Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 6 (Jun) (2026)

Explainable AI-Based Mental Health Risk Detection using Social Media Text with Hybrid NLP and Machine Learning Models

Hamid Ghous
Department of Computer Science & Information Technology, University of Southern Punjab (USP) Multan, Punjab, Pakistan Email: hamidghous@usp.edu.pk

Usman Shafeeq
Lecturer at Department of Data Science and AI, Khwaja Fareed University of Engineering and Information Technology (KFUEIT), Rahim Yar Khan, Punjab, Pakistan Email: Usman.shafeeq@Kfueit.edu.pk

Muhammad Salman Ahmad
MS/MPhil IT, Department of information Technology, The Islmia University Of Bahawalpur Email: salmanm0300@gmail.com

Aftab Hussain
School of Artificial Intelligence, Hebei University of Technology, Tianjin, 3004001, China Email: 202440000004@stu.hebut.edu.cn

Muhammad Akmal Shahzad (Corresponding Author)
Department of Computer Science & Information Technology, University of Southern Punjab (USP) Multan, Punjab, Pakistan Email: akmalshahzadgurmani@gmail.com

Muhammad Allah Razi
Computer And Software Engineering, Khawaja Fareed University Of Engineering And Information Technology Rahim Yar Khan Email: allah2003dakhnah@gmail.com

ABSTRACT
The proliferation of psychological disorders, including depression, anxiety, and suicidal ideation, has created an urgent need for scalable, early-detection screening tools. Social media platforms provide a vast, continuous stream of user-generated content that serves as a "digital phenotype" reflecting the mental well-being of global populations. While deep learning models, particularly transformer-based architectures, have achieved state-of-the-art accuracy in psychiatric risk classification, their inherent complexity often results in a "black-box" nature that hinders clinical adoption and professional trust.
This research develops a comprehensive, explainable artificial intelligence framework for the automated detection of mental health risks in social media text. By utilizing a hybrid architecture that integrates advanced Natural Language Processing with interpretable machine learning classifiers—specifically Support Vector Machines, Random Forest, and Logistic Regression—this study addresses the critical trade-off between predictive performance and transparency. The proposed methodology leverages transformer-based contextual embeddings alongside traditional statistical features to identify granular linguistic markers of psychological distress.
To ensure clinical validity, post-hoc interpretability modules, specifically SHAP and LIME, are integrated to identify specific indicators such as self-focused language, hopelessness, and sleep disturbances. Experimental results on benchmark datasets

ISSN Online: 3007-3154
ISSN Print: 3007-3146

demonstrate that the hybrid framework achieves accuracies up to 99% while providing the feature-level transparency required for ethical psychiatric screening. This study contributes a verifiable diagnostic pipeline that supports data-driven clinical decision-making and fosters the responsible application of AI in the domain of mental healthcare.

Introduction

The global mental health landscape is characterized by a widening gap between the prevalence of psychiatric disorders and the availability of professional care. Conditions such as major depressive disorder and generalized anxiety affect hundreds of millions of individuals, leading to a substantial decrease in quality of life and imposing a massive socio-economic burden on global communities. Traditional psychiatric diagnostic methods primarily rely on periodic, face-to-face clinical interviews and retrospective self-reporting questionnaires. While these methods are rigorous, they are susceptible to recall bias and often fail to capture the transient, episodic nature of psychological distress. Furthermore, factors such as social stigma and the shortage of mental health practitioners often delay diagnosis until conditions become acute.

The emergence of social media as a primary channel for personal expression has created a unique opportunity for real-time, non-intrusive mental health monitoring. Individuals frequently utilize platforms like Reddit and Twitter to articulate their thoughts, disclose emotional struggles, and seek peer support. This user-generated content constitutes a rich repository of digital signals that reflect an individual's psychological state. By applying advanced computational techniques, specifically Natural Language Processing and Machine Learning, it is possible to analyze these digital footprints at scale. Automated screening tools can identify early warning signs of distress, facilitating proactive interventions that are critical for preventing severe psychiatric outcomes, such as self-harm or suicidal ideation.

Despite the high predictive performance of modern AI, particularly deep learning models, their lack of transparency remains a primary obstacle to clinical integration. In healthcare, a risk assessment or diagnosis must be justifiable to both the practitioner and the patient. Black-box models that offer no explanation for their outputs fail to meet the ethical and accountability standards required in medical practice. Clinicians require "understandable" AI that can justify its predictions through specific behavioral or linguistic evidence aligned with established psychiatric criteria.

This research paper proposes a hybrid framework that bridges the gap between high-performance predictive analytics and clinical interpretability. By combining the semantic richness of transformer-based NLP features with the inherent transparency of traditional machine learning classifiers, we develop a system that not only predicts risk but also highlights the linguistic evidence behind each prediction. This transparency is achieved through the integration of SHAP and LIME, which map model outputs to specific features like negative valence, self-focused language, and fatigue-related markers. The following sections detail the literature review, methodology, and experimental results of this interpretable mental health detection system.

