A Comparative Evaluation of Machine Learning and Deep Learning Models for Detecting Depressive behavior on X

Ainun Najwa Mohd Pauzi, Abdul Syukor Mohamad Jaya*

Fakulti Kecerdasan Buatan dan Keselamatan Siber, Universiti Teknikal Malaysia Melaka Hang Tuah Jaya, 76100 Durian Tunggal, Melaka

***Corresponding Author**

DOI: https://doi.org/10.47772/IJRISS.2026.100900616

Received: 24 September 2026; Accepted: 01 October 2026; Published: 17 October 2026

ABSTRACT

Early detection of mental health concerns through social media platforms like X offers significant potential for proactive intervention. However, identifying nuanced emotional and psychological distress from short, informal texts remains a complex challenge. This study aims to bridge this gap by developing and comprehensively evaluating a spectrum of machine learning, deep learning, and transformer-based models for detecting depressive behavior from X posts. A robust natural language processing (NLP) pipeline was established, encompassing data cleaning, text preprocessing, feature extraction, and classification. Four distinct architectures were investigated: Logistic Regression, Linear Support Vector Machine (SVM), CNN-LSTM, and Fine-Tuned Distil BERT. Traditional models utilized TF-IDF representations, while deep learning and transformer models employed token-based representations. Model performance was rigorously benchmarked using accuracy, precision, recall, and F1-score. Experimental results reveal that all models achieved high efficacy with accuracies exceeding 96%. Notably, the Fine-Tuned Distil BERT secured the best overall performance with an accuracy of 97.46% and an F1-score of 96.68%, whereas Linear SVM demonstrated exceptional sensitivity with the highest recall of 97.05%. These findings demonstrate that transformer-based models and advanced NLP pipelines significantly enhance the reliability of automated mental health screening on social media, providing a scalable foundation for intelligent, text-based psychological monitoring systems.

Keywords: Depression Detection, Natural Language Processing (NLP), Transformer Models, Machine Learning, Social Media Analytics

INTRODUCTION

Depression represents one of the most pervasive mental health conditions globally, severely impairing individuals' psychological well-being, interpersonal relationships, and daily functioning. Traditional psychological evaluations rely heavily on self-reported questionnaires and clinical interviews, which are often limited by low reporting rates, social stigma, and late-stage diagnosis. The widespread adoption of social media platforms particularly X (formerly Twitter) has introduced unprecedented opportunities for unobtrusive, passive mental health screening. Millions of users daily share unfiltered thoughts, personal struggles, and emotional states, generating rich longitudinal text streams that reflect underlying psychological states.

Despite the promise of social media analytics, identifying depressive signals in microblogging platforms presents formidable computational hurdles. Microblog posts are brief, informal, and heavily laden with slang, abbreviations, typing errors, and contextual ambiguity. Furthermore, standard automated text processing pipelines frequently struggle with negation structures (e.g., 'not happy', 'no motivation', 'never fine'). Typical stopword removal routines risk eliminating negation words, fundamentally altering the emotional polarity of the text and triggering high misclassification rates.

To overcome these challenges, this study formulates a rigorous, comparative machine learning framework for text-based depression detection. The primary objective of this research is to establish a specialized NLP pre-

processing pipeline that eliminates social media noise while explicitly preserving critical negation terms and sentiment indicators. Besides, this study develops, trains, and fine-tunes four distinct modeling paradigms: Logistic Regression, Linear Support Vector Machine (SVM), a hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) network, and a fine-tuned DistilBERT transformer model. The scope of this investigation centers on binary text classification (depressed vs. non-depressed) applied to an extensive dataset of 174,466 preprocessed X posts, establishing empirical benchmarks across traditional statistical models, deep recurrent architectures, and state-of-the-art pretrained transformers.

LITERATURE WORK

Public Health Context and Prevalence of Youth Mental Health Disorders

Depression represents a severe global public health concern, yet a significant proportion of cases remain undiagnosed due to a lack of awareness, societal stigma, and limited access to professional mental health services. As highlighted in [1], mental health disorders are increasingly prevalent among youth. Left untreated, depressive conditions can precipitate severe long-term consequences, including social withdrawal, risky behavior, chronic physical health deterioration, and suicidal ideation. These diagnostic barriers underscore the critical need for scalable, non-invasive early detection mechanisms that can identify individuals in need of intervention.

