

Duangporn Janon* and Pornpimol Chaiwuttisak

64605045@kmitl.ac.th*, and pornpimol.ch@kmitl.ac.th

Department of Statistic, Faculty of Science, King Mongkut's Institute of Technology
Ladkrabang (KMITL), Bangkok Thailand 10520

*Corresponding author E-mail: 64605045@kmitl.ac.th

Received: April 22, 2026 **Revised:** June 9, 2026 **Accepted:** June 24, 2026

Citation reference :

Janon, D., Chaiwuttisak, P.(2026). Deep Learning Models for Early Detection of Depression from Social Media Text. **International Journal of Industrial Education and Technology**, 8(1), 82-92.

ABSTRACT

Depression is a major global health concern. Recent studies suggest that language used on social media may reveal early signs of psychological distress before individuals seek formal clinical support. The objective of this study was to develop and evaluate transformer-based natural language processing models for detecting depressive risk signals in long-form Reddit posts using the ERISK2025 dataset. A traditional TF-IDF representation combined with Logistic regression served as the baseline and was compared with BERT-base and the emotion-aware Emotion-RoBERTa model under identical experimental conditions. Model performance was assessed using accuracy, risk-class recall, macro-F1, AUC, and confusion-matrix-based error distribution. Emotion-RoBERTa achieved the highest overall accuracy (0.76) and the strongest sensitivity to high-risk users with a risk-class recall of 0.91 and the lowest false-negative rate among the tested models. BERT-base showed moderate performance. The TF-IDF baseline was weakest at identifying high-risk cases. These findings indicate that emotion-aware language representations can improve early detection of depression risk in social media text and may identify vulnerable users more reliably than traditional text-representation methods. The results also indicate that model evaluation in this setting should rely on clinically meaningful metrics rather than accuracy alone. Future work should consider larger datasets together with temporal patterns in user posts and ethical constraints related to real-world mental health screening.

Keywords: Depression detection, Reddit, eRisk2025, TF-IDF; BERT-base, Emotion-RoBERTa

I. INTRODUCTION

Over the past decade, social media has become deeply integrated into everyday life. Large proportions of the global population now depend on online platforms for communication, information exchange, and self-expression (Chaffey, 2025, Online; Pew Research Center, 2025, Online). In Thailand, national statistics report high levels of internet and mobile phone use, particularly after the COVID-19 pandemic shifted many activities into digital environments (National Statistical Office of Thailand, 2025, p. 33). During periods of social distancing, social media was widely used to share experiences, concerns, and uncertainty, making these platforms a primary space for emotional expression and social support (Tsao et al., 2021, p. 184; Handayani et al., 2023, p. 8). Platform features such as status updates, comments, and reactions allow users to express feelings quickly and respond to others in real time, creating conditions where emotional content spreads easily (Steinert & Dennis, 2022, pp. 11-13; Chen et al., 2022, p. 61). Prior studies show that online language can reflect psychological conditions, including depression, social exclusion, and bullying among adolescents (Ophir et al., 2019, pp. 3-4), and emotional changes can be observed in online discussions during recovery from substance use (Yang et al., 2023, p. 2). These findings suggest that social media data may reveal signs of distress before individuals seek formal help.

Depression is a major global health problem, affecting a large number of people and contributing substantially to disability and suicide worldwide (WHO, 2025a, Online). Recent estimates indicate that depressive and anxiety disorders account for a significant share of disability-adjusted life years and have increased steadily over the past decades (Fan et al., 2025, pp. 8-10). Many cases remain undiagnosed or untreated, leading to functional impairment and economic burden (Yoch & Sirull, 2021, Online; Proudman et al., 2021, p. 620). Social and economic pressures, including chronic stress, loss, and inequality, are associated with the onset and progression of depressive symptoms (Reddy, 2010, pp. 1; Moreno-Agostino et al., 2021, p. 241). In Thailand, mental health reports show rising levels of stress and depression risk, especially among young people, while access to mental health services remains limited (WHO, 2025b, Online; Bangkok Post, 2022, Online; Munthuli et al., 2023, p. 3; National Health Security Office, 2023, Online). Stigma toward mental illness also discourages help-seeking behaviour (Pruksarungruang & Rhein, 2022, p. 2). In such conditions, individuals often express emotional distress online before contacting health professionals, making social media a useful source of early indicators.

