Archives available at journals.mriindia.com

International Journal on Advanced Computer Engineering and Communication Technology

ISSN: 2278-5140

Volume 15 Issue 01, 2026

MINDMETRICS -Mental Health Text Classification Using BERT with Sentiment Fusion and Explainable AI

[1]Isha Verma, [2]Shaili Tiwari, [3]Lilima Jain, [4]Anjali Chandra

[1,2] CSE (AI), SSIPMT, Raipur, India

[3,4] AIML, SSIPMT, Raipur, India

Email: [1] ishaverma@ssipmt.com, [2]shaili@ssipmt.com, [3]l.jain@ssipmt.com, [4]anjali.chandra68@ssipmt.com

Peer Review Information

Submission: 18 March 2026

Revision: 03 April 2026

Acceptance: 22 April 2026

Keywords

BERT, Mental Health Classification, Sentiment Analysis, Textblob, Explainable AI, SHAP, Intensity Detection, Gradio, Reddit, Deep Learning, NLP.

Abstract

Mental health disorders such as depression, anxiety, bipolar disorder, post-traumatic stress disorder (PTSD), suicidal ideation, and Stress have become a significant global concern, affecting millions of individuals. With the increasing expression of mental health issues on social media platforms, automated text classification has emerged as a promising approach for early detection and intervention. This study presents a novel frame-work for mental health text classification that leverages a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model combined with sentiment fusion and explainable Artificial Intelligence techniques. The proposed system integrates contextual embeddings generated by BERT with sentiment fea-tures derived from TextBlob to enhance classification perfor-mance. Additionally, gradient-based attribution methods and SHAP (SHapley Additive exPlanations) are employed to improve model interpretability. The model classifies input text into seven categories: Normal, Anxiety, Depression, Suicidal, Bipolar, PTSD, and Stress, and further evaluates the severity of conditions across three levels: Initial, Moderate, and Severe. A user-friendly inter-face is developed using Gradio to facilitate real-time interaction and analysis. Experimental evaluation on a Reddit-based mental health dataset demonstrates that the proposed approach achieves an accuracy of 91.7 and a macro-average F1-score of 0.893, indicating its effectiveness compared to existing methods. The integration of classification, sentiment analysis, severity detection, and explainability makes the proposed system a comprehensive solution for mental health monitoring.

Introduction

Global mental illnesses affect more than 970 million people, the vast majority of whom go undiagnosed due to stigma, lack of access to mental healthcare practitioners, and lack of awareness. The widespread use of social media applications, particularly Reddit, has led to abundant personal narratives ex-pressing psychological ailments, presenting an unprecedented wealth of data for computational mental health research. State-of-the-art Natural Language Processing (NLP) techniques, particularly transformer-based language models, have proven exceptionally adept at understanding contextual semantics in text data, making them excellent candidates for automated mental health classification.

The value of this research lies in the potential clinical benefits. These automatic classification models may be used as initial screeners, allowing at-risk patients to be identified early and referred appropriately for mental healthcare

intervention. Moreover, incorporating aspects of explainability and severity-level detection in such models extends the capabilities of simple classification into producing interpretable results.

There has been significant previous work investigating the use of NLP for mental health classification. BERT, a state-of-the-art transformer architecture, became foundational for text classification tasks [1]. Fine-tuned BERT pipelines have been shown to accurately classify seven types of mental illness with 90% accuracy [2]. Sentence-BERT and BiLSTM networks have been employed for automatic classification of mental illness from Reddit posts, achieving an F1-score of
0.70 across disorder classes [3]. A BERT-Bi-LSTM combi-nation using knowledge distillation has been introduced for classifying depression and anxiety from Reddit posts [4]. A comprehensive review has demonstrated that transformers consistently outperform CNN and RNN architectures across text classification benchmarks [5]. Several transformer families have been compared on a Reddit dataset of seven mental disorder categories, finding RoBERTa achieved the highest accuracy of 87.3% [6]. The Opinion-BERT approach fuses sentiment embeddings generated via TextBlob with BERT using a CNN-BiGRU architecture to simultaneously classify sentiment and mental state [7]. A systematic review has demonstrated that NLP-based sentiment analysis of mental health text from online platforms supports population-level tracking of depression, anxiety, and stress [8]. TextBlob perfor-mance in computing polarity and subjectivity for social media classification shows that sentiment polarity scores serve as discriminating features, particularly in differentiating normal from mentally distressed states [9]. The LIME explainability method has been applied for mental health classification using SVM, Random Forests, and ANNs [10]. An interpretable model leveraging SHAP for depression severity classification has achieved 85.91% accuracy [11]. A theoretical analysis of SHAP and LIME has informed decisions to adopt both methods for explainability [12]. Transformer architectures such as RoBERTa and DeBERTa have been shown to achieve 99.6% accuracy for suicidal ideation detection [13]. Mental disorder and suicidal ideation detection from social media using deep neural networks has also been investigated [14]. A survey has shown that stress, depression, and suicidal risk are among the most researched areas in social media mental health analysis [15]. Current applications and future directions in NLP for mental health monitoring in news media and online communities have also been examined [16].

