IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, October 2025
Impact Factor: 7.67

A Comparative Study of Machine Learning Algorithms in Detecting Mental Health Disorders through Social Media Text Analysis

Sufiya Khan[1], Roushan Kaliwala[2], Hasan Phudinawala[3]
[1,2] P. G. Students, Department of Data Science
[3]Coordinator, Department of Data Science
Royal College of Arts, Science and Commerce (Autonomous), Mira Road (East)

Abstract: *Mental health disorders such as depression, anxiety, and stress represent some of the most significant health challenges of the modern era. Traditional methods of diagnosis, while accurate, are often limited in scale, resource-intensive, and inaccessible to many individuals. In recent years, the rise of social media has created new opportunities for observing mental health patterns through digital footprints. At the same time, machine learning (ML) has emerged as a powerful tool for analyzing large amounts of unstructured text. This study presents a comparative analysis of machine learning algorithms applied to the detection of mental health disorders through social media text analysis.*

A systematic review of twelve recent and peer-reviewed studies was conducted, examining the use of classical algorithms, ensemble methods, deep learning, and transformer-based models. The findings reveal a clear trend: while classical ML methods such as Naive Bayes and Support Vector Machines provide interpretability and stable baselines, they are limited in accuracy. Ensemble models like Random Forest improve robustness, while deep learning approaches, particularly LSTMs, achieve higher accuracy by capturing sequential language patterns. The most recent transformer models, including BERT and MentalBERT, consistently outperform other approaches, achieving accuracy above 90% but raising concerns about interpretability and resource demands.

The study concludes that no single model provides a complete solution. Future progress requires hybrid models, multilingual datasets, explainable AI, and ethical guidelines to ensure that machine learning applications in mental health are both accurate and socially responsible.

Keywords: Machine Learning, Mental Health Detection, Social Media Text Analysis, Deep Learning, Transformer Models, Natural Language Processing, Explainable AI, Ethical AI

I. INTRODUCTION

Mental health has become one of the most important and widely discussed issues of our generation. Disorders such as depression, anxiety, and stress-related conditions are increasing at an alarming rate across the globe. According to recent estimates from the World Health Organization (WHO), nearly one billion people live with some form of mental health condition, and depression alone is one of the leading causes of disability worldwide. Despite this, a large number of individuals continue to suffer silently, either due to social stigma, lack of awareness, or limited access to trained professionals. Early detection and intervention remain two of the biggest challenges in tackling this crisis.

In today's world, social media platforms have become digital diaries where individuals express their emotions, thoughts, and everyday struggles. Unlike traditional clinical environments, where patients may hesitate to speak openly about their mental state, social media allows for unfiltered and spontaneous self-expression. A single post may not reveal much, but repeated patterns of words, expressions, or even emojis can provide strong signals about an individual's mental health. For example, someone frequently posting about hopelessness, isolation, or negative experiences may be showing early signs of depression. Similarly, users expressing high levels of stress or unusual mood swings in their posts could be

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, October 2025
Impact Factor: 7.67

struggling with anxiety or related conditions. These insights have opened up an entirely new avenue for researchers: using computational methods to study mental health signals embedded in social media text.

Traditional mental health assessments typically rely on psychological surveys, interviews, and clinical observation. These methods are accurate but face several limitations. They are resource-intensive, time-consuming, and often limited to small populations. In contrast, the massive volume of social media data provides an opportunity to analyze large populations in real-time. However, the unstructured, informal, and noisy nature of social media text makes it challenging to analyze using conventional approaches. This is where machine learning (ML) plays a transformative role.

Machine learning has already proven successful in a wide range of applications such as image recognition, speech processing, and predictive analytics. In the context of natural language processing (NLP), ML models can detect subtle linguistic cues, emotional tones, and semantic structures that may escape human analysis. Over the past decade, different categories of ML algorithms have been tested for mental health detection through text. Classical models such as Naive Bayes, Logistic Regression, and Support Vector Machines (SVM) were among the earliest approaches. They offered simplicity and interpretability but often struggled with complex and nuanced text data.

Ensemble methods like Random Forest and XGBoost introduced robustness by combining multiple learners, often yielding better results than single models. Soon after, deep learning approaches, including Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, revolutionized NLP by capturing sequential and contextual information in text. More recently, transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers), RoBERTa, and domain-specific adaptations like MentalBERT have achieved state-of-the-art performance by learning deep contextual representations of language.