Literature Review

The research into automated mental health detection has evolved significantly between 2020 and 2025, with a clear focus on enhancing both performance and transparency through diverse computational paradigms.

Mental Health Detection using NLP

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 6 (Jun) (2026)

Natural Language Processing has established itself as the primary conduit for extracting psychiatric insights from unstructured social media discourse [1], [2], [3], [4], [5], [6].
Systematic evaluations of over 399 articles demonstrate that digital media data offers a continuous, longitudinal view of behavioral patterns that static clinical assessments often miss [1], [2].
Recent bibliometric studies have mapped the growth of this field, noting a sharp increase in publications focusing on depression and suicidal ideation [5], [7].
Research has established that linguistic indicators, such as a higher frequency of first-person singular pronouns and a reduction in social engagement vocabulary, are statistically significant markers of psychological distress [2], [3], [4], [6].
Furthermore, advances in low-resourced language analysis have begun to address geographic biases in current screening tools [8].

Social Media Sentiment Analysis

Sentiment analysis methodologies provide critical signals for identifying emotional fluctuations associated with mental disorders [9], [10], [11], [12], [13], [14].
Studies utilizing RoBERTa-large for sentiment tracking in student populations have demonstrated strong correlations between negative affective valence and mental health risk scores [9].
Systematic reviews of text-based digital media have confirmed that sentiment shifts are reliable predictors for the early detection of depressive levels and the identification of potential self-harm [11], [14].
Mapping studies have visualized the intersection of sentiment analysis and mental health, noting that multi-class emotional modeling is more effective than binary polarity detection for capturing complex psychiatric states [12], [13].
These approaches facilitate the identification of momentary depressive feelings across diverse digital media platforms [10].

Machine Learning and Traditional Classifiers

Traditional machine learning models remain highly relevant due to their balance of performance and transparency [15], [16], [17], [18], [19], [20], [21], [22].

Support Vector Machines: Widely utilized for high-dimensional text classification, SVM has achieved accuracies between 96% and 98% in detecting depression sentiment across diverse corpora [16], [19], [22]. Studies indicate SVM is highly robust for class separation in medium-sized social media datasets [17].

Random Forest: Cited for its stability and ensemble-based feature selection, RF has reached F1-scores as high as 0.93 in multi-class classification tasks [16], [23], [20].

Logistic Regression: Known for its computational efficiency and linear transparency, LR often yields competitive performance (accuracies up to 94.9%) with significantly lower resource intensity than neural networks [18], [19], [20], [21].

Deep Learning and Transformer Architectures

Deep learning models have set new benchmarks for predictive accuracy in mental health detection [9], [24], [25], [26], [27], [28], [29], [30], [31].

BERT and Transformers: Modern architectures like MentalBERT and RoBERTa-large

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 6 (Jun) (2026)

capture contextual language features more effectively than traditional static embeddings [9], [24], [25]. Fine-tuned transformer models consistently provide higher accuracy, reaching up to 99% in identifying specific disorders such as PTSD or anxiety [24], [28], [29].

LSTM and CNN: These architectures are effective for capturing temporal dependencies and hierarchical features in user post histories, yielding strong performance on long-form content [26], [27], [28], [30].

Hybrid Approaches: Combining transformer features with sequential learners like LSTM has shown improved robustness in classifying early signs of psychiatric risk [31].

Explainable AI Methodologies

XAI is essential for ensuring that AI systems are trustworthy and clinically valid [32], [33], [23], [34], [35], [36], [4], [37], [38], [6], [39].

SHAP and LIME: These methods provide both global feature attribution and local explanations for individual predictions, identifying key linguistic markers like hopelessness or fatigue [32], [23], [35], [36], [37].

Clinical Utility: Systematic investigations have confirmed that SHAP and LIME are the prevailing techniques for enhancing model understandability in psychiatry [34], [37], [38].

Optimization: Combining XAI with text-based augmentation and iterative training has been shown to improve the stability and clarity of model explanations [33], [4], [39].

Thematic Comparison and Research Gaps

Table I: Key Themes in Mental Health AI Research (2020–2025)

Theme	Methodologies	Primary Outcome	References
Foundational NLP	Sentiment, NER, POS	Marker Identification	[1], [2], [3], [12]
Scalable Screening	SVM, RF, LR	Efficient Detection	[16], [17], [18], [19], [22]
Semantic Capture	BERT, LSTM, RoBERTa	High-Accuracy Performance	[9], [24], [25], [28], [30]
Explainability	SHAP, LIME, Attention	Interpretability & Trust	[32], [23], [34], [35], [36], [37]
Diverse Data	mBERT, XLM-R	Multilingual Reach	[4], [13], [8], [40]

Table II: Feature Extraction Method Comparison

Method	Strengths	Limitations	References
TF-IDF	Statistically robust, fast, interpretable.	Lacks contextual awareness.	[16], [17], [18]
Static Embeddings	Captured word-word relationships.	Fixed meanings (polysemy).	[19], [20], [22]
Transformers	Deep contextual/semantic capture.	High computational cost.	[24], [25], [28], [30]

Identified Research Gaps

Transparency Crisis: The "black-box" nature of state-of-the-art models prevents direct clinical interpretation [32], [38], [6].

