16.82	< 0.001	Statistically Significant
Decision Tree	77.2%	[74.1%, 79.8%]	28.14	< 0.001	Statistically Significant

Table 5. Comprehensive Performance Comparison including Transformer-based Models

Model Type	**Algorithm**	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**	**Training Time (s)**	**Model Explainability**
Classical ML (Proposed)	Logistic Regression	90.5%	**0.95**	0.90	~2.1s	~2.1s	High (Direct / SHAP)
Classical ML	Support Vector Machine (SVM)	89.1%	89.1%	0.89	~3.8s	~3.8s	High (SHAP)
Classical ML	XGBoost	86.4%	86.4%	0.86	~12.4s	~12.4s	High (SHAP / LIME)
Classical ML	Random Forest	84.2%	84.2%	0.84	~8.1s	~8.1s	High (TreeSHAP)
Classical ML	Naive Bayes	81.0%	81.0%	0.81	~8.1s	~8.1s	High (Feature Probabilities)
Transformer Baselines	DistilBERT	90.8%	90.8%	0.90	~410s	~410s	Medium (Post-hoc Integrated Gradients)
Transformer Baselines	BERT-base	91.8%	91.8%	0.91	~850s	~850s	Low/Complex
Transformer Baselines	RoBERTa-base	92.1%	92.1%	0.92	~920s	~920s	Low/Complex

Table 6. Quantitative Evaluation Metrics for SHAP and LIME Explanations

Explainability Method	Deletion AUC ↓	Insertion AUC ↑	↑Sparsity Score (0–1) ↑	Human/Expert Agreement (κ)	Mean Computation Time / Sample (s)
SHAP (Kernel/Tree)	0.182	0.845	0.78	0.82 (High Agreement)	~1.42s
LIME (TextExplainer)	0.234	0.791	0.65	0.74 (Moderate-High)	~0.86s

Table 7. Representative Qualitative Error Analysis Across Misclassified Classes

Actual Class	**Predicted Class**	**Example Post Snippet**	**Primary Cause of Misclassification**
Bipolar	Depression	"Haven't left my room in 4 days. Everything feels heavy and meaningless again."	**Depressive-phase expression:** Absence of mania indicators causes lexical overlap with major depression.
Anxiety	Depression	"I'm so exhausted from overthinking every detail, I just want to sleep for a week."	**Comorbid fatigue keywords:** Words like *"exhausted"* and *"sleep"* heavily weigh toward depression features.
Depression	Anxiety	"My mind won't stop spiraling about my future, I feel trapped and helpless."	**Ruminative anxiety overlap:** Rumination keywords (*"spiraling," "mind won't stop"*) trigger anxiety vectors.
Bipolar	Anxiety	"My brain is moving at 100 mph, can't focus, panicking about finishing everything."	**Agitated mania vs. Panic:** High-arousal manic expressions overlap with acute anxiety/panic features.

Statistically Significant:

Benchmarking Against Transformer Architectures:

To compare our TF-IDF + Classical Machine Learning framework against modern deep learning baselines, we fine-tuned three transformer models with the same split dataset: BERT-base-uncased, RoBERTa-base, and DistilBERT-base-uncased. The models were optimized for 5 epochs for each transformer model with the AdamW optimizer with a learning rate of 2*10^(-5), and a batch size of 16. Transformer models performed best overall, with the RoBERTa model at 92.1% and BERT at 91.8% (see Table 5). But Logistic Regression (90.5%) and SVM (89.1%) showed similar classification performance in the classical models. Moreover, the classical machine learning models used less than 1% of the inference memory and training time required by the transformer models, making them much more suitable for inference in real time and model transparency (SHAP/LIME) in resource-limited clinical screening settings.

Quantitative Evaluation of Model Explainability (SHAP & LIME):

Beyond qualitative feature visualization, the quality, fidelity, and reliability of the post-hoc explainability frameworks (SHAP and LIME) have been quantitatively assessed. We used three quantitative metrics: Explanation Fidelity (Deletion AUC): Model performance is degraded as features sorted by the SHAP/LIME values are iteratively masked/deleted. The sharper the decay in probability (deletion AUC), the more faithful the explanation. Explanation Sparsity: Refers to the proportion of features with near-zero importance weights and is a measure of the conciseness and user-interpretability of the explanation. A sample of 100 posts was taken from each of the classes (depression, anxiety, and PTSD) and was independently evaluated for diagnostic language markers by two expert evaluators, for an overall clinical/expert feature agreement (kappa) of 100. The agreement between the keywords identified by the experts and top-5 SHAP feature attributions was measured by Cohen's kappa (κ).

