Automated Detection of Suicidal Ideation on Social Media Using Hybrid Recurrent Neural Networks

N Sai Satwik Reddy*, V Venkata Alluri Rohith*, V Poorna Muni Sasidhar Reddy*,
Y Shashank Reddy*, Sachin Kumar S*, Neethu Mohan*, K P Soman*
*Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India
{satwikreddy987, vennaalluri1, vpoornareddy2004, ysrpersom}@gmail.com,
{s_sachinkumar, m_neethu}@cb.amrita.edu, kp_soman@amrita.edu

***Abstract*—Suicide ranks as the fourth leading cause of death globally, highlighting it as a pressing mental health concern that requires immediate attention. Both the United Nations Sustainable Development Goals and the World Health Organization's Global Mental Health Action Plan have set a target to reduce the global suicide rate by one-third by the year 2030. Examining suicide notes or related posts on social media enables early detection of suicidal ideation, facilitating timely intervention and support. This study introduces two innovative Recurrent Neural Network (RNN)-based approaches for automated detection of suicidal ideation in social media content. Leveraging the context of diverse interactions on platforms like Reddit, the first architecture employs a Long Short-Term Memory (LSTM) network to decipher intricate patterns, while the second utilizes a Gated Recurrent Unit (GRU) for computational efficiency. Evaluation on a dataset sourced from Reddit posts yields promising results, with the LSTM-based approach achieving 94.30% accuracy and the GRU-based approach achieving 93.20%. Therefore, the proposed hybrid RNN models offer promising prospects for the development of robust systems capable of identifying and intervening in potential cases of suicidal ideation online.**

***Index Terms*—Suicide, Social media, Natural language processing, Deep learning, Recurrent neural networks, Long short-term memory, Gated recurrent unit.**

I. Introduction

In recent years, the intersection of mental health and technology has been one of the highly scrutinized issues that acts as a basis for better assessment and management of depression and suicide ideation. A chronic mood disorder called depression, dominated by extended, almost permanent feelings of sadness and hopelessness, is very often seen together with the tragic reality of suicidal ideation [1]. Suicide, the extremely tragic act of intentionally terminating one's life, has a tremendous impact not only on the life of one who acts but also on families and communities around them [2].

According to the World Health Organization (WHO) report, suicidal thoughts affect individuals across various age groups, with a higher risk observed among 15–29-year-olds. Identifying the prevalence of suicides within specific demographics is vital for targeted intervention strategies [3]. The multifaceted nature of suicidal thoughts involves factors like mental health disorders, societal pressures, and personal crises. WHO highlights that more than 700,000 people commit suicide and die annually, intensifying the critical necessity for comprehensive global initiatives to address mental health challenges. The report also underscores that 77% of global suicides occur in low- and middle-income countries, signifying the need for tailored interventions in diverse socio-economic contexts. Common methods, such as the ingestion of pesticides, hanging, and firearms, further emphasize the urgency of addressing this global public health issue [4].

The practice of leaving a note before committing suicide raises questions about communication, seeking understanding, and the potential for intervention [5]. Exploring the role of social media in this context becomes pertinent, as individuals often express their emotions and struggles online. The posts shared by individuals on social media platforms can serve as a valuable resource for detecting suicidal tendencies. Analyzing and understanding the linguistic patterns indicative of depression and suicidal thoughts in online communication is crucial [6]. Detecting suicide tendencies, especially through text analysis on social media platforms, is essential in the contemporary digital age. Early identification can facilitate timely intervention and support, potentially saving lives [7].

In this study, two effective Recurrent Neural Network (RNN)-based architectures for suicidal ideation detection are proposed. The proposed methods are evaluated on a suicide-related text dataset obtained from social posts on Reddit. The first proposed architecture utilizes a Long Short-Term Memory (LSTM) network to discern patterns and nuanced expressions related to suicidal ideation within the diverse and dynamic context of social media interactions. The second architecture employs a Gated Recurrent Unit (GRU), which is computationally more efficient than LSTMs while maintaining comparable performance. By leveraging these advanced RNN-based methods, our study aims to contribute to the ongoing efforts in developing reliable and sensitive tools for identifying early signs of suicidal ideation on online platforms.

The rest of the work is organized as follows: Section II discusses the related literature, and Section III demonstrates the procedure for data preparation. The proposed methodology is detailed in Section IV, followed by the results in Section V. Subsequently, Section VI explains the conclusion and future scope.