Recent research has explored the use of natural language processing and machine learning to identify depressive signals in social media text. Patterns of sadness, hopelessness, and loneliness in online posts have been used to predict depression with promising results. Social media data are suitable for early detection because they are abundant, time-stamped, and reflect real-world communication. Advances in deep learning have improved the ability to extract emotional features from large text corpora, but several challenges remain. Social media language is informal and often includes slang, spelling variation, sarcasm, and code-switching, which can obscure emotional meaning (Alaramma et al., 2023, p. 77). Emojis and special characters also add complexity to interpretation (Alqahtani et al., 2023, p. 325). Many earlier studies relied on short texts, small datasets, or simple binary labels, which limit the detection of subtle psychological states (Salas-Zárate et al., 2022, p. 2; Ahmed et al., 2022, p. 5; Bao et al., 2025, p. 1). Long-form platforms such as Reddit provide richer narratives that allow observation of emotional change over time, and the eRisk dataset offers longitudinal posts labelled using validated self-report instruments for research on early risk detection (Boettcher, 2021, pp. 2-3; Bao et al., 2025, p. 1). These characteristics make long-form data suitable for



testing transformer-based models designed to capture contextual and emotional information in text (Triantafyllopoulos et al., 2023, p. 3; Bokolo & Liu, 2023, p. 9; Zhang et al., 2022, pp. 46-47).

Based on these considerations, this study investigates the effectiveness of transformer-based classification models for detecting depressive risk signals in long-form Reddit posts using the ERISK2025 dataset. The research compares three model conditions under identical experimental settings. The first condition is a traditional baseline that combines TF-IDF features with Logistic regression. The second is the general-purpose BERT-base transformer. The third is the emotion-aware Emotion-RoBERTa model. The comparison gives particular attention to performance measures that are relevant for early screening. These include recall for high-risk users and reduction of false-negative predictions. In addition, the study applies a multi-metric evaluation framework, including accuracy, macro-F1, recall, and AUC, in order to provide a more clinically meaningful interpretation of model performance in early depression risk detection.

II. LITERATURE REVIEW

Recent studies have examined the use of natural language processing and machine learning for detecting depression from textual data, particularly from social media where users often express emotions and personal experiences. Automated text analysis has been proposed as an alternative to traditional screening methods based on interviews and questionnaires, allowing large volumes of text to be analysed for linguistic patterns associated with depressive symptoms (Teferra et al., 2024, pp. 1-2). Transformer-based language models such as BERT and related architectures have been widely used because they capture contextual relationships more effectively than feature-based approaches, and domain-adapted models trained on mental-health corpora have shown improved ability to detect emotional cues in informal language (Mao & Han, 2025, p. 6). Comparative studies also report that traditional machine learning algorithms, including logistic regression and support vector machines, can perform competitively depending on dataset size, class balance, and feature representation, which makes multi-metric evaluation necessary for reliable comparison (Ding et al., 2025, p. 1; Mao & Han, 2025, pp. 8-9). Other research shows that combining contextual embeddings with additional linguistic or behavioural information, as well as applying balancing techniques, can improve classification performance when analysing social-media text (Kerasiotis et al., 2024, p. 1; Qasim et al., 2025, p. 4; Zhang et al., 2024, pp. 6-8).

