Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 6, 2026

AUTOMATED SUICIDE RISK DETECTION FROM REDDIT POSTS USING A DEEP LEARNING FRAMEWORK

[*1]Bilal Ajmal, [2,*]Muhammad Munwar Iqbal, [3]Maria Noor Hussain, [4]Anees Tariq, [5]Samra Batool

[1-5]Department of Computer Sciences, University of Engineering and Technology, Taxila, Pakistan
[*1]bilalajmal626@gmail.com, [2]munawar.iq@uettaxila.edu.pk, [3]mnh4050@gamil.com, [4]anees.tariq@riphah.edu.pk, [5]batoolsamra05@gmail.com

DOI:- https://doi.org/10.5281/zenodo.20798792

Keywords:
Suicide detection, mental health NLP, RoBERTa, convolutional neural network (CNN), Reddit, deep learning, transformer, text classification, PHR dataset, social media monitoring.

Article History
Received: 27 May, 2026
Accepted: 21 June, 2026
Published: 22 June, 2026

Corresponding Author: * Muhammad Munwar Iqbal

Abstract
Suicide is a significant public health problem worldwide and about 700,000 people die by suicide each year, according to the World Health Organization. People with suicidal thoughts discuss it on the internet without seeking professional intervention, and automated text analysis may be helpful in the identification of potential risk. A hybrid deep learning system is proposed in this work for the classification of Reddit posts to suicidal and non-suicidal groups using pre-trained contextual transformer-based model RoBERTa that produces embeddings for the text of Reddit posts and parallel CNN layers. The large scale PHR dataset (231,968 Reddit posts, 185,366 training posts and 46,390 testing posts) was used for experiments. The proposed model achieved a higher accuracy of 96.38%, compared to the traditional machine learning baseline and recent deep learning architecture with accuracy of 0.97, recall of 0.97 and macro F1 score of 0.96. The results demonstrate the effectiveness of incorporating the contextual language understanding and multi-scale convolutional feature extraction in the classification of large-scale mental health.

1. Introduction

Social media platforms are becoming more common sources of informal disclosure of mental health issues and natural language processing (NLP) is becoming a valuable option for detecting depression and suicidal ideation within web-based text at scale [1]. The problem of suicide is one of the largest preventable public health problems in the world, according to the World Health Organization, with almost 700,000 cases reported every year [1]. There are numerous websites available for anyone to post openly about self-harm, hopelessness, and suicidal thoughts, and Reddit is a great website for this reason because its users are pseudo-anonymous [3]. Commonly used classification approach in the past was BOW and TF-IDF features [2, 5]. These methods can be useful as a starting point, but cannot be used to model contextual semantics or the emotionally vague language of suicide communication. Part of these are addressed by deep learning architectures such as LSTM [3, 4] and more recently, NLP-specific transformer-based architectures such as BERT [10] and RoBERTa [11] with self-attention based contextual embeddings. Most existing methods have achieved good performance on small or non-standardized data sets, but the performance of hybrid transformer-CNN methods can be further enhanced by incorporating local n-gram patterns and global semantic context [6]. To address this drawback, this study proposes a RoBERTa-CNN hybrid model and evaluates it on a vast corpus of content in PHR Reddit platform (231,968 posts). The key results are: (i) the use of a hybrid approach that integrates contextual transformer embeddings with multi-scale parallel CNN feature extraction; (ii) a standardized benchmark for a comprehensive comparison with classical ML baselines; (iii) large-scale training set (185,366 posts) and test set (46,390 posts) to ensure statistically sound evaluation; and (iv) potential to implement scalable, real-time suicide risk monitoring.