II. Related Work

Several machine learning (ML) and deep learning (DL) based methods have been proposed to detect suicidal ideation from text data over the past few years.

[8] addressed the surge in suicide-related cases on social media, proposing two models (C-BiGRU-MHA-CNN and L-BiLSTM-MHA-CNN) to detect negative emotions in suicide-related texts. Achieving accuracies, around 98.12% on Reddit and 97.68% on CEASE dataset, the study emphasizes the significance of identifying emotions in suicide-related texts and their impact on mental health, decision-making, and individuals' well-being. A blend of bidirectional long short-term memory (Bi-LSTM) and bidirectional encoder representations from transformers (BERT) is used to detect suicidal tendencies in an individual from their social media activity [9]. In [10], fastText embedding is utilized for contextual analysis of the text, followed by classification using an XGBoost classifier. However, this approach achieved a very low accuracy of 78%. A hybrid DL architecture, which includes a convolutional neural network (CNN), Long Short-Term Memory (LSTM), and fasttext embedding, is proposed for detecting depressive thoughts in [11]. In [12], the CNN-2 Layer LSTM outperformed the conventional CNN-LSTM, demonstrating dominant performance, but it also adds an extra layer of complexity. [13] presents a comprehensive survey of the natural language processing techniques used to detect and prevent suicidal tendencies through text modality. ML models such as random forest, neural network, bagging tree, and XGBoost classifiers are utilized to identify suicide-related text from the portal messages of the patients in the study presented in [14].

In [15], a lightweight attention-based LSTM-CNN model was developed for detecting depressive Bangla social media texts and achieved an accuracy of 94.3%. The model demonstrated robustness and cross-language performance, outperforming classical machine learning models, ensemble approaches, transformers, and existing architectures. The text mining approach for identifying depression and suicidal ideation is explored in [16]. A two-stage architecture with the utilization of a natural language processing approach for filtering keywords and logistic regression for the detection task is presented in [17]. [18] used machine learning on Weibo data to identify linguistic characteristics related to depression and suicidal ideation, finding that specific features significantly predicted depression and suicidal thoughts. [19] employs natural language processing and deep learning on German medical documentation, utilizing Electronic Health Records (EHRs) to detect suicide attempts through structured embeddings and a combined LSTM-CNN approach, aiming for quantitative, automated early detection and potential advancements in mental health care. In [20], synthetic suicide-related text data, generated with large language models (LLMs), is fed into selected state-of-the-art methods centered around BERT architectures and tested on real-time data. The explainability challenge posed by the black-box nature of transformer-based models is addressed using SHAP and LIME in [21]. The proposed methodologies bring key advantages to suicidal ideation detection. Firstly, the LSTM-based architecture excels in capturing intricate patterns and nuanced expressions within the dynamic context of social media. Secondly, the GRU-based approach offers computational efficiency without compromising performance, ensuring real-time applicability. Lastly, by employing these advanced RNN-based methods, our study contributes to the development of reliable tools for the early identification of suicidal ideation in online platforms.

III. Data Preparation

Reddit is a social media platform and online community where users can share and discuss content on various topics through posts and comments. The dataset comprises posts obtained from the "SuicideWatch" subreddits on the Reddit platform, utilizing the Pushshift API. Posts from "SuicideWatch" created between December 16, 2008, and January 2, 2021, were included in the dataset. Each post retrieved from "SuicideWatch" is annotated with the labels "suicide". Table I presents a summary of the dataset, showing the distribution of data points between non-suicide (Class 0) and suicide (Class 1).

TABLE I
Dataset Summary

Labels	Class	Number of instances
0	non-suicide	116037
1	suicide	116037

Text pre-processing is very essential in any NLP task to reduce redundancy and complexity. The removal of stop words involves eliminating common words (e.g., "and," "the," "is") that often carry little semantic meaning. Conversion to lower case standardizes the text by transforming all characters to lowercase, ensuring consistent comparisons. Removing emails entails the extraction of email addresses from the text. The removal of HTML tags involves stripping off any HTML markup present in the text, leaving only the plain text content. Eliminating special characters involves excluding symbols, punctuation, and non-alphanumeric characters. Removing accented characters involves replacing or stripping diacritical marks from letters, ensuring uniformity and facilitating text processing tasks. These techniques collectively contribute to the cleanliness and uniformity of textual data, enhancing the efficiency of subsequent analyses or classification models. Examples of text samples corresponding to non-suicide and suicide classes are shown in Table II.