Taken together these studies establish that transformer models can detect depressive language in social media text and that domain adaptation tends to improve sensitivity to emotional cues. Two methodological questions remain less settled. The first concerns long-form posts because extended writings place greater demands on context handling than the short texts examined in much prior work. The second concerns emotionally nuanced modelling because the added value of emotion-aware pretraining over general contextual representation is not yet clear under directly comparable conditions. Simpler classifiers also remain competitive under some data conditions which further complicates model selection. The present study addresses these unresolved points by comparing a traditional baseline with general-purpose and emotion-aware transformers on long-form Reddit text under identical conditions and by evaluating the models with multiple metrics.

The present research is based on previous findings comparing traditional classifiers and transformer-based models for depression detection, particularly studies showing that contextual and emotion-aware representations may improve the identification of risk-related language in social-media text, while simpler models may still provide stable performance under certain data conditions (Kuzmin et al., 2025, p. 34535). Research using long-form data from platforms



such as Reddit indicates that extended text provides richer emotional context, which supports the use of transformer architectures for risk detection (Qasim et al., 2025, pp. 3-4; Zhang et al., 2024, p. 3). Prior work also shows that evaluation based only on accuracy may not reflect true model behaviour, especially when class imbalance and false-negative errors are present, and therefore multiple metrics should be used for reliable comparison (Bokolo & Liu, 2023, p. 11). Based on these findings, the present study proposes three hypotheses: that an emotion-aware transformer model will achieve higher recall for the depressive risk class than a general-purpose transformer model (H1), that the emotion-aware model will produce fewer false-negative predictions in long-form text classification (H2), and that differences in model performance will be more clearly identified when multiple evaluation metrics such as accuracy, macro-F1, recall, and AUC are used instead of accuracy alone (H3).

III. RESEARCH METHODOLOGY

This study employed a quantitative experimental design with a comparative approach to evaluate the performance of deep learning-based natural language processing models for detecting depressive signals in social media text. A cross-sectional framework was used to analyse patterns of emotional expression within a defined time period and to ensure a fair comparison between models under identical conditions (Teferra et al., 2024, p. 12). The data source was the ERISK2025 benchmark dataset, which contains 11,042 English-language text instances derived from longitudinal Reddit user writings for early mental health risk research. Each instance was labeled as risk or non-risk. To maintain reproducibility and class balance, the dataset was divided using a stratified procedure into 8,833 instances for training and 2,209 instances for testing, preserving the original distribution (risk = 55.4%, non-risk = 44.6%). This corresponds to an approximately 80:20 train/test partition in which the test set was held out from model development and used only for final evaluation. Included texts were required to be written in English, contain interpretable emotional content, and exceed a minimum length, while posts consisting only of links, symbols, advertisements, duplicated content, or very short messages without clear emotional meaning were excluded to reduce noise during model training (Salas-Zárate et al., 2022, p. 4; Ahmed et al., 2022, p. 6).

Data processing and analysis were performed using Python. Text preprocessing and model implementation were conducted with the HuggingFace Transformers library together with PyTorch. A TF-IDF + Logistic Regression classifier was implemented as a traditional machine-learning baseline using the same stratified train/test partition to provide a direct comparison with the transformer-based models. Two transformer-based language models were evaluated: BERT-base (uncased) as a general-purpose pretrained model and Emotion-RoBERTa as a model pretrained on mental health-related text. Both models were adapted for binary classification using the same classification head to allow direct comparison (Bokolo & Liu, 2023, p. 7; Qasim et al., 2025). The ERISK2025 dataset was constructed from longitudinal Reddit posts provided by users who participated voluntarily in standardized mental health assessments. Risk labels were assigned from validated self-report questionnaires, including the Beck Depression Inventory (BDI), rather than clinical diagnosis, so the labels indicate elevated depressive risk rather than confirmed disorder (Bao et al., 2025, p. 2). Therefore, the labels used in this study represent depressive risk identified through screening instruments rather than confirmed clinical diagnoses. Text data were cleaned, deduplicated, and standardised before tokenisation. Inputs were prepared using padding, truncation, and attention masks, with a fixed maximum sequence length of 256 tokens. Texts longer than this limit were truncated, and shorter texts were padded to ensure consistent batch processing.



