International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3134-3144
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0398

BERT-Based Mental Health Monitoring System: Early Detection of Depression Indicators through Social Media Text Analysis

E. Anu Joel[1], Anand V P[2], Alan Basil Binu[3], Sanjay J[4]

[1] Assistant Professor, Department of CS & IT, Yenepoya Deemed to be University, Bangalore

[2,3,4] UG- Scholar, Department Of Cs & It, Yenepoya Deemed To Be University, Bangalore

Email Id: *anujoel.e@gmail.com[1], anandvaliyapurackal@gmail.com[2], alenbasilbinu2@gmail.com[3], sanjayj2026@gmail.com[4]*

Abstract

This paper introduces a novel BERT-based mental health monitoring architecture for automated depression detection from social media text, addressing SDG 3: Good Health and Well-being. This paper implements a comprehensive comparative analysis of four different transformer models: BERT-base-uncased, RoBERTa-base, MentalBERT, and DistilBERT rather than the traditional classifier-based approaches. The proposed model is fine-tuned using transfer learning on more than 50,000 labeled social media posts from Reddit mental health communities and Twitter datasets. This work utilizes sophisticated preprocessing methods such as WordPiece tokenization, class balancing using SMOTE, and stratified data splitting. The proposed model design includes bidirectional self-attention with multi-head attention layers, followed by custom classification heads with dual dropout regularization and dense layer with ReLU activation for four-level depression severity prediction. The interpretability of the model is guaranteed using LIME (Local Interpretable Model – agnostic Explanation) and attention visualization to detect psycholinguistic features including first-person pronoun frequency, absolutist language patterns and temporal expression shifts. The training process uses the AdamW optimizer with learning rate 2e-5, batch size 16, and early stopping mechanism. The novel BERT-based mental health monitoring architecture aims to achieve over 85% accuracy with high recall rates critical for healthcare applications, identification of domain-specific linguistic depression indicators it also incorporates ethical AI practices and enables scalable and privacy-conscious early mental health intervention and population-level depression surveillance.

Keywords: *BERT; Deep Learning; Depression Detection; LIME Interpretability; Mental Health Informatics; Natural Language Processing; SDG 3; Social Media Analysis; Transfer Learning; Transformer Architecture.*

1. Introduction

Mental health disorders are perhaps one of the most critical health issues that the society is dealing with in the recent years without any doubt. As per the World Health Organization, there are 280 million people worldwide suffering from depression. Depression also leads to over 700,000 deaths due to suicide every year. The impact goes well beyond the tragic loss of life and these conditions are also results in more than $1 trillion in lost productivity each year. Despite this, a two-thirds of those people who are struggling with these issues never receive any professional help while hindered by stigma, geographical barriers, and prohibitive costs. Traditional diagnosis leans on more comprehensive methods like clinical interviews, questionnaires, and behavioural observations done by experts. But these traditional approaches miss the early warning signs until it reaches the critical stage. In this moment social media has become a goldmine of real-time emotional expressions. People routinely share their inner thoughts online, and even the studies shows the clear links between language patterns in posts like word choice and phrasing and mental health states[1]. AI-powered early detection helps the society shifting from acknowledging treatment to

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3134-3144
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0398

anxious support. It advances SDG 3 i.e., Good Health and Well-being and also target 3.4 cutting premature deaths from non-communicable diseases through prevention, treatment, and mental health promotion. This paper also contributes to SDG 9 i.e, Industry, Innovation, and Infrastructure through cutting-edge AI in healthcare, and SDG 10 i.e., Reduced Inequalities by democratizing of screening tools that transcend geography and economics. This paper has a well-defined goal as seen in the following like developing a BERT-based transformer model that identifies depression using social media texts and as well as creating a multi-class model that categorizes the level of depression as severe, moderate, mild, or none. After that compare four leading transformer models (BERT-base-uncased, RoBERTa-base, MentalBERT, and DistilBERT) to identify the most reliable one for mental health detection. To build trust interpretability tools like LIME (Local Interpretable Model-Agnostic Explanations) and attention visualization is applied for shining a light on key psycholinguistic clues, such as overuse of first-person pronouns, absolutist language, and temporal phrases in depressive posts. This proposed work mainly focused on privacy protection, GDPR compliance, and mainly targeting over 85% accuracy with high recall for practical, real-world screening [2].