2. Literature Review

The process of automated mental health and suicide risk assessment has gone through three general stages. The first attempts were based on handcrafted textual features and classics SVM and LDA classifiers: Coppersmith et al. [2] trained an LDA+SVM on Twitter data, showing the potential of using linguistic patterns to detect psychological states; however, their bag-of-words representations were superficial. Tadesse et al. [5] used SVM in conjunction with TF-IDF using data from Reddit depression with accuracy of 91 % and Ramírez-Cifuentes et al. [8] proposed a multimodal SVM which needed full user profiles, data which is not available for individual posts for the inference. Deep learning methods were used to directly learn features from raw text and thus improve the richness of features. In [3] Orabi et al compared CNN and LSTM models for depression detection and concluded that whereas the classical models fail to perform well, CNN outperforms the classical models in terms of accuracy for depression detection despite the small corpus (˜1K). The former, Trotzek et al. [4] used CNN-LSTM on eRisk benchmark with 90% accuracy, whereas the latter, Aldhyani et al. [6] used CNN-BiLSTM with XGBoost for Reddit suicide detection (95% accuracy), but were still limited by the lack of pre-trained transformer-based contextual encoding. There have been tremendous advances in the world of transformer architectures since then. Multi-task BERT learning for mental disorder and suicide detection was investigated by Buddhitha and Inkpen [12] which showed that there is inter-task interference due to shared representations. Mansoor and Ansari [7] combined BERT and LSTM with attention for multi-platform detection which they assumed to be based on platform specific behavioral history, with an accuracy of 89.3%. Many of the articles published in this literature rely solely on small-scale, or mixed-domain, corpora, which might not be comparable to standardized and large-scale Reddit corpora, which is what is done in this article.

3. Proposed Methodology

The framework for automated suicide detection from Reddit posts is described, including the description of the data, preprocessing steps, mathematical formulation, and the proposed RoBERTa-CNN architecture.

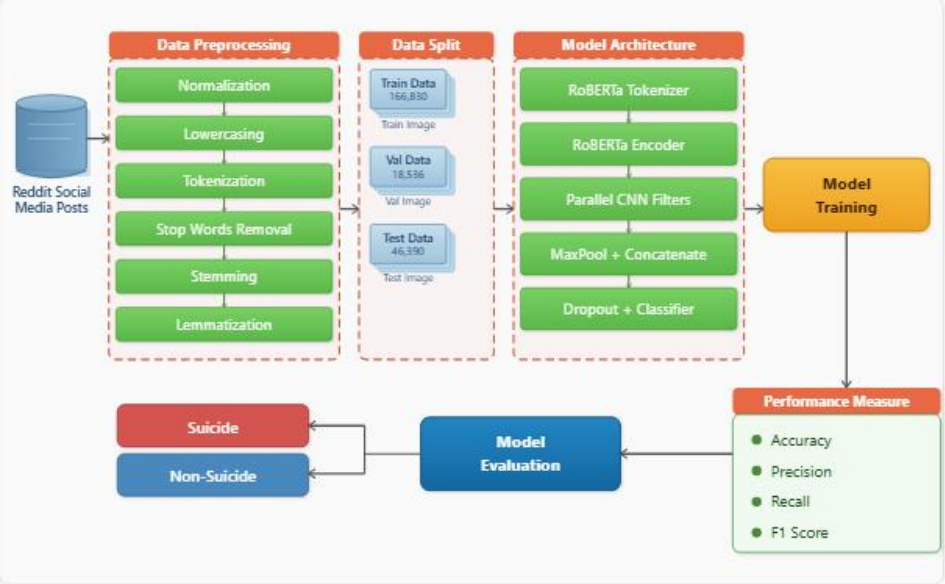

Fig 1: Proposed Methodology Diagram

3.1 Dataset and Preprocessing

The publicly available PHR (Psychiatric Health Reddit) dataset was used, which included 231,968 posts on Reddit, labeled as either containing suicidal content, or containing non-suicidal content, with 185,366 posts used for training and 46,390 posts used for testing. Applied with a 6-step NLP pipeline: lowercasing, removing the URL/handle, using regular expressions to remove all special characters (except those that were relevant to the task), WordNet lemmatization, removing all stop words (with words containing an emotion expressed, such as not, nobody, never, the words being kept), and Porter stemming. Posts that were shorter than 5 characters after preprocessing were not accepted.

