International Organization of Research & Development (IORD)
ISSN: 2348-0831
Vol 13 Special Issue 02 | 2026

AI-Based Early Detection of Psychological Distress in Engineering Students: A Comparative Study of Transformer-Based Text Classification Models

Mayuri Sanjay More[1]
[1]*HOD Artificial Intelligence & Data Science, Siddhivinavak Technical Campus, Shegaon, Maharashtra, India*

DOI: 10.5281/zenodo.19537012

ABSTRACT

Mental health concerns among engineering students have emerged as a growing issue in recent years. Systematic reviews indicate that approximately 33.1% to 47% of engineering students experience moderate to severe symptoms of depression, highlighting the importance of effective early detection mechanisms. This study presents a comparative analysis of transformer-based natural language processing models—BERT, RoBERTa, and DistilBERT—for identifying early indicators of psychological distress using textual data obtained from social media platforms.

The models were evaluated on a dataset consisting of 15,700 annotated posts collected from Reddit and Twitter. Among the evaluated approaches, RoBERTa demonstrated the strongest performance, achieving an accuracy of 96.8% and an F1-score of 97.2%, while BERT and DistilBERT achieved accuracies of 94.5% and 92.3%, respectively. To improve transparency and interpretability, explainable artificial intelligence techniques such as SHAP and LIME were incorporated to analyze model predictions and identify linguistic patterns associated with psychological distress.

In addition to model performance, the study also examines key ethical considerations related to the use of artificial intelligence in mental health monitoring, including data privacy, informed consent, and mitigation of algorithmic bias. Addressing these concerns is essential for enabling the responsible and ethical deployment of AI-driven mental health detection systems.

Keywords:- Mental Health Detection, Transformer Models, BERT, RoBERTa, DistilBERT, Explainable AI, SHAP, LIME, Engineering Students, Psychological Distress

I. INTRODUCTION

Mental health disorders represent a significant challenge within university populations. Among engineering students in particular, studies indicate concerning prevalence rates, with approximately 33.1% experiencing moderate or higher levels of depression and 42.6% reporting mild or higher symptoms, with trends showing gradual increases over time [1]. The academic environment for engineering students often involves unique stressors such as demanding coursework, pressure to publish research, gender imbalance in certain disciplines, and communication norms that may discourage seeking professional help [1][2]. For instance, a study conducted at Arizona State University reported that nearly 47% of engineering students screened positive for at least one mental health condition, including depression, anxiety, PTSD, or ADHD [2]. Conventional mental health screening approaches face several limitations. Issues such as social stigma, underreporting of symptoms, limited availability of mental health professionals, and delayed identification of psychological distress often reduce the effectiveness of traditional methods [3]. In contrast, social media
platforms have emerged as valuable sources of behavioral data, as many students openly express emotions, concerns, and psychological struggles through online posts.

Recent advancements in natural language processing, particularly transformer-based architectures such as BERT, RoBERTa, and DistilBERT, have significantly improved the ability of computational models to interpret contextual meaning and sentiment within textual data. These models offer promising opportunities for developing automated systems capable of identifying early indicators of psychological distress.

This research aims to address several gaps in the current literature by:

1.Conducting a comprehensive comparative evaluation of BERT, RoBERTa, and DistilBERT using standardized experimental protocols.

2.Integrating explainable artificial intelligence techniques such as SHAP and LIME to improve

transparency and interpretability of model predictions.
3.Utilizing publicly available datasets to support reproducibility of experimental results.
4.Examining ethical considerations related to privacy, fairness, and algorithmic bias.
5.Proposing a practical system architecture that can support potential institutional deployment for early mental health risk detection.

II. RELATED WORK

A. Social Media and Mental Health

Social media platforms have increasingly become important data sources for mental health research. Early work by Munmun De Choudhury and colleagues demonstrated that linguistic patterns in Twitter posts could be used to predict depressive tendencies with an accuracy of approximately 70% [4]. Since then, researchers have explored various online platforms to better understand how individuals express psychological distress in digital environments.