According to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), clinical depression is characterized by persistent low mood, loss of interest in daily activities, reduced energy, and impaired cognitive concentration lasting for at least two consecutive weeks. Unlike temporary emotional fluctuations, untreated depressive states severely disrupt an individual's long-term social, academic, and occupational functioning. Public health reports indicate escalating prevalence rates of depression across various age demographics, particularly among youth and adolescents. Despite this growing burden, traditional diagnostic approaches rely heavily on self-reported questionnaires and voluntary clinical evaluations, which frequently fail to reach affected individuals due to deep-seated social stigma, lack of mental health literacy, and systemic healthcare inaccessibilities.

To bridge this critical diagnostic gap, health informatics researchers have increasingly turned to social media analytics as a non-invasive, continuous monitoring alternative. Digital platforms such as X (formerly Twitter) host billions of global users who regularly post real-time updates regarding their daily moods, life stressors, and emotional struggles. Due to the perceived online anonymity and informal nature of microblogging, individuals often demonstrate a higher willingness to express vulnerability and psychological distress on social media than in formal face-to-face clinical consultations. These spontaneous disclosures function as a digital diary, embedding key linguistic markers, such as elevated negative sentiment, expressions of loneliness, and self-referential language that allow natural language processing systems to detect subtle behavioral indicators of depression at a population scale.

Theoretical Foundations of Artificial Intelligence in Text Analysis

To bridge the gap in traditional clinical screening, computational methods leverage Artificial Intelligence (AI) to automate psychological monitoring. Drawing from the foundational definition [2], AI systems aim to imitate human thought, reasoning, and decision-making processes to solve complex problems with minimal human intervention. Within this broad discipline, methods range from expert systems and fuzzy logic to artificial neural networks and logic programming. Because each algorithmic paradigm presents unique strengths and operational trade-offs, careful selection is required to optimize analytical performance and maximize the efficacy of automated screening systems.

The ubiquity of microblogging platforms like X (formerly Twitter) provides a continuous, real-time reflection of user emotion and psychological state. As demonstrated in [3], user-generated text on social media can be systematically analyzed to predict depressive behavior. By applying Natural Language Processing (NLP) techniques to extract underlying linguistic markers, unstructured posts can be converted into structured indicators capable of reliably classifying mental health conditions.

Text-based depression detection has evolved rapidly from early lexicon matching and sentiment scoring to advanced deep learning and transformer representations. Early studies established that psychological distress manifests distinctly in written language through elevated use of first-person singular pronouns ('I', 'me', 'my'), negative emotion words ('sadness', 'loneliness', 'hopeless'), and social isolation themes [4-5]. However, key challenges persist regarding linguistic ambiguity, imbalanced data, and informal syntax.

Machine Learning and Feature Engineering

Evaluating machine learning models across social media datasets requires balancing accuracy, interpretability, and computational efficiency. As reviewed in [6], pattern recognition and machine learning algorithms provide a versatile foundation for extracting diagnostic signals from large volumes of user-generated content. In specialized applications such as high-risk ideation detection, a study investigated multi-modal and feature-based techniques to identify suicide ideation on social platforms [7]. Their evaluation demonstrated that traditional algorithms, such as Logistic Regression, could achieve high classification performance (reaching 87% accuracy), illustrating the strength of baseline machine learning classifiers when paired with effective feature representation.

Converting cleaned text into numerical representations that machine learning models can process is a crucial stage in depression detection. A study examined sentiment analysis techniques on X/Twitter data, utilizing tools such as TextBlob to calculate polarity (positive, negative, neutral) and subjectivity scores [8]. Integrating these affective metrics enables classifiers to capture emotional context and subtle linguistic nuances that improve classification accuracy.

Classical machine learning algorithms, notably Support Vector Machines (SVM) and Logistic Regression, have demonstrated strong competitive performance when coupled with Term Frequency-Inverse Document Frequency (TF-IDF) feature representations. Alphonsa Sini and Sherly [9] demonstrated that combining Random Forest feature selection with Synthetic Minority Over-sampling Technique (SMOTE) yielded high classification accuracy across depression, anxiety, and stress datasets. Similarly, Chiong [10] showed that combining sentiment lexicons with n-gram Bag-of-Words features enabled Support Vector Machines and Multi-Layer Perceptrons to identify subtle depressive cues without relying on explicit keywords.