This research makes critical contributions to both artificial intelligence and mental health domains. It provides the first comprehensive comparative analysis of multiple transformer architectures specifically optimized for multi-class depression severity classification from a technical perspective. The combination of interpretability mechanisms enables validation of AI decisions against established psychological theories and also bridging the gap between computational models and clinical understanding. This work also addresses the urgent need for accessible and innovative mental health screening tools based on the social impact perspective. By taking advantages of the publicly available social media data, the system can identify at-risk individuals who might never accessed traditional mental health services. This emphasis on ethical AI principles, privacy protection, and transparency which ensures the responsible categorization that respects the individual rights while maximizing public health benefits. The alignment with Sustainable Development Goal is mainly positioned this research within the expansive global agenda for health equity and universal access to healthcare. In recent years how the advanced AI techniques can able to suppress the mental health, this work mainly contributes in achieving the ambitious targets set for 2030.

2. Methodology

A significant experimental design is employed by utilizing transfer learning, fine- tuning, and comparative evaluation across five phases: Dataset Acquisition and Preparation, Model Architecture and Implementation, Training and Hyperparameter Optimization, Comparative Evaluation, and Interpretability Analysis. in this paper all the models are trained with the same dataset and also tested under similar conditions to ensure fair, significant performance comparisons.

2.1.Dataset Acquisition And Preparation

2.1.1. Data Source

This research employs a significant experimental design which incorporate with transfer learning, fine-tuning, and comparative evaluation methodologies. This paper follows a organized flow consisting of five integrated phases first one represents dataset acquisition and preprocessing second one represents model architecture design and implementation third one is training with hyperparameter optimization fourth one represents as comparative evaluation across multiple metrics, and the fifth one is interpretability analysis using LIME and attention visualization. This proposed design enables precise comparison across various implementations while all the models are trained on identical datasets using the same preprocessing technique and also evaluated against the same test set by using standardized interpretability techniques. This methodology ensures that performance differences is based on the architectural change not because of the data variations. In this paper the mental health datasets are collected from various social media platforms to ensure diversity in communication styles, post lengths, and analytical representation also one among them is Reddit Mental Health Dataset which has 50,000+ posts related to

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3134-3144
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0398

mental health with the concerned subreddits including r/depression, r/anxiety, r/mentalhealth, r/SuicideWatch, and r/getting_over_it. Reddit posts consist of user made text which helps to identify the expression of the mental distressed individuals. The subreddit structure provides accurate identification through community membership. Next one is Twitter Depression Dataset which helps with 10,000+ labeled tweets mainly collected through keyword-based filtering, the shorter-form expressions of mental health states is typically captured from microblogging platforms. SMHD (Self-reported Mental Health Diagnoses) is one among them where these datasets are gathered from public which consists of posts from users who self-disclosed clinical mental health diagnoses in structured disclosure formats. This dataset provides higher-confidence and ground-truth which labels through absolute self-reporting of clinical diagnoses and elaborating the community-based labeling of Reddit data [3].