TABLE II
Sample Data from Each Class

Text	Class
Finally 2020 is almost over... So I can never hear "2020 has been a bad year" ever again. I swear to fucking God it's so annoying	non-suicide
It ends tonight.I can't do it anymore. I quit.	suicide

IV. Methodology

This section provides a detailed explanation of the information pertaining to the layers of the two hybrid recurrent neural networks (RNNs) proposed in this study. The proposed methodology is depicted in Fig. 1.

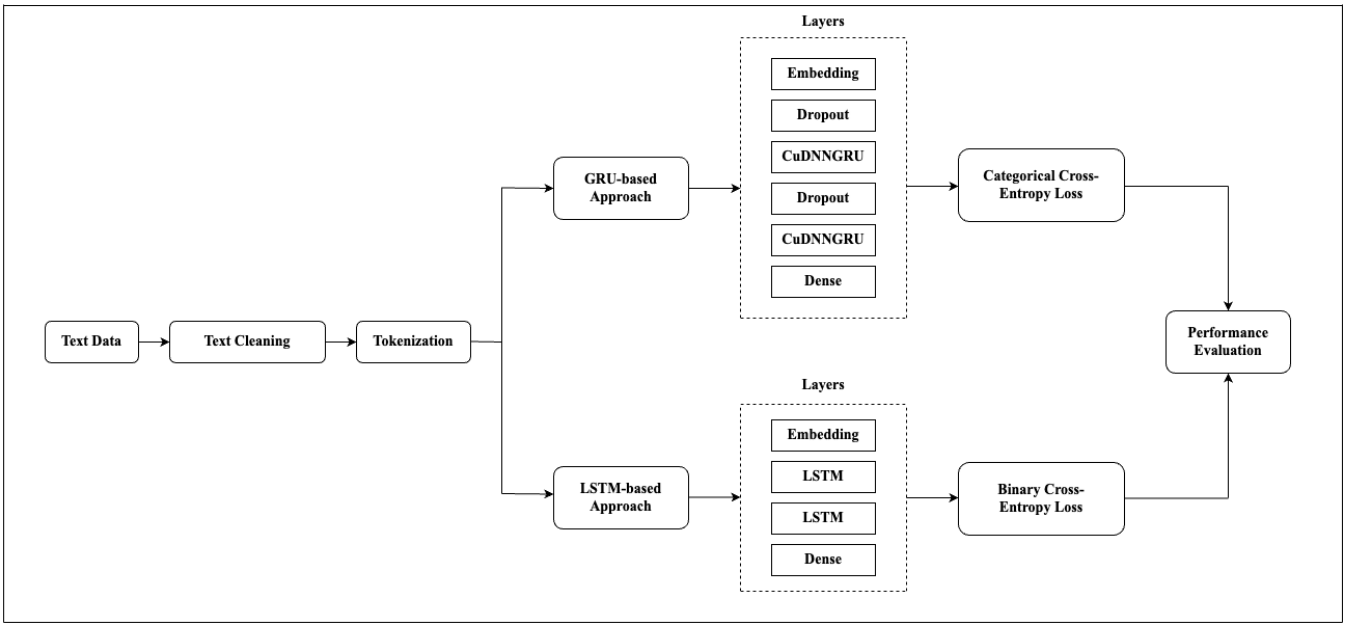

Fig. 1. A schematic overview of the RNN-based methods for detecting suicidal ideation.

A. GRU-Based Approach

The methodology involves creating a text classification model using a pre-trained word embedding technique. Initially, the input texts are tokenized, converting them into sequences of integers, with a limited vocabulary of the top 10,000 words. These sequences are then padded or truncated to ensure a uniform length of 200. The model employs pre-trained GloVe word embeddings, obtained from the Gensim library, to initialize an embedding matrix for the Tokenizer's vocabulary. This layer is designed to be non-trainable, preventing the adjustment of weights during training. Dropout layers are employed to mitigate overfitting by randomly deactivating a fraction of input units. The core of the network comprises two CuDNNGRU layers, facilitating fast and efficient processing of sequential data. The first CuDNNGRU layer returns sequences, capturing temporal dependencies, while the subsequent one encapsulates the sequence information into a final output. A Dense layer with a softmax activation function is employed to yield class probabilities.