Model development and evaluation used the same stratified partition of 8,833 training instances and 2,209 testing instances described above. The train/test split was generated with the class label used for stratification and `random_state = 42` specified for the split procedure. Separating data into distinct training and evaluation sets follows established practice in text-classification research (Kowsari et al., 2019, p. 4; Wang, 2023, pp. 93-96). Both models were fine-tuned with the AdamW optimiser, learning rate 2×10^{-5} , batch size 16, and 3 training epochs. A linear learning-rate scheduler with warmup (ratio = 0.1) was applied, and all settings were kept identical for both models to maintain a controlled comparison. Model performance was evaluated on the test set using accuracy, recall, F1-score, AUC, and confusion-matrix-based error distribution, with macro or weighted F1 reported when class imbalance occurred, and confusion matrices were used to analyse classification errors (Bokolo & Liu, 2023, p. 12; Teferra et al., 2024, p. 9).

IV. RESULTS

The ERISK2025 dataset used in this study contained 11,042 text instances. To ensure fair evaluation, the data were divided using a stratified split into 8,833 documents for training and 2,209 documents for testing while preserving the original label distribution. The proportion of texts labelled as risk was 55.4%, whereas 44.6% were labelled as non-risk. Model performance on the held-out test set was reported using accuracy, risk-class recall, macro-F1, AUC, and confusion-matrix-based error distribution to provide both threshold-dependent and threshold-independent assessment. Each evaluation metric was linked to one of the study hypotheses. Recall for the risk class served as the primary indicator for H1 because it measures the proportion of high-risk users that a model identifies. False-negative counts from the confusion matrices served as the indicator for H2 because they record missed high-risk cases directly. The combined use of accuracy, risk-class recall, macro-F1, AUC, and error distribution addressed H3 because performance differences among models are identified more clearly with multiple metrics than with accuracy alone. Table 1 presents the dataset summary and class distribution.

Table 1: Dataset summary and class distribution

Item	Count
Total documents	11,042
Train set	8,833
Test set	2,209
Non-risk	4,925 (44.6%)
Risk	6,117 (55.4%)

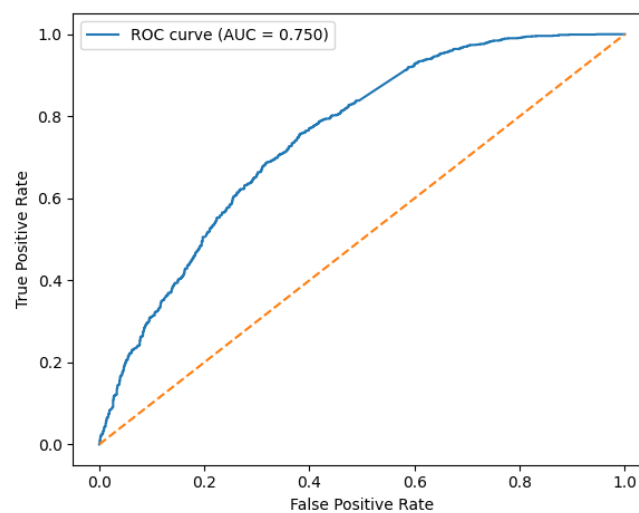
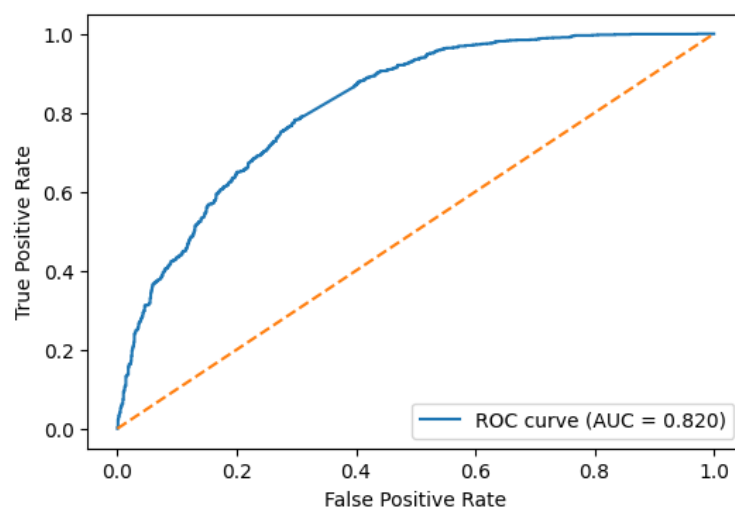
The performance of the evaluated models on the test set is shown in Table 2, including the TF-IDF + Logistic regression baseline, BERT-base, and Emotion-RoBERTa. Evaluation focused on accuracy, macro-F1, recall for the risk class, and AUC. Emotion-RoBERTa achieved the highest recall for the risk class (0.91), followed by BERT-base (0.88) and the baseline model (0.81). Overall accuracy values were 0.69 for TF-IDF + Logistic regression, 0.75 for BERT-base, and 0.76 for Emotion-RoBERTa. The hypothesis mapping is therefore interpreted from the main held-out test-set results in Table 2 and the error-distribution results in Table 3, rather than from a separate additional comparison table.