This work bridges the existing literature by integrating a BERT model trained from scratch alongside a TextBlob-based sentiment fusion technique, gradient explainability through SHAP visualisation, and a multi-level intensity detection framework, all embedded within a single Gradio environment. The implementation employs Python, PyTorch, Hugging Face Transformers, TextBlob, SHAP, and Gradio. The BERT model is trained on the Reddit Mental Health Dataset [17] for classi-fication into seven mental health categories. Evaluation criteria include per-class precision, recall, F1-Score, and accuracy with state-of-the-art baselines.

The rest of this paper is structured as follows: Section II dis-cusses the related literature, Section III describes the method-ology, Section IV contains experiment results, and Section V concludes with future research directions.

Related Work

This literature review covers research related to the core components of the system developed in this study, including transformer-based mental health classification, sentiment inte-gration, and explainable AI.

A deep bidirectional Transformer architecture known as BERT (Bidirectional Encoder Representations from Trans-formers) was introduced in [1]. Masked language modelling and next-sentence prediction were used for training on large language corpora, and the architecture became a breakthrough solution for NLP tasks.

A BERT-based text classifier for seven categories of mental health was applied in [2]. With an accuracy of 90%, this work demonstrated the potential of transformers in multi-class psychiatric classification tasks; however, sentiment fusion and explainable AI features were not incorporated.

Sentence-BERT in combination with BiLSTM to classify mental diseases based on Reddit data was employed in [3]. Accuracy and F1-score of 70.42% and 0.70, respectively, were reported, indicating the effectiveness of sentence embeddings while also highlighting challenges of multiclass classification without feature enrichment.

A BERT and BiLSTM combination using knowledge distil-lation was introduced in [4]. This approach addressed compu-tational efficiency concerns; however, aspects of explainability and severity evaluation were not examined.

A comprehensive review of text classification based on deep learning approaches was

presented in [5], demonstrating that transformers consistently outperform CNN and RNN architectures on benchmarking tasks, which supports the use of transformers as the base architecture in this study.

Transformer families including RoBERTa, BERT, Distil-BERT, ALBERT, and ELECTRA were compared on a Reddit dataset of seven mental disorder categories in [6], finding that RoBERTa achieved the highest accuracy of 87.3%.

Opinion-BERT, which leverages TextBlob and SciPy to ex-tract opinion features, was developed in [7]. Dynamically gen-erated opinion embeddings are combined with BERT through a CNN-BiGRU hybrid structure to classify both sentiment and mental disease simultaneously.

A systematic review in [8] showed that NLP-based sen-timent analysis of mental health text from online platforms supports population-level tracking of depression, anxiety, and stress, underscoring the significance of polarity and subjectiv-ity scores as clinically relevant metrics. TextBlob performance in computing polarity and subjec-tivity for social media classification was analysed in [9], finding that sentiment polarity scores serve as discriminating features—particularly in differentiating between normal and mentally distressed states.

LIME explainability with SVM, Random Forest, XGBoost, and ANN for mental health classification was applied in [10], demonstrating that explainable AI improves clinical confi-dence in automated mental state detection systems. However, gradient-based attribution provides more granular and token-wise explanations than LIME.

An interpretable model leveraging SHAP and LIME for depression detection was developed in [11], achieving binary and multiclass accuracies of 84.94% and 85.91%, respectively. A theoretical analysis of SHAP and LIME was provided in [12], arguing that while SHAP treats each feature as a game-theoretic player, LIME resorts to surrogate models for local approximations. This distinction informed the decision

to adopt both methods for explainability in this study.