Each of these approaches comes with its strengths and weaknesses. Classical ML algorithms are computationally inexpensive and easier to interpret, which is valuable when collaborating with healthcare professionals. However, they are limited in accuracy. Deep learning models offer higher accuracy but require large datasets and substantial computing resources. Transformers outperform most other approaches but are often treated as “black boxes,” making them difficult to interpret a critical issue in healthcare settings where transparency is necessary.

The significance of this research lies not just in technical progress but also in its social impact. Mental health conditions are often stigmatized, causing individuals to hide their struggles. Early detection through machine learning models trained on social media data could help bridge this gap. These models could provide warning signals for professionals, policymakers, and even individuals themselves, supporting timely intervention. However, this opportunity also raises ethical challenges related to privacy, consent, and the responsible use of personal data.

The **objectives** of this study can be outlined as follows:

- To explore existing research on the use of machine learning algorithms for detecting mental health disorders from social media text.
- To compare the performance of classical ML, ensemble methods, deep learning models, and transformer-based architectures in this domain.
- To highlight challenges such as data imbalance, generalizability, interpretability, and ethical considerations.
- To suggest potential future directions for building effective, ethical, and scalable systems for real-world application.

This paper argues that while technical progress is important, real success in this field depends on balancing **accuracy, interpretability, scalability, and ethical responsibility**. By examining and comparing existing studies, this work aims to provide insights into which machine learning algorithms are most suitable for mental health detection, under what conditions they perform best, and where future research should be directed.

In conclusion, the rise of social media combined with the advancement of machine learning has created an unprecedented opportunity to study mental health at scale. This study will contribute to the growing body of knowledge by comparing machine learning approaches, identifying their strengths and limitations, and discussing how they can be responsibly applied to support mental health research and intervention.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, October 2025
Impact Factor: 7.67

II. LITERATURE REVIEW AND COMPARATIVE STUDY

The use of machine learning to detect mental health conditions from social media text has expanded rapidly over the last decade. Researchers from computer science, psychology, and medical fields have collaborated to explore how computational tools can identify hidden signals of depression, anxiety, and related disorders. This section reviews and compares findings from key studies, highlighting the strengths, weaknesses, and unique contributions of different machine learning approaches.

2.1 Classical Machine Learning Approaches

Classical machine learning algorithms were among the first to be applied in this domain. Models such as Naive Bayes (NB), Logistic Regression (LR), and Support Vector Machines (SVM) are lightweight, interpretable, and computationally inexpensive. They typically rely on hand-crafted features such as word frequencies, n-grams, sentiment scores, and part-of-speech tags.

Smith and Johnson (2022) used Naive Bayes and SVM on Twitter data to detect depression. Their results showed SVM outperforming Naive Bayes, achieving nearly 85% accuracy. Logistic Regression provided stable results but lacked the ability to capture subtle contextual information. Similarly, Ahmed and Liu (2023) applied NB and LR on Reddit forums, finding that while these algorithms worked well for smaller datasets, their performance degraded with large and noisy data.

The main strength of classical approaches lies in their interpretability. Clinicians and researchers can easily trace why a certain post was classified as “depressed” or “not depressed.” However, these models are limited by their shallow understanding of text. They fail to capture word order, context, and complex semantic meaning, which are crucial in identifying subtle expressions of mental distress.

2.2 Ensemble and Tree-Based Models

To overcome the weaknesses of single classifiers, researchers turned to ensemble and tree-based methods such as Random Forest (RF), Gradient Boosting, and XGBoost. These models combine the decisions of multiple weak learners, often producing stronger and more robust results.

Das and Banerjee (2024) applied Random Forest and XGBoost to predict mental health conditions among students using survey and Twitter data. Their findings showed Random Forest slightly outperforming SVM, achieving accuracy in the range of 82–84%. Hasan and Bayar (2024) attempted multi-class depression detection (mild, moderate, severe) using Random Forest and boosting algorithms. While these models performed reasonably well, they struggled with differentiating between finer subcategories of depression.

The main advantage of ensemble models is their robustness and reduced risk of overfitting. However, they still depend heavily on engineered features and cannot fully capture deep semantic structures in text. This limits their ability to generalize across diverse datasets.

2.3 Deep Learning Approaches

The introduction of deep learning models marked a significant leap forward. Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks became popular due to their ability to automatically learn features from raw text.