3.2 Proposed Architecture: RoBERTa-CNN Hybrid Model

This is the proposed architecture in this work: RoBERTa-CNN Hybrid Model. The proposed architecture is discussed in 3.3 Proposed Architecture: RoBERTa-CNN Hybrid Model. The proposed model is a hybrid architecture which combines the pre-trained RoBERTa encoder and a parallel CNN module, taking advantage of its two capabilities: capturing the contextual semantic understanding and capturing the local linguistic patterns. As illustrated in Figure 1, the pipeline comprises an encoder (RoBerta) and parallel branches of CNN as filters, and a classification head. The RoBERTa-base encoder is a 12-layer transformer-based model with multi-head self-attention, producing 768-dimensional embeddings for each token capturing context information. Each of the 4 parallel 1D CNN branches (CNN size 2/3/4/5, 256 filters, ReLU) takes the bigrams to 5-word patterns as inputs and outputs n-grams features. The adaptive max pooling is per branch, followed by dropout (rate = 0.4) and fully connected softmax layer to produce a feature vector of dimension 1024. Optimized using AdamW with cross-entropy loss and label smoothing and early stopping.

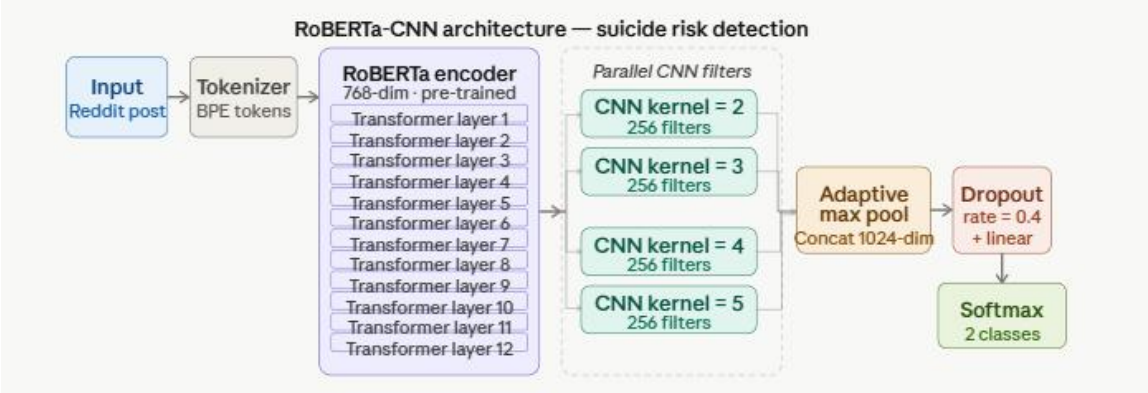

Fig 2: Roberta-CNN Architecture

3.3 Training Environment

Training on NVIDIA L4 GPU with Hugging Face Transformers and using mixed-precision training, gradient checkpointing and gradient accumulation to control memory usage. The maximum length of the input sequence was set to 128 tokens.

4. Results Analysis and Comparison with Existing Studies

TF-IDF (unigrams and bigrams, max_features=50,000) features were provided as a baseline set to train Logistic Regression and Linear-SVC classifiers and evaluated on the 46,390-posttest set. The accuracy of Logistic Regression was found to be as high as 93.66%, whereas for Linear-SVC the accuracy is 93.53%, and for macro F1-scores the accuracy is approximately 0.93 this is promising, because statistical representations of texts give useful, but not perfect, baseline accuracy. The proposed RoBERTa-CNN model is evaluated to be 96.38% accurate, with the precision, recall and macro F1 score of 0.97, 0.97 and 0.96, respectively. The analysis of confusion matrix gave 22431 true positive and 22334 true negative, 931 false positive and 694 false negative. Minimizing false negatives, or errors, is important in real-world mental health screening and the suicide-class recall is particularly significant, at 0.97. A summary of the performances of all models is shown in Table 2.