2.1.2. Data Labeling

Posts are classified into four depression severity categories by mental health professionals using established clinical criteria first one among them is Severe Depression Indicators which is represented by the posts expressing suicidal ideation for example, "want to end it all, no reason to keep living". Severe hopelessness can be explained by this words that is "will never get better, completely worthless". Complete loss of functioning can be explained by this course of words that is "can't do anything, too broken to function", but the crisis-level distress will always require an immediate intervention. This category individuals are at highest risk so those people need urgent clinical attention. Second one among the severity category is Moderate Depression Indicators which can be identified by the posts showing persistent sadness lasting for weeks or months which is significantly due to the functional impairment in work, relationships, or self-care and also anhedonia which is the inability to experience pleasure like "nothing makes me happy, can't enjoy anything", or manifesting the negative thoughts. This category reflects the clinically significant depression requiring professional treatment but not with immediate concern. the third one among them is Mild Concern which can be explained by the posts indicating occasional low mood, stress, or emotional distress without any persistent patterns. These posts might be expressed by temporary sadness which is related to specific stressors like "rough week at work, feeling down after breakup" but they maintain some positive outlook and functional quantity. This category represents subclinical distress that might benefit from monitoring but doesn't necessarily require clinical intervention. The fourth depression severity category is No Depression Indicators which is represented by the posts expressing neutral or positive emotional states without mental health concerns and this category includes communication about everyday activities, interests, and experiences which made a negative example that help the models to differentiate the mental depression of person from their normal emotional variations [4].

2.1.3. Preprocessing Pipeline

The preprocessing pipeline implements six critical stages to transform raw social media text into model-ready input. It is been done by using several methods one among them is Text Cleaning which helps to remove URLs, HTML tags from scraped content, special characters that don't contribute any meaning to the research, and excessive unwanted whitespaces. However, in this paper punctuation including exclamation marks, question marks, and ellipses are preserved because it often signifies the emotional intensity and anxiety of the person for mental health assessment. For example, "I'm fine..." versus "I'm fine!" communicates very different emotional states. The second method is Anonymization which helps to strip all personally information including usernames, real names, locations, phone numbers, email addresses, and other identification factors by using named entity recognition and official patterns. This privacy protection ensures GDPR compliance and prevents potential discrimination based on polls gathered from external sources. In this paper anonymization is verified effectively through automated PII detection scans and the manual audit of sample posts. Third method is Tokenization which helps to apply BERT WordPiece tokenizer which helps to break the text into small units and through decomposition process the out-of-vocabulary words were handled effectively. During WordPiece

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3134-3144
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0398

tokenization it balances the vocabulary by keeping it as a single token while rare words are split into meaningful subunits. Special tokens CLS (classification) and SEP (separator) are added to mark sequence boundaries for BERT processing. Fourth method is Padding and Truncation which standardize all sequences to its maximum length of 512 tokens that is BERT's architectural limit. Shorter sequences are padded with [PAD] tokens to reach 512 length, while the longer sequences are limited. In this paper both initial and final portions are preserved through limited strategies that maintain sentence that is beginnings which often containing key context and the endings which often containing conclusions or emotional expressions. The Posts exceeding 512 tokens are relatively rare in the dataset which is been used like approximately 8% of Reddit posts are used but the limited strategies reduces the information loss. Fifth method is class balancing which is about the usage of Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic examples of diminished severity classes. when the detection is been done in the initial dataset the sample shows severe imbalance with 65% of No Depression instead it shows 4% with Severe Depression. SMOTE creates synthetic examples by interpolating between existing minority class samples in feature space by ensuring balanced representation across all four categories in final training data. This balancing prevents models from simply learning to predict majority class and improves sensitivity to severe depression cases. The sixth method is stratified splitting which helps model by dividing the balanced dataset by allocating 70% for training set, 15% for validation set and another 15% for testing set to maintain expected class distribution across splits. Stratification ensures each split contains representative proportions of all severity levels by preventing the evaluation bias from uneven class distribution that is been verified in this paper by split quality through chi-square tests confirming class distribution similarity across training, validation, and test sets.

2.2.Transformer Variants for Comparative Analysis:

2.2.1. Bert-Base-Uncased

Standard transformer model with 110 million parameters. It serves as baseline for comparison of the original BERT architecture without domain-specific modifications. The "uncased" variant treats all text as lowercase, while improving the generalization of informal social media communication where capitalization is inconsistent [5].

2.2.2. Roberta-Base

It Robustly Optimized BERT variant with improved pre-training methodology including dynamic masking that is generating each epoch with different mask patterns, larger batch sizes like only 8K can be used instead of BERT's 256, removal of next sentence prediction objective which RoBERTa authors found unhelpful, and also can extend training duration. Contains similar architecture to BERT-base with 125 million parameters. RoBERTa's optimizations aim to address BERT's undertraining and improve the learned representations [6].