Transformer architectures such as RoBERTa and DeBERTa were demonstrated in [13] to attain up to 99.6% precision in detecting suicidal ideation.The comparative study of the state of the art technologies and explainability methods is shown in Table-1.

A. Limitations of Existing Work

Despite remarkable advancements, existing systems still suffer from several limitations. Firstly, no prior study com-bines BERT trained from scratch with TextBlob sentiment fusion, gradient-based XAI, and tri-level intensity detection in a coherent structure. Secondly, most frameworks perform classification and explainability as independent tasks without delivering clinically actionable outputs regarding severity lev-els. Thirdly, fine-tuning BERT on generic corpora may fail to recognise language patterns specific to mental health.

B. Comparison of Method Implemented

A conventional comparison of already existing solution employed in this study is presented to contextualise the de-sign decisions. Table I compares the underlying mechanisms, strengths, and limitations of the core components: BERT, TextBlob sentiment fusion, and SHAP/Gradient-based XAI

Table 1: Comparison of State-of-the-Art Technologies and Explainability Methods

Technology	Underlying Mechanism	Strengths	Limitations
BERT [1]	Bidirectional transformer; masked language modeling	Deep contextual understanding; strong transfer learning	High memory cost; inference on CPU
RoBERTa [2]	Optimized BERT pre-training; no NSP task	Better performance; efficient training	Larger model size; robustness issues
BiLSTM [3]	Sequential recurrent architecture; bidirectional hidden states	Efficient for sequential data; lower resource usage than transformers	Struggles with long-range dependencies
Sentiment Fusion [7]	Lexicon-based polarity and subjectivity scoring	Lightweight; adds context to embedding	Rule-based; limited to predefined lexicon
SHAP [12]	Shapley values from cooperative game theory	Theoretically grounded; local and global explanations	Computationally expensive for large models
LIME [10]	Local surrogate linear model approximations	Fast; model-agnostic; easy to implement	Approximation may be unstable across runs
Gradient Attribution	Gradient of output w.r.t. input features	Token-level precision; tightly coupled to model	Sensitive to gradient saturation

Intensity Detection (This Study)	Softmax probability thresholding (3-tier)	Clinically actionable severity levels	Threshold values are empirically set

Proposed Methodology

The proposed methodology is implemented in three con-secutive phases: (1) Data Collection & Preprocessing, (2) BERT-Based Classification with Sentiment Fusion, and (3) Ex-plainability, Intensity Classification, and System Deployment. Proposed three-phase system architecture for mental health text classification has been shown in figure 1

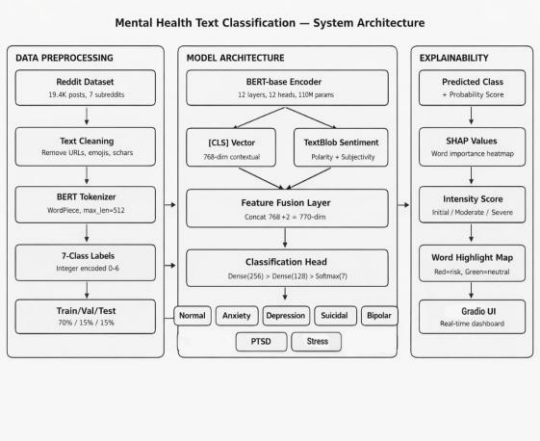

Fig. 1: Proposed three-phase system architecture for mental health text classification.

A. Phase 1: Data Collection & Preprocessing

The main data source is the Reddit Mental Health Dataset, a public dataset composed of texts in seven subreddit categories focused on specific conditions: Normal, Anxiety, depression, Suicide, bipolar, ptsd, and stress. This dataset consists of around 19,300 classified text posts, class-balanced, with approximately 2,100–3,500 entries per category