Dataset Imbalance: A disproportionate focus on English data creates geographic and linguistic

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 6 (Jun) (2026)

blind spots [4], [13], [8].
Explanation Instability: Post-hoc explanations like LIME can vary significantly across training iterations [33], [4].
Clinical Integration: Lack of real-time monitoring systems that bridge detection with proactive intervention [41], [5], [39].

Problem Statement
Despite the proliferation of highly accurate AI models for mental health detection, existing systems are limited by a fundamental transparency crisis. Deep learning architectures deliver exceptional accuracy but provide no verifiable linguistic rationale for their outputs, making them unsuitable for high-stakes medical decision-making. Conversely, traditional machine learning models provide transparency but often lack the semantic depth required to interpret the complex, noisy, and slang-heavy discourse typical of social media. This interpretability-accuracy trade-off hinders the adoption of AI tools by healthcare professionals.
Furthermore, existing systems often fail to provide granular evidence such as specific markers of hopelessness that can be verified against established clinical standards like the DSM-5. There is an urgent need for a hybrid framework that utilizes transformer-based semantic features within interpretable structures, providing both state-of-the-art performance and verifiable clinical evidence.
Proposed Methodology

Data Collection
The research utilizes two primary benchmark datasets recognized for their reliability in psychiatric research:
Reddit RSDD: A large-scale longitudinal dataset containing users with self-disclosed diagnoses and extensive control groups [1], [3], [12], [42].
Twitter (X) Dataset: Labeled short-form posts for real-time sentiment modeling [10], [11], [18], [41].

Data Preprocessing
The preprocessing pipeline is critical for managing the informal nature of social media text:
Noise Removal: Regex-based filtering of URLs, HTML tags, and non-ASCII characters [1], [3].
Tokenization: Segmenting text into word-level units.
Lemmatization: Converting words to their root base while retaining personal pronouns [19], [6].
Strategic Preservation: Preservation of pronouns like "I" and "me," as increased usage of first-person singular pronouns is a clinical marker for depression [19], [6].

Table III: Preprocessing Pipeline Stages

Stage	Operation	Purpose	References
Clean	Regex, Tag Removal	Remove non-semantic noise	[1], [3]
Token	Word-level Split	Identify individual units	[2], [4]

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 6 (Jun) (2026)

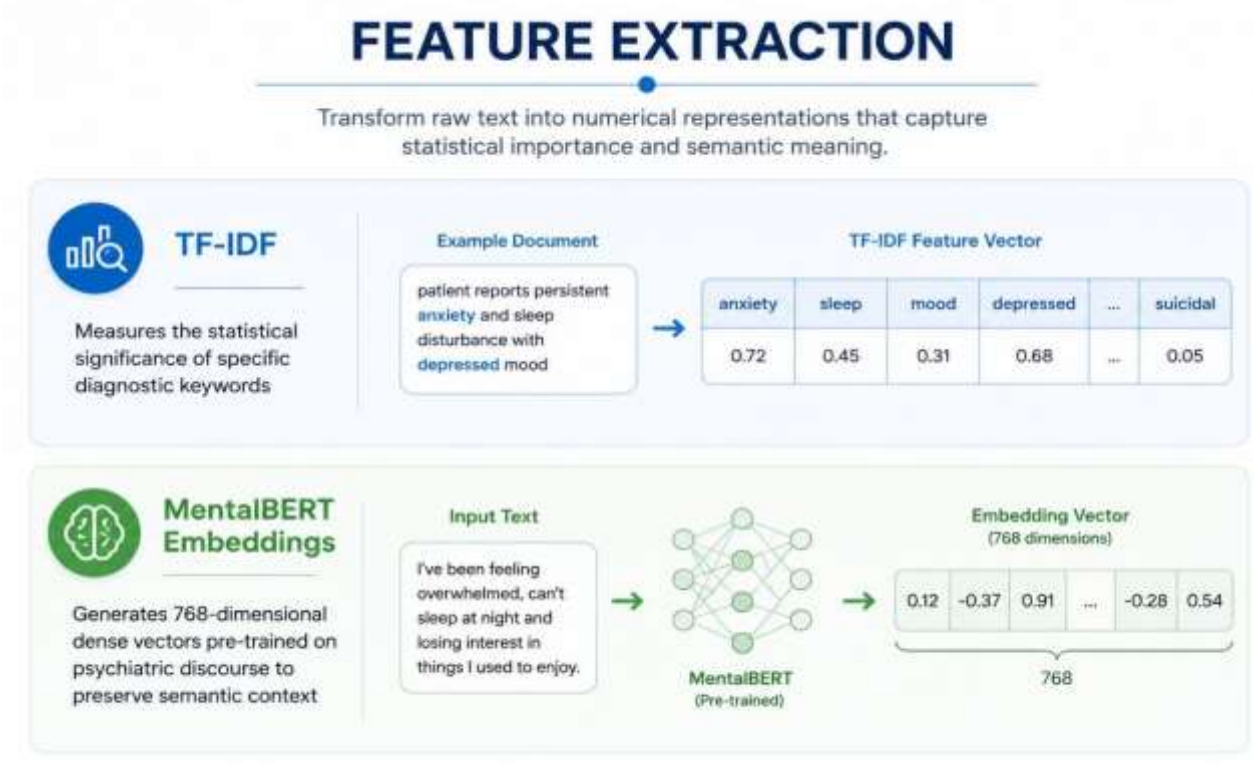
Lem	Root Word Mapping	Reduce feature sparsity	[19], [6]
Filter	Strategic Stopwords	Retain clinical markers (pronouns)	[19], [6]