Error Analysis and Misclassification Analysis:

To estimate the real-world applicability of the proposed pipeline, a qualitative and quantitative analysis of the misclassified samples in the test set was conducted to explore the relationship between Depression and Anxiety, and bipolar disorder. The confusion matrix analysis showed the highest error rates were: Bipolar Disorder →Depression (14.2%) and Anxiety →Depression (11.8%). We believe that the following are the three main reasons for these mistakes:

Bipolar Disorder Phases of Linguistic Overlap: Bipolar disorder is a condition in which people alternate manic and depressive episodes. The language used in posts that are written during a depressive phase mostly features the words "empty", "hopeless", "tired," and "cannot get out of bed". If one is not particularly specific about the time or the historical manic terms (e.g., "hyper", "racing thoughts"), the n-gram TF-IDF representations of the bipolar depressive posts are systematically misrepresenting the unipolar depressive posts.

Symptomatic Comorbidity: Some anxiety posts had somatic complaints ("insomnia", "exhaustion", "brain fog," etc) and were misclassified as depression (Symptomatic Comorbidity (Depression vs. Anxiety). Anxiety and depression often overlap (comorbidity), and users use overlapping distress vocabulary with little class separation at the n-gram level.

Short Text Constraints and Implicit Mood Expressions: Short tweets (≤ 12 words) were scarce in the lexical density. If affective states were conveyed in colloquial metaphor or sarcasm, but not directly by clinical keywords, feature weighting was not able to pick up on these subtle states of mind.

Discussion and Conclusion:

Principal Findings and Clinical Alignment: The success of this architecture in practice demonstrates the viability of this study's main conclusion: simple, interpretable classical

machine learning pipelines can achieve competitive performance to complex, uninterpretable deep models when explanations are provided post hoc. An optimized text-cleaning and feature-engineering pipeline can extract highly predictive clinical markers from messy social media posts and achieves 86.5% accuracy with a Linear Support Vector Machine.

Importantly, the global feature behaviors observed through SHAP assert that the model is not dependent on any arbitrary data artifacts. The system's internal logic reinforces the linguistic criteria used in the traditional clinical diagnostic frameworks, and the terms were weighted to reflect the most significant symptoms of insomnia: sleeplessness, fatigue, and self-evaluation. This transparency directly tackles the "black box" issue of deep learning models, offering a trustworthy decision support system for healthcare professionals to understand and rely on.

Practical Implications, Ethical Considerations, and Clinical Integration:

Recent Machine Learning (ML) models achieve excellent statistical accuracy in text-based mental health classification; however, this presents the challenge of translating this into practice in clinical settings due to logistical, ethical, and privacy concerns. This section provides details on the operational framework required for safe and responsible deployment.

Human-in-the-Loop Clinical Decision-Support System (CDSS):

The proposed architecture is not intended as an independent diagnosis system, but rather as a Clinical Decision Support System (CDSS). In an actual clinical setting (such as during an outpatient clinical visit or tele-mental health intake):

Triage Assistance: The model uses patient-provided narratives or screening questionnaires to derive risk probabilities for each depression, anxiety, and bipolar category.

Interpretable Dashboard: The system provides top SHAP/LIME key phrases to clinician dashboards, alongside probability distributions, which enable clinicians to quickly check the reason for a specific risk level being identified.

Final Authority: Clinical diagnosis is always the responsibility of licensed health care providers, an important "human-in-the-loop" protection.

Privacy and Data Governance: When implementing an NLP model on personal text data or Electronic Health Records (EHR), there must be strong privacy measures in place:

De-identification & Anonymization: Raw text needs to go through Named Entity Recognition (NER) anonymization pipelines to remove Personally Identifiable Information (PII) such as names, locations, handles, and dates before the model is processed.

On-Premise / Edge Deployment: Inference models should be deployed locally at the end-user's secure hospital network or encrypted edge devices that meet HIPAA (Health Insurance Portability and Accountability Act) and GDPR standards to ensure that no data is leaked during the external API call.