Table 2: Performance comparison across models

Model	Accuracy	Precision	Recall	Macro-F1	AUC
TF-IDF + Logistic regression	0.69	–	0.81 (Risk)	0.67	0.75
BERT-base	0.75	–	0.88 (Risk)	0.73	0.82
Emotion-RoBERTa	0.76	–	0.91 (Risk)	0.74	0.82

Receiver Operating Characteristic (ROC) curves for the three models are shown in Figures 1–3. The baseline model produced an AUC of 0.75, indicating moderate discrimination ability. The BERT-base model achieved an AUC of 0.82, showing improved separability between risk and non-risk instances. Emotion-RoBERTa also produced an AUC of 0.82 with a curve closer to the upper-left corner, indicating stronger sensitivity across thresholds.

**Figure 1:** ROC curve of the TF-IDF + Logistic regression model on the test set**Figure 2:** ROC curve of the BERT-base model on the test set

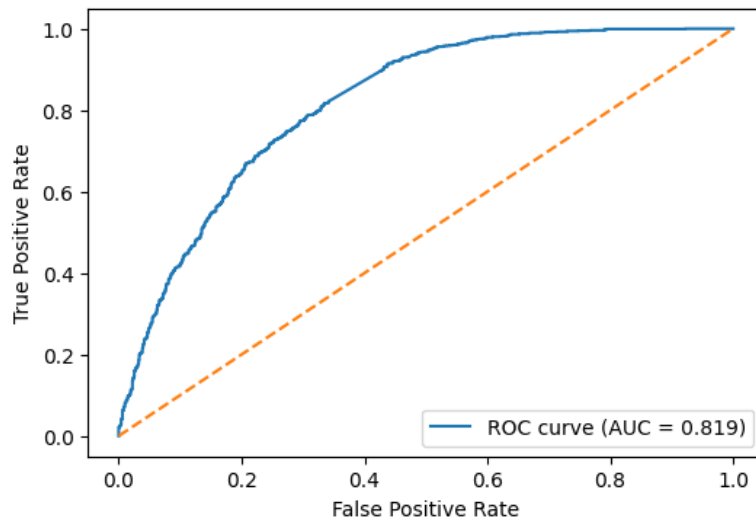


Figure 3: ROC curve of the emotion-RoBERTa model on the test set

Error distribution across models is summarised in Table 3, including true positives, false positives, false negatives, and true negatives. The baseline model produced 228 false negatives and 458 false positives. BERT-base reduced false negatives to 143 and false positives to 407. Emotion-RoBERTa produced the lowest number of false negatives (109) with 431 false positives. These results show differences in classification outcomes across models and provide additional information on prediction behaviour in the test set.