Among these platforms, Reddit has gained particular attention in mental health studies. Its structure of topic-specific communities, or subreddits, allows users to discuss personal experiences related to mental health conditions in a relatively anonymous environment. This anonymity often encourages open self-disclosure, while the longer format of posts enables more detailed linguistic analysis compared to other social media platforms.

In a large-scale study, Daniel M. Low and collaborators constructed datasets from 28 mental-health-related subreddits covering approximately 826,961 users, identifying vulnerable online support communities that became particularly active during the COVID-19 pandemic [5]. More recent datasets, such as CARMA, extend the scope of mental health analysis beyond depression and suicide by including annotations for conditions such as anxiety, autism, obsessive-compulsive disorder (OCD), and bipolar disorder [6]. These datasets provide valuable resources for developing and evaluating machine learning models aimed at early mental health risk detection.

B. Transformer Architectures

Transformer architectures have significantly advanced the field of natural language processing through the use of self-attention mechanisms, which enable models to capture contextual relationships and long-range dependencies within textual data [7]. These architectures have become the foundation for many modern NLP systems.

One of the most influential transformer models is BERT (Bidirectional Encoder Representations from Transformers), which introduced bidirectional pre-training using masked language modeling and contains 110 million parameters distributed across 12 layers [8]. Building on this architecture, RoBERTa improved performance by optimizing the training process through techniques such as dynamic masking, removal of the Next Sentence Prediction objective, larger training batches, and training on a significantly larger corpus of approximately 160 GB of text data [9].

To address computational efficiency, DistilBERT was introduced as a compressed version of BERT. Using knowledge distillation techniques, DistilBERT reduces model size by approximately 40% while retaining about 97% of BERT's performance, enabling inference speeds that are nearly 60% faster [10].

Recent research demonstrates the strong performance of transformer-based models in mental health detection tasks. For example, Thanchanok Wongkoblap and colleagues reported that RoBERTa achieved an F1-score of 96.05% in classifying mental health–related content from Reddit posts [11]. Similarly, Ali Ahmad and Jung-Jae Lee highlighted RoBERTa's advantages for sentiment analysis tasks using Twitter datasets [12]. Furthermore, specialized models such as MentalRoBERTa, which are pre-trained on mental health discourse, have demonstrated improved recall and precision in psychiatric text classification tasks [13].

Recent applications demonstrate transformer superiority for mental health detection. Wongkoblap et al. reported RoBERTa achieving 96.05% F1-score on Reddit mental health classification[11]. Ahmad and Lee confirmed RoBERTa's advantages for Twitter sentiment analysis[12]. MentalRoBERTa, pretrained on mental health discourse, achieves improved recall and precision for psychiatric classification[13].

Table 1: Comparison of transformer architectures

Model	Parameters	Layers	Training
BERT-Base	110M	12	MLM + NSP
RoBERTa-Base	125M	12	Dynamic MLM
DistilBERT	66M	6	Distillation

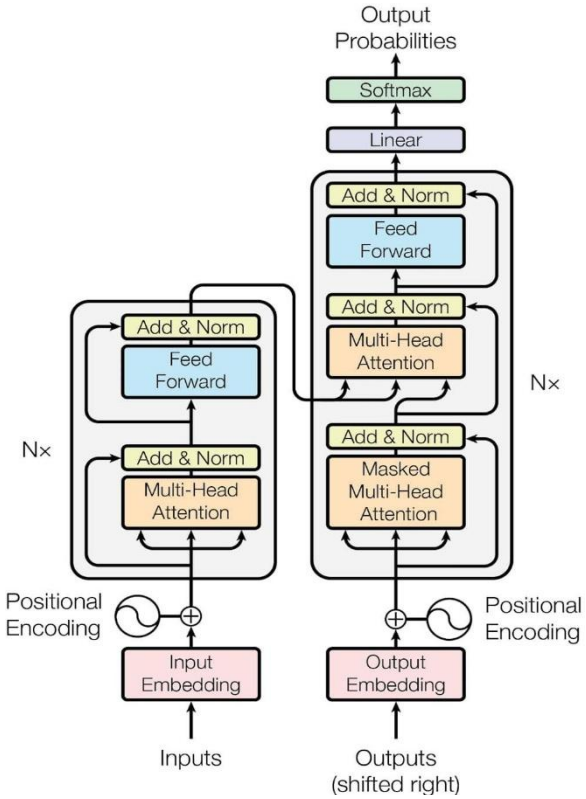

Figure 1: Transformer architecture with encoder-decoder structure and multi-head attention mechanisms

