DOI: 10.2478/ijasitels-2025-0008

Achieving Clinical Reliability in Suicide Risk Detection: A Low-Resource Benchmark of RoBERTa vs. DistilBERT

Stefania-Eliza Berghia[1]*, Adrian Barglazan*[1]

[1]Computer Science and Electrical and Electronics Engineering Department, Faculty of Engineering, "Lucian Blaga" University of Sibiu, Romania
{stefania.berghia, adrian.barglazan} @ulbsibiu.ro

Abstract

Detecting suicidal ideation in social media text is a critical public health objective, demanding high-accuracy, deployable models. This study addresses the challenge of achieving clinical reliability within severe hardware constraints. We conduct a comparative fine-tuning benchmark of two Transformer models, DistilBERT and RoBERTa, for binary classification of suicidal risk. The models were optimized on a balanced, 10,000-sample subset of the Reddit Suicide Detection Dataset under CPU-only, low-resource constraints. Prediction robustness is achieved through a simple selection process that chooses the output with the highest confidence score from the two models. RoBERTa achieved a peak F1-Score of 97.4% and 97.40% accuracy, substantially outperforming DistilBERT (94.1% F1-Score). Crucially for safety-critical applications, error analysis confirmed RoBERTa's ethical superiority by achieving a 40% reduction in False Negatives compared to DistilBERT. The validated framework establishes that high-performance, ethically robust risk detection is feasible under resource limitations, enabling the safe integration of the final classification system with a constrained LLaMA 3 conversational module for proactive support.

Keywords: suicidal ideation detection, transformer models, clinical NLP

1 Introduction

Suicide represents a major global public health concern, with early identification of suicidal ideation remaining a persistent challenge. Many individuals who experience severe psychological distress do not actively seek clinical support, yet they often express early warning signs in digital environments. Social media platforms such as Reddit provide an open, large-scale stream of unsolicited textual data, which has contributed to the development of digital phenotyping approaches for understanding mental states from online behaviour. Language expressed in communities like r/SuicideWatch offers valuable indicators of emotional crises, while general subreddits provide contrasting baseline content.

These characteristics create an opportunity to develop non-intrusive, scalable, language-based systems capable of detecting risk signals in real time. However, effective deployment in practice requires models that are not only accurate but also computationally accessible and clinically reliable, particularly in low-resource settings where specialized hardware is not available.

This study addresses these challenges by providing two key contributions. First, it presents a rigorous low-resource benchmark comparing two prominent Transformer-based language models, Distilled Bidirectional Encoder Representations from Transformers (DistilBERT) and Robustly Optimized BERT Pretraining Approach (RoBERTa), for suicidal ideation detection using a balanced 10,000-sample subset of the Suicide Detection Dataset. Second, it examines model errors with particular attention to false negatives, the most consequential error type in suicide-risk detection, as these represent missed high-risk cases. Finally, the work demonstrates a practical application through a web-based detect-and-support pipeline that integrates the hybrid model with an empathetic conversational agent based on the Large Language Model Meta AI, version 3, 8-billion parameters (LLaMA3-8B).

2 Related work

Transformer-based architectures form the foundation of modern text classification. [1] introduced the Bidirectional Encoder Representations from Transformers (BERT) model, which leverages masked language modeling and bidirectional self-attention to capture contextual dependencies. Building on this foundation, [2] proposed DistilBERT, a compressed version that maintains most of BERT’s performance while reducing inference cost. RoBERTa was introduced in [3] and improves BERT by applying dynamic masking, removing the Next Sentence Prediction objective, and scaling training with larger batch sizes and corpora.

Transformers have also enabled progress in multimodal mental health detection. [4] presented the Multimodal Hierarchical Attention (MHA) model, which integrates Convolutional Neural Network (CNN)-based textual features, ResNet-18 visual representations, temporal metadata, and emotion lexicon signals using intra- and inter-modality attention mechanisms. Evaluated on a clinically validated Weibo dataset of 10,000 users, the model achieved 89.7% accuracy, outperforming Term Frequency–Inverse Document Frequency (TF-IDF) and Support Vector Machine (SVM), CNN, Bidirectional Long Short-Term Memory (BiLSTM), Bidirectional Encoder Representations from Transformers (BERT), and RoBERTa.