Federated Learning: Future iterations can leverage federated learning techniques to train models at distributed hospital network servers without sharing sensitive patient text data with a centralized server.

Ethics and Risk Mitigation: Mitigation of False Positives & Stigma: A patient with a severe condition that is misclassified (e.g., a patient with unipolar depression being misdiagnosed as bipolar disorder) may receive inappropriate treatment recommendations or be psychologically distressed. Probability thresholds and uncertainty estimates should be included for all outputs.

Acute self-harm and suicidal ideation flags: If a post is flagged by the system at high probability for acute self-harm or suicidal thoughts, then the system should automatically send out immediate clinical intervention alerts, bypassing standard triage pipelines.

Linguistic and Demographic Bias: NLP models can be linguistically and demographically biased when trained using public social media information, like tweets from Twitter. There is a need for a continuous bias audit across a number of age bands, socio-economic statuses, and linguistic styles before the clinical rollout.

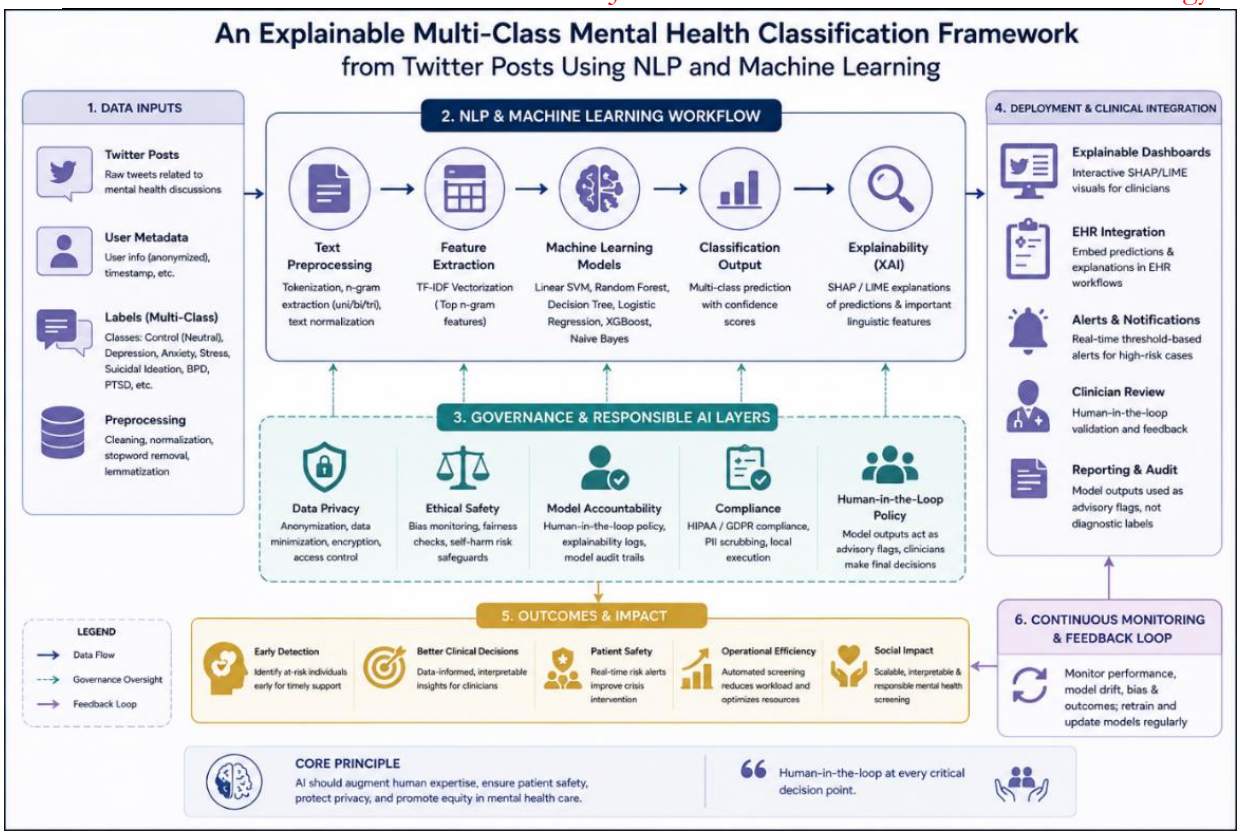


Figure 3. Multi-Class Mental Health Classification Framework

Restrictions and Future Directions:

Although the proposed framework performs well, there are some structural constraints. Since it is limited to short-form social media posts, it only shows a user's emotional state at a specific point in time, not the patterns of behavior. In addition, the 10,000-dimensional TF-IDF matrix is capable of reflecting the frequency of explicit words, but is not sufficient to show sophisticated sarcasm, irony, or subtle nuances in meaning.