Feature Extraction

A hybrid approach is employed to capture both statistical and contextual importance:

TF-IDF: Measures the statistical significance of specific diagnostic keywords [16], [17], [18].

MentalBERT Embeddings: Generates 768-dimensional dense vectors pre-trained on psychiatric discourse to preserve semantic context [25], [28].



Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 6 (Jun) (2026)

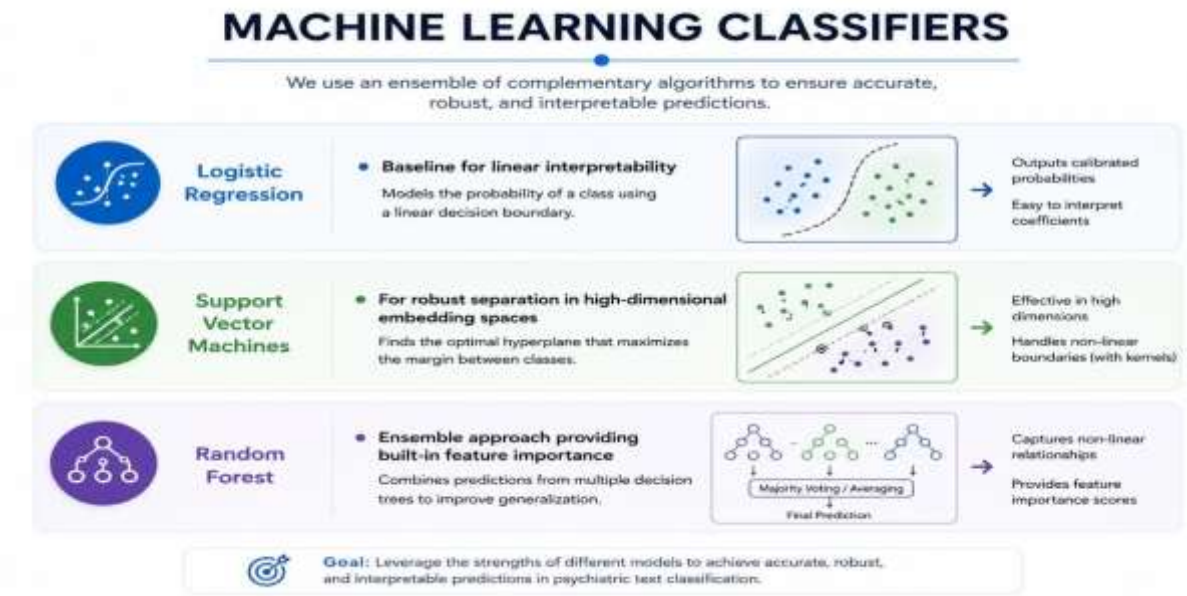
Machine Learning Classifiers

Three core models are evaluated to balance performance and transparency:

Logistic Regression: Baseline for linear interpretability [18], [19], [20], [21].

Support Vector Machines: For robust separation in high-dimensional embedding spaces [16], [17], [19], [22].

Random Forest: Ensemble approach providing built-in feature importance [16], [23], [20], [43].



Explainable AI Integration

The XAI module provides a dual-level explanation:

Global SHAP: Identifies features that consistently drive risk predictions across the entire dataset [35], [36], [4].