Deep Learning and Hybrid Architectures

To capture sequential semantics and local contextual structures, deep learning models have been widely applied. Amanat [5] highlighted the efficacy of Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) layers with attention mechanisms, achieving top-tier accuracy on social media comments. CNN-LSTM hybrid architectures leverage 1D convolutional layers to extract local n-gram feature representations, which are subsequently passed into LSTM units to learn temporal sequential dependencies [11-12].

Transformer Models in Mental Health

Transformer-based architectures, such as BERT and DistilBERT, represent the current frontier in NLP. Through self-attention mechanisms, transformers generate dynamic contextual embeddings where word meanings are computed relative to the surrounding context rather than fixed static vectors. Pretrained models fine-tuned on social media text have consistently outperformed traditional classifiers in capturing indirect distress cues, sarcasm, and complex negation [13-14].

Table 1. Comparative Summary of Related Literature in Text-Based Depression Detection

Study	Preprocessing	Feature Extraction	Models Evaluated	Key Findings & Notable Metrics
[9]	Data cleaning, train-test split, SMOTE balancing	Random Forest feature selection	SVM, ANN, XGBoost	SVM achieved 99.32% accuracy for depression;

				SMOTE effective for class balance.
[11]	URL & stopword removal, tokenization, stemming	Linguistic features, sentiment, LDA topics	RNN, LSTM, Attention-LSTM	Attention-LSTM reached 99.66% accuracy on textual online comments.
[10]	Tokenization, negation detection, lemmatization	BOW, unigrams, bigrams, trigrams	Logistic Regression, SVM, MLP, DT	Negation detection enhanced model ability to spot distress without explicit keywords.
[12]	Emoji/punctuation removal, lowercasing, lemmatization	Attention-based feature embeddings	Deep Learning with Attention	Evolutionary feature extraction improved classification on social media big data.
[16]	Text, emoji, and image preprocessing	Multimodal text & emoticon features	Hybrid SVM + Random Forest	Multimodal integration achieved F-measure of 0.877 on X platform data.

METHODOLOGY

The overall research workflow follows a structured 7-stage framework: dataset collection, data cleaning, custom NLP preprocessing, feature extraction, stratified data splitting, model development and training, and rigorous performance evaluation.

Dataset Description

The primary corpus utilized in this study is a benchmark Kaggle dataset comprising 232,052 raw records of user-generated X microblogs. Following rigorous cleaning and deduping, the refined dataset contains 174,466 high-quality textual samples. The class distribution consists of 67,253 depressed posts (38.55%) and 107,213 non-depressed posts (61.45%), establishing a realistic, moderately imbalanced evaluation environment.

Data Cleaning and NLP Pre-Processing Pipeline

To transform noisy, unstructured microblogs into standardized textual representations without losing vital emotional signals, a multi-stage pre-processing pipeline was developed:

- **Text Normalization & Noise Removal:** Removal of raw URLs, user mentions (@username), hashtags, HTML entities, non-ASCII special characters, and numeric digits while converting all text to lowercase.
- **Contraction Expansion:** Systematic expansion of colloquial contractions (e.g., 'can't' → 'cannot', 'won't' → 'will not', 'didn't' → 'did not') to standardize negation phrases.
- **Custom Negation-Preserving Stopword Filtering:** Standard NLTK stopword lists were modified to retain all crucial negation tokens (not, no, never, cannot, nor, neither, lack) to prevent reversing the polarity of distress indicators.
- **Lemmatization:** WordNet Lemmatization reduced inflected word forms to their canonical dictionary root while retaining syntactic context.

Feature Extraction Strategies

Two distinct feature representation strategies were implemented to accommodate traditional statistical models and deep/transformer neural networks:

i. TF-IDF Vectorization: Employed for Logistic Regression and Linear SVM. Word unigram and bigram features (n-gram range 1–2) were extracted with sublinear TF scaling to produce sparse numeric feature matrices.

ii. Token-Based Sequence Representation: For CNN-LSTM and DistilBERT models, textual samples were tokenized into integer sequence vectors. A maximum sequence length of 128 tokens was specified for CNN-LSTM, while DistilBERT utilized WordPiece tokenization with input IDs and attention masks padded to sequence length 512.