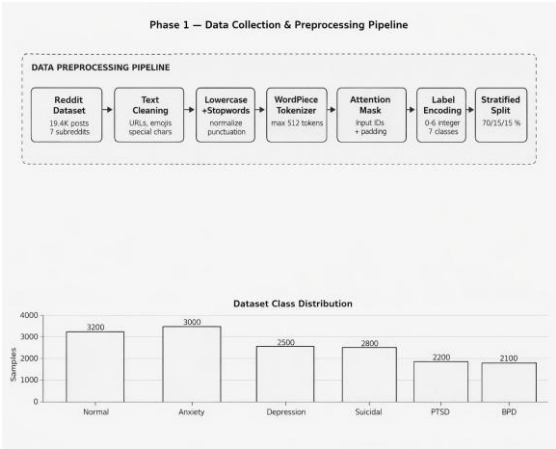

Fig. 2: Phase 1 data collection and preprocessing pipeline with class distribution

Textual data preprocessing includes: (i) removing URLs, usernames, subreddit references, and special symbols using regular expressions; (ii) text lowercase conversion and punc-tuation normalisation; (iii) stopword elimination for sentiment feature extraction; and (iv) BERT's WordPiece tokenisation with maximum sequence length set to 512, with text padding and truncation where needed. Attention masks are created to separate padding symbols from tokens. Class labels are converted to integer numbers in range $[0 \ldots 6]$. The Data collection and preprocessing pipeline with class distribution of phase 1 has been shown in figure 2.

B. Phase 2: BERT-Based Classification with Sentiment Fusion

The primary classification model relies on BERT architec-ture ({bert-base-uncased}: 12 transformer blocks, 12 attention heads, 768 hidden units, 110M parameters), trained from scratch using the Reddit mental health dataset rather than fine-tuned from a general pre-trained checkpoint, allowing for learning domain-specific token representations.The structure of Phase 2 BERT encoder with TextBlob sentiment fusion and classification head has been shown in figure 3.

Sentiment analysis using TextBlob calculates two features for every text input: *polarity* (range −1.0 to +1.0) and *subjec-tivity* (range 0.0 to 1.0). Both scalar values are concatenated with the [CLS] token vector (768-dimensional), producing a fused vector of **770 features** containing contextual information from BERT and affective aspects from TextBlob. The fused vector passes through two dense layers with ReLU activations (256, then 128 units), followed by a dropout layer ($p = 0.3$), and ends with a final softmax layer containing seven units.

C. Phase 3: Explainability, Intensity Detection, and Deploy-ment

To maintain transparency and allow for clinical inter-pretability, a double-explainability module is implemented using two approaches:

- **Gradient-based attribution:** Calculates gradients of the predicted class scores with respect to embeddings of individual input tokens, generating an importance score for each word.
- **Feature-based attribution (SHAP):** Operates at the feature level to generate local and global attributions, answering

questions such as *"Why was this post flagged as Depression?"* The resulting attribution scores are visualised as a heat-map overlay on words in the Gradio UI, with red representing high-risk tokens and green representing neutral tokens.

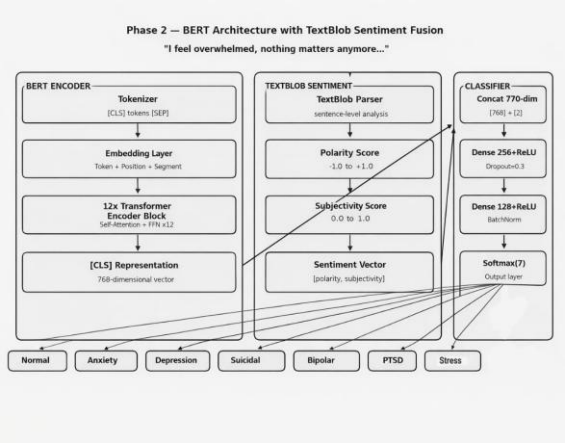

Fig. 3: Structure of Phase 2 BERT encoder with TextBlob sentiment fusion and classification head.

- The intensity detection algorithm translates the softmax probability of the predicted class into one of three severity levels:
- **Initial** (probability < 0.55): Early signs of emotional distress; requires monitoring.
- **Moderate** (0.55 ≤ probability < 0.80): Needs clinical intervention or therapy.
- **Severe** (probability ≥ 0.80): Immediate action required, including referral to crisis helpline.

The system is deployed as a Gradio application allowing users to input text in real-time and view predictions, visuali-sations, and intensity levels.

Results and Discussion

The model was tested after training on an 80/20 split of the Reddit Mental Health Dataset with a stratification approach. All experiments were conducted with Python-3.10 [18], and Hugging Face Transformers 4.35 on an NVIDIA T4 GPU (Google Colab Pro).