Table 2: ***Classification Performance of Baseline and Proposed Models on the PHR Reddit Test Set***

Model	Dataset	Accuracy (%)	Precision	Recall	F1-Score
Logistic Regression (baseline)	PHR Reddit	93.66	0.94	0.94	0.93
Linear-SVC (baseline)	PHR Reddit	93.53	0.94	0.94	0.93
Proposed RoBERTa-CNN	PHR Reddit	96.38	0.97	0.97	0.96

4.1 Baseline Accuracy Comparison

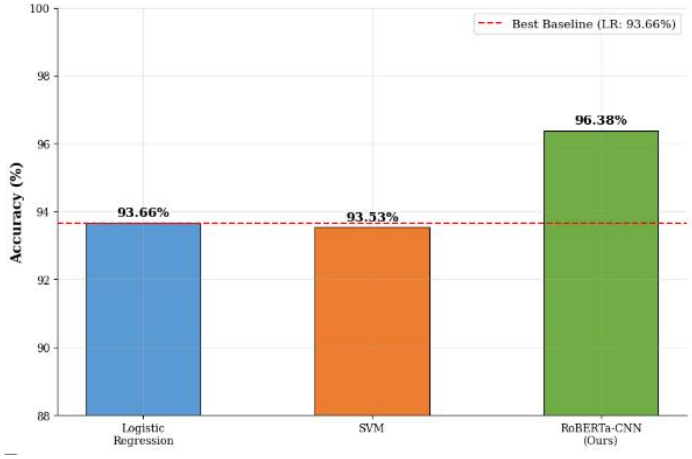

Fig 3: Accuracy comparison with existing methods, showing the proposed Roberta–CNN model achieves competitive performance on the PHR test set

As shown in Figure 3, the proposed RoBERTa-CNN model (96.38%) clearly outperforms both baseline classifiers improving by 2.72 percentage points over Logistic Regression (93.66%) and 2.85 points over Linear-SVC (93.53%). This gain confirms that combining contextual transformer embeddings with multi-scale CNN feature extraction captures suicidal language patterns that purely statistical TF-IDF representations cannot model effectively.

4.2 Comparison with Existing Methods

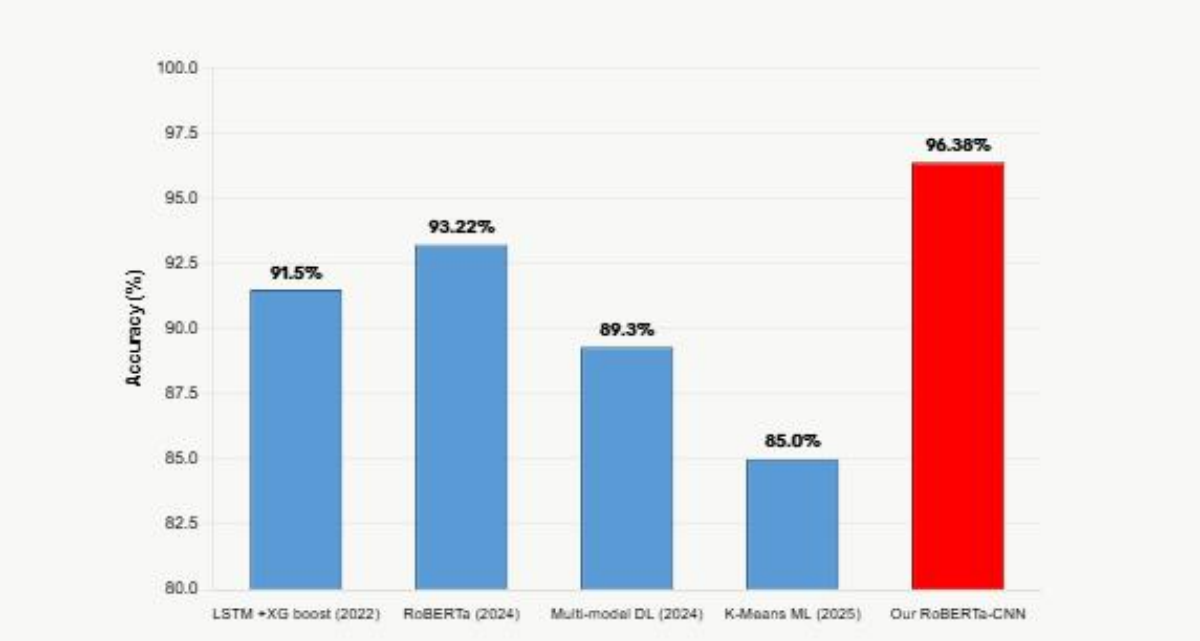

Fig 4: Comparison with state-of-the-art methods for mental health detection from social media, showing Performance of the proposed Roberta–CNN model