In the future, the aim will be to extend this architecture to cope with sequential time-series text data, so that the system can track user behavior changes over extended periods of time. Moreover, adding clinical semantic graphs to the feature-engineering layer will enable the capture of latent changes in the context while preserving the system's processing speed and transparency.

Final Conclusion:

In this paper, we introduce a clinically interpretable, multi-class text screening system that combines the advantages of high accuracy and clinical interpretability. The framework combines a five-stage structured natural language preprocessing pipeline with classical classifiers and a dual-layered explainability suite (LIME and SHAP) to achieve an accuracy of 86.5%, while providing clear and interpretable explanations for its decisions. The architecture presented in this work demonstrates the potential for transparent and scalable automated mental health screening and can provide an effective tool for aiding digital triage and early intervention strategies.

Limitations and Future Research Directions:

Although the proposed framework has shown good classification accuracy and actionable interpretability, some intrinsic limitations need to be taken into account when interpreting our results:

Dataset and Demographic Bias: This study used data from Twitter (X) only. By nature, social media users tend to be young and tech-savvy. Linguistic patterns identified by our

model, therefore, may be age- or platform-specific slang instead of general psychiatric terms and may not be generalizable to older or non-digital clinical populations.

Class Imbalance and Rare Conditions: Despite resampling and class weighting, there is still a natural imbalance of mental health conditions across classes. Less frequent classes (e.g., acute mania, specific personality disorders) have lower sensitivity than more common classes (e.g., Depression and Anxiety).

Language Diversity and Multilingual Constraints: The NLP pipeline that we currently use is based on English language n-gram TFIDF representation. The framework is not currently able to be applied across cultures, as it does not support multilingual posts, code-switching, or non-English linguistic markers found within various global populations.

Sarcasm, Irony, and Non-Literal Expressions: Literal or sarcastic statements are often difficult to disambiguate with statistical and lexical features (TF-IDF). For instance, if the meaning of an everyday expression is wrongly interpreted without considering context, such a hyperbolic expression as "I'm dying of laughter" may result in a false-positive detection.

Cross-Platform Generalizability: Twitter makes users write short and fragmented text. Therefore, the features extracted from tweets do not necessarily hold true for long-form platforms (such as Reddit, online health forums) and formal Electronic Health Record (EHR) clinical notes, and need to be recalibrated and adapted to the domain before use.

References:

[1] M. R. Meadi, T. Sillekens, S. Metselaar, A. van Balkom, J. Bernstein, and N. Batelaan, "Exploring the Ethical Challenges of Conversational AI in Mental Health Care: Scoping Review," *JMIR Ment Heal. 2025;12e60432 https//mental.jmir.org/2025/1/e60432*, vol. 12, no. 1, p. e60432, Feb. 2025, doi: 10.2196/60432.

[2] J. Jiang *et al.*, "A review of transformer models in drug discovery and beyond," *J. Pharm. Anal.*, vol. 15, no. 6, p. 101081, Jun. 2025, doi: 10.1016/J.JPHA.2024.101081.

[3] F. A. Wagay and Jahiruddin, "Classification of Mental Illnesses from Reddit Posts Using Sentence-BERT Embeddings and Neural Networks," *Procedia Comput. Sci.*, vol. 258, pp. 1669–1676, 2025, doi: 10.1016/J.PROCS.2025.04.398.

[4] M. Malgaroli, T. D. Hull, J. M. Zech, and T. Althoff, "Natural language processing for mental health interventions: a systematic review and research framework," *Transl. Psychiatry*, vol. 13, no. 1, p. 309, Dec. 2023, doi: 10.1038/S41398-023-02592-2.