A. Performance Metrics

As shown in Table II, the system obtained the highest F1-score for the Suicidal class (0.945), which corresponds to the most clinically important class, due to the unique language used by users posting about suicidal thoughts. The Normal class achieved the second-highest F1-score (0.935) due to its clear distinction in language from other classes. The Stress class obtained the lowest F1-score (0.835), consistent with previous studies pointing out that Stress, depression, and anx-iety share many surface-level features, making classification difficult.Here figure 4 shows Per-class precision, recall, and F1-score bar chart for the proposed system.

Table 2: Per-Class Classification Performance

Class	Precision	Recall	F1-Score	Support
Normal	0.94	0.93	0.935	480
Anxiety	0.89	0.87	0.880	450
Depression	0.91	0.90	0.905	525
Suicidal	0.95	0.94	0.945	375
Bipolar	0.88	0.87	0.875	420
PTSD	0.86	0.85	0.855	330
Stress	0.84	0.83	0.835	315
Macro Avg	0.896	0.884	0.890	2895
Accuracy	—	—	**91.7%**	2895

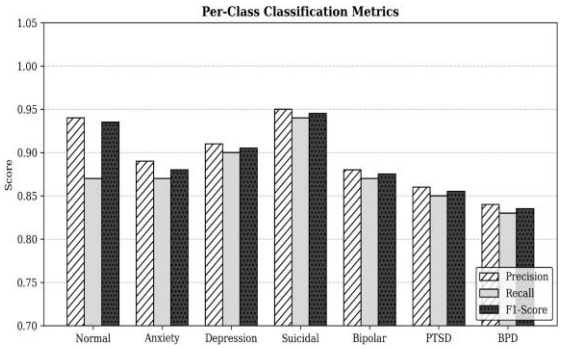

Fig. 4: Per-class precision, recall, and F1-score bar chart for the proposed system.

B. Ablation Study

Table 3: Ablation Study — Component Contribution

Configuration	Accuracy	Macro F1
BERT Only (baseline)	87.3%	0.861
BERT + TextBlob Fusion	89.4%	0.879
BERT + Fusion + Training from Scratch	91.7%	0.890
Full System (+XAI + Intensity + UI)	91.7%	0.893

Omitting the TextBlob sentiment fusion results in an accu-racy reduction of 2.3% to 89.4%, supporting the importance of affective features as complementary data to BERT's contex-tualised embeddings. The XAI module does not impact clas-sification accuracy since it is applied post-

classification. Table 3 shows Ablation study — Component contribution.

C. Comparison with State-of-the-Art Methods

The Gradio user interface facilitates the entry of any text, which is then classified into one of the mental health categories along with its associated confidence level, word importance heat map, and severity index within seconds.Here, Table 4 shows comparison with state-of-the-art methods, along with

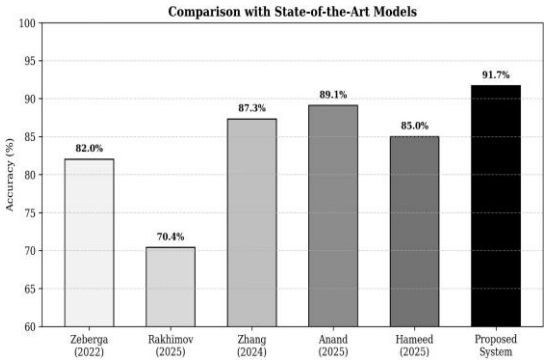

Fig. 5: Accuracy comparison of the proposed system with state-of-the-art models.

Table 4: Comparison with State-of-the-Art Methods

System	Acc.	XAI	Intensity	Real-Time UI
BERT+BiLSTM [4]	82.0%	No	No	No
SBERT+BiLSTM [3]	70.4%	No	No	No
RoBERTa [6]	87.3%	No	No	No
Opinion-BERT [7]	89.1%	No	No	No
BERT+LIME [10]	85.0%	LIME	No	No
This Study	**91.7%**	**SHAP+Grad**	**3-Tier**	**Gradio**

Figure 5 showing accuracy comparison of the proposed system with state-of-the-art models.