TABLE III
HYPERPARAMETERS USED IN GRU-BASED APPROACH

Component	Hyperparameter	Value
Embedding Layer	Embedding dimension	100
	Number of words	10,000
CuDNNGRU Layers	CuDNNGRU units	100
Dropout	Dropout rate	0.2
Dense Layer	Output units	2
	Activation function	Softmax
Model Compilation	Loss function	Categorical Crossentropy
	Optimizer	Adam
	Epochs	25

B. LSTM-Based Approach

This approach involves the preprocessing of text data for a binary classification task and the subsequent construction and training of an LSTM neural network model using TensorFlow and Keras. The text is tokenized, creating a vocabulary with a specified size, and sequences are generated for the training, validation, and test datasets. These sequences are then padded or truncated to ensure uniform length. Subsequently, the data is converted into NumPy arrays. The LSTM model is composed of an embedding layer to map the tokenized words into continuous vectors, followed by two LSTM layers for capturing sequential dependencies in the data. The final dense layer employs a sigmoid activation function for binary classification.

TABLE IV
HYPERPARAMETERS USED IN LSTM-BASED APPROACH

Component	Hyperparameter	Value
Embedding Layer	Embedding dimension	8
	Number of words	16,004
LSTM Layers	LSTM units	28
Dense Layer	Output units	1
	Activation function	Sigmoid
Model Compilation	Loss function	Binary Crossentropy
	Optimizer	Adam
	Epochs	10

C. ML-Based Approach

In the detection of suicidal ideation using machine learning, traditional features such as n-grams and bag of words are frequently employed as inputs to classifiers. N-grams capture the contiguous sequences of n items from a given text, effectively preserving local word order and context, which can be crucial

for identifying nuanced expressions of suicidal thoughts. The bag of words approach, on the other hand, represents text by the frequency of its constituent words, disregarding grammar and word order, thus providing a straightforward yet powerful method for feature extraction. When fed into machine learning classifiers, these features enable the models to discern patterns and associations within textual data, facilitating the accurate detection of suicidal ideation through linguistic analysis.

V. Results and Discussion

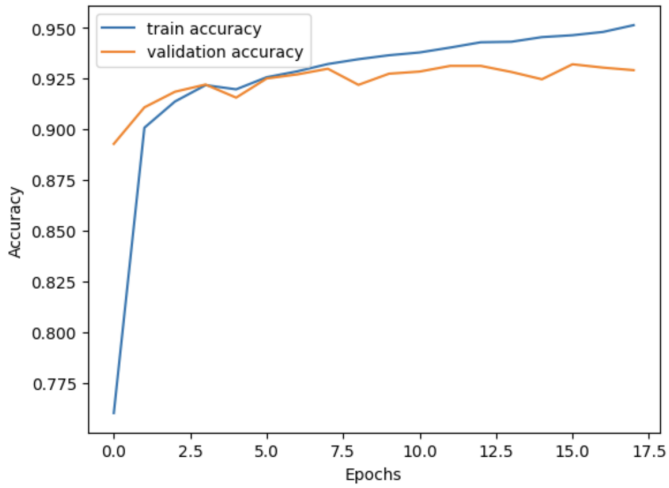

Fig. 2. Illustration of the training and validation accuracy plot for GRU-based approach.

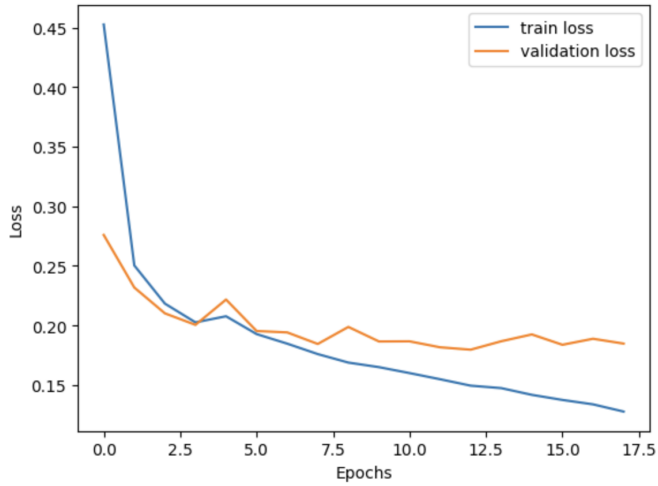

Fig. 3. Illustration of the training and validation loss plot for GRU-based approach.