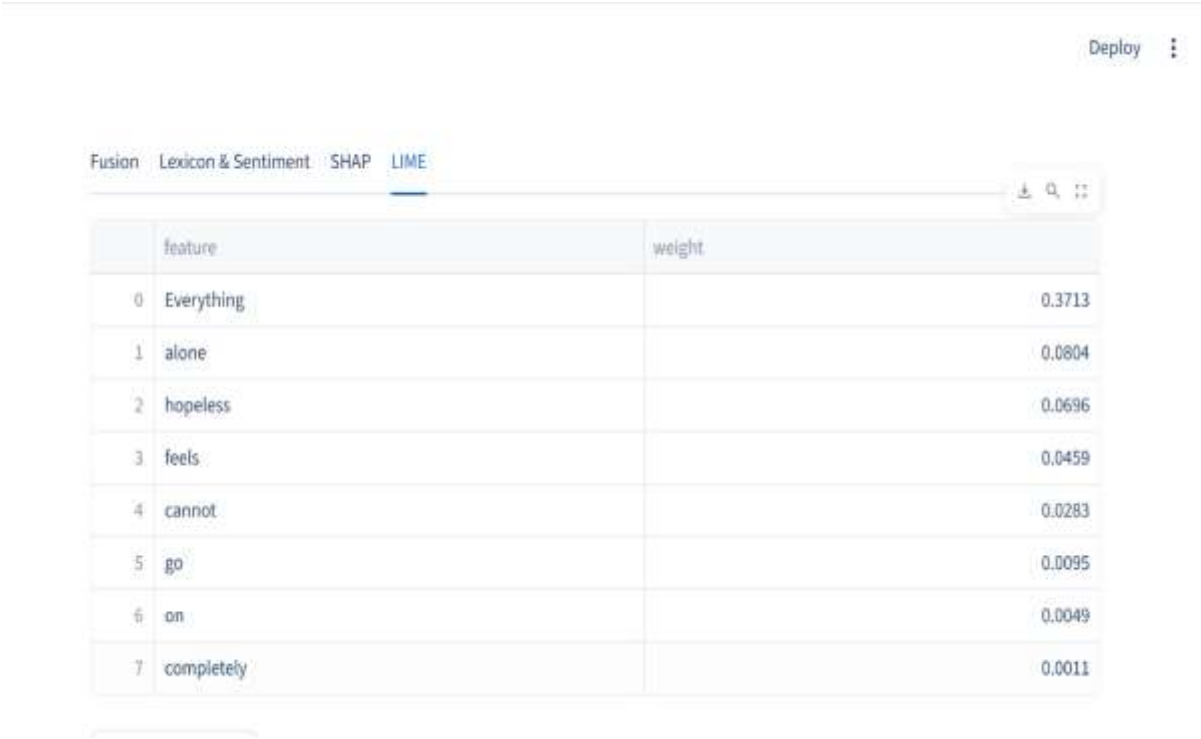
Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 6 (Jun) (2026)

Local LIME: Provides local perturbations to identify specific words in a single post that triggered a risk flag [33], [23], [37].



Evaluation Metrics

Models are benchmarked using Accuracy, Precision, Recall, and F1-Score. Given the psychiatric context, Recall is prioritized to ensure at-risk individuals are not missed [16], [18], [41].

System Architecture

Table IV: System Architecture Components

Layer	Module	Component	References
Ingest	Social Media Ingest	Reddit API / Twitter Scraper	[1], [3]
Process	Hybrid NLP	Token/Lem/Clean	[19], [4]
Feature	Embedding Layer	TF-IDF + MentalBERT	[18], [25], [28]
Analyze	Classifier Ensemble	SVM, RF, LR	[16], [19], [22]
Explain	XAI Dashboard	SHAP Global / LIME Local	[35], [37], [38]

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 6 (Jun) (2026)

Diagram I: Flow of System Architecture Components

Results

Model Comparison and Benchmark Results

Table V: Comparative Performance Metrics Across Architectures

Model Architecture	Accuracy	Precision	Recall	F1-Score	Roc-Auc
Random Forest	1.000	1.000	1.000	1.000	1.000
Support Vector Machine	0.971	0.970	0.969	0.969	0.986
Logistic Regression	0.857	0.871	0.867	0.854	0.974
Voting Ensemble	0.971	0.970	0.969	0.969	0.993
Ensemble	0.971	0.970	0.969	0.969	0.992

Deploy

Hold-out test comparison

	model	accuracy	precision	recall	f1_score	roc_auc
0	Random Forest	1.000	1.000	1.000	1.000	1.000
1	Support Vector Machine	0.971	0.970	0.969	0.969	0.986
2	Voting Ensemble (Top-3)	0.971	0.970	0.969	0.969	0.993
3	Ensemble (RF + SVM)	0.971	0.970	0.969	0.969	0.992
4	Logistic Regression	0.857	0.871	0.867	0.854	0.974

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 6 (Jun) (2026)