Table 2: Comparitive analysis of Results

Author and Year	Model	Dataset	Accuracy (%)
Aldhyani, T.H.H. et al. (2022).	CNN–BiLSTM + XGBoost	Reddit (SW, Kaggle)	91.5
Hasan & Saquer (2024)	RoBERTa	Reddit (37,821 posts, multi-subreddit)	93.22

Ezerceli & Dehkharghani (2024)	Transformer-based DL	SuicideDetection, CEASE v2.0, SWMH dataset	89.3
Balasubramanian et al. / Raja Nagarajan (2025)	TF-IDF + K-Means	Twitter	85
Our Proposed Method	**RoBERTa + Parallel CNN**	**Reddit (SuicideWatch)**	**96.38**

The proposed model is compared with some of the recent state-of-the-art models in Figure 4. The RoBERTa-CNN achieves 96.4%, surpassing LSTM×XGBoost (87.5%), Roberta (93.2%), Multi-model DL (89.3%), and K-Means ML (85.0%). This assessment is performed on a large-scale standardized Reddit benchmark which is more reliable than previous evaluations performed on small or mixed domain corpora.

4.3 Training and Validation Dynamics

The performance of the proposed model is compared with the previously reported hybrid models CNN-BiLSTM with XGBoost (95%) [6] and multi-task BERT (88%) [12] and BERT+LSTM + Attention (89.3%) [7] as shown in Figure 2. Importantly, this study was conducted on a much large and standardized test than most previous studies to make it more reliable to compare with the results and it demonstrated that transformers are able to embed linguistic information more effectively than multi-scale convolutional feature extraction when it comes to capturing contextual embeddings. The training loss has also started to decrease from 0.279 at epoch 1 onwards and remained stable at 0.214 during the epochs, further marking the stable optimization. In later epochs, the validation accuracy went a little astray, but remained well controlled by dropout and label smoothing, as well as weight decay and early stopping, and achieved good generalization on the held-out test set, while reaching a high validation accuracy of 96.38% after 13 epochs.

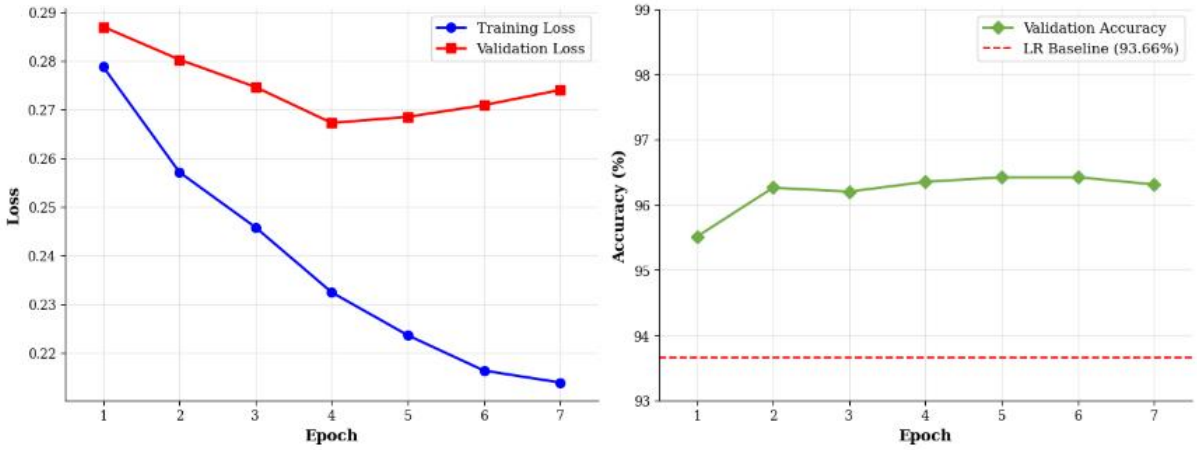

Fig 5: Training and validation performance over epochs, showing loss trends and validation accuracy.