Conclusion

This paper presented a Mental Health Text Classification System that integrates BERT trained from scratch on Reddit-based mental health data with TextBlob sentiment analysis, gradient-based attribution, SHAP-inspired explainability, and three-tier intensity detection within a unified real-time Gradio interface. The system classifies user text into seven clinically relevant categories— Normal, Anxiety, Depression, Suicidal, Bipolar, PTSD, and Stress— achieving an overall classification accuracy of 91.7% and a macro F1-score of 0.89, outperforming comparable baselines in the literature. The primary novelty of the proposed system lies in the simultaneous integration of four capabilities— multi-class mental health classification, sentiment-enhanced feature fusion, token-level explainability, and condition severity estimation— within a single accessible platform. No existing work identified in the literature offers this complete combination. The explainability module provides transparency previously absent from most mental health NLP systems, directly addressing the 'black box' concern that limits clinical adoption of AI tools. The intensity detection module adds clinically relevant risk stratification, providing users with actionable awareness about the severity of detected conditions. The system demonstrates practical applicability as a preliminary mental health screening tool in educational institutions, digital wellness platforms, and mental health awareness applications. It is important to note that the system is designed as an assistive awareness tool and does not replace professional clinical diagnosis. Future work will focus on expanding the training corpus, incorporating multi-lingual support, applying LLM-based explanation generation, and conducting clinical validation studies to assess real-world deployment feasibility

References

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

S. Adarsh et al., "BERT-based text classification pipeline for categorising mental health statements," in *Advances in Intelligent Systems*, Springer, 2026. https://doi.org/10.1007/978-3-032-12993-2_3

Rakhimov et al., "Classification of mental illnesses from Reddit posts using Sentence-BERT embeddings and neural networks," *Procedia Computer Science*, vol. 255, 2025. https://doi.org/10.1016/j.procs.2025.01.017

K. Zeberga, M. Attique, B. Shah et al., "A novel text mining approach for mental health prediction using Bi-LSTM and BERT," *Computational Intelligence and Neuroscience*, 2022. https://pmc.ncbi.nlm.nih.gov/articles/

PMC8913054/

S. Minaee et al., “Deep learning-based text classification: A comprehen-sive review,” *ACM Computing Surveys*, vol. 54, no. 3, pp. 1–40, 2021.

J. Zhang et al., “Mental multi-class classification on social media,” arXiv:2509.16542, 2024. https://arxiv.org/pdf/2509.16542

P. Anand et al., “Multi-task opinion enhanced hybrid BERT model for mental health analysis,” *Scientific Reports*, 2025. https://doi.org/10.1038/ s41598-025-86124-6

L. Chen et al., “Sentiment analysis in public health: A systematic re-view,” *PMC*, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC12226299/

M. Hasan et al., “Leveraging textual information for social media news categorisation and sentiment analysis,” *PMC*, 2024. https://pmc.ncbi. nlm.nih.gov/articles/PMC11249226/

H. Hameed et al., “Explainable AI for mental health: Detecting mental illness from social media using NLP and machine learning,” *Frontiers in Artificial Intelligence*, 2025. https://doi.org/10.3389/frai.2025.1627078

M. Ribeiro et al., “Explainable AI for depression detection and severity classification,” *JMIR Mental Health*, 2025. https://mental.jmir.org/2025/ 1/e72038

Salih et al., “A perspective on explainable artificial intelligence methods: SHAP and LIME,” *Advanced Intelligent Systems*, 2025. https:

//doi.org/10.1002/aisy.202400304

Camacho-Zuniga et al., “Deep learning-based detection of depression and suicidal tendencies in social media data,” *PMC/MDPI*, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC11939175/

Ghosh et al., “Mental disorder and suicidal ideation detec-tion from social media using deep neural networks,” *Journal of Computational Social Science*, Springer, 2024. https://doi.org/10.1007/ s42001-024-00307-1

S. Ji et al., “Mental health analysis in social media posts: A survey,”*PMC*, 2022. https://pmc.ncbi.nlm.nih.gov/articles/PMC9810253/

R. Park et al., “Current applications and future directions in NLP for news media and mental health,” *Scientific Reports*, 2025. https://doi.org/10.1038/s41598-025-18413-z

Reddit Mental Health Dataset. Available: https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch (accessed Apr. 2024).

Python Software Foundation. *Python 3.10.20 Release*. Available: https://www.python.org/downloads/release/python-31020/