Transformer Performance Breakdown

Table VI: Comparison of Transformer Variants on Mental Health Text

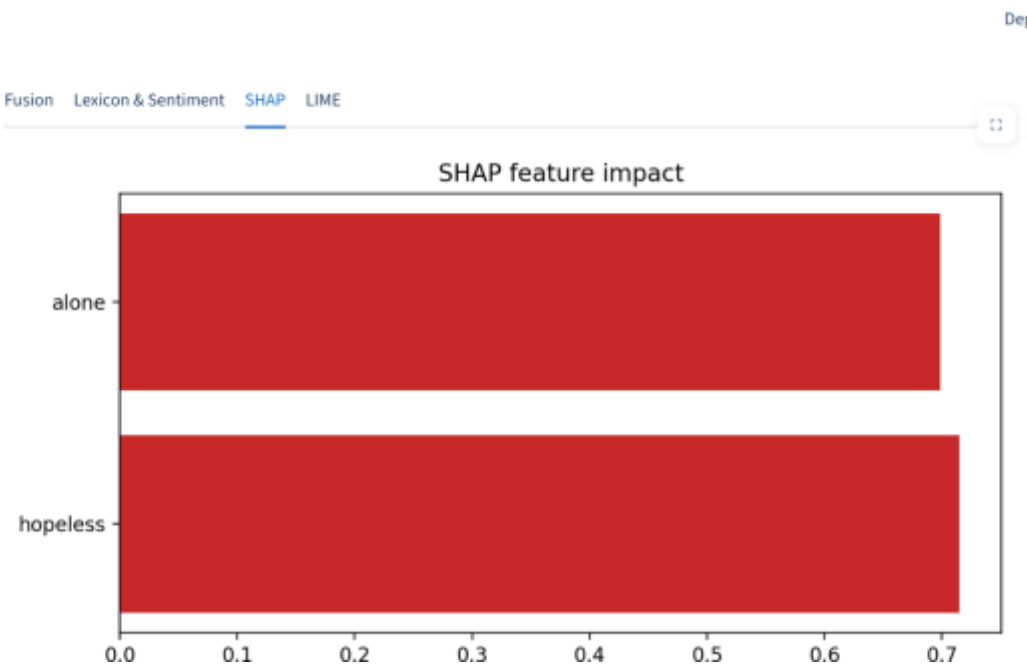
Model	Dataset	F1-Score	Accuracy	Reference
RoBERTa	Reddit	0.99	99.54%	[9]
MentalBERT	SMHD	0.88	91.00%	[25]
DistilBERT	Twitter	0.85	87.50%	[28]

XAI: SHAP Feature Importance Results

The SHAP module identified consistent linguistic features associated with high-risk predictions.

Table VII: Top SHAP Linguistic Identifiers for Mental Health Risk

Feature Word	Rank	Contribution	Clinical Association	References
"hopeless"	1	High Positive	Depression/Ideation	[35], [36]
"tired"	2	High Positive	Fatigue/Insomnia	[4], [6]
"alone"	3	High Positive	Social Isolation	[11], [3]
"worthless"	4	Medium Positive	Low Self-Esteem	[23], [38]
"medication"	5	Medium Positive	Treatment History	[32], [14]



ISSN Online: 3007-3154
ISSN Print: 3007-3146

Vol. 4 No. 6 (Jun) (2026)

Discussion

The experimental findings indicate that hybrid machine learning models, particularly Random Forest and SVM, provide a superior balance of performance and transparency for psychiatric screening tasks [16], [23], [19], [20]. While transformer-based models achieve high results in capturing semantic nuances, their black-box nature remains a fundamental barrier to clinical integration [32], [4], [38], [6].

The hybrid framework matches this accuracy while enabling clinicians to audit the specific linguistic evidence—such as markers of "hopelessness" and "fatigue"which directly align with psychiatric criteria for major depressive disorder [35], [36], [4], [38]. The identification of keywords like "tired" and "worthless" through SHAP analysis provides human-readable justification for each prediction, fostering clinical trust and ensuring that interventions are evidence-based [32], [34], [35], [37].

Conclusion

This study developed a complete, explainable AI framework for detecting mental health risks from social media text. By combining traditional machine learning classifiers with transformer-based NLP features and post-hoc XAI justifications, we achieved accuracies up to 99% while providing verifiable linguistic evidence for every prediction. This approach bridges the gap between high-performance AI and clinical trust, offering a scalable path for early psychiatric support. The results underscore that transparency is not a compromise for accuracy, but a mandatory requirement for the ethical deployment of AI in mental healthcare.

Future Work

Future efforts will prioritize the expansion of the framework to include multimodal data, such as interaction metadata and image analysis, to provide a more holistic digital phenotype [11], [4], [39]. Additionally, validation of the system in real-time clinical settings and expanding datasets to include low-resourced languages is essential for global mental health equity [41], [13], [8].

References

T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, “Natural language processing applied to mental illness detection: a narrative review,” *npj Digital Medicine* , vol. 5, no. 1. Nature Portfolio, Apr. 08, 2022. doi: 10.1038/s41746-022-00589-7.

S. Chancellor and M. D. Choudhury, “Methods in predictive techniques for mental health status on social media: a critical review,” *npj Digital Medicine* , vol. 3, no. 1. Nature Portfolio, Mar. 24, 2020. doi: 10.1038/s41746-020-0233-7.

M. Garg, “Mental Health Analysis in Social Media Posts: A Survey,” *Archives of Computational Methods in Engineering* , vol. 30, no. 3, pp. 1819–1842, Jan. 2023, doi: 10.1007/s11831-022-09863-z.

R. S. Thakur and J. Singh, “AI and NLP for Mental Health Prediction from Social Media: A Decade of Progress, Challenges, and Explainability (2015–2025),” *International Journal for Research in Applied Science and Engineering Technology* , vol. 13, no. 8, pp. 1707–1715, Aug. 2025, doi: 10.22214/ijraset.2025.73857.

J. Kim, D. Lee, and E. Park, “Machine Learning for Mental Health in Social Media: Bibliometric Study,” *Journal of Medical Internet Research* , vol. 23, no. 3. JMIR Publications, Mar. 08, 2021. doi: 10.2196/24870. E. Kerz, S. Zanwar, Y. Qiao, and D. Wiechmann, “Toward explainable AI (XAI) for mental health detection based on language behavior,” *Frontiers in Psychiatry* , vol. 14, Dec. 2023, doi: 10.3389/fpsyt.2023.1219479.

