Analyzing Multiple Data Sources for Suicide Risk Detection: A Deep Learning Hybrid Approach

Saraf Anika, Swarup Dewanjee, Sidratul Muntaha
Computer Science & Engineering, East Delta University, Chattogram-4209, Bangladesh

***Abstract*—In the current digital landscape, social media's extensive user-generated content presents a unique opportunity for identifying emotional distress signals. With suicide rates on the rise, this study takes aid of Natural Language Processing (NLP) and Sentiment Analysis to detect suicide risk. Centering primarily around deep learning (DL) architectures, including Convolutional Neural Network (CNN), Bidirectional Gated Recurrent Unit (Bi-GRU) and their combined hybrid BiGRU-CNN model, the research incorporates machine learning (ML) for comparative analysis through multisource datasets from Reddit and Twitter. The methodology commenced with data pre-processing, followed by exploring word embedding techniques. This research included an analysis of both Word2Vec variants as well as pretrained GloVe embeddings, where Skip-Gram paired with Adam optimizer showed superior results. For thorough evaluation, Receiver Operating Characteristic (ROC) curves, Confusion Matrix and Accuracy-Loss graphs were utilized. Furthermore, generalizability of employed models was testified and evaluated by in-depth inspections. The process was accomplished by activating manual input test, cross dataset test and k-fold cross validation procedures. In the course of scrutinizing, the proposed BiGRU-CNN model outperformed the traditional DL and ML models with consistent and reliable performance. Correspondingly, the proposed model achieved accuracies of 93.07% and 92.47% on the respective datasets which advocate its potential as a tool for the early detection of suicidal thought.**

***Keywords*—BiGRU-CNN hybrid; multisource dataset; word embeddings; NLP; sentiment analysis; cross-dataset testing**

I. INTRODUCTION

Suicide is a deeply sensitive and significant issue that demands empathy and understanding. The World Health Organization (WHO) carried out a research investigation in 2021 that declares suicide as a prominent and relentless global cause of death. According to [1] and [2], annually 0.7 million people die by suicide marking it as fourth leading cause of death. Reportedly, majority of the victims fall under the age group of 15 to 29 years. Over the past 15 years, the world witnessed a 24% increase in the overall suicide rate as per The National Institute of Mental Health [3]. In the pandemic year of 2020 suicides increased by 10% from previous year in India and reached a record high of 1, 53, 052 [4].

Suicidal thoughts can arise from various factors like mental health issues, societal pressures, isolation and hopelessness. Social media platforms provide a practical angle for analyzing mental health indicators as individuals tend to often express their emotions and struggles there. Traditional methods for assessing suicide risk are often limited by their time-consuming nature and inability to detect immediate risks. This has led to a paradigm shift towards employing Artificial Intelligence (AI) for text analysis as AI promises enhanced accuracy and efficiency. Human emotions from textual data now can be deciphered by automated system with the help of evolving areas of NLP [5] and Sentiment Analysis [6].

This research sought to detect early signs of suicidal thoughts by developing an automated system. The study proceeded with collecting data from social media platform like Reddit and Twitter. Consequently, the effectiveness of DL architectures was evaluated, particularly CNN, Bi-GRU along with their combined hybrid BiGRU-CNN model. These models were chosen due to their ability of extracting local features and capturing sequential, contextual information from text. ML models including Linear Support Vector Classifier, Logistic Regression, Decision Tree and AdaBoost were selected. Additionally, the study conducted a performance comparison of the proposed hybrid model against these traditional ML and DL algorithms to establish a benchmark. Metrics such as accuracy, precision, recall and F1-scores were evaluated to assess the model's performance.

By addressing the research objective, this study aims to create an automated system that scans social media content to identify early signs of suicidal thoughts and provide a tool for suicide risk detection. This system ought to underscore the significant implications for helping people struggling with suicidal thoughts. The proposed model is anticipated to aid in timely interventions to support individuals who are exhibiting signs of distress on social media platforms.