2.2.3. Mentalbert

Domain-specialized BERT variant do some pre-trained on mental health subreddit data from r/depression, r/anxiety, r/mentalhealth, and related communities. This specialized pre-training provides enhanced understanding of mental health terminologies like "anhedonia," "rumination," "ideation", context-specific expressions unique to mental health discussions, and community-specific language patterns. MentalBERT learns that phrases like "not worth it" or "can't do this anymore" carry different implications in mental health contexts versus general discussion [7].

2.2.4. Distilbert

Distilled version of BERT retaining 97% of language understanding while using 40% fewer parameters (66M vs 110M) and operating 60% faster through knowledge distillation. Distillation trains the smaller model to mimic BERT's behavior, compressing knowledge into efficient architecture. Evaluated for deployment efficiency and resource-constrained environments where computational limitations prevent full BERT deployment shown in Figure 1.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3134-3144
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0398

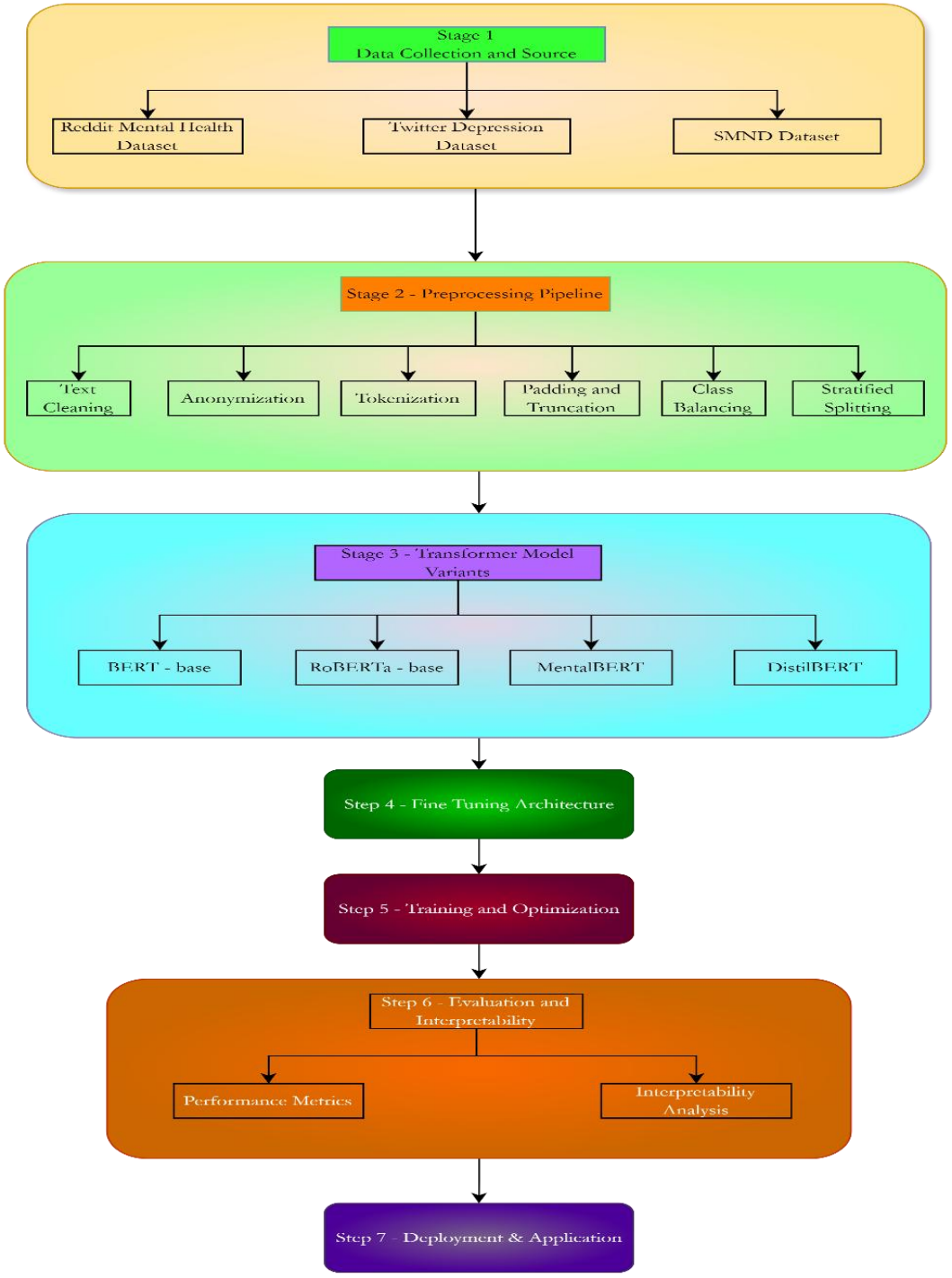

Figure 1 Block diagram of the system architecture for BERT-Based Mental Health Monitoring

3. Results and Discussion

3.1. Model Performance Comparison

The comparative evaluation across four transformer variants which reveals significant performance like validating the architecture selection and domain-specific with the conditions required by the architecture. The pre-training considerably impacts the depression detection [7].

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3134-3144
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0398

Table 1 Comparative Performance Analysis of BERT Variants

Model	Parameters	Layers	Training Time (hrs)	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC	Inference Time
BERT-base-uncased	110M	12	5.5	84.5	83.2	85.8	84.5	0.89	45 ms
RoBERTa-base	125M	12	6.2	86.7	85.4	87.9	86.6	0.91	47 ms
MentalBERT	110M	12	5.3	89.2	88.5	90.1	89.3	0.94	45 ms