Table 3: Summary of FN / FP of three models

Model	TP	FP	FN	TN
Baseline (TF-IDF + LR)	996	458	228	527
BERT-base	1081	407	143	578
Emotion-RoBERTa	1115	431	109	554

V. CONCLUSION AND DISCUSSION

The study evaluated the effectiveness of different natural language processing approaches for detecting depressive risk in social media text using the ERISK2025 dataset. The comparison included a TF-IDF + Logistic regression baseline, the BERT-base model, and the Emotion-RoBERTa model, with performance assessed using accuracy, recall, macro-F1, AUC, and error distribution. The results showed that Emotion-RoBERTa achieved the highest recall for the risk class and the lowest number of false-negative predictions, indicating stronger sensitivity to high-risk language than the other models. The results also show that model selection should not rely on accuracy alone. The TF-IDF + Logistic regression model remains a useful traditional baseline because it provides an interpretable comparison point. However, BERT-base improved over the baseline in accuracy, risk-class recall, macro-F1, and AUC, while Emotion-RoBERTa provided the strongest risk-class recall and the lowest false-negative count. For depression-risk screening, model selection should therefore consider risk-class recall, false negatives, macro-F1, AUC, and error distribution rather than overall accuracy alone. The analysis also confirmed that evaluating models with multiple metrics provided a clearer understanding of performance than relying on accuracy alone, particularly when class imbalance and safety-related errors were considered. Error distribution analysis further showed that the emotion-aware model minimized missed high-risk cases, whereas the baseline model produced the highest number



of false negatives. These findings indicate that model evaluation in depression risk detection should consider both aggregate metrics and error patterns to obtain reliable results.

The findings indicate that the emotion-aware transformer model provided stronger performance in detecting depressive risk signals than the general-purpose transformer architecture, as Emotion-RoBERTa showed higher recall for the risk class and a more balanced macro-F1 score, reflecting improved sensitivity to psychologically vulnerable users. This result is consistent with previous studies reporting that transformer architectures can capture richer semantic and affective information from text, while domain-adapted or emotionally trained models may better detect subtle psychological expressions in social media language (Mao & Han, 2025, pp. 1-3; Kuzmin et al., 2025, p. 34542). The comparison between Logistic Regression and BERT also suggests that higher model complexity does not always result in more stable classification performance. Although BERT-base achieved better overall evaluation scores, the TF-IDF + Logistic Regression baseline remained competitive and provided consistent results under the same experimental conditions. This finding supports previous studies indicating that simpler classifiers may remain effective when feature representations are informative and class distributions are relatively stable. Therefore, model selection in mental health NLP should consider not only predictive performance but also interpretability, computational efficiency, and deployment requirements. In addition, the results confirm that multi-metric evaluation provides a more reliable interpretation of model behaviour than accuracy alone, because recall, macro-F1, and AUC revealed differences in class detection and error distribution that were not visible from accuracy values, which is consistent with methodological recommendations that high-stakes NLP applications should rely on multiple performance indicators to capture differences in sensitivity, stability, and error patterns (Mao & Han, 2025, p. 5).

SUGGESTION

Future research should extend the present findings by using larger and more diverse datasets, including longitudinal user histories, to improve robustness and support earlier identification of depressive risk patterns. Additional work should examine alternative model architectures and hybrid approaches, such as temporal or hierarchical transformers and ensemble methods that combine semantic and emotional features, in order to improve detection stability and sensitivity beyond the performance observed in this study. Validation across different platforms, languages, and cultural contexts is also recommended to assess generalisability and reduce dependence on a single dataset, with evaluation under conditions that resemble practical screening use. Further studies should also consider issues related to model interpretability, error balance, and ethical use, including the need to protect user privacy and to apply automated detection as a supportive tool rather than a clinical diagnosis.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the faculty members and academic experts from the Department of Statistics and Business Analytics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, for their valuable guidance, support, and encouragement throughout the course of this research.