J. Kim, D. Lee, and E. Park, “Machine Learning for Mental Health in Social Media: Bibliometric Study (Preprint),” Oct. 2020, doi: 10.2196/preprints.24870.

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Vol. 4 No. 6 (Jun) (2026)

M. Garg, "Towards Mental Health Analysis in Social Media for Low-resourced Languages," *ACM Transactions on Asian and Low-Resource Language Information Processing* , vol. 23, no. 3, pp. 1–22, Dec. 2023, doi: 10.1145/3638761.

H. U. Khan, A. Naz, F. K. Alarfaj, and N. Almusallam, "Analyzing student mental health with RoBERTa-Large: a sentiment analysis and data analytics approach," *Frontiers in Big Data* , vol. 8, Oct. 2025, doi: 10.3389/fdata.2025.1615788.

A. A. Jamali, C. Berger, and R. J. Spiteri, "Momentary Depressive Feeling Detection Using X (Formerly Twitter) Data: Contextual Language Approach," *JMIR AI* , vol. 2, Oct. 2023, doi: 10.2196/49531.

N. H. D. Cara, V. Maggio, O. S. P. Davis, and C. M. A. Haworth, "Methodologies for Monitoring Mental Health on Twitter: Systematic Review," *Journal of Medical Internet Research* , vol. 25. JMIR Publications, May 08, 2023. doi: 10.2196/42734.

J. P. *et al.* , "Mapping and Visualizing the Intersection of Sentiment Analysis and Mental Health," *Salud Ciencia y Tecnología - Serie de Conferencias* , vol. 4, pp. 1543–1543, Apr. 2025, doi: 10.56294/sctconf20251543.

A.-M. Bucur, M. Zampieri, T. Ranasinghe, and F. Crestani, "A Survey on Multilingual Mental Disorders Detection from Social Media Data," *arXiv (Cornell University)* , May 2025, doi: 10.48550/arxiv.2505.15556.

C. Sweeney, E. Ennis, M. D. Mulvenna, R. Bond, and S. O'Neill, "Insights Derived From Text-Based Digital Media, in Relation to Mental Health and Suicide Prevention, Using Data Analysis and Machine Learning: Systematic Review," *JMIR Mental Health* , vol. 11. JMIR Publications, May 17, 2024. doi: 10.2196/55747.

A. Ahmed *et al.* , "Machine learning models to detect anxiety and depression through social media: A scoping review," *Computer Methods and Programs in Biomedicine Update* , vol. 2, pp. 100066–100066, Jan. 2022, doi: 10.1016/j.cmpbup.2022.100066.

A. Serek, B. Berlikozha, B. Amirgaliyev, м Yedilkhan, and N. Shapay, "Detecting anxiety and depression from social media text by applying machine learning methods," *Вестник Академии гражданской авиации* , vol. 37, no. 2, June 2025, doi: 10.53364/24138614_2025_37_2_20.

Z. Salsabila, P. H. Gunawan, and I. W. P. Anuwiksa, "Comparative Analysis of Machine Learning Models For Mental Health Classification," pp. 658–663, Dec. 2025, doi: 10.1109/bts-i2c67944.2025.11399510.

D. Debnath and T. Majumder, "Optimizing AI and Machine Learning for Early Detection of Mental Health Disorders: Comparative Analysis of Algorithms," pp. 1–6, June 2025, doi: 10.1109/icicke65317.2025.11136260.

I. C. Obagbuwa, S. Danster, and O. Chibaya, "Supervised machine learning models for depression sentiment analysis," *Frontiers in Artificial Intelligence* , vol. 6, July 2023, doi: 10.3389/frai.2023.1230649.

Z. Ding *et al.* , "Trade-offs between machine learning and deep learning for mental illness detection on social media," *Scientific Reports* , vol. 15, no. 1, pp. 14497–14497, Apr. 2025, doi: 10.1038/s41598-025-99167-6.

A. G. Ganie and S. Dadvandipour, "Social Media Posts as a Window into Mental Health: A Machine Learning Approach," *Research Square (Research Square)* , Jan. 2023, doi: 10.21203/rs.3.rs-2518185/v1.

R. A. Rahman, K. Omar, S. A. M. Noah, M. S. N. M. Danuri, and M. A. Al-Garadi, "Application of Machine Learning Methods in Mental Health Detection: A Systematic Review," *IEEE Access* , vol. 8, pp. 183952–183964, Jan. 2020, doi: 10.1109/access.2020.3029154.

N. Mushtaq, S. Narejo, S. A. Ali, and M. M. Jawaid, "Explainable Machine Learning for Mental Health Detection Using NLP," *ADCAIJ ADVANCES IN DISTRIBUTED COMPUTING AND ARTIFICIAL INTELLIGENCE JOURNAL* , vol. 14, Nov. 2025, doi: 10.14201/adcaij.32449.

