ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 3, March 2026

A Comprehensive Depression Identification Strategy for Medical Students Using Psychometric, Computational, and Clinical Methods

[1] Madina Amiri, [2] Sundresan Perumal, [3] Norita MD Norwawi, [4] Nurazah Ismail

[1] [2] [3] Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM), Bandar Baru Nilai, Malaysia
[4] Psychiatry Unit, Faculty of Medicine and Health Sciences, Universiti Sains Islam Malaysia (USIM), Bandar Baru Nilai, Malaysia
Corresponding Author Email: [1] madina.amiri1001@raudah.usim.edu.my, [2] sundresan.p@usim.edu.my, [3] norita@usim.edu.my,
[4] azah.ismail@usim.edu.my

Abstract*— Post-COVID-19 public health has witnessed a significant rise in depression, anxiety, suicidal ideation, and other mental health disorders. Although machine learning (ML) techniques and sentiment analysis have been widely applied to detect depressive symptoms from social media, major gaps remain in misclassification, limited clinical validation, and cultural bias. This study proposes a Human-in-the-Loop (HITL) integrative framework combining the Patient Health Questionnaire-9 (PHQ-9), sentiment analysis of social media text, and psychiatrist-led clinical interviews to enhance depression detection accuracy. Data were collected from 29 medical students at Universiti Sains Islam Malaysia (USIM), whose social media posts were analyzed. Statistical analyses including chi-square test, Fisher's exact test, multinomial logistic regression, agreement analysis, and confusion matrix were applied. Results show significant disagreement between automated sentiment detection and clinical interviews ($\chi^2 = 33.908$, $p = 0.00001$), demonstrating that AI-based detection alone is insufficient. The findings emphasize the necessity of clinical validation and HITL integration for reliable depression detection.

.

Keywords*— Clinical Interviews, Depression Detection, Human-in-the-Loop, Machine Learning, PHQ-9, Sentiment Analysis.

I. INTRODUCTION

Depression is a major global mental health concern affecting over 280 million individuals worldwide (Duffy et al., 2023). The COVID-19 pandemic further intensified psychological distress, leading to increased suicide ideation and depressive symptoms (Movahed et al., 2025). Medical students represent a high-risk population due to academic pressure, emotional stress, and professional expectations (Olum et al., 2020).

With the expansion of social media platforms such as Facebook, Instagram, and Threads, indivals increasingly express emotional distress online. Researchers have leveraged machine learning and Natural Language Processing (NLP) to detect depressive indicators from user-generated content (Kolliakou et al., 2020). However, studies frequently report misclassification and lack of clinical validation (Milintsevich et al., 2024).

This study addresses these limitations by integrating:

- PHQ-9 standardized screening
- Sentiment analysis of social media posts
- Psychiatrist-led clinical interviews
- Human-in-the-Loop (HITL) validation

II. LITERATURE REVIEW

A. Machine Learning in Depression Detection

Supervised learning algorithms such as Support Vector Machines (SVM), Random Forest (RF), and Neural Networks (NN) have been widely applied for depression classification (Radhoush et al., 2023). Yang et al. (2020) and Shatte et al. (2020) demonstrated promising results using SVM on Facebook and Reddit datasets.

However, Kim et al. (2022) reported only 68.3% accuracy using PHQ-9 guided NLP models, indicating significant misclassification risks.

Unsupervised methods like K-means clustering and Affinity Propagation have also been explored (Joshi & Patwardhan, 2020; Aragon et al., 2023). Yet, these models struggle with noisy social media data and cultural expression variability.

B. Clinical Assessment and PHQ-9

PHQ-9 is a validated depression screening instrument with severity cutoffs:

- 0–4: Minimal
- 5–9: Mild
- 10–14: Moderate
- 15–19: Moderately Severe
- 20–27: Severe

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 3, March 2026

Clinical interviews remain the gold standard for diagnosis (Baydili et al., 2025a). However, most AI studies lack psychiatric verification.

C. Human-in-the-Loop (HITL)

HITL integrates human expertise into ML decision-making to reduce misclassification and improve interpretability (Baydili et al., 2025b). Previous studies recommend clinical validation but rarely implement it.

This study operationalizes HITL by incorporating psychiatrist assessment alongside AI analysis.

III. METHODOLOGY

This study employed a quantitative cross-sectional design to evaluate the integration of PHQ-9, sentiment analysis, and clinical interviews for depression detection among medical students. The research framework was based on a Human-in-the-Loop (HITL) approach, where automated sentiment analysis was validated through standardized psychometric assessment and psychiatrist evaluation. The purpose was to reduce misclassification and improve diagnostic reliability.