Data Division and Evaluation Framework

The dataset was divided into an 80% training set (139,572 samples) and a 20% testing set (34,894 samples) using stratified random sampling to preserve exact class ratios across splits. Models were evaluated across four core performance metrics: Accuracy, Precision, Recall, and F1-score.

Data Division and Evaluation Framework

Four classification architectures spanning statistical machine learning, recurrent deep learning, and transformer models were developed and benchmarked:

Logistic Regression (LR)

Logistic Regression serves as a strong baseline statistical model. It models the probability of depressive class assignment using a sigmoid activation function applied to a linear combination of TF-IDF feature weights with L2 regularization penalty.

Linear Support Vector Machine (SVM)

Linear SVM constructs an optimal high-dimensional decision hyperplane that maximizes the margin separating depressed and non-depressed TF-IDF feature vectors. SVM is particularly effective in high-dimensional sparse text spaces and robust against overfitting.

CNN-LSTM Deep Learning Architecture

The hybrid CNN-LSTM network combines localized feature extraction with sequential temporal modeling. The architecture comprises:

- **Embedding Layer:** Maps tokenized word indices into dense continuous vector space.
- 1D Convolutional Layer (Conv1D): Applies 128 filters with kernel size 5 to extract local n-gram patterns and phrase-level emotional cues.
- **Max Pooling Layer:** Downsamples feature maps while preserving dominant feature activations.
- **LSTM Layer:** 64 recurrent LSTM units capture long-range temporal dependencies and sentence-level semantic flow, followed by a dense sigmoid classification output.

Fine-Tuned DistilBERT Transformer Model

DistilBERT is a lightweight, distilled transformer model retaining 95% of BERT's language understanding capabilities with 40% fewer parameters. The fine-tuning architecture and hyperparameter configurations are summarized in Table 2.

Table 2. Distil BERT Transformer Fine-Tuning Hyperparameters

Hyperparameter / Configuration	**Setting / Value**
Base Pretrained Architecture	distilbert-base-uncased
Vocabulary Size	30,522 tokens
Transformer Encoder Layers	6 layers
Attention Heads	12 multi-head self-attention heads
Hidden Dimension	768 units
Maximum Sequence Length	512 tokens
Learning Rate & Optimizer	2×10^{-5} (AdamW Optimizer)
Batch Size & Epochs	Batch Size = 16, Epochs = 2
Classification Objective	Single-label binary classification

RESULT AND DISCUSSION

This section presents exploratory data insights, benchmark classification results, confusion matrix evaluations, error analyses, and actionable model comparative findings.

Dataset Exploratory Insights

Exploratory Data Analysis (EDA) revealed notable linguistic divergences between classes. Word frequency, cloud visualizations, and N-gram extractions confirmed that depressed posts exhibit a strong prevalence of self-referential negative emotion bigrams ('want die', 'alone like', 'not want', 'no one', 'suicidal thoughts'). Conversely, non-depressed posts featured general social expressions ('anyone want', 'feel like', 'let us', 'play video game'). Furthermore, n-gram analysis underscored that negation phrases ('not know', 'not want', 'not even') were heavily prevalent across both classes, reinforcing the necessity of retaining negation terms during stopword removal.

Benchmark Model Efficiency

Table 3. Benchmark Model Performance Metrics on Test Set (%)

Model Architecture	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Logistic Regression (TF-IDF)	97.06%	96.32%	96.04%	96.18%
Linear SVM (TF-IDF)	97.14%	95.59%	97.05%	96.31%
CNN-LSTM Deep Learning	96.58%	95.84%	95.27%	95.55%
Fine-Tuned DistilBERT	97.46%	97.13%	96.24%	96.68%

All four trained architectures achieved excellent classification performance on the unseen 34,894-sample test set, with overall accuracies exceeding 96%. Table 3 presents a comparative performance evaluation of the four model architectures on the test set, demonstrating that all evaluated models achieved high efficacy with overall accuracies exceeding 96%. Among them, the Fine-Tuned DistilBERT model secured the best overall performance, registering the highest accuracy at 97.46%, precision at 97.13%, and F1-score at 96.68%. This

superiority is attributed to its self-attention mechanism, which effectively handles bidirectional context and complex linguistic expressions. In comparison, the traditional machine learning models paired with TF-IDF vectorization proved exceptionally competitive, matching or outperforming the hybrid CNN-LSTM deep learning network, which recorded an accuracy of 96.58%. Notably, the Linear SVM achieved the highest recall among all architectures at 97.05%. This high sensitivity makes Linear SVM particularly valuable for real-world screening scenarios where minimizing false negatives or failing to identify individuals in psychological distress is paramount.