VI. Conclusion

In this paper, we address the problem of detecting suicidal ideation through the social media activity of individuals on Reddit, proposing two hybrid RNNs as solutions for the same. Both approaches were evaluated using a social media text

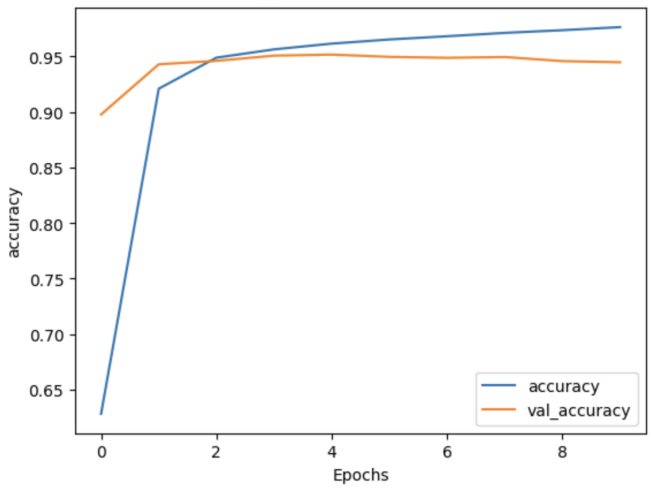

Fig. 4. Illustration of the training and validation accuracy plot for LSTM-based approach.

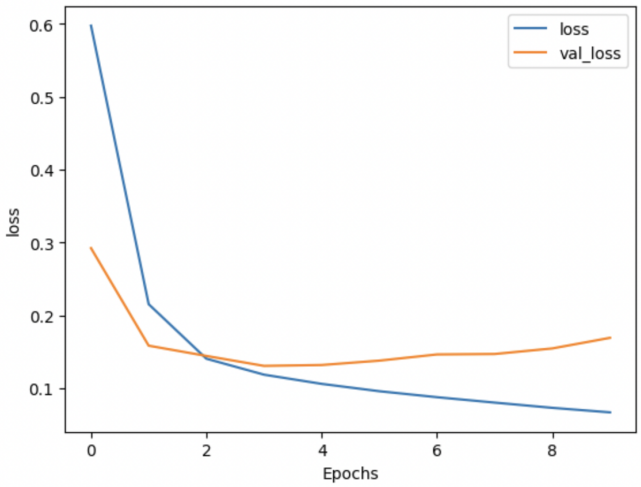

Fig. 5. Illustration of the training and validation loss plot for LSTM-based approach.

dataset obtained from the Reddit platform. As part of the pre-processing, various text cleaning techniques are employed to eliminate redundancy caused by irrelevant information. The initial approach, a GRU-based method, achieved a 93.20% accuracy, while the LSTM-based approach outperformed the former, achieving 94.30% accuracy. The LSTM-based ap-

TABLE V
Class-wise Performance Metrics for GRU and LSTM-based Approaches

Class	Accuracy	Precision	F1-score	Recall
GRU-based Approach				
Non-Suicide	0.926	0.940	0.930	0.930
Suicide	0.938	0.930	0.930	0.940
LSTM-based Approach				
Non-Suicide	0.953	0.930	0.940	0.950
Suicide	0.933	0.950	0.940	0.930

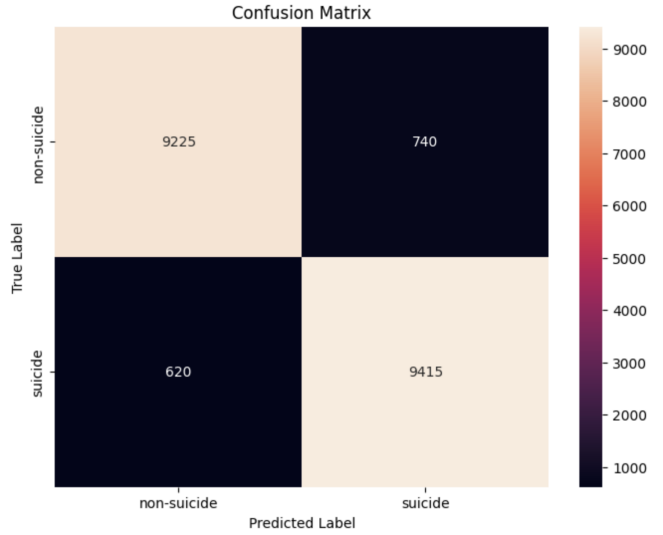

Fig. 6. Illustration of the confusion matrix obtained through an GRU-based approach.