A related multimodal approach was introduced in [5] through the Emotion-based Reinforcement Attention Network (ERAN). ERAN uses a TextCNN emotion extractor, a BiLSTM encoder, and a reinforcement-learning agent that selects emotionally salient posts. Evaluated on the MDD dataset, it obtained 90.6% accuracy, outperforming Naive Bayes, SVM, Long Short-Term Memory (LSTM), BiLSTM-Attention, BERT-base, and RoBERTa-base. Ablation experiments confirmed the importance of both the emotion extractor and the reinforcement-learning–based post-selection mechanism.

Complementing these models, [6] introduced the Multimodal Hierarchical Attention Graph Convolutional Network with Vision Transformer (MHA-GCN_ViT), which integrates Graph Convolutional Networks (GCN) for modeling EEG inter-channel dependencies with a Vision Transformer (ViT) for spectral audio features. On the MODMA dataset, MHA-GCN_ViT achieved 89.03% accuracy, outperforming SVM, CNN, and earlier multimodal frameworks.

Research specifically addressing suicidal ideation detection is more limited. [7] compared multiple Transformer-based models on annotated Reddit posts, reporting that RoBERTa achieved 90.5% accuracy and an 89.3% F1-score, outperforming LSTM-based architectures through improved long-range contextual modeling. A complementary approach is presented in [8], which combines Linguistic Inquiry and Word Count (LIWC) psycholinguistic features with word embeddings and Extreme Gradient Boosting (XGBoost). While effective, feature-engineered pipelines require extensive manual preprocessing relative to Transformer-based models, which learn contextual features directly from text.

3 Proposed methodology

The proposed system for suicidal ideation detection is built around a supervised text-classification framework that combines modern Transformer architectures with a lightweight hybrid decision mechanism. The methodological pipeline comprises four stages: dataset preparation, text preprocessing, model fine-tuning, and the integration of a confidence-based hybrid inference strategy designed to enhance prediction robustness in low-resource environments.

3.1 Dataset description

This study utilizes the Suicide Detection Dataset [9], a publicly available corpus on the Kaggle platform. The dataset was curated by crawling posts from the Reddit social media platform. Its composition is binary, consisting of posts from two distinct subreddits:

- r/SuicideWatch: A community providing support for individuals in suicidal crises. Posts from this subreddit are labeled as 'suicide'.
- r/teenagers: A general-interest subreddit for teenagers to discuss daily life. Posts from this subreddit are labeled as 'non-suicide'.

The original dataset contains 232,074 posts, equally balanced between the two classes.

3.2 Corpus curation for low-resource environments

Training large Transformer models on a dataset of over 230,000 samples requires significant computational resources, specifically high-VRAM GPUs. A primary objective of this study was to develop a high-performance model under significant hardware constraints (a laptop with an Intel Core i7-1065G7 CPU, 8 GB RAM, and no dedicated GPU).

To this end, a representative subset of 10,000 samples was randomly selected from the full dataset. This approach serves a dual purpose: first, it makes the experiment feasible on the available hardware, and second, it provides a valuable, reproducible benchmark for other researchers working in similarly low-resource environments. The use of random_state=42 during sampling ensures that this exact subset can be replicated for future studies.

After selection and cleaning, the final subset comprised 9,996 posts. The random sampling process preserved the near-perfect class balance of the original corpus, as detailed in Table 1.

Table 1. Class distribution within the selected dataset

Class	Samples	Percentage	Source
Suicidal	5021	50.25%	Subreddit SuicideWatch

Non-suicidal	4975	49.75%	Subreddit teenagers
Total	9996	100%	Reddit via Pushshift API

3.3 Text preprocessing

Prior to fine-tuning the models, each Reddit post underwent a standardized preprocessing sequence designed to reduce noise while preserving semantically relevant content. Text normalization involved lowercasing all characters, removing hyperlinks, email addresses, and platform-specific tokens such as user mentions or hashtags. Superfluous whitespace and repeated characters were trimmed, and non-essential symbols were filtered to produce cleaner textual input.

After normalization, posts were converted into model-ready representations using the tokenizers associated with each model. DistilBERT employs the WordPiece tokenizer inherited from BERT, whereas RoBERTa uses a Byte-Pair Encoding (BPE) tokenizer. Both tokenizers generated input_ids and attention_mask tensors, with sequences truncated or padded to a maximum length of 32 tokens, an empirically determined value based on the dataset's typical post lengths.