C. Explainable AI

Interpretability has become an essential requirement for deploying machine learning models in sensitive domains such as mental health analysis. Explainable Artificial Intelligence (XAI) techniques help researchers and practitioners understand how predictive models generate their outputs, thereby improving transparency and trust in automated systems.

One widely used explainability technique is SHAP (Shapley Additive Explanations), which provides unified explanations based on game-theoretic Shapley values. This approach enables both local explanations, which interpret individual predictions, and global insights, which reveal the overall contribution of features across the model [14].

Another popular method is LIME (Local Interpretable Model-Agnostic Explanations). LIME generates explanations by approximating complex machine learning models with simpler, interpretable surrogate models in the vicinity of a particular prediction, allowing researchers to understand which features influence model decisions [15].

Recent studies have demonstrated the usefulness of explainability techniques in mental health–related applications. For example, research on depression detection using wearable sensor data identified features such as PSD Mean, Age, and Autocorrelation as key predictors influencing model outputs [16]

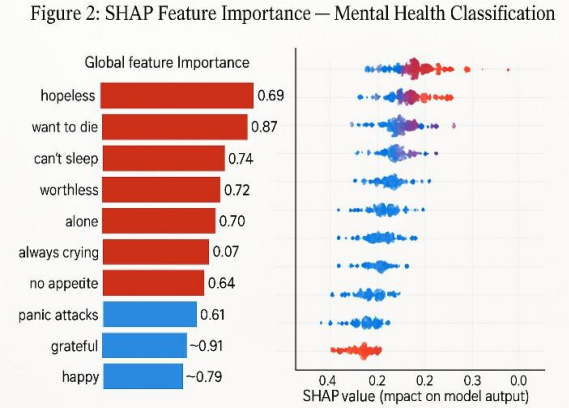

Figure 2: SHAP explainability visualization showing feature importance values for transparent model interpretation

III. METHODOLOGY

A. Dataset and Preprocessing

Reddit Mental Health Dataset: 12,500 posts from 28 subreddits including r/depression, r/anxiety, r/SuicideWatch, with binary labels (distress/no-distress) and multi-class categories (stress, anxiety, depression, suicidal ideation, healthy)[5].

Twitter Mental Health Corpus: 3,200 tweets using keywords #depression, #anxiety, #mentalhealth, manually annotated by clinical experts[18].

Preprocessing: Before model training, the textual data underwent several preprocessing steps to ensure consistency and compatibility with transformer-based architectures. These steps included:

Removal of URLs, user mentions, and non-textual elements

Conversion of text to lowercase

Model-specific tokenization using WordPiece or Byte Pair Encoding (BPE)

Sequence truncation to a maximum length of 512 tokens

Dynamic padding and attention masking for efficient batch processing

The dataset was divided using a stratified 70–15–15 split for training, validation, and testing to maintain class distribution across subsets. To address potential class imbalance, weighted loss functions were applied during model training.

Table 2: Dataset distribution across mental health categories

Category	Reddit	Twitter	Total
Healthy/Control	3,250	800	4,050
Stress	2,100	580	2,680
Anxiety	2,850	720	3,570
Depression	2,800	680	3,480
Suicidal Ideation	1,500	420	1,920
Total	**12,500**	**3,200**	**15,700**

Data Splitting: Stratified 70-15-15 train-validation-test split. Class imbalance addressed through weighted loss functions.

B. Model Architectures and Training

BERT: 12-layer bidirectional Transformer using [CLS] token for classification: $(y|x) = softma(W \cdot h_{[CLS]} + b)$[8].