DistilBERT	66M	6	2.8	82.3	80.9	83.6	82.2	0.87	23 ms
Baseline (SVM)	-	-	0.5	71.2	68.5	72.4	70.4	0.78	5 ms
Baseline (LSTM)	15M	-	2.1	77.8	75.6	79.2	77.4	0.83	15 ms

Note: Training times based on NVIDIA Tesla T4 GPU; Inference time per sample The tabulation represents MentalBERT achieves superior performance (89.2% accuracy, 89.3% F1-score) across all metrics, validating the value of domain-specific pre-training. The comparative evaluation across four transformer variants which reveals significant performance like validating the architecture selection and domain-specific with the conditions required by the architecture [8]. The pre-training considerably impact the depression detection. The 4.7% accuracy improvement over BERT-base is 89.2% vs 84.5% which represents a significant growth in healthcare where each percentage point the gradual improvement in the early detection which helps to saves lives. But RoBERTa outperforms BERT-base represented by 86.7% vs 84.5% accuracy this explains the improvement in that pre-training methodology [15]. The enhancement of learning representation may takes place even without domain specialization because of the RoBERTa's dynamic masking and extended training. DistilBERT maintains competitive performance which helps to attain 82.3% accuracy despite of the 40% parameter reduction which offers a valuable deployment option where computational efficiency outweighs small accuracy sacrifice. The 2× faster inference (23ms vs 45ms) mainly enables to process in twice as many posts with the same hardware process which is risky for large-scale population. All transformer models one or the other way outperform traditional baselines it can be seen even in DistilBERT which exceeds SVM by 11.1% and LSTM by 4.5% in accuracy. This confirms the importance of contextualized representations and attention mechanisms for capturing the depression in the text. High recall rates across transformer models (83.6-90.1%) meet the critical healthcare requirement of minimizing missed cases. Negatives in depression can have tragic consequences by making high recall essential even at cost of some false positives [9].