3.4 Transformer architectures

The models used in this study are based on the Transformer encoder architecture introduced by Vaswani et al. [4], which replaces recurrence and convolution with a multi-head self-attention mechanism capable of modeling long-range dependencies in parallel. This design is particularly suitable for analyzing user-generated content on social media, where emotional cues may appear at any point in the text and where conventional grammatical patterns are frequently absent. Because the suicidal ideation detection task is a sequence-classification problem, only the encoder component of the Transformer is required.

The Transformer encoder is composed of stacked layers, each containing a multi-head self-attention sublayer followed by a position-wise feed-forward network. These layers are surrounded by residual connections and layer normalization. During fine-tuning for classification, a special prepended token—[CLS] for BERT-based models and <s> for RoBERTa—is used to summarize the meaning of the entire input sequence. The final hidden state corresponding to this token serves as the input to the classification head.

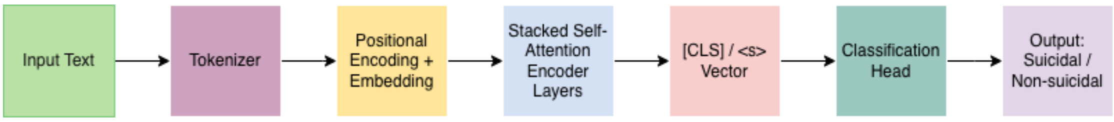
Input Text
Tokenizer
Positional Encoding + Embedding
Stacked Self-Attention Encoder Layers
[CLS] / <s> Vector
Classification Head
Output: Suicidal / Non-suicidal

Figure 1. Transformer fine-tuning workflow

3.4.1 Model A: Fine-Tuned DistilBERT

The first model evaluated in this study is DistilBERT [2], a compressed version of BERT that reduces the depth of the encoder to six layers and totals approximately 66 million parameters. This compact architecture is well suited for CPU-only environments, offering substantially faster inference while retaining most of BERT's semantic modeling capabilities.

To adapt DistilBERT for binary classification, a dropout layer and a linear transformation nn.Linear(768, 2) were added on top of the [CLS] token representation.

The model was fine-tuned using the AdamW optimizer with a learning rate (LR) of 5e-5 for two epochs, following a linear warm-up schedule. A batch size of eight samples was selected as the largest configuration compatible with CPU memory constraints.
This training configuration ensures that the evaluation reflects DistilBERT's representational capacity under realistic low-resource conditions.

3.4.2 Model B: Fine-Tuned RoBERTa

The second model is the RoBERTa [3], a Transformer encoder with twelve layers and approximately 125 million parameters. RoBERTa differs from BERT-style models through several pretraining refinements, such as dynamic masking, removal of the Next Sentence Prediction objective, and the use of larger corpora, which together yield more robust contextual embeddings.
RoBERTa was fine-tuned using the same classification head as DistilBERT—an nn.Dropout layer followed by nn.Linear(768, 2)—to ensure that any performance differences arise strictly from the encoder representations rather than classifier complexity. The model was trained using the AdamW optimizer with a LR of 2e-5 for four epochs, also employing a linear warm-up scheduler.
This lower LR reflects RoBERTa's greater parameter count and results in more stable convergence during CPU-based fine-tuning.

3.4.3 Inference-time prediction selection

Although both DistilBERT and RoBERTa achieve strong results, their strengths differ: DistilBERT offers faster, more efficient inference, while RoBERTa provides richer contextual representations. To take advantage of both in the deployed system, a lightweight inference-time strategy is employed, designed to choose the prediction from the model that is most certain about its own result.
This selection strategy does not rely on traditional ensemble methods such as voting, averaging, or training an additional meta-classifier. Instead, both models process the input text in parallel, and each produces a probability distribution over the two classes. The mechanism identifies the highest probability value (i.e., the model's internal confidence) from each distribution. The final label is assigned by simply selecting the prediction from the model with the larger confidence score.
This approach adds only a single comparison after the standard forward passes of the two models, introducing virtually no computational overhead. Despite its simplicity, this selection method provides a practical robustness advantage by dynamically choosing between a faster model and a deeper, more expressive one on a case-by-case basis. This allows the system to flexibly exploit DistilBERT's efficiency and RoBERTa's representational power without additional training or architectural modifications.

4 Experiments and results

The models were trained on the 72% training set, with hyperparameters tuned against the 8% validation set. The final, optimized models were then evaluated a single time on the 20% held-out test set. This allocation (20% of the total dataset) was purposefully set high to ensure a more robust and statistically reliable evaluation of critical metrics—particularly False Negatives (FN)—for this high-stakes task.