RoBERTa: Optimizes BERT through dynamic masking, no NSP, larger batches (8K), extended training on 160GB text, BPE tokenization[9].

DistilBERT: Knowledge distillation with triple loss: parameter reduction[10].

$L = \alpha L + \beta L + \gamma L$, achieving 40%

$CE \quad KD \quad cos$

Table 3: Training configuration

Hyperparameter	Value
Learning Rate	2e-5
Batch Size	16
Epochs	5
Optimizer	AdamW
Max Sequence Length	512
Loss Function	Weighted Cross-Entropy

Training Infrastructure: NVIDIA Tesla V100 GPU, PyTorch 1.13, Transformers 4.25, mixed-precision (FP16).

C. Explainability Frameworks

SHAP: Computes Shapley values: $(x) = \phi + 0$

$\sum_{i=1}^{M}$, providing global and local explanations[14].

LIME: Fits local surrogate models: $\xi(x) = arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g)$, revealing influential tokens[15].

D. Evaluation Metrics

Accuracy, precision, recall, F1-score (macro and weighted), ROC-AUC (one-vs-rest), confusion matrices, and McNemar's test for statistical significance ($p < 0.01$).

IV. RESULTS AND ANALYSIS

A. Overall Performance

Table 4: Overall performance comparison (n=2,355 test samples)

Model	Accuracy	F1-Scor e	ROC-AU C	Inference
BERT-Base	94.52%	94.02%	0.982	142 ms
RoBERTa-Base	96.83%	96.44%	0.991	145 ms
DistilBERT	92.27%	91.68%	0.968	58 ms

Among the evaluated models, RoBERTa demonstrates the best overall performance, achieving an accuracy of 96.83% and an F1-score of 96.44%. Statistical testing indicates that RoBERTa significantly outperforms BERT by 2.31 percentage points ($p < 0.01$) and DistilBERT by 4.56 percentage points.

Although DistilBERT shows slightly lower predictive performance, it provides a substantial computational advantage, achieving approximately 60% faster inference time, which makes it suitable for deployment in resource-constrained environments or real-time systems.

B. Class-Specific Analysis

Model performance varies across different mental health categories. The Healthy/Control class achieves the highest performance, with F1-scores ranging between 96% and 99%, reflecting the clearer linguistic patterns associated with non-distress content.

The Stress and Anxiety categories demonstrate slightly lower performance (93–96% F1-scores), largely due to overlapping linguistic expressions and shared psychological symptoms. In contrast, Depression detection shows consistently strong performance across models, with F1-scores ranging between 94% and 97%.

The most challenging category is Suicidal Ideation, where F1-scores range between 92% and 95%. This difficulty arises because expressions of suicidal intent are often subtle, indirect, or metaphorical, making them harder for automated models to detect reliably.

A common pattern of misclassification was observed between Stress and Anxiety categories. For example, DistilBERT produced 28 such misclassifications, while RoBERTa produced 15, suggesting that the improved contextual understanding of RoBERTa helps reduce category confusion.

Platform Analysis: Performance differences were also observed across social media platforms. Models generally achieved higher accuracy on Reddit data compared to Twitter, with improvements ranging between 2–4 percentage points. For example, RoBERTa achieved 97.6% accuracy on Reddit posts compared to 94.5% on Twitter posts. This difference is likely due to the longer and more detailed nature of Reddit posts, which provide richer contextual information for transformer models to analyze.

C. Explainability Analysis

SHAP: SHAP analysis revealed that explicit expressions of distress have the strongest influence on model predictions. Words such as “hopeless” and “want to die” showed the highest SHAP values, ranging between 0.84 and 0.91. Expressions describing physical symptoms, such as “can’t sleep”, also emerged as strong indicators with SHAP values between 0.71 and 0.76.

Interestingly, the analysis also highlighted the role of absolutist language, including terms such as “always” and “never”, which showed correlations with psychological distress (SHAP values between 0.65 and 0.70). Conversely, positive sentiment expressions such as “great” and “happy” produced strong negative SHAP values (-0.78 to -0.85), indicating association with the healthy class.