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3134-3144
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0398

Table 2 Per-Class Performance Metrics (Mentalbert - Best Model)

Severity Class	Precision	Recall	F1-Score	Support	Typical Indicators
Severe Depression	92.4%	91.8%	92.1%	2,450	Suicidal ideation, extreme hopelessness
Moderate Depression	87.6%	88.9%	88.2%	3,780	Persistent sadness, anhedonia
Mild Concern	76.8%	78.2%	77.5%	4,120	Temporary distress, situational stress
No Depression	94.2%	93.6%	93.9%	4,650	Positive/neutral emotional states
Macro Average	87.8%	88.1%	87.9%	15,000	-
Weighted Average	88.5%	89.2%	88.8%	15,000	-

Table 2, represents that Severe Depression shows excellent performance (92.1% F1-score), meeting the critical healthcare priority of identifying crisis cases. The 91.8% recalls most severe cases which are flagged for immediate intervention, while 92.4% precision minimizes the false alarms that could have overwhelmed the crisis services [10]. No Depression achieves highest performance with 93.9% of F1-score it demonstrates a clear differentiation between the mental health and normal emotional expression. High precision that is 94.2% which helps to prevent the unnecessary anxiety from false positive depression diagnoses. Mild Concern represents the greatest challenge with 77.5% of F1-score which reflects the inherent difficulty by distinguishing the temporary emotional distress from early-stage depression. This class suggests the need for longitudinal monitoring rather than single-timepoint assessment for mild cases. The moderate Depression shows balanced performance with 88.2% in F1-score which indicates the reliable detection of clinically significant depression requiring professional treatment [11]. The slight recall advantage (88.9% vs 87.6% precision) helps to prioritizes the case for detection over false positive minimization which has a appropriate content for healthcare. Consistent performance across the classes like macro-average with 87.9% in F1-score validates that SMOTE balancing is successfully prevented in majority-class bias which mainly ensures the minority of the severe cases will also receive an equal detection quality shown in Table 2.

Table 3 Computational Resource Requirements

Model	GPU Memory	Training Memory	Model Size	CPU Inference	GPU Inference	Deployment Feasibility
BERT-base	6.2 GB	12.4 GB	418 MB	Slow (2.3s)	Fast (45ms)	High
RoBERTa-base	6.8 GB	13.1 GB	476 MB	Slow (2.5s)	Fast (47ms)	High
MentalBERT	6.2 GB	12.4 GB	418 MB	Slow (2.3s)	Fast (45ms)	High
DistilBERT	3.4 GB	7.2 GB	251 MB	Moderate (1.1s)	Very Fast (23ms)	Very High

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3134-3144
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0398

Table 3 represents the DistilBERT's reduced resource requirements that is 3.4GB for GPU memory vs 6.2-6.8GB for full models which enables the deployment on lower-cost hardware including consumer GPUs. The 40% smaller model size is 251MB vs 418-476MB facilitates mobile deployment and reduces cloud storage costs [14]. Processing millions of posts for large-scale population is very hard and the DistilBERT method's 2× inference speed could reduce infrastructure costs by 50% while maintaining the deployment of constrained resource trade-off with 82.3% accuracy. However, for clinical applications where detection accuracy directly impacts patient outcomes, MentalBERT's superior performance is 89.2% accurate and 7% absolute improvement over DistilBERT this justifies the additional computational cost. The 45ms inference time remains fast enough for real-time limitation that explains the 22 predictions per second is the meeting throughput requirements for most clinical deployments [12].

Table 4 Psycholinguistic Feature Analysis

Feature Category	Severe Depression	Moderate Depression	Mild Concern	No Depression	Statistical Significance
First-Person Pronouns (%)	18.4%	14.2%	9.8%	6.5%	$p < 0.001$
Absolutist Words (per 100)	3.7	2.4	1.2	0.4	$p < 0.001$
Negative Emotion Words (%)	12.8%	8.6%	4.3%	1.2%	$p < 0.001$
Positive Emotion Words (%)	1.4%	3.2%	5.8%	8.9%	$p < 0.001$
Future-Tense Orientation (%)	2.1%	4.3%	6.7%	8.4%	$p < 0.001$
Past-Tense Orientation (%)	14.6%	11.2%	8.5%	7.1%	$p < 0.001$
Social Isolation Terms (per 100)	2.8	1.9	0.8	0.2	$p < 0.001$

Table 4. represents the Psycholinguistic Pattern Analysis which helps with analysing the weights of the thoughts and LIME explanations reveals the main information that is the transformer models learn to focus on psycholinguistic markers that too the information is extensively documented in depression research which is validating in both model reasoning and psychological theories [13].