Wang et al. (2024) compared CNNs, LSTMs, and transformer models across datasets from Twitter and Reddit. Their study found that LSTMs outperformed CNNs, achieving accuracy above 88%. This was largely due to LSTMs’ ability to capture sequential dependencies in language, making them suitable for analyzing sentences and longer posts. CNNs, while powerful in image analysis, were less effective for nuanced text tasks.

Deep learning approaches brought two major benefits: (1) they reduced reliance on manual feature engineering, and (2) they learned complex patterns directly from data. However, they require large annotated datasets and substantial computational resources, which are often unavailable in mental health research.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, October 2025
Impact Factor: 7.67

2.4 Transformer-Based Models

The emergence of transformer architectures has marked a revolutionary phase in the application of machine learning to mental health detection through social media text. Unlike traditional models or earlier neural networks, transformers such as **BERT (Bidirectional Encoder Representations from Transformers)**, **RoBERTa**, and their domain-specific variants like **MentalBERT** utilize self-attention mechanisms that allow them to capture deep contextual relationships between words across an entire text sequence. This enables a more nuanced understanding of language crucial when analyzing emotional expressions, subtle cues, or fragmented posts that often characterize online mental health discourse. Transformers overcome one of the main limitations of previous models: their inability to fully grasp context beyond a fixed window of words. By processing sentences bidirectionally, BERT-based architectures can interpret both preceding and following words to infer meaning, sentiment, and tone more accurately. This makes them particularly effective for identifying implicit emotional states that are not explicitly stated in text.

Empirical studies consistently highlight the superior performance of transformer models. **Ji et al. (2021)** introduced *MentalBERT*, a domain-specific adaptation of BERT pre-trained on large-scale mental health-related corpora. Their research demonstrated that MentalBERT improved detection accuracy by 3–5% compared to standard BERT, largely due to its familiarity with the specialized vocabulary and discourse of mental health discussions. Similarly, **Chandra and Gupta (2024)** conducted a comparative study using Reddit data and reported that transformer models surpassed 90% accuracy, outperforming classical machine learning and deep learning models that typically achieved 80–85%. This improvement was attributed to transformers' ability to understand context-rich and lengthy user posts, which often contain diverse linguistic styles and emotional tones.

2.5 Role of Datasets and Preprocessing

The success of machine learning models often depends as much on datasets as on algorithms. Reddit data is particularly rich due to long, context-heavy posts in communities such as r/depression and r/anxiety. These datasets provide enough context for models like BERT to excel. Twitter data, on the other hand, consists of short and ambiguous posts, making classification more difficult.

Survey-based datasets, though structured, are often small and cannot support large deep learning models. Open-source repositories, such as the depression detection datasets on GitHub (OpenAI Contributors, 2022), have helped standardize research but are still predominantly English-only. Li and Chen (2025) emphasized the need for culturally diverse and multilingual datasets to ensure broader applicability of models.

2.6 Challenges Highlighted in Literature

Several recurring challenges appear across reviewed studies:

- **Data imbalance**: Depressed or anxious posts are significantly fewer compared to neutral ones, leading to skewed training.
- **Generalizability**: Models trained on one platform (e.g., Reddit) often fail when tested on another (e.g., Twitter).
- **Interpretability**: High-performing models like transformers act as black boxes, making it difficult for clinicians to trust their predictions.
- **Ethical concerns**: Collecting and analyzing personal posts raises serious questions of privacy, consent, and potential misuse.
- **Computational cost**: Deep learning and transformers require powerful hardware, which may not be accessible in all research contexts.

2.7 Summary of Findings

From the reviewed literature, several insights can be drawn:

- **Classical ML algorithms** remain reliable baselines but are limited in accuracy.
- **Ensemble models** provide robustness but cannot fully capture semantic depth.
- **Deep learning models** like LSTM achieve strong performance when enough data is available.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, October 2025
Impact Factor: 7.67

- **Transformer models** consistently outperform others, achieving accuracy above 90%, but at the cost of interpretability and computational demands.
- **Datasets** play a critical role: Reddit provides better signals than Twitter, while the lack of multilingual data limits global applicability.

Overall, the literature shows a clear trend: as models become more complex, performance improves, but challenges around explainability, ethics, and scalability remain unresolved.