The core contribution of this work was in implementing the proposed hybrid BiGRU-CNN model, employing both ML and DL models, experimenting upon multiple datasets, conducting trials on various word-embedding techniques and performing cross-dataset test by taking advantage of multi-source data. Moreover, this work assessed and verified the potential of the generalization ability of the models through cross-validation and manual dataset creation. The study demonstrated the superiority of the proposed model through these meticulous process for suicide detection across both datasets. These processes ensure a significant improvement over existing research that often overlook or insufficiently emphasize these vital facets.

The structure of our research paper is methodically organized into following key sections: Section II reviews the relevant existing work. Section III thoroughly describes the employed methodology and explains carried out experiments. Section IV presents the outcomes of the various architectures

tested as well includes an in-depth analysis. Section V provides an interpretation of the discussion. The paper concludes with Section VI, summarizing the findings. Conclusively, Section VII suggests future research.

II. Related Work

We explored the extensive research on depression and suicide risk detection. To capture insights within this domain, we studied researches working on a wide range of methods and data sources.

The research outlined in study [7] explored potential suicidal thoughts in 49,178 tweets using text preprocessing techniques and feature extraction methods. Various DL techniques were trained including LSTM, Bi-LSTM, GRU, Bi-GRU, and a hybrid CNN-LSTM to check the proficiency. Upon evaluation, the Bi-LSTM model stands out with a high accuracy of 93.6% in handling the nature of tweet data.

In study [8], the authors proposed a hybrid model combining CNN with Bi-LSTM to detect depression from Twitter data. The proposed model outperformed traditional RNN and CNN models, achieving a remarkable accuracy of 94.28%.

Another study utilized Reddit data to test the LSTM-CNN hybrid model through the use of Word2Vec embedding techniques. The outshining results confirmed the model's usefulness in text classification [9].

The authors of study [10] predicted text data-based depression with proposed RNN-LSTM techniques, using one-hot approach. Considered data was collected from Kaggle and processed via stemming and lemmatization. The proposed technique proved its ability with a commendable accuracy, excelling other methods like Naive Bayes, Support Vector Machine (SVM), CNN, and Decision Trees.

The investigator's work in study [11] employed various text representation techniques like TF-IDF and Word2Vec, along with a combination of DL (CNN-BiLSTM) and ML (XGBoost) algorithms for text classification.

In contrast to other research, the study [12] applied text mining techniques and numerous algorithms to categorize Cantonese YouTube comments for suicide risk. The paper handled data imbalance using re-sampling and focal loss methods, resulting in g-mean scores of 84.3% and 84.5% for the LSTM model. The best performing model demonstrates the potential for effective automatic suicide risk detection in social media content.

While comparing with SVM, CNN, LSTM and LSTM-CNN combined model, the findings of [13] showcased the efficacy of LSTM-attention-CNN model with 90.3% accuracy. The researchers extracted the Reddit dataset with the assistance of Reddit API to train their employed models.

These previous researches have made significant contributions to the field of suicide detection. However, several limitations exist including reliance on single dataset that may limit the generalizability of their findings and limited exploration of word embedding techniques. Most studies have solely focused on either ML or DL model for their analysis, thereby overlooking the potential benefits of integrating both approaches.

Building upon the strengths of previous research in suicide detection, this study aimed to address key limitations through a comprehensive approach. To get a broader understanding of how people express themselves in different contexts, this research used multiple datasets from various platforms, like Reddit and Twitter. Instead of relying solely on training and testing within the same dataset, a novel cross-dataset testing methodology was implemented. Here, the model was trained on one dataset and then tested on others. This innovative testing process helps in understanding the model's adaptability in real-world applications. Additionally, the study went beyond the commonly used Word2Vec technique and also incorporated pre-trained GloVe embeddings to allow the model to capture meaning from the text.

By addressing these limitations and employing these approaches, our research intends to contribute meaningfully to the advancement of suicide detection and mental stability observation.

III. Experimental Methodology

A. Origin of Data

This study operated on multi datasets for the purpose of ensuring diverse and comprehensive collection of textual content related to mental health and suicidal thought. The Reddit dataset was obtained from the "Suicide Watch" section on Kaggle.[1] It consists of a balanced collection of 232,074 posts which was equally divided with 116,037 posts each in the 'suicide' and 'non-suicide' categories. Additionally, the Twitter dataset was collected from a GitHub repository.[2] It comprises 9,119 tweets and categorized as 5,121 non-suicidal (0) and 3,998 suicidal (1) tweets. The deliberate selection of data from multiple platforms was critical for the models in understanding the expressions of suicidal thought. Table I carries a general overview about the utilized datasets by presenting few sample texts along with their assigned class labels.