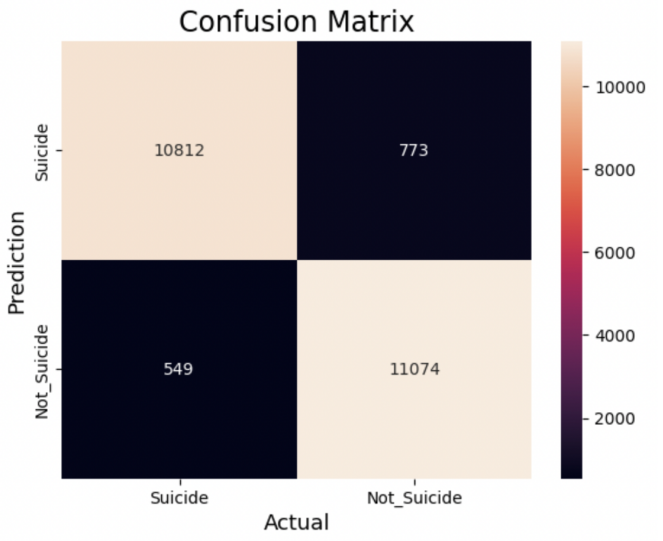

Fig. 7. Illustration of the confusion matrix obtained through an LSTM-based approach.

proach outperformed the GRU-based approach due to its superior ability to handle long-term dependencies and complex sequential patterns. Overall, this study contributes to the ongoing efforts to promote early intervention and support for individuals in distress within the digital landscape.

TABLE VI
PERFORMANCE METRICS OF THE PROPOSED METHODS

Method	Accuracy	Precision	F1-score	Recall
GRU-based	0.9320	0.9271	0.9326	0.9382
LSTM-based	0.9430	0.9347	0.9436	0.9527
LR-n-gram	0.9402	0.9395	0.9402	0.9414
LR-bag-of-words	0.9300	0.9293	0.9302	0.9297

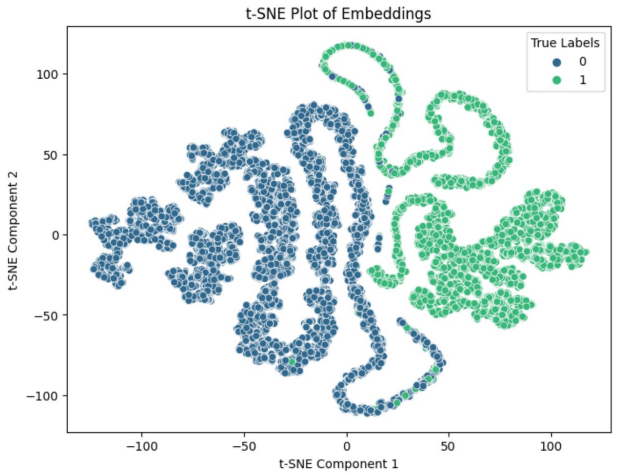

Fig. 8. t-SNE Visualization of the test data.

The future scope of this work involves fine-tuning hyperparameters for enhanced model performance, exploring explainability aspects to provide a better understanding of the black box nature of deep learning networks, and delving into ensemble-based approaches despite time-complexity challenges. These directions promise to improve the performance of the current framework and pave the way for the development of mental health chatbots and automated identification of suicidal ideations from social media monitoring.

REFERENCES

[1] M. Matero, A. Idnani, Y. Son, S. Giorgi, H. Vu, M. Zamani, P. Limbachiya, S. C. Guntuku, and H. A. Schwartz, "Suicide risk assessment with multi-level dual-context language and bert," in *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pp. 39–44, 2019.

[2] J. Robinson, G. Cox, E. Bailey, S. Hetrick, M. Rodrigues, S. Fisher, and H. Herrman, "Social media and suicide prevention: a systematic review," *Early intervention in psychiatry*, vol. 10, no. 2, pp. 103–121, 2016.

[3] R. Radhakrishnan and C. Andrade, "Suicide: an indian perspective," *Indian journal of psychiatry*, vol. 54, no. 4, pp. 304–319, 2012.