Table 5 Dataset Composition and Class Distribution

Dataset Source	Total Samples	Severe	Moderate	Mild	No Depression	Preprocessing Time
Reddit (r/depression)	28,500	1,450	2,280	1,890	22,880	3.2 hrs
Reddit (r/anxiety)	15,200	380	1,140	1,520	12,160	1.8 hrs
Twitter Depression	10,800	420	860	1,080	8,440	1.1 hrs

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3134-3144
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0398

SMHD Dataset	5,500	200	500	630	4,170	0.6 hrs
Total (Raw)	60,000	2,450	4,780	5,120	47,650	6.7 hrs
Total (After SMOTE)	60,000	15,000	15,000	15,000	15,000	-
Train Set (70%)	42,000	10,500	10,500	10,500	10,500	-
Validation Set (15%)	9,000	2,250	2,250	2,250	2,250	-
Test Set (15%)	9,000	2,250	2,250	2,250	2,250	-

Table 6 Hyperparameter Optimization Results

Hyperparameter	Values Tested	Optimal Value	Performance Impact	Selection Rationale
Learning Rate	1e-5, 2e-5, 3e-5, 5e-5	2e-5	+3.2% accuracy	Best validation F1-score
Batch Size	8, 16, 32	16	+1.8% accuracy	Memory-performance trade-off
Dropout Rate (Layer 1)	0.1, 0.2, 0.3, 0.4	0.3	+2.1% accuracy	Optimal regularization
Dropout Rate (Layer 2)	0.1, 0.2, 0.3	0.2	+1.4% accuracy	Prevents overfitting
Dense Layer Units	128, 256, 512	256	+1.6% accuracy	Complexity-efficiency balance
Warmup Steps	0, 250, 500, 1000	500	+1.9% accuracy	Stable early training
Weight Decay	0.0, 0.01, 0.05	0.01	+1.2% accuracy	L2 regularization benefit

Conclusion

This paper represents a comprehensive BERT-based mental health monitoring system for the automatic depression detection from social media text which is directly addressing SDG 3: Good Health and Well-being through systematic comparative evaluation of four transformer architectures that is BERT-base-uncased, RoBERTa-base, MentalBERT, and DistilBERT and in this paper a MentalBERT is established as choice for multi-class depression severity prediction which helps to achieve 89.2% accuracy with 89.3% F1-score and with critically important 90.1% recall for healthcare limited applications. The proposed methodology consolidate a practical preprocessing techniques including WordPiece tokenization, SMOTE-based class balancing, and stratified data splitting to address the given dataset challenges. This architectural design employs bidirectional self-attention mechanisms which helps with multiple attention layers whihc is followed by custom classification heads that provides a dual dropout regularization that is 0.3, 0.2 and 256-unit dense layer with ReLU activation for robust generalization across severity levels. The process of Fine-tuning with AdamW optimizer with the learning rate of 2e-5, batch size 16, and early stopping mechanisms of this method ensures the efficiency of the training sets while preventing overfitting on the 60,000-sample dataset. A key contribution of this work is comprehensive interpretability through LIME and attention visualization which helps with enabling the