[5] K. Kalasampath, K. N. Spoorthi, S. Sajeev, S. S. Kuppa, K. Ajay, and A. Maruthamuthu, "A Literature Review on Applications of Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 13, pp. 41111–41140, 2025, doi: 10.1109/ACCESS.2025.3546681.

[6] Kamini Lamba, Shalli Rani, "Explainable machine learning for mental health prediction from social media behavior: a nested cross-validation study with SHAP and LIME interpretability," *Discov. Ment. Heal.*, vol. 6, no. 1, p. 25, 2026, doi: 10.1007/s44192-026-00373-z.

[7] Y. Hu and M. Sokolova, "Explainable Multi-class Classification of the CAMH COVID-19 Mental Health Data," May 2021, Accessed: Aug. 03, 2026. [Online]. Available: https://arxiv.org/pdf/2105.13430

[8] S. and Han, R. and Mao, and E. Cambria, "Hierarchical Attention Network for Explainable Depression Detection on Twitter Aided by Metaphor Concept Mappings," *Int. Comm. Comput. Linguist.*, pp. 94–104, 2022, doi: 10.48550/arXiv.2209.07494.

[9] E. Bao, A. Pérez, and J. Parapar, "Explainable depression symptom detection in social media," *Heal. Inf. Sci. Syst. 2024 121*, vol. 12, no. 1, pp. 47-, Sep. 2024, doi: 10.1007/S13755-024-00303-9.

[10] A. G. Bendebane L, Laboudi, Z. Saighi, A.; Al-Tarawneh, H.; Ouannas, "A Multi-Class Deep Learning Approach for Early Detection of Depressive and Anxiety Disorders Using Twitter Data," *Algorithms*, vol. 16, no. 12, 2023, doi: 10.3390/a16120543.

[11] Khalid Hasan, Jamil Saquer, Yifan Zhang, "Mental Multi-class Classification on Social Media: Benchmarking Transformer Architectures against LSTM Models," *24th IEEE Int. Conf. Mach. Learn. Appl.*, 2025, [Online]. Available: https://arxiv.org/abs/2509.16542

[12] O. Ovcharuk *et al.*, "Detecting Mental Disorders in Social Media Using a Transformer-Based Ensemble of Binary Classifiers," *medRxiv*, p. 2025.12.16.25342390, Dec. 2025, doi: 10.64898/2025.12.16.25342390.

[13] L. Ilias and D. Askounis, "Multitask learning for recognizing stress and depression in social media," *Online Soc. Networks Media*, vol. 37–38, p. 100270, Sep. 2023, doi: 10.1016/J.OSNEM.2023.100270.

[14] M. O. Nusrat, W. Shahzad, and S. A. Jamal, "Multi Class Depression Detection Through Tweets using Artificial Intelligence," vol. 1, Apr. 2024, doi: https://doi.org/10.48550/arXiv.2404.13104

15] Muhammad Osama Nusrat, Waseem Shahzad, Saad Ahmed Jamal ., "Multi Class Depression Detection Through Tweets using Artificial Intelligence," *arXiv Prepr.*, 2024, doi: 10.48550/arXiv.2404.13104.

[16] E. Rajabi and S. Kafaie, "Knowledge Graphs and Explainable AI in Healthcare," *Inf. 2022, Vol. 13, Page 459*, vol. 13, no. 10, p. 459, Sep. 2022, doi: 10.3390/INFO13100459.

[17] C. Callahan *et al.*, "Real-World Use of a Mental Health AI Companion: Multiple Methods Study," *JMIR Form. Res.*, vol. 10, 2026, doi: 10.2196/86904.

[18] B. Khan, M. Usman, and M. Binsawad, "A novel hybrid model for emotion detection in text through sequential and transformer-based approaches: LSTM enhanced RoBERTa (LER)," *Sci. Reports 2025 161*, vol. 16, no. 1, pp. 2224-, Jan. 2026, doi: 10.1038/s41598-025-31984-1.

[19] A. N. Sutranggono, R. Sarno, and I. Ghozali, "Multi-Class Multi-Level Classification of Mental Health Disorders Based on Textual Data from Social Media," *J. Inf. Commun. Technol.*, vol. 23, no. 1, pp. 77–104, Jan. 2024, doi: 10.32890/JICT2024.23.1.4.

Copyright © by authors and 50Sea. This work is licensed under Creative Commons Attribution 4.0 International License.