TABLE I. Sample Text of Suicidal and Non-suicidal Context

Dataset	Text	Class
Reddit	It ends tonight. I can't do it anymore. I quit.	Suicide
	Its almost 2 am Why am I tired?? It's so early damn	Non-suicide
Twitter	i am looking for someone to talk to all i want to do is die	1
	attempting a poetry essay listening to jessie rose and feeling fat	0

B. Data Preprocessing

Real-world data often comes in a messy and unorganized form with mistakes and irrelevant details. This complexity and inconsistency can negatively impact in model's ability and lead to inaccurate predictions. Pre-processing is a fundamental step that converts those raw data into structured format and positively influence the efficacy of the models.

[1] https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch
[2] https://github.com/laxmimerit/twitter-suicidal-intention-dataset

In our study, the process initiated by removing accented characters and expanding contractions to standardize text expressions. Successively stopwords, symbols URLs, digits, and special characters was discarded to eliminate noise from data. Further refinement included lemmatization, correction of spelling mistakes, and word lengthening. Irreverent words and posts with no content were excluded to eliminate potential noise and bias in the datasets. With the help of a sample text, the process is depicted in Fig. 1. Here red box contains the raw text and cleaned text is presented in the green box.

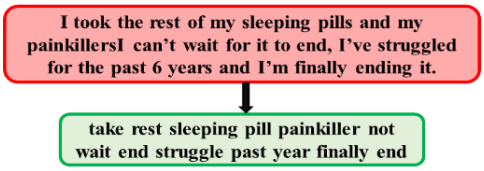

Fig. 1. Sample of preprocessing technique.

Reddit dataset consisted an unnamed column that was irrelevant to this task and was dropped in the cleaning step. After performing data cleaning, Reddit dataset exhibited a class imbalance with 'suicide' posts at 38.9% and 'non-suicide' at 61.1%. The imbalance arose due to original dataset had several 'suicide' labelled rows without any corresponding texts. Consequently, under-sampling was introduced to tackle the potential bias toward the majority class and improve model's performance. Collectively, these preparations were pursued to enhance the model's capability for accurate understanding and detection of the texts.

C. Word Embeddings

Word embedding [14] techniques provide a means of converting textual data into numerical form so that semantic meanings and relationships between words can be captured. This paper exploited both approach of Word2Vec, namely Skip-Gram and Continuous Bag of Words (CBOW) methods to turn words into useful vectors and trained on cleaned-up text from Twitter and Reddit. Moreover, GloVe (global vectors for word representation) embeddings was explored as well, specifically the GloVe 6B dataset collected from online.[3] The pre-trained models such as GloVe present an extensive vocabulary with vectors trained on a vast corpus, therefore, offer a profound source of semantic data.

D. System Overview with Classifiers

The core of our methodology was characterized by the deployment of both ML and DL models along with several word embedding techniques. The architecture in Fig. 2 summarized this entire process, highlighting the systematic and data-driven approach of our research where BiGRU-CNN was denoted as "Hybrid" and AdaBoost was as "Ensemble" methods.

After collecting and pre-processing the data, cleaned data was split into training, validation and testing sets. Following that, training and validation datasets underwent word embedding process through tokenization, effectively transforming the textual data into vector representations. Subsequently, these vectorized data were utilized to train and validate the models. Based on the validation and testing accuracies, manual hyper-parameter tuning was performed. Moreover, the models went under Cross-validation to understand their performance against different subsets of the data. To analyze the effectiveness of the models against unseen data, manual text input was provided for predicting. On top of that, cross-data testing strategy was implemented to ensure versatility, where models were trained on one dataset and tested against the other. Upon performance comparison, the best model was then chosen as proposed model.