[4] X. Li, F. Chen, and L. Ma, "Exploring the potential of artificial intelligence in adolescent suicide prevention: current applications, challenges, and future directions," *Psychiatry*, pp. 1–14, 2024.

[5] H.-S. Choi and J. Yang, "Innovative use of self-attention-based ensemble deep learning for suicide risk detection in social media posts," *Applied Sciences*, vol. 14, no. 2, p. 893, 2024.

[6] R. Kancharapu and S. N. Ayyagari, "Suicidal ideation prediction based on social media posts using a gan-infused deep learning framework with genetic optimization and word embedding fusion," *International Journal of Information Technology*, pp. 1–17, 2024.

[7] L. S. Khoo, M. K. Lim, C. Y. Chong, and R. McNaney, "Machine learning for multimodal mental health detection: A systematic review of passive sensing approaches," *Sensors*, vol. 24, no. 2, p. 348, 2024.

[8] D. Kodati and R. Tene, "Identifying suicidal emotions on social media through transformer-based deep learning," *Applied Intelligence*, vol. 53, no. 10, pp. 11885–11917, 2023.

[9] S. Devika, M. Pooja, M. Arpitha, and R. Vinayakumar, "Bert-based approach for suicide and depression identification," in *Proceedings of Third International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2022*, pp. 435–444, Springer, 2023.

[10] S. Ghosal and A. Jain, "Depression and suicide risk detection on social media using fasttext embedding and xgboost classifier," *Procedia Computer Science*, vol. 218, pp. 1631–1639, 2023.

[11] V. Tejaswini, K. Sathya Babu, and B. Sahoo, "Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 1, pp. 1–20, 2024.

[12] B. Priyamvada, S. Singhal, A. Nayyar, R. Jain, P. Goel, M. Rani, and M. Srivastava, "Stacked cnn-lstm approach for prediction of suicidal ideation on social media," *Multimedia Tools and Applications*, pp. 1–22, 2023.

[13] A. Arowosegbe and T. Oyelade, "Application of natural language processing (nlp) in detecting and preventing suicide ideation: A systematic review," *International Journal of Environmental Research and Public Health*, vol. 20, no. 2, p. 1514, 2023.

[14] A. R. Bhandarkar, N. Arya, K. K. Lin, F. North, M. J. Duvall, N. E. Miller, and J. L. Pecina, "Building a natural language processing artificial intelligence to predict suicide-related events based on patient portal message data," *Mayo Clinic Proceedings: Digital Health*, vol. 1, no. 4, pp. 510–518, 2023.

[15] T. Ghosh, M. H. Al Banna, M. J. Al Nahian, M. N. Uddin, M. S. Kaiser, and M. Mahmud, "An attention-based hybrid architecture with explainability for depressive social media text detection in bangla," *Expert Systems with Applications*, vol. 213, p. 119007, 2023.

[16] A. Sedano-Capdevila, M. Toledo-Acosta, M. L. Barrigon, E. Morales-González, D. Torres-Moreno, B. Martínez-Zaldivar, J. Hermosillo-Valadez, E. Baca-García, F. Aroca, A. Artes-Rodriguez, *et al.*, "Text mining methods for the characterisation of suicidal thoughts and behaviour," *Psychiatry research*, vol. 322, p. 115090, 2023.

[17] A. Swaminathan, I. López, R. A. G. Mar, T. Heist, T. McClintock, K. Caoili, M. Grace, M. Rubashkin, M. N. Boggs, J. H. Chen, *et al.*, "Natural language processing system for rapid detection and intervention of mental health crisis chat messages," *NPJ Digital Medicine*, vol. 6, no. 1, p. 213, 2023.

[18] W. Pan, X. Wang, W. Zhou, B. Hang, and L. Guo, "Linguistic analysis for identifying depression and subsequent suicidal ideation on weibo: machine learning approaches," *International journal of environmental research and public health*, vol. 20, no. 3, p. 2688, 2023.

[19] A. Korda, "Suicide prediction with natural language processing of electronic health records," *medRxiv*, pp. 2023–09, 2023.

[20] H. Ghanadian, I. Nejadgholi, and H. Al Osman, "Socially aware synthetic data generation for suicidal ideation detection using large language models," *IEEE Access*, 2024.

[21] A. Malhotra and R. Jindal, "Xai transformer based approach for interpreting depressed and suicidal user behavior on online social networks," *Cognitive Systems Research*, vol. 84, p. 101186, 2024.