Error Analysis: Confusion Matrix Insights

Contextual understanding is especially critical in detecting depression since expressions of emotion can vary greatly. As an illustration in Figure 1 and 2, the same word can be used both in the depressed and non-depressed texts but have different meanings in the entire sentence construction. To evaluate model utility in clinical or real-world screening scenarios, error analysis was performed on false negative (FN) and false positive (FP) predictions. Specifically, Linear SVM produced only 397 false negatives out of 13,451 actual depressed test samples, achieving the highest sensitivity with a recall of 97.05%, where missed cases primarily involved posts utilizing implicit distress expressions, metaphorical language, or sub-threshold emotional cues without explicit clinical keywords. Concurrently, Linear SVM recorded 602 false positives out of 21,443 non-depressed samples, which stemmed predominantly from casual conversations discussing mental health topics hypothetically, sarcastic posts, or song lyrics containing depression-associated terms.

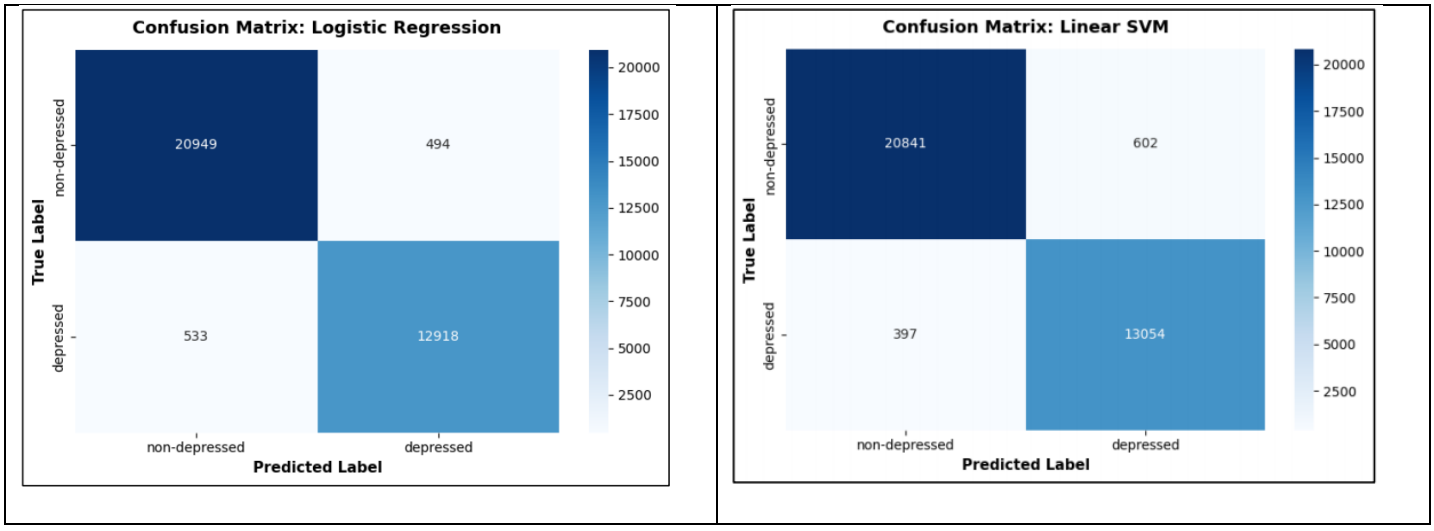


Figure 1. Confusion Matrix for (a) Logistic Regression, (b) Linear SVM

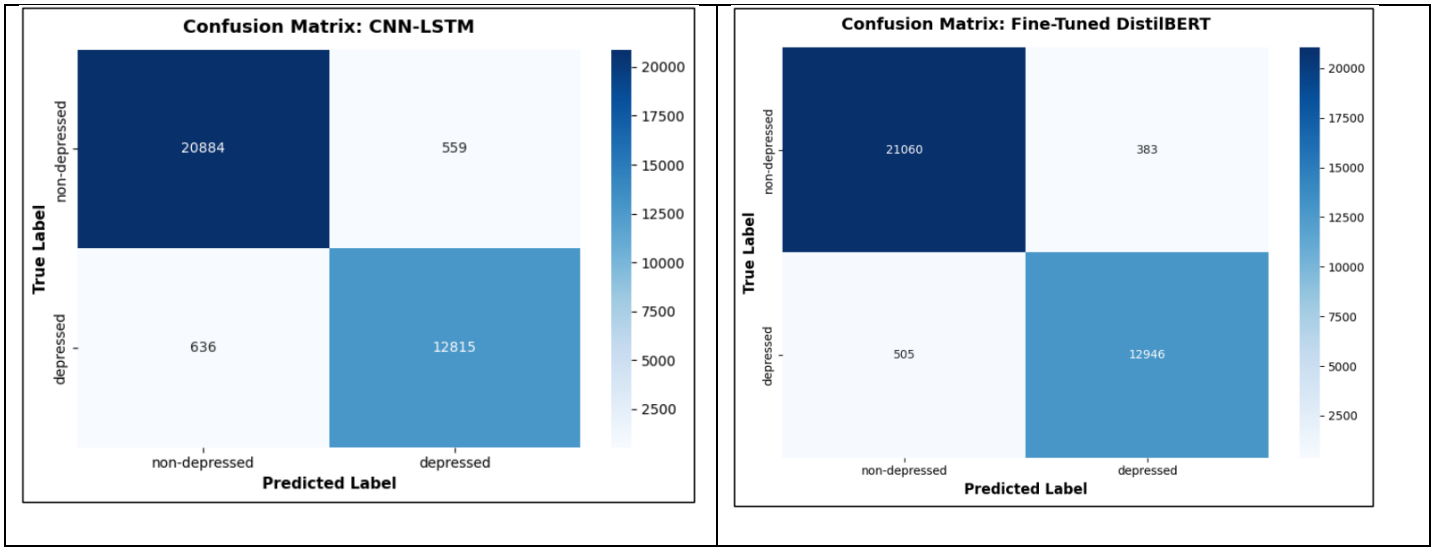


Figure 2. Confusion Matrix for (a) CNN-LSTM, (b) Fine-Tuned DistilBERT

DISCUSSION AND MODEL SELECTION

The experimental findings highlight key trade-offs between accuracy, sensitivity, and computational overhead. For overall efficacy, Fine-Tuned DistilBERT is selected as the top overall performing model, achieving the highest accuracy (97.46%), precision (97.13%), and F1-score (96.68%). Its self-attention mechanism