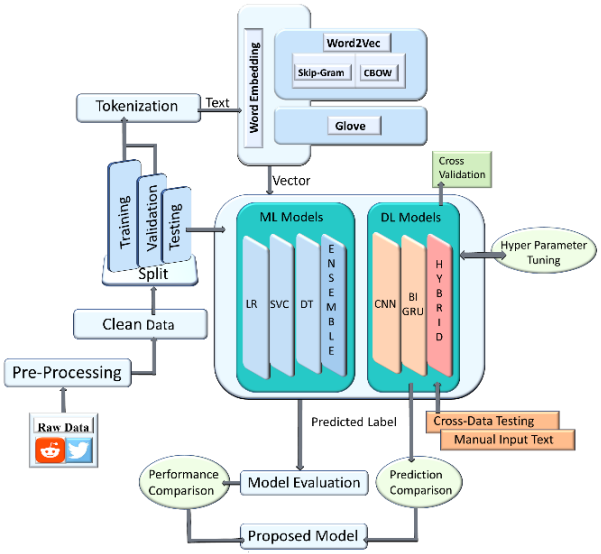

Fig. 2. Architecture of the proposed system.

1) BiGRU-CNN: The proposed model architecture depicted in Fig. 3 combined the features of Bi-GRU and CNN with a view to improve the capability of classifying suicidal text.

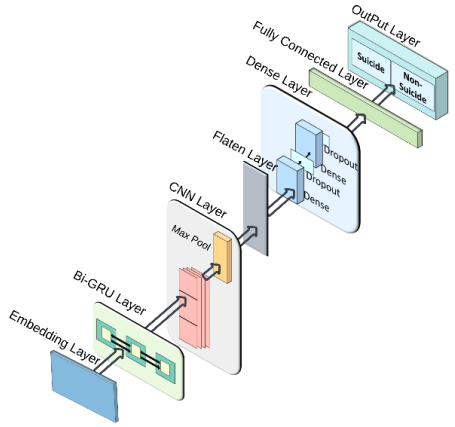

Fig. 3. Proposed model architecture.

While implementing the hybrid model, initially an embedding layer was established that converted text into dense vectors through word embedding. Building on this, a Bi-GRU layer intricately processed these embeddings, capturing the textual context from both forward and backward directions. Thus, it formed a more elaborate representation of the input sequence. The subsequent Conv1D layer applied convolutional

[3] https://nlp.stanford.edu/projects/glove/

operations with multiple filters which proficiently identified and extracted key local features and patterns that were indicative of suicidal thought. Thereafter, dimensionality reduction was achieved through a MaxPooling layer, followed by flattening for dense layer compatibility. Ultimately, dense layers with dropout regularization guided to the output layer, where a sigmoid activation function was employed to compute the probability of suicidal thought.

The layers and parameters along with their corresponding values are tabularized in Table II.

TABLE II. PARAMETERS USED IN PROPOSED MODEL

Layers	Parameters	Values
Embedding	Embedding Dimension	100
Bi-GRU	Units	64
	Dropout Rate	0.25
Conv1D	Filters	128
	Kernel Size	3
	Activation Function	ReLU
MaxPooling1D	Pool Size	2
Dense	Units	128, 64
	Activation Function	ReLU
	Regularization	L2 (0.01)
Dropout	Dropout Rate	0.5
Output	Activation Function	Sigmoid
	Optimizer	Adam
	Loss Function	Binary Cross-entropy
	Batch Size	Twitter Dataset (32)
		Reddit Dataset (128)
	Epoch	20

In our study, we Initialized embedding layer with weights and embedding dimension from the pre-trained Word2Vec and GloVe models. It acts as:

$$Embedded(i) = EmbeddingMatrix[i] \quad (1)$$

Here, i is the word index and $EmbeddingMatrix$ is a matrix of shape containing the learned word embeddings.

The Bi-GRU augments the Gated Recurrent Unit (GRU) by processing data in opposite directions. Prior to exploring the Bi-GRU, we will examine the equations that govern a standard GRU.