4.1 Hyperparameter Optimization

The fine-tuning process involved a systematic sensitivity analysis of the LR and the number of training epochs to identify the optimal configuration for maximum generalization and stability for both DistilBERT and RoBERTa models.

4.1.1 DistilBERT Optimization

The optimization study for DistilBERT, an efficient 6-layer architecture, focused on finding the optimal configuration to maximize performance while preventing premature overfitting. Sensitivity analysis, visualized in Fig. 2, confirmed that performance peaked at a higher LR of 5e-5, with metrics declining for lower rates (e.g., 2e-5). Analysis of epoch count, shown in Fig. 3, revealed that the model achieved maximum generalization after only 2 epochs of training, with subsequent epochs leading to performance degradation attributable to overfitting. The final configuration adopted for the model used a LR of 5e-5, trained for two epochs with a batch size of 8 and a maximum sequence length of 32 tokens.

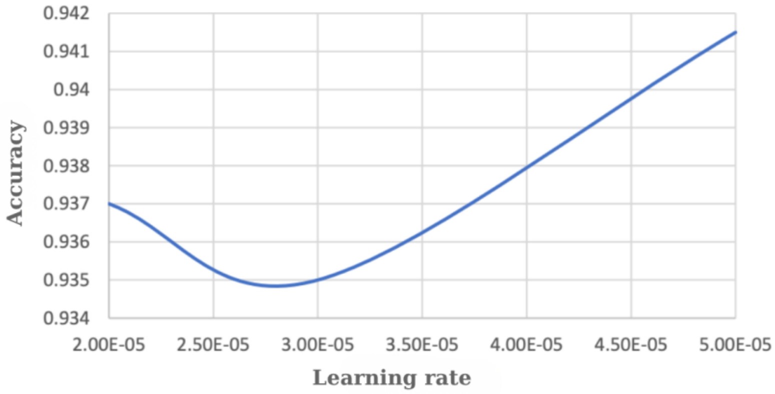

Figure 2. Influence of LR on model accuracy

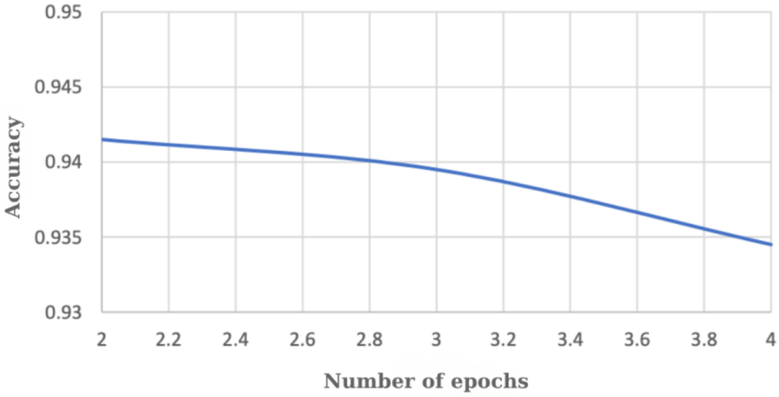

Figure 3. Influence of number of epochs on model accuracy

4.1.2 RoBERTa Optimization

Given RoBERTa's deeper, 12-layer architecture and superior representational power, a slower fine-tuning strategy was required. The hyperparameter analysis demonstrated an inverse relationship between LR and performance: metrics maximized at the lowest tested rate, 2e-5, confirming the necessity of slower, meticulous parameter updates to ensure stable convergence as shown in Fig. 4. Unlike DistilBERT, RoBERTa benefited from extended training, showing continuous metric improvement and reaching peak performance after 4 epochs (Fig. 5). This confirmed that the deeper model required more steps to capture nuanced semantic patterns without exhibiting premature overfitting. The final configuration adopted for the model used a LR of 2e-5, trained for four epochs with a batch size of 8 and a maximum sequence length of 32 tokens.