effectively resolved bidirectional context and nuanced linguistic expressions. In terms of clinical screening utility, the Linear SVM proved exceptionally competitive, achieving the highest Recall (97.05%). In public health applications where minimizing false negatives (missed individuals in distress) is paramount, Linear SVM offers an outstanding, low-latency screening alternative. When evaluating the broader machine learning versus deep learning paradigm, traditional machine learning models equipped with TF-IDF vectorization performed on par with, or even superior to, the hybrid CNN-LSTM network, which achieved an accuracy of 96.58%. This performance parity confirms that a rigorous, high-quality natural language processing preprocessing pipeline combined with TF-IDF n-grams can effectively capture significant domain-specific signals without requiring computationally demanding deep learning architectures.

CONCLUSION

This research comprehensively evaluated machine learning, deep learning, and transformer architectures for automated depression detection on X social media microblogs. By designing a custom NLP preprocessing pipeline that preserves critical negation structures and leveraging a dataset of 174,466 records, all evaluated models demonstrated superior predictive performance exceeding 96% accuracy.

Fine-Tuned DistilBERT achieved state-of-the-art overall efficacy (97.46% accuracy, 96.68% F1-score), establishing the value of deep contextual transformer embeddings. Concurrently, Linear SVM achieved the highest recall (97.05%), demonstrating that lightweight statistical models remain highly effective for screening applications prioritizing sensitivity.