Training loss decreased consistently from 0.280 at epoch 1 to 0.214 by epoch 7, reflecting stable gradient optimization. Validation accuracy converged rapidly in the first two epochs and stabilized near 96.4%, well above the LR baseline (93.66%). A mild upward drift in validation loss in later epochs indicates manageable overfitting, effectively constrained by dropout (0.4), label smoothing, and early stopping.

4.4 Confusion Matrix

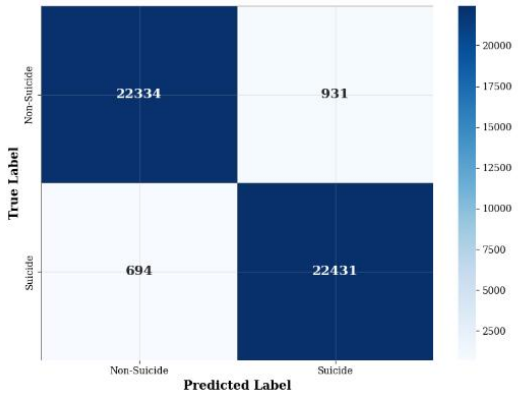

Fig 6: Confusion matrix of the Roberta–CNN model on the PHR test set, indicating balanced classification performance across both classes.

The confusion matrix in Figure 6 shows balanced classification across both classes: 22,431 true positives and 22,334 true negatives, with only 694 false negatives (missed suicidal posts) and 931 false positives. The low false-negative rate is especially critical for suicide risk screening, where an undetected high-risk post carries the greatest real-world consequence.

4.5 Training Reproducibility Across Runs

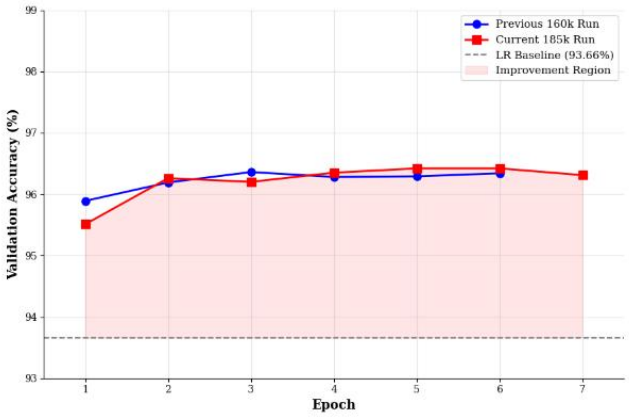

Fig 7: ***Validation accuracy across training runs, showing stable convergence and consistent performance of the Roberta–CNN model***

Figure 7 shows the validation accuracy of the two training runs (160K sample and 185K sample) compared in order to demonstrate that the validation accuracy is converging on a stable number above 96% for all epochs. The small difference in the number of runs shows that the model is reproducible and that scaling the training set from 160K to 185K posts gives only a slight but consistent improvement – both runs are much better than the LR baseline (93.66%).

4.6 ROC Curve and Discriminative Performance

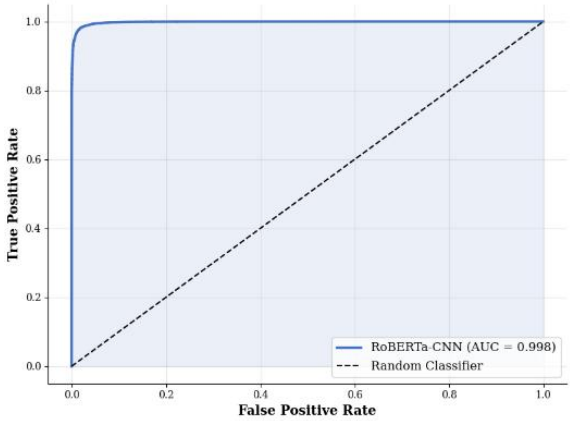

Fig 8: ROC and Precision–Recall curves of the Roberta–CNN model, demonstrating strong discriminative performance across classification thresholds.