REFERENCES

- Chaffey, D. (2025). Global social media statistics research summary. *Smart Insights*. <https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/>.
- Pew Research Center. (2025). Social media fact sheet. <https://www.pewresearch.org/internet/fact-sheet/social-media/>.
- National Statistical Office of Thailand. (2025). *Survey on the use of information and communication technology in households, 2025 (Quarter 2)*. Ministry of Digital Economy and Society. https://www.nso.go.th/nsoweb/storage/survey_detail/2025/20250903131427_86572.pdf.
- Tsao, S. F., Chen, H., Tisseverasinghe, T., Yang, Y., Li, L., & Butt, Z. A. (2021). What social media told us in the time of COVID-19: A scoping review. *The Lancet Digital Health*, 3(3), 175–194.
- Handayani, P. W., Zagatti, G. A., Kefi, H., & Bressan, S. (2023). Impact of social media usage on users' COVID-19 protective behavior: Survey study in Indonesia. *JMIR Formative Research*, 7(1).
- Steinert, S., & Dennis, M. J. (2022). Emotions and digital well-being: On social media's emotional affordances. *Philosophy & Technology*, 35(2), 36.
- Chen, J. C. J., Yan, Y. Y. Y., & Leach, J. L. J. (2022). Are emotion-expressing messages more shared on social media? A meta-analytic review. *Review of Communication Research*, 10.
- Ophir, Y., Asterhan, C. S., & Schwarz, B. B. (2019). The digital footprints of adolescent depression, social rejection and victimization of bullying on Facebook. *Computers in Human Behavior*, 91, 62–71.
- Nesi, J., Rothenberg, W. A., Bettis, A. H., Massing-Schaffer, M., Fox, K. A., Telzer, E. H., & Prinstein, M. J. (2022). Emotional responses to social media experiences among adolescents: Longitudinal associations with depressive symptoms. *Journal of Clinical Child & Adolescent Psychology*, 51(6), 907–922.
- Yang, G., King, S. G., Lin, H. M., & Goldstein, R. Z. (2023). Emotional expression on social media support forums for substance cessation: Observational study of text-based Reddit posts. *Journal of Medical Internet Research*, 25.
- WHO. (2025a). Depressive disorder (depression). <https://www.who.int/news-room/fact-sheets/detail/depression>.
- Fan, Y., Fan, A., Yang, Z., & Fan, D. (2025). Global burden of mental disorders in 204 countries and territories, 1990–2021: Results from the global burden of disease study 2021. *BMC Psychiatry*, 25(1), 486.
- Yoch, M., & Sirull, R. (2021). New global burden of disease analyses show depression and anxiety among the top causes of health loss worldwide. Institute for Health Metrics and Evaluation. <https://www.healthdata.org/news-events/insights-blog/acting-data/new-global-burden-disease-analyses-show-depression-and>.
- Proudman, D., Greenberg, P., & Nellesen, D. (2021). The growing burden of major depressive disorders (MDD): Implications for researchers and policy makers. *Pharmacoeconomics*, 39(6), 619–625.
- Reddy, M. S. (2010). Depression: The disorder and the burden. *Indian Journal of Psychological Medicine*, 32(1), 1–2.