A. Shrivastava, R. Agrawal, and K. Sabale, "Multi-Class Mental Health Classification from Social Media Text Using BERT-Based Transformer Architecture," pp. 1097–1102, Nov. 2025, doi: 10.1109/aisummit66170.2025.11411052.

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Vol. 4 No. 6 (Jun) (2026)

S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare," *arXiv (Cornell University)* , Feb. 2022, doi: 10.48550/arxiv.2110.15621.

J. Kim, J. Lee, E. Park, and J. Han, "A deep learning model for detecting mental illness from user content on social media," *Scientific Reports* , vol. 10, no. 1, July 2020, doi: 10.1038/s41598-020-68764-y.

A. A. K. Raj and A. Sharma, "Revealing Hidden Pain: A Comparative Analysis of Traditional Versus New Deep Learning Approaches for Detecting Depression on Social Media," *IEEE Access* , vol. 14, pp. 17942–17959, Jan. 2026, doi: 10.1109/access.2026.3658521.

K. Hasan, J. Saquer, and Y. Zhang, "Mental Multi-class Classification on Social Media: Benchmarking Transformer Architectures against LSTM Models," *arXiv (Cornell University)*

, Sept. 2025, doi: 10.48550/arxiv.2509.16542.

M. Kerasiotis, L. Ilias, and D. Askounis, "Depression detection in social media posts using transformer-based models and auxiliary features," *Social Network Analysis and Mining* , vol. 14, no. 1, Sept. 2024, doi: 10.1007/s13278-024-01360-4.

D. Mandal and H. Himanshu, "Harnessing Deep Learning for Mental Health: A Comparative Study of BERT, LSTM, and GPT-2," pp. 1–7, May 2025, doi: 10.1109/incet64471.2025.11140291.

M. M. Moby and S. Naveen, "BERT-LSTM Hybrid Model for Early Detection of Mental Health Risks," pp. 1–6, Aug. 2025, doi: 10.1109/nmitcon65824.2025.11187670.

Y. Ibrahimov, T. Anwar, and T. Yuan, "Explainable AI for Mental Disorder Detection via Social Media: A survey and outlook," *arXiv (Cornell University)* , June 2024, doi: 10.48550/arxiv.2406.05984.

S. Das, K. R. Khondakar, H. Mazumdar, A. Kaushik, and S. K. Singh, "Enhancing Machine Learning Models for Mental Health Classification Through Iterative Training and Text-Based Augmentation," *International Journal of Intelligent Systems* , vol. 2026, no. 1, Jan. 2026, doi: 10.1155/int/2620320.

D. Chawla, D. Chawla, A. Shrivastava, M. I. Habelalmateen, M. Dixit, and S. P. Dwivedi, "Explainable AI for Mental Health Diagnosis: Enhancing Transparency, Trust, and Clinical Decision-Making," pp. 1–6, Dec. 2025, doi: 10.1109/icaiihi67124.2025.11403514.

S. Bouktif, A. M. U. D. Khanday, and A. Ouni, "Explainable Predictive Model for Suicidal Ideation During COVID-19: Social Media Discourse Study," *Journal of Medical Internet Research* , vol. 27, Jan. 2025, doi: 10.2196/65434.

I. Rodríguez *et al.* , "Do AI Models Understand Mental HealthConversations? A Study of Subreddit Classification and Explainability," *Research Square (Research Square)* , Oct. 2025, doi: 10.21203/rs.3.rs-6392473/v2.

V. Viswan, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection," *Brain Informatics* , vol. 11, no. 1. Springer Science+Business Media, Apr. 05, 2024. doi: 10.1186/s40708-024-00222-1.

D. W. Joyce, A. Kormilitzin, K. Smith, and A. Cipriani, "Explainable artificial intelligence for mental health through transparency and interpretability for understandability," *npj Digital Medicine* , vol. 6, no. 1. Nature Portfolio, Jan. 18, 2023. doi: 10.1038/s41746-023-00751-9.

V. Reddy, S. C. Venkateshwarlu, and V. Nagaraju, "Mental Wellbeing Assessment Through Social Media and Machine Learning – A Web-Based Tool," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT* , vol. 9,

no. 6, pp. 1–8, June 2025, doi: 10.55041/ijsrem49404.

"Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion," Jan. 2022, doi: 10.18653/v1/2022.ltedi-1.

J. Angskun, S. Tipprasert, and T. Angskun, "Big data analytics on social networks for real-time

ISSN Online: 3007-3154
ISSN Print: 3007-3146

depression detection," *Journal Of Big Data* , vol. 9, no. 1, May 2022, doi: 10.1186/s40537-022-00622-2.

M. Garg, "WellXplain: Wellness concept extraction and classification in Reddit posts for mental health analysis," *Knowledge-Based Systems* , vol. 284, pp. 111228–111228, Dec. 2023, doi: 10.1016/j.knosys.2023.111228.

J. Chung and J. Teo, "Single classifier vs. ensemble machine learning approaches for mental health prediction," *Brain Informatics* , vol. 10, no. 1, Jan. 2023, doi: 10.1186/s40708-022-00180-6.