The study involved 29 medical students from Universiti Sains Islam Malaysia (USIM). Participation was voluntary, and informed consent was obtained prior to data collection. Ethical approval was secured to ensure confidentiality and compliance with psychiatric research standards. All participants were active users of Facebook, Instagram, and Threads, and only English-language posts were analyzed to maintain linguistic consistency.

Data collection was conducted in three stages. First, participants completed the Patient Health Questionnaire-9 (PHQ-9), which provided standardized depression severity scores. Based on PHQ-9 results, 20 participants were categorized as mildly depressed, 6 as moderately depressed, and 3 as severely depressed.

Second, selected social media textual posts were collected from participants. The text data were preprocessed using Natural Language Processing techniques, including cleaning, tokenization, and normalization. Sentiment analysis was then applied to classify posts into positive, negative, and neutral sentiment categories.

Third, each participant underwent a clinical interview conducted by a psychiatrist. The clinical assessment served as the gold standard for comparison. This step ensured that automated predictions were evaluated against expert psychiatric judgment, operationalizing the HITL framework.

For statistical analysis, descriptive statistics were used to summarize PHQ-9 scores, sentiment results, and clinical classifications. Chi-square tests and Fisher's Exact Test were applied to examine associations among the three methods. A confusion matrix was constructed to identify misclassification patterns. Additionally, multinomial logistic regression was performed to assess whether sentiment scores significantly predicted PHQ-9 depression categories.

The results showed significant deviation from expected agreement ($\chi^2 = 33.908$, $p = 0.00001$), indicating that sentiment analysis alone was insufficient for reliable depression detection. This finding supports the necessity of integrating clinical validation within AI-based systems.

IV. RESULTS

Descriptive analysis showed that most participants were classified as mildly depressed according to PHQ-9 (n = 20), followed by moderate (n = 6) and severe (n = 3) categories. Sentiment analysis of social media posts produced varying proportions of positive, negative, and neutral sentiment scores across participants. However, the distribution of sentiment polarity did not consistently align with PHQ-9 severity levels.

Clinical interviews conducted by the psychiatrist classified the majority of participants as either mild or non-depressed, with fewer cases identified as moderate or severe. When comparing PHQ-9 results with clinical interviews, discrepancies were observed in several cases, particularly within moderate and severe categories.

The confusion matrix revealed misclassification patterns between sentiment-based predictions and clinical diagnosis. Several participants categorized as moderate or severe by PHQ-9 did not exhibit strongly negative sentiment patterns in their textual posts. Conversely, some participants with higher negative sentiment percentages were not clinically diagnosed with severe depression.

Chi-square analysis indicated a strong deviation from expected agreement among PHQ-9, sentiment analysis, and clinical interviews ($\chi^2 = 33.908$, $p = 0.00001$). Fisher's Exact Test further supported the presence of significant differences between sentiment polarity and PHQ-9 categories. Multinomial logistic regression analysis demonstrated that sentiment scores were not strong standalone predictors of depression severity. Overall, the findings highlight inconsistency between automated text-based detection and psychiatric evaluation.

V. DISCUSSION

The findings demonstrate that sentiment analysis alone is insufficient for accurate depression detection. Although negative sentiment may reflect emotional distress, it does not necessarily correspond to clinical depression severity. This may be due to cultural expression differences, emotional masking, indirect language use, or sarcasm in social media posts.

The observed misclassification patterns align with previous studies reporting annotation and prediction limitations in AI-based mental health detection. Automated models often struggle to capture contextual meaning, behavioral cues, and psychological nuances that trained clinicians identify during interviews. Therefore, relying solely on sentiment polarity can lead to both false positives and false negatives.

The integration of PHQ-9 screening and clinical interviews within a Human-in-the-Loop framework improved interpretability and reliability. The psychiatrist's evaluation provided contextual understanding beyond textual analysis, reducing over-reliance on algorithmic outputs. This supports the argument that AI systems should function as supportive screening tools rather than standalone diagnostic mechanisms.

Furthermore, the small but clinically validated dataset in this study emphasizes quality over quantity. While large datasets may improve computational performance, they do not guarantee diagnostic accuracy without expert validation. Thus, clinical involvement remains essential in AI-assisted mental health assessment.