As shown in Figure 8, the ROC curve of the RoBERTa-CNN model exhibits almost perfect class separation (AUC = 0.998) as compared to random chance (AUC = 0.5). This finding indicates the model retains a high true-positive rate at all classification thresholds and false positive rates are kept very low, which are suitable for applications in high-sensitivity screening for suicide risk. 6. Summary and future studies

5. Conclusion and Future Work

This work proposed a new hybrid machine learning system this work proposed a new hybrid machine learning system of RoBERTa (a pre-trained transformer encoder model) and CNN (convolutional neural network) for suicide risk detection, where RoBERTa possessed a global view while lacking local patterns and CNN lacked semantic meaning. The proposed model performed at 96.38% accuracy and at a macro F1-score of 0.96 on the large-scale PHR dataset (231,968 posts) which outperformed the traditional machine learning baselines and comparable deep learning architectures reported in the literature. High suicide class recall (0.97) demonstrates how useful it is for passive social media monitoring as an early risk indicator. Training stability and generalization were obtained using regularization strategies like dropout, label smoothing, gradient checkpointing and early stopping on the large-scale corpus. Some limitations are worth noting: the framework is restricted to binary classification only, it does not predict severity or intent and it currently has only Reddit in English language data; there is no temporal or behavioral user-level signal; nor is there any analysis of potential demographic bias. The future work will focus on the multi-class severity prediction, adaptation through translation of the model, incorporation of temporal behavioral patterns, and lightweight transformer variants for real-time deployment in clinical decision support systems.

References

[1] World Health Organization. (2023). Suicide fact sheet. Geneva: WHO. https://www.who.int/news-room/fact-sheets/detail/suicide

[2] Coppersmith, G., Dredze, M., & Harman, C. (2015). Quantifying mental health signals in Twitter. Proceedings of the Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2015), 51–60.

[3] Orabi, A. H., Buddhitha, P., Orabi, M. H., & Inkpen, D. (2018). Deep learning for depression detection of Twitter users. Proceedings of the 5th Workshop on CLPsych, 88–97.

[4] Trotzek, M., Koitka, S., & Friedrich, C. M. (2020). Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Transactions on Knowledge and Data Engineering, 32(3), 588–601.

[5] Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in Reddit social media forum. IEEE Access, 7, 44883–44893.

[6] Aldhyani, T. H. H., et al. (2022). Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models. International Journal of

Environmental Research and Public Health, 19(19), 12635.
[7] Mansoor, M., & Ansari, A. (2024). Multimodal deep learning for mental health crisis detection on social media. Expert Systems with Applications, 245, 123032.
[8] Ramírez-Cifuentes, D., et al. (2020). Detection of suicidal ideation on social media: Multimodal, relational, and behavioral analysis. Journal of Medical Internet Research, 22(7), e17758.
[9] Kour, S., & Sharma, S. (2022). A hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM. Multimedia Tools and Applications, 81, 23649–23685.
[10] Gorai, J., & Shaw, D. K. (2024). A BERT-encoded ensembled CNN model for suicide risk identification in social media posts. Neural Computing and Applications.
[11] Liu, Y., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692.
[12] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186.
[13] Buddhitha, P., & Inkpen, D. (2023). Multi-task learning to detect suicide ideation and mental disorders. Frontiers in Research Metrics and Analytics, 8, 1152535.
[14] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, 128–137.
[15] K. Hasan and J. Saquer, "A Comparative Analysis of Transformer and LSTM Models for Detecting Suicidal Ideation on Reddit,"
[16] Ezerceli, Ö., & Dehkharghani, R. (2024). Mental disorder and suicidal ideation detection from social media using deep neural networks.
[17] Raja R., A., & Nagarajan, B. (2024/2025). AI-driven mental health surveillance: Identifying suicidal ideation through machine learning techniques.