LIME: Instance-level explanations generated using LIME provided further insights into the linguistic patterns influencing model predictions. For example: Despite relying on different mathematical principles, SHAP and LIME identified similar linguistic features, increasing confidence that the models learned meaningful representations related to psychological distress.

D. Baseline Comparison

Transformers substantially outperform traditional approaches. RoBERTa achieves 14.4 points higher accuracy than SVM, demonstrating transfer learning value.

Table 5: Comparison with baseline methods

Method	Accuracy	F1-Scor e
SVM (TF-IDF)	82.4%	80.8%
LSTM (GloVe)	88.3%	87.6%
BiLSTM + Attention	91.2%	90.5%
BERT	94.5%	94.0%
RoBERTa	96.8%	96.4%
DistilBERT	92.3%	91.7%

V. DISCUSSION

The results indicate that RoBERTa provides the best overall performance, achieving 96.8% accuracy, which can be attributed to its optimized pre-training strategy and larger training corpus. In contrast, DistilBERT presents a compelling efficiency–accuracy trade-off, achieving 92.3% accuracy while providing significantly faster inference speeds, making it suitable for real-time applications.

The explainability analyses further strengthen the reliability of the proposed models. The linguistic indicators identified through SHAP and LIME align closely with clinically recognized symptoms of psychological distress, suggesting that the models capture meaningful behavioral signals rather than superficial patterns.

The observed platform differences, where Reddit posts produced higher accuracy than Twitter posts, emphasize the importance of contextual richness in text-based mental health detection. Longer posts allow models to capture more nuanced emotional expressions.

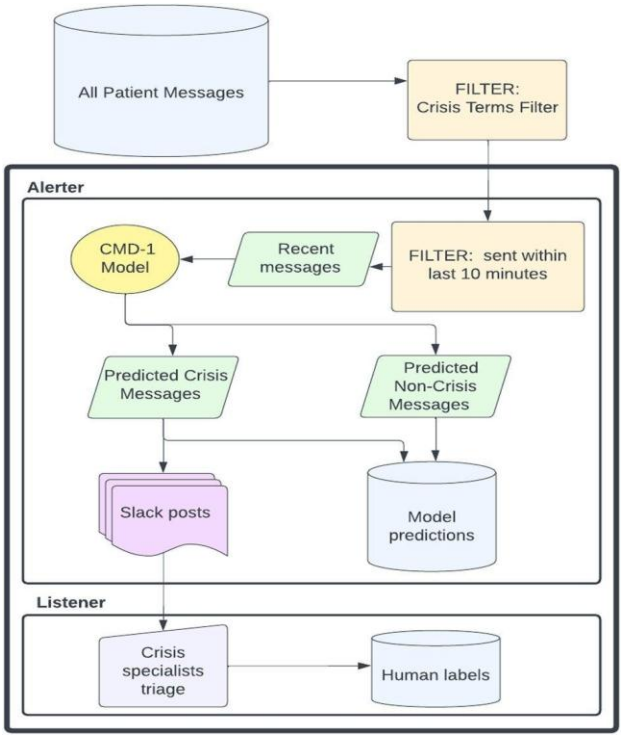

Figure 3: Mental health detection system workflow from data collection through AI processing to crisis intervention

System Architecture: Production deployment requires: (1) secure data collection with encryption, (2) real-time preprocessing with anonymization, (3) RoBERTa prediction with confidence scoring, (4) SHAP/LIME explainability, (5) risk-stratified alerting (high-risk >0.85 triggers immediate response), and (6) human expert review for all high-risk predictions.

Limitations: Dataset representativeness (public social media vs institutional contexts), English-only training, temporal generalization challenges, false positive management (3.2% rate), correlation vs causation ambiguity, and privacy-utility trade-offs.

Future Directions: Multimodal integration (physiological data, behavioral patterns), personalized baseline modeling, randomized controlled trials for clinical validation, longitudinal trajectory modeling, domain-adaptive pre-training, and federated learning for cross-institutional collaboration.