VI. CONCLUSION

This study proposed and evaluated an integrative framework combining PHQ-9, sentiment analysis, and clinical interviews for depression detection among medical students. The findings revealed significant disagreement between automated sentiment analysis and psychiatric diagnosis, confirming that AI-based models cannot independently provide reliable depression detection.

The statistical results, particularly the significant chi-square value ($\chi^2 = 33.908$, $p = 0.00001$), demonstrate the necessity of clinical validation in mental health AI systems. Sentiment analysis may serve as an early screening indicator, but it should not replace standardized assessment tools or psychiatric expertise.

The Human-in-the-Loop approach proved valuable in enhancing reliability and reducing misclassification. Future research should expand the sample size, incorporate multimodal data such as behavioral and speech features, and develop culturally adaptive models. Ultimately, effective depression detection systems must balance technological advancement with clinical responsibility.

REFERENCES

[1] Adarsh, V., Arun Kumar, P., Lavanya, V., & Gangadharan, G. R. (2023). Fair and Explainable Depression Detection in Social Media. Information Processing and Management, 60(1). https://doi.org/10.1016/j.ipm.2022.103168Al-Amer, R., et al. (2025). Depression among medical students in the Arab region: A systematic review.

[2] Alyoubi, R., Alaslani, W., Alabdullatif, N., Albukhari, A., Sindi, S., Al Kaibi, Z., Alshahrani, F., Al Mubarak, H., & Alzahrani, K. T. (2024). CAHIERS MAGELLANES-NS 5982 PREVALENCE OF DEPRESSION AMONG MEDICAL STUDENTS IN SAUDI ARABIA. https://doi.org/10.6084/m9.figshare.2632599 Aragon, M., et al. (2023). Emotion clustering for depression detection.

[3] Baydili, İ., Tasci, B., & Tasci, G. (2025a). Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection. Behavioral Sciences, 15(3). https://doi.org/10.3390/bs150303 52 73

[4] Baydili, İ., Tasci, B., & Tasci, G. (2025b). Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection. Behavioral Sciences, 15(3). https://doi.org/10.3390/bs15030 352

[5] Duffy, D., Richards, D., Hisler, G., & Timulak, L. (2023). Implementing internet delivered cognitive behavioural therapy for depression and anxiety in adults: A systematic review (Preprint). Journal https://doi.org/10.2196/47927 of Medical Internet Research.

[6] Joshi, D., & Patwardhan, D. M. (2020). An analysis of mental health of social media users using unsupervised approach. Computers in Human Behavior Reports, 2(October), 100036. https://doi.org/10.1016/j.chbr.2020.100036

[7] Kim, N. H., Kim, J. M., Park, D. M., Ji, S. R., & Kim, J. W. (2022). Analysis of depression in social media texts through the Patient Health Questionnaire-9 and natural language processing. Digital Health, 8. https://doi.org/10.1177/2055 2076221114204

[8] Milintsevich, K., Sirts, K., & Dias, G. (2024). Your Model Is Not Predicting Depression Well And That Is Why: A Case Study of PRIMATE Dataset. http://arxiv.org/abs/2403.00438

[9] Movahed, M., et al. (2025). Suicide risk in post-COVID mental health crisis.

[10] Olum, R., Nakwagala, F. N., & Odokonyero, R. (2020). Prevalence and Factors Associated with Depression among Medical Students at Makerere University ,.

[11] Radhoush, S., Whitaker, B. M., & Nehrir, H. (2023). An Overview of Supervised Machine Learning Approaches for Applications in Active Distribution Networks. Energies, 16(16). https://doi.org/10.3390/en16165972

[12] Shatte, A. B. R., Hutchinson, D. M., Fuller-tyszkiewicz, M., & Teague, S. J. (2020). Social Media Markers to Identify Fathers at Risk of Postpartum Depression : A Machine Learning Approach. 00(00), 1–8. https://doi.org/10.1089/cyber.2019.07 46

[13] Yang, Y., Yang, Y., Pan, A., Xu, Z., Wang, L., & Zhang, Y. (2022). Identifying Depression in Parkinson ' s Disease by Using Combined Diffusion Tensor Imaging and Support Vector Machine. 13(June), 1–8. https://doi.org/10.3389/ fneur.2022.878691

[14] Zhao, X., Lyu, Y., Wang, D., & Tang, B. (2025). Predicting Depression in Screening Interviews from Interactive http:// arxiv.org/abs/2502.12204 Multi-Theme Collaboration.