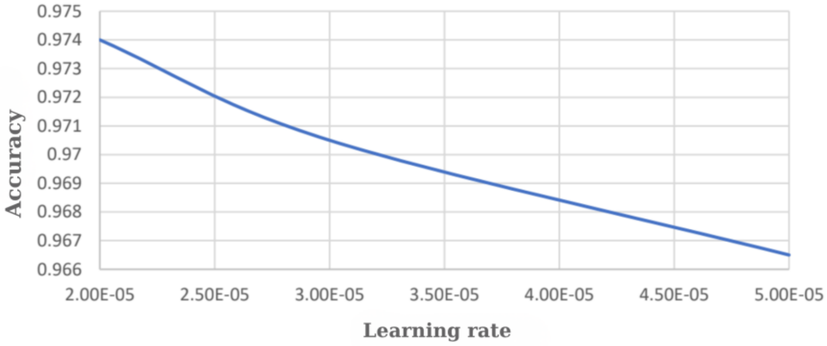

Figure 4. Influence of LR on model accuracy

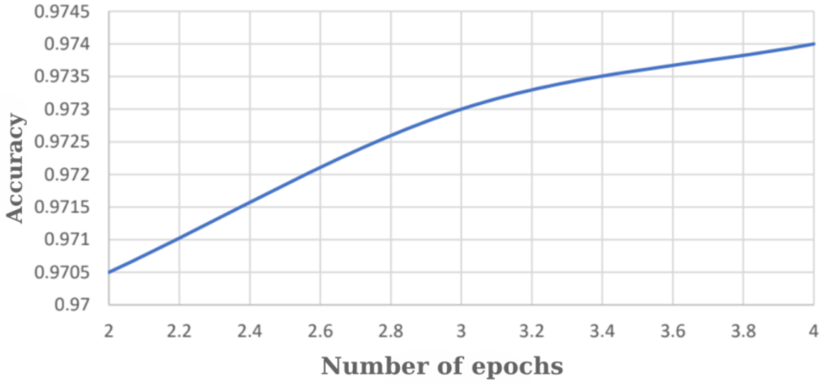

Figure 5. Influence of number of epochs on model accuracy

4.2 Performance evaluation

The two optimally configured models were evaluated on the N=2,000 test set, confirming their complementary roles and establishing the basis for the hybrid fusion mechanism.

A head-to-head comparison confirmed RoBERTa's substantial superiority in both generalization and critical error reduction, essential for this high-stakes task. The performance metrics are summarized in Table 2.

Table 2. Comparative performance DistilBERT vs RoBERTa

Model	Accuracy	Precision	Recall	F1 Score
DistilBERT	94.15%	94.30%	95.20%	94.10%
RoBERTa	97.40%	97.70%	97.10%	97.40%

RoBERTa consistently outperformed DistilBERT across all generalized metrics, achieving a peak accuracy of 97.40% and an F1-Score of 97.4% (3.3 percentage points higher than DistilBERT's 94.1% F1-score). The RoBERTa model demonstrated superior recall (97.1%) and precision (97.7%), indicating exceptional capability in both detecting actual risk messages and minimizing false warnings. This performance validates RoBERTa's efficacy and ethical superiority, establishing it as the high-accuracy anchor of the hybrid system. This exceptional performance is largely attributed to the meticulous data curation and selection strategy on the perfectly balanced subset, which allowed the model to learn highly specific, decisive linguistic features efficiently.

4.3 Error analysis

The design philosophy of the implemented hybrid system prioritizes the minimization of FN above all other metrics, aligning with the highest ethical mandate for suicide risk detection systems. An FN represents a failure to flag a high-risk message, potentially leading to catastrophic consequences. Conversely, a False Positive (FP) is tolerable, resulting only in the unnecessary offer of supportive resources (such as the LLaMA 3 chatbot access).

The DistilBERT model achieved balanced True Positives (TP) and True Negatives (TN), but generated 48 critical FNs—missed opportunities for intervention—which represented the primary vulnerability of deploying DistilBERT as a standalone system. The 69 FPs were considered acceptable, reflecting a conservative ethical approach that prioritizes safety over nuisance warnings. The full breakdown of its classification errors on the test set is presented in Table 3.

Table 3. DistilBERT confusion matrix

	Predicted Suicide	Predicted Non-Suicide
Actual Suicide	TP = 957	FN = 48
Actual Non-Suicide	FP = 69	TN = 926

The RoBERTa model demonstrated high reliability by substantially minimizing the critical error rate. As detailed in Table 4, it reduced the number of FN to only 29 and the FP to 23. This means RoBERTa achieved a 40% reduction in the critical FN rate compared to DistilBERT (48 FN vs. 29 FN), validating its ethical superiority in this high-stakes task. The low FP rate further highlights RoBERTa's precision in distinguishing genuine risk from non-suicidal emotional expression.