Research Contributions

This study makes two primary contributions: first, it formulates a specialized, negation-preserving natural language processing preprocessing framework specifically tailored to handle noisy microblog text; and second, it delivers an exhaustive comparative evaluation across classical machine learning, deep recurrent networks, and pretrained transformers under uniform experimental conditions.

Study Limitations and Future Work

This study is subject to specific limitations: (1) Binary classification simplifies psychological distress into discrete classes, whereas clinical depression exists along a continuous severity spectrum; (2) Transformer models entail substantial computational requirements for real-time deployment; and (3) Text-only analysis excludes multimodal cues like images or user network interactions.

Future research will expand this framework into multi-class depression severity grading (mild, moderate, severe), incorporate multimodal features (text, emoticons, and network metadata), and evaluate cross-platform domain adaptability across diverse social media platforms.

ACKNOWLEDGEMENT

We would like to thank you to UTeM for financial and facilities support for this study.

REFERENCES

1. Samsudin, S., Ismail, R., Mohd Daud, S. N., & Yahya, S. (2024). The prevalence and underlying factors of mental health disorders in youth. Journal of Affective Disorders Reports, 15, 100680.
2. Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.

3. Biswas, S., & Hasija, Y. (2022). Predicting depression through social media. Lecture Notes on Data Engineering and Communications Technologies, 112, 45-58.
4. Cavazos-Rehg, P. A., Krauss, M. J., Sowles, S., Connolly, S., Rosas, C., Miller, M. B., & Bierut, L. J. (2016). A content analysis of depression-related tweets. Computers in Human Behavior, 54, 351-357.
5. Gu, L., Li, M., & Li, Y. (2026). Linguistic markers of depression and emergent self-stigma in online self-disclosures: A mixed-methods study on Chinese social media. Journal of Affective Disorders, 395, 112-123.
6. Adugna, T. D., Ramu, A., & Haldorai, A. (2024). A review of pattern recognition and machine learning. Journal of Machine and Computing, 4(1), 12-25.
7. Chatterjee, M., Kumar, P., Samanta, P., & Sarkar, D. (2022). Suicide ideation detection from online social media: A multi-modal feature based technique. International Journal of Information Management Data Insights, 2(2), 100095.
8. Ahmed, M., Ansari, M., Qadir, A., & Hashmi, D. (2025). Depression detection using machine learning techniques on X/Twitter data. International Journal of Information Technology and Computer Engineering, 13, 225–232.
9. Alphonsa Sini, P. J., & Sherly, K. K. (2023). Early detection of anxiety, depression and stress among potential patients using machine learning and deep learning models. Proceedings of the 2nd International Conference on Computational Systems and Communication (ICCSC), 104-110.
10. Chiong, R., Budhi, G. S., Dhakal, S., & Cambria, E. (2021). Combining sentiment lexicons and content-based features for depression detection. IEEE Intelligent Systems, 36(6), 99-105.
11. Amanat, A., Rizwan, M., Javed, A. R., Abdelhaq, M., Alsaqour, R., Pandya, S., & Sujatha, R. (2022). Deep learning for depression detection from textual data. Electronics (Switzerland), 11(5), 780.
12. Prabhakar, K., & Kavitha, V. (2024). An evolutionary approach for depression detection from Twitter big data using a novel deep learning model with attention based feature learning mechanism. Automatika, 65(2), 512-524.
13. Tejaswini, V., Babu, K. S., & Sahoo, B. (2024). Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(1), 1-22.
14. Patkar, U. C., Sharma, R., & Singh, K. (2024). Object and sentiment pattern recognition using machine learning and deep learning. International Journal of Intelligent Systems and Applications in Engineering, 12(1s), 415-422.
15. Turuba, R., Johnson, P., & Smith, L. (2025). A summative content analysis of mental health-related content on social media platforms. Digital Health, 11, 1-14.
16. Angskun, T., Tipprasert, S., Thippongtorn, A., & Angskun, J. (2024). D_2X: Depression detection system through X using hybrid machine learning. IEEE Access, 12, 172820–172831.
17. Tejaswini, V., Babu, K. S., & Sahoo, B. (2024). Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(1), 1-22.