- Moreno-Agostino, D., Wu, Y. T., Daskalopoulou, C., Hasan, M. T., Huisman, M., & Prina, M. (2021). Global trends in the prevalence and incidence of depression: A systematic review and meta-analysis. *Journal of Affective Disorders*, 281, 235-243.
- WHO. (2025b). Mental health and social connection in Thailand.
<https://www.who.int/thailand/news/feature-stories/detail/mental-health-and-social-connection-in-thailand>.
- Bangkok Post. (2022). Teens face pandemic woes.
<https://www.bangkokpost.com/thailand/general/2314054/teens-face-pandemic-woes>.
- Munthuli, A., Pooprasert, P., Klangpornkun, N., Phienphanich, P., Onsuwan, C., Jaisin, K., & Tantibundhit, C. (2023). Classification and analysis of text transcription from Thai depression assessment tasks among patients with depression. *PLoS ONE*, 18(3).
- National Health Security Office. (2023). Mental health hotline integrated into the UCS.
<https://eng.nhso.go.th/view/1/DescriptionNews/Mental-health-hotline-integrated-into-the-UCS/511/EN-US>.
- Pruksarungruang, J., & Rhein, D. (2022). Depression literacy: An analysis of the stigmatization of depression in Thailand. *SAGE Open*, 12(4).
- Alaramma, S. K., Habu, A. A., Ya'u, B. I., & Madaki, A. G. (2023). Sentiment analysis of sarcasm detection in social media. *Gadua Journal of Pure and Allied Sciences*, 2(1), 76-82.
- Alqahtani, A., Alhenaki, L., & Alsheddi, A. (2023). Text-based sarcasm detection on social networks: A systematic review. *International Journal of Advanced Computer Science and Applications*, 14(3).
- Salas-Zárate, R., Alor-Hernández, G., Salas-Zárate, M. P., Paredes-Valverde, M. A., Bustos-López, M., & Sánchez-Cervantes, J. L. (2022). Detecting depression signs on social media: A systematic literature review. *Healthcare*, 10(2).
- Ahmed, A., Aziz, S., Toro, C. T., Alzubaidi, M., Irshaidat, S., Serhan, H. A., & Househ, M. (2022). Machine learning models to detect anxiety and depression through social media: A scoping review. *Computer Methods and Programs in Biomedicine Update*, 2.
- Bao, E., Pérez, A., & Parapar, J. (2025). ReDSM5: A Reddit dataset for DSM-5 depression detection. In *Proceedings of the ACM International Conference on Information and Knowledge Management* (pp. 6323-6327).
- Boettcher, N. (2021). Studies of depression and anxiety using Reddit as a data source: Scoping review. *JMIR Mental Health*, 8(11).
- Triantafyllopoulos, I., Paraskevopoulos, G., & Potamianos, A. (2023). Depression detection in social media posts using affective and social norm features. *arXiv preprint arXiv:2303.14279*.
- Bokolo, B. G., & Liu, Q. (2023). Deep learning-based depression detection from social media: Comparative evaluation of ML and transformer techniques. *Electronics*, 12(21).
- Teferra, B. G., Rueda, A., Pang, H., Valenzano, R., Samavi, R., Krishnan, S., & Bhat, V. (2024). Screening for depression using natural language processing: Literature review. *Interactive Journal of Medical Research*, 13(1).
- Mao, H., & Han, Q. (2025). Applications of transformer-based language models for depression detection: A scoping review. *Journal of Integrated Social Sciences and Humanities*.
- Ding, Z., Wang, Z., Zhang, Y., Cao, Y., Liu, Y., Shen, X., & Dai, J. (2025). Trade-offs between machine learning and deep learning for mental illness detection on social media. *Scientific Reports*, 15(1), 14497.



- Kerasiotis, M., Ilias, L., & Askounis, D. (2024). Depression detection in social media posts using transformer-based models and auxiliary features. *Social Network Analysis and Mining*, 14(1), 196.
- Qasim, A., Mehak, G., Hussain, N., Gelbukh, A., & Sidorov, G. (2025). Detection of depression severity in social media text using transformer-based models. *Information*, 16(2), 114.
- Zhang, Z., Zhu, J., Guo, Z., Zhang, Y., Li, Z., & Hu, B. (2024). Natural language processing for depression prediction on Sina Weibo. *JMIR Mental Health*, 11.
- Kuzmin, G., Strepetov, P., Stankevich, M., Chudova, N., Shelmanov, A., & Smirnov, I. (2025). Exploring large language models for detecting mental disorders. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 34523-34547).
- Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4), 150.
- Wang, M. (2023). Research on text classification method based on NLP. *Advances in Computer, Signals and Systems*, 7(2).

The content, as well as the use of lan gauge in the article, is the responsibility of the author.