Table 4. RoBERTa confusion matrix

	Predicted Suicide	Predicted Non-Suicide
Actual Suicide	TP = 976	FN = 29
Actual Non-Suicide	FP = 23	TN = 972

The prediction selection mechanism serves as a strategic deployment of the system's conservative ethical bias. By dynamically adopting the prediction with the highest certainty, the system ensures that RoBERTa's superior sensitivity—demonstrated by its significantly lower FN rate—drives the final decision in ambiguous or high-stake cases. This reliable and ethically conservative classification enables the safe and responsible activation of the constrained LLaMA 3 support module. This ensures advanced conversational aid is only triggered when a robust, high-certainty risk signal is present.

5 Conclusions and further work

The experimental results definitively validate the efficacy of the comparative methodology for suicide risk detection from social media text. The RoBERTa architecture demonstrated substantial superiority, achieving an F1-score of 97.4% and, most significantly, a 40% reduction in the critical FN rate compared to the computationally lighter DistilBERT model. This FN reduction confirms the ethical requirement for prioritizing deep, robust architectures over sheer computational speed in high-stakes mental health applications.

A core contribution of this work is the validation of a resource-efficient fine-tuning methodology that achieved state-of-the-art-level performance on a constrained dataset subset, utilizing only CPU resources, demonstrating feasibility for low-resource environments.

Finally, the resultant hybrid system, integrated with an ethically constrained, prompt-guided LLaMA 3 conversational module via the Groq API, establishes a viable, safe, and responsible prototype for an early intervention, detect-and-support pipeline in digital mental health.

To transition the prototype into a fully robust and clinically relevant tool, several key development directions are necessary.

The most immediate required improvement is the scaling of the training process. The next phase of research must utilize advanced GPU computational environments to fine-tune the optimal RoBERTa model configuration on the full original dataset of 232,074 examples to validate robustness across a wider and noisier linguistic distribution.

Further research should focus on multimodal and linguistic expansion. This includes integrating non-textual features such as temporal patterns, posting frequency, and interaction metadata to improve predictive accuracy, as shown in related research, as well as expanding the data ingestion pipeline to handle inputs from multiple social media platforms and developing a variant capable of processing texts in Romanian.

For practical implementation, clinical validation and real-time intervention capabilities are essential. Future work must involve formal collaborations with mental health professionals to perform clinical evaluation, establish calibrated risk thresholds based on standard diagnostic criteria, and integrate professional feedback for continuous improvement. Developing a secure native mobile application component capable of real-time monitoring and implementing automated, location-aware emergency alerts (such as integration with national emergency services like 112) would transform the system into a complete tool for proactive suicide prevention.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,"* arXiv preprint arXiv:1810.04805, 2018.

[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, *"DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,"* arXiv preprint arXiv:1910.01108, 2019.

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, *et al.*, *"RoBERTa: A Robustly Optimized BERT Pretraining Approach,"* arXiv preprint arXiv:1907.11692, 2019.

[4] J. Li, Z. An, W. Cheng, J. Zhou, F. Zheng, and B. Hu, *"MHA: a multimodal hierarchical attention model for depression detection in social media," Health Information Science and Systems*, vol. 11, no. 1, pp. 1–13, 2023.

[5] B. Cui, J. Wang, H. Lin, Y. Zhang, L. Yang, and B. Xu, *"Emotion-Based Reinforcement Attention Network for Depression Detection on Social Media: Algorithm Development and Validation," JMIR Medical Informatics*, vol. 10, no. 8, e37818, 2022.

[6] X. Jia, J. Chen, K. Liu, Q. Wang, and J. He, *"Multimodal depression detection based on an attention graph convolution and transformer," Mathematical Biosciences and Engineering*, vol. 22, no. 3, pp. 652–676, 2025.

[7] K. Hasan and J. Saquer, "A comparative analysis of transformer and LSTM models for detecting suicidal ideation on Reddit," arXiv preprint arXiv:2411.15404, 2024.

[8] A. Izmaylov, S. Malmasi, and A. Yates, *"Combining psychological theory with language models for suicide risk detection,"* in *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 2405–2414, 2023.

[9] K. Nikhileswar, D. Vishal, L. Sphoorthi, and S. Fathimabi, "*Suicide Ideation Detection in Social Media Forums*," in *Proceedings - 2nd International Conference on Smart Electronics and Communication*, ICOSEC 2021, doi: 10.1109/ICOSEC51865.2021.9591887, 2021.