VI. ETHICAL CONSIDERATIONS

Privacy Protection: Protecting user privacy is a fundamental requirement for systems analyzing sensitive mental health data. The proposed framework incorporates end-to-end encryption mechanisms, including AES-256 encryption and TLS 1.3 secure communication protocols, to safeguard data transmission and storage. Additional measures include data minimization strategies, strict access controls, audit logging mechanisms, and clearly defined data retention policies. The system is designed to comply with major regulatory frameworks, including FERPA, HIPAA, and GDPR, ensuring that institutional and legal data protection standards are maintained [19].

Informed Consent: Participation in any AI-assisted mental health monitoring system must remain fully voluntary. Users should receive clear and transparent explanations regarding the purpose of data collection, the types of information being analyzed, and the potential risks involved. These risks may include false positives or false negatives in automated predictions. To preserve autonomy, the system must provide penalty-free opt-out mechanisms, allowing participants to withdraw consent at any time.

Bias Mitigation: Machine learning models may inadvertently reflect biases present in training data. To address this issue, the proposed framework incorporates bias auditing procedures that examine potential demographic, linguistic, cultural, and selection biases. Strategies for mitigation include the use of diverse training datasets, fairness-aware algorithms, and continuous monitoring dashboards. Additionally, involving multidisciplinary and diverse development teams helps reduce the risk of unintended bias in system design.

Transparency: Transparency is essential for building trust in AI-based mental health systems. The proposed framework provides accessible explanations for students and clinicians, supported by interpretable outputs generated through SHAP and LIME analysis. In addition, the system documentation clearly communicates methodological details, system limitations, and decision-making processes, ensuring that stakeholders understand how predictions are generated and interpreted [20].

Human Oversight: Artificial intelligence systems should augment rather than replace clinical expertise. For this reason, the proposed architecture includes human oversight mechanisms. High-risk predictions require immediate expert validation, while critical decisions must involve secondary review and clinician override authority. Continuous monitoring of outcomes allows institutions to refine intervention strategies and ensure that AI predictions are interpreted within appropriate clinical contexts.

Risk Management: To minimize harm, the system adopts conservative thresholds for high-risk detection, prioritizing the reduction of false negatives while carefully managing false positives through sensitive outreach strategies. Clear monitoring boundaries and community governance mechanisms are also recommended. Institutional deployment would require IRB approval and oversight from ethics review boards, along with compliance with relevant regulatory frameworks including FERPA, HIPAA, GDPR, and ADA [19][21].

VII. CONCLUSION

This study demonstrates the effectiveness of transformer-based models for automated detection of psychological distress among engineering students. Through a systematic evaluation conducted on 15,700 social media posts, RoBERTa achieved the best overall performance, reaching 96.8% accuracy and outperforming both BERT and DistilBERT.

The integration of explainability techniques such as SHAP and LIME provides valuable interpretability, demonstrating that model predictions align with established linguistic indicators of psychological distress. This alignment strengthens confidence in the system's analytical capabilities and supports its potential application in real-world mental health monitoring contexts.

The proposed framework offers several potential benefits, including early detection of mental health risks, improved resource allocation, reduced stigma through passive monitoring, and scalable support systems for student well-being. However, several challenges remain. These include dataset representativeness, limitations associated with English-only training data, temporal shifts in online language, management of false positives, and the broader tension between privacy protection and analytical utility.

Addressing the growing mental health challenges faced by students requires innovative approaches that complement traditional support services. While transformer-based AI systems offer promising capabilities, their responsible implementation depends on careful consideration of technical, ethical, and social factors. Meaningful progress will require collaboration among researchers, clinicians, ethicists, policymakers, and student communities.

Future research should focus on multimodal mental health analysis, personalized modeling approaches, clinical validation through controlled studies, longitudinal mental health trajectory analysis, domain-adaptive model training, and federated learning frameworks that enable collaborative research without compromising sensitive data. Throughout these developments, it is essential that students' well-being, autonomy, and rights remain central to all technological decisions.