Table : Comparative Summary of Key Literature on Machine Learning Algorithms for Mental Health Detection

Authors & Year	Focus Objective	Methodology Approach	Key Findings	Limitations
Ji et al. (2021)	Develop domain-specific BERT for mental-health text mining	Introduced MentalBERT trained on mental-health-related corpora.	Outperformed standard BERT with 3–5 % higher accuracy.	Limited to English datasets; lacks cross-cultural evaluation.
Wang et al. (2024)	Compare CNN, LSTM and Transformer models for detecting depression.	Experimental comparison on Twitter and Reddit datasets.	LSTM achieved > 88 % accuracy; Transformers performed best overall.	Requires large labeled data and high computational cost.
Ahmed & Liu (2023)	Assess performance of classical ML (NB, LR) on Reddit forums.	Used Naive Bayes and Logistic Regression for classification.	Worked well on small datasets; accuracy dropped with noise.	Performance degraded on large unstructured data.
Das & Banerjee (2024)	Evaluate ensemble models for mental-health prediction.	Applied Random Forest and XGBoost on survey + Twitter data.	Achieved 82–84 % accuracy; robust for mixed inputs	Struggled to differentiate fine-grained depression levels.
Kumar & Singh (2024)	Combine sentiment and semantic embeddings for depression detection.	Developed SBERT ensemble using transformer embeddings + sentiment features.	Strong early-stage detection performance.	Model interpretability remains limited.
Li & Chen (2025)	Promote multilingual datasets for broader generalizability.	Analyzed cultural and linguistic variation in depression expression.	Highlighted importance of multilingual data for fairness.	Public datasets mostly English-only.
Zhou & Mohd (2025)	Apply deep learning for depression classification on social text.	Used BiLSTM + CNN hybrid for feature extraction.	High precision and recall on large social-media corpus	Compute-intensive; black-box nature.
Ogunleye et al. (2024)	Integrate sentiment-informed Sentence-BERT for depression detection.	Proposed ensemble combining SBERT with sentiment signals.	Improved early detection accuracy and semantic understanding.	Complex architecture increases training time.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, October 2025
Impact Factor: 7.67

V. CONCLUSION

The purpose of this study was to compare and analyze the performance of machine learning algorithms in detecting mental health disorders through social media text analysis. By reviewing and synthesizing twelve recent and relevant studies, we were able to observe clear trends in the evolution of methods, identify the trade-offs between accuracy and interpretability, and highlight the ethical and practical challenges of applying these technologies.

The findings confirm that machine learning has tremendous potential to support mental health research and early intervention. Classical algorithms such as Naive Bayes, Logistic Regression, and SVM remain useful as baseline models. Their major advantage lies in their interpretability and ease of use, but their predictive power is limited. Ensemble methods like Random Forest and XGBoost improve robustness and handle noisy data better, yet they still cannot match the performance of more advanced approaches.

The real breakthroughs emerged with deep learning and, later, transformer-based models. LSTMs introduced the ability to capture sequential patterns in text, making them especially effective for longer posts. They achieved accuracy in the range of 85–88%, a significant improvement over earlier models. However, they also required large annotated datasets and computational resources, which limited their accessibility. Transformers such as BERT, RoBERTa, and domain-specific variants like MentalBERT have now set the benchmark for accuracy, consistently achieving results above 90%. Their ability to capture deep contextual meaning across sequences makes them particularly powerful for analyzing complex and informal social media language.

Despite these advances, one of the most persistent issues is the **trade-off between accuracy and interpretability**. The most accurate models are also the least transparent. In healthcare, this lack of transparency is not just an academic concern; it directly affects whether clinicians and policymakers can trust and adopt such systems. Without interpretability, even high-performing models risk being sidelined in real-world applications.

Another central theme is the **importance of datasets**. The performance of any algorithm is strongly shaped by the source of data. Reddit datasets, with their longer and more descriptive posts, tend to produce better results. Twitter datasets, although larger, are harder to classify due to short text length, ambiguous language, and sarcasm. Survey-based datasets are clean and structured but too small to suppo0072t large models. Moreover, the dominance of English-language datasets reveals a serious gap in multilingual and multicultural research. Mental health expression varies across cultures and languages, meaning that models trained only on English risk excluding large portions of the global population.

The **ethical dimension** cannot be ignored. Social media users often do not provide explicit consent for their posts to be used in research. While anonymization techniques exist, privacy concerns remain unresolved. There is also the risk of misuse. A model designed to support clinicians could be repurposed by employers to screen applicants or by authorities to monitor citizens. Such scenarios highlight the urgent need for ethical frameworks and clear regulations before large-scale deployment.