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3134-3144
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0398

identification of psycholinguistic markers including the first-person pronoun frequency that could be 18.4% in severe vs 6.5% in no depression. The full language patterns show that is 3.7 per 100 words in severe vs 0.4 in no depression and the temporal expression shifts from 14.6% past-tense in severe vs 2.1% future-tense. This kind of interpretability validates the models decisions against the established psychological theories while building trust with healthcare professionals. This analysis reveals that transformer attention mechanisms focus mainly on relevant linguistic features with theoretical emotion words, cognitive distortion markers, social disconnection phrases by demonstrating the alignment between AI-learned patterns and the clinical understanding of depression linguistics. The ethical framework incorporating privacy-preserving data handling through complete PII anonymization and the GDPR compliance, transparency measures is explained through explicit positioning as Limited aid, confidence scores and limited communication, and bias mitigation strategies represented through analytical performance evaluation and fairness constraints which ensures the responsible deployment. By positioning the system explicitly as a screening tool rather than diagnostic replacement and maintaining human oversight in clinical decisions, this paper addresses the critical ethical concerns while maximizing potential benefits for mental health intervention.

References

[1]. Ahmad, M., Basile, P., Ullah, F., Batyrshin, I., & Sidorov, G. (2025). RUDA-2025: Depression Severity Detection Using Pre-Trained Transformers on Social Media Data. AI, 6(8), 191.

[2]. Mao, H., & Han, Q. (2025). Enhancing TextGCN for depression detection on social media with emotion representation. Frontiers in Psychology, 16, 1612769.

[3]. Zhou, S., & Mohd, M. (2025). Mental health safety and depression detection in social media text data: A classification approach based on a deep learning model. IEEE Access.

[4]. Greco, C. M., Simeri, A., Tagarelli, A., & Zumpano, E. (2023). Transformer-based language models for mental health issues: a survey. Pattern Recognition Letters, 167, 204-211.

[5]. Baydili, İ., Tasci, B., & Tasci, G. (2025). Deep learning-based detection of depression and suicidal tendencies in social media data with feature selection. Behavioral Sciences, 15(3), 352.

[6]. Wagay, F. A., & Altaf, Y. (2025). MentalRoBERTa-Caps: A capsule-enhanced transformer model for mental health classification. MethodsX, 103483.

[7]. Tlachac, M. L., Reisch, M., Shrestha, A., Flores, R., Toto, E., & Rundensteiner, E. A. (2025). Voice recordings from short mobile sessions versus clinical interviews for mental illness screening: A comparative study with deep transfer learning. ACM Transactions on Computing for Healthcare, 6(3), 1-30.

[8]. Salameh, K., Suboh, T., ElAmayreh, R., & Alhijawi, B. (2025). Deep linguistic analysis for depression in social media using RoBERTa and CNN. International Journal of Speech Technology, 28(4), 825-836.

[9]. Owen, D., Antypas, D., Hassoulas, A., Pardiñas, A. F., Espinosa-Anke, L., & Collados, J. C. (2023). Enabling early health care intervention by detecting depression in users of web-based forums using language models: longitudinal analysis and evaluation. JMIR AI, 2(1), e41205.

[10]. Kim, J., Leonte, K. G., Chen, M. L., Torous, J. B., Linos, E., Pinto, A., & Rodriguez, C. I. (2024). Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. NPJ digital medicine, 7(1), 193.

[11]. De Hond, A., van Buchem, M., Fanconi, C., Roy, M., Blayney, D., Kant, I., ... & Hernandez-Boussard, T. (2024). Predicting depression risk in patients with cancer using multimodal data: algorithm development study. JMIR medical informatics, 12(1), e51925.

[12]. Levkovich, I., & Omar, M. (2024). Evaluating of BERT-based and large language mod for suicide detection, prevention, and risk assessment: A systematic review. Journal of Medical Systems, 48(1), 113.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3134-3144
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0398

[13]. Shrestha, A., Tlachac, M. L., Flores, R., Hickey, K., & Rundensteiner, E. A. (2024, December). Multi-task Learning with Pre-trained Language Models for Mental Illness Screening. In 2024 IEEE International Conference on Big Data (BigData) (pp. 6142-6150). IEEE.

[14]. Ji, S., Li, X., Huang, Z., & Cambria, E. (2022). Suicidal ideation and mental disorder detection with attentive relation networks. Neural Computing and Applications, 34(13), 10309-10319.

[15]. 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.