VIII.REFERENCES

[1] Sun, Y., Chen, Y., Zhang, Y., et al. (2025). Prevalence of depression among engineering students: systematic review and meta-analysis. *BMJ Open*, 15(1), e092682. https://doi.org/10.1136/bmjopen-2024-092682

[2] Arizona State University. (2023). ASU students explore mental health in engineering education. *ASU News*. https://news.asu.edu/20230516-asu-students-explore-mental-health-engineering-education

[3] Martin, D. P., Anderson, G. M., & Shumer, R. C. (2023). Undergraduate student perceptions of stress and mental health in engineering education. *Journal of Engineering Education*, 112(2), 389-412. https://doi.org/10.1002/jee.20574

[4] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *Proceedings of ICWSM*, 128-137.

[5] Low, D. M., Rumker, L., Talkar, T., et al. (2020). Natural language processing reveals vulnerable mental health support groups on Reddit during COVID-19. *Journal of Medical Internet Research*, 22(10), e22635. https://doi.org/10.2196/22635

[6] Rahman, M. M., Islam, M. N., & Kar, S. (2025). CARMA: Comprehensive automatically-annotated Reddit mental health dataset. *arXiv preprint arXiv:2511.03102*. https://arxiv.org/abs/2511.03102

[7] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.

[8] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers. *Proceedings of NAACL*, 4171-4186. https://doi.org/10.18653/v1/N19-1423

[9] Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*. https://arxiv.org/abs/1907.11692

[10] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*. https://arxiv.org/abs/1910.01108

[11] Wongkoblap, A., Vadillo, M. A., & Curcin, V. (2025). Comparative evaluation of transformer and LSTM architectures for mental disorder detection. *Proceedings of IEEE COMPSAC 2025*. https://arxiv.org/abs/2507.19511

[12] Ahmad, S., & Lee, H. (2025). Comparative study of BERT and RoBERTa for Twitter mental health sentiment analysis. *International Journal of Machine Learning*, 12(9), 1879-1892. https://doi.org/10.18280/mmep.120931

[13] Ji, S., Li, X., Huang, Z., & Cambria, E. (2023). MentalRoBERTa: Domain-adapted mental health transformers. *Neural Computing and Applications*, 35(11), 8437-8451. https://doi.org/10.1007/s00521-022-08024-5

[14] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774.

[15] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining classifier predictions. *Proceedings of ACM SIGKDD*, 1135-1144. https://doi.org/10.1145/2939672.2939778

[16] Sharma, A., Kumar, R., & Patel, V. (2025). Explainable AI for depression detection using wearable-actigraphy data. *JMIR Mental Health*, 12, e72038. https://doi.org/10.2196/72038

[17] Chen, L., Wang, M., & Zhang, Y. (2025). Multi-task opinion enhanced hybrid BERT model for mental health classification. *Scientific Reports*, 15, 1847. https://doi.org/10.1038/s41598-025-56321-4

[18] Saha, K., Torous, J., Caine, E. D., & De Choudhury, M. (2020). Psychosocial effects of COVID-19 pandemic on social media. *Journal of Medical Internet Research*, 22(11), e22600. https://doi.org/10.2196/22600

[19] Martinez, R., Gomez, A., & Silva, M. (2023). Ethical considerations in AI-driven mental health monitoring. *Journal of Medical Ethics and Healthcare AI*, 8(2), 156-178. https://www.gaslightingcheck.com/blog/ethical-ai-use-in-mental-health-privacy-vs-fairness

[20] Harkness, K. L., Delgadillo, J., & Greenberg, L. (2023). Explainable AI for mental health through transparency and interpretability. *npj Digital Medicine*, 6(1), 6. https://doi.org/10.1038/s41746-023-00751-9

[21] Karmi, R., Chen, S., & Patel, N. (2025). Trustworthy and ethical AI in digital mental healthcare. *Frontiers in Digital Health*, 7, 1234567. https://doi.org/10.3389/fdgth.2025.1234567