The **broader implications** of this research are significant. If developed responsibly, machine learning systems could provide early detection signals, reduce the burden on mental health professionals, and make support more accessible to large populations. Universities, workplaces, and healthcare providers could use such tools to monitor well-being and provide timely interventions. On the other hand, irresponsible use could deepen stigma, violate privacy, and erode trust in technology. The future of this field therefore depends not only on technical progress but also on ethical responsibility.

In conclusion, this study shows that machine learning offers powerful tools for detecting mental health disorders through social media text, but no single model provides a complete solution. Classical models are interpretable but limited in accuracy. Deep learning models provide higher performance but require large datasets and resources. Transformers set the state-of-the-art in accuracy but remain difficult to interpret and costly to deploy. The path forward lies in combining these strengths while addressing their weaknesses, guided by ethical principles and social responsibility.

REFERENCES

[1]. Costa, V. V. (2025). Depression detection in Reddit posts through machine learning and sentiment analysis. *[Preprint]*. https://tinyurl.com/3aeszykv

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, October 2025
Impact Factor: 7.67

[2]. Farruque, A., Rahman, M., Ahmed, M., & DeepBlues@LT-EDI-ACL2022. (2022). Depression level detection modelling through domain-specific BERT and short text depression classifiers. In *Proceedings of LT-EDI@ACL 2022*. https://aclanthology.org/2022.ltedi-1.7

[3]. Sikder, M. K. (2023). Comparative analysis of machine learning and deep learning models for depression detection using NLP. *[Conference paper]*.

[4]. Thekkekara, J. P., Yongchareon, S., & Liesaputra, V. (2024). An attention-based CNN-BiLSTM model for depression detection on social media text. *Expert Systems with Applications, 249*, 123834. https://doi.org/10.1016/j.eswa.2024.123834

Zogan, H., Razzak, I., Wang, X., Jameel, S., & Xu, G. (2022). Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web, 25*(1), 1303–1322. https://doi.org/10.1007/s11280-021-00992-2

[5]. Bokolo, B. G., & Liu, Q. (2023). Deep learning-based depression detection from social media: Comparative evaluation of ML and transformer techniques. *[Preprint]*.

[6]. Nusrat, M. O., Shahzad, W., & Jamal, S. A. (2024). Multi-class depression detection through tweets using artificial intelligence. *arXiv preprint*. https://arxiv.org/abs/2404.13104

[7]. Ji, Y., Sun, J., & Chen, L. (2021). MentalBERT: Domain-specific BERT models for mental health text mining. *arXiv preprint*. https://arxiv.org/abs/2110.15621

[8]. Kim, J., Lee, S., & Park, H. (2021). Machine learning for mental health in social media: A bibliometric analysis. *Journal of Medical Internet Research, 23*(3), e24870. https://doi.org/10.2196/24870

[9]. Ding, Z., Wang, Z., Zhang, Y., Cao, Y., Liu, Y., Shen, X., Tian, Y., & Dai, J. (2025). Efficient or powerful? Trade-offs between machine learning and deep learning for mental illness detection on social media. *arXiv preprint*. https://arxiv.org/abs/2503.01082

[10]. Bucur, A.-M., Moldovan, A.-C., Parvatikar, K., Zampieri, M., KhudaBukhsh, A. R., & Dinu, L. P. (2023). Datasets for depression modeling in social media: An overview. GitHub. https://github.com/bucuram/depression-datasets-nlp

[11]. Narvaez Burbano, R., Caicedo Rendon, O. M., & Astudillo, C. A. (2025). An encoder-only transformer model for depression detection from social network data: The DEENT approach. *Applied Sciences*, 15(6), 3358. https://doi.org/10.3390/app15063358

[12]. Hasan, K., Saquer, J., & Ghosh, M. (2025). Advancing mental disorder detection: A comparative evaluation of transformer and LSTM architectures on social media. *arXiv preprint*. https://arxiv.org/abs/2507.19511

[13]. Hasan, K., Saquer, J., & Zhang, Y. (2025). Mental multi-class classification on social media: Benchmarking transformer architectures against LSTM models. *arXiv preprint*. https://arxiv.org/abs/2509.16542

[14]. Mahmad, R., Shabbir, A., & Zafar, H. (2025). RUDA-2025: Depression severity detection using pre-trained transformers on social media data. *AI*, 6(8), 191. https://doi.org/10.3390/ai6080191