For a single GRU cell, the following equations define its operation. Here z_t represents the update gate, r_t denotes the resent gate, σ signifies sigmoid activation function, $\tilde{h}_t$ stands the potential hidden state for the current node in the hidden layer, h_t is the hidden state at time t, and h_{t-1} is the input at time is the hidden state of the previous time step. w and u are weight matrices.

$$z_t = \sigma\,(w_{zx}x_t + u_{zh}h_{t-1}) \quad (2)$$

$$r_t = \sigma\,(w_{rx}x_t + u_{rh}h_{t-1}) \quad (3)$$

$$\tilde{h}_t = \tan\,(w_{hx}x_t + r_t * u_{hh}h_{t-1}) \quad (4)$$

$$h_t = (1 - z_t) * \tilde{h}_t + z_t * h_{t-1} \quad (5)$$

In a Bi-GRU, these operations are performed in two separate GRUs: one processing the sequence from start to end ($h_t^{forward}$) and the other from end to start ($h_t^{backward}$). The final hidden state (h_t) for each time step is a concatenation of these two directional hidden states:

$$h_t = [\, h_t^{forward};\, h_t^{backward} \,] \quad (6)$$

Each GRU in the Bi-GRU has its own set of parameters, and they are trained to capture temporal dependencies in both directions of the input sequence.

Following the Bi-GRU layer, the Conv1D layer applies a 1D convolution operation with $relu$ activation function applied afterwards. $K(u)$ embodies the value of kernel at position u and $I(i+u)$ is the input feature at position $i+u$.

$$F(i) = \sum_{u=0}^{Kernel\_size-1} I(i+u).K(u) \quad (7)$$

$$relu(x) = \max(0, x) \quad (8)$$

After that, max pooling operation was performed over the 1D input. For input feature map F, pooling size $poolsize$, and output feature map P, the max pooling operation at position (i) is:

$$P(i) = \max_{0 \le u < poolsize} F(i+u) \quad (9)$$

The Flatten layer simply reshaped the output to a single dimension. Dense layers performed linear transformation followed by $relu$ activation, with $L2\ Regularization$ applied to the weights w:

$$Flatten(F) = Reshape\ to\ 1D(F) \quad (10)$$

$$y = relu(w_x + b) + L2\ Regularization \quad (11)$$

To conclude the model, a sigmoid activation function was used to provide the probability of the text indicating suicidal ideation.

The proposed model's ability to understand both the context and specific language signs in social media data makes it suited for identifying potential signs of suicidal thought.

Beyond the proposed model, our study explored additional models to comprehensively analyze suicide risk, including:

2) CNN: Designited for its intriguing ability to detect local patterns and features in textual data [15]. The arhitecture consisted of a 100-dimentioal embedding layer to encode words, followed by a Conv1D layer with 128 filters of size 3 and for feature extraction ReLU activation function was applied. To prevent overfitting, the model included MaxPooling1D and Dropout layers. Other configurations remained similar to the proposed model. CNN model was utilized through the layers and parameters that are shown in Table III.

TABLE III. OPERATIONAL DETAILS OF CNN MODEL

Layers	Parameters	Values
Embedding	Embedding Dimension	100
Conv1D	Filters	128
	Kernel Size	3
	Activation Function	ReLU
MaxPooling1D	Pool Size	2
Output	Activation Function	Sigmoid

3) Bi-GRU: Chosen for its proficiency in capturing contextual information from text sequences in both forward and backward directions [16]. The Bi-GRU model also employs a 100-dimentional embedding layer simmilar to CNN and proposed model. It incorporated two BiGRU layers with 128 and 64 units and recurrent connections for bidirectional sequence processing. The dense and dropout layers, optimizer, loss function, batch sizes, and epochs are retained as in the CNN configuration. Table IV illustrates the values of parameters used in Bi-GRU model.

TABLE IV. OPERATIONAL DETAILS OF BI-GRU MODEL

Layers	Parameters	Values
Embedding	Embedding Dimension	100
Bi-GRU Layer	Units	128,64
	Dropout	0.25
	Recurrent Dropout	0.25
Output	Activation Function	Sigmoid

4) Linear Support Vector Classifier (Linear SVC): Employed for its effectiveness in binary classification of textual data. In our context, the below mentioned decision function was used to predict the new data points classes that were fed into the Linear SVC model.

$$f(x) = sign\,(w.x + b) \tag{12}$$

where, $f(x)$ is the decision function that determines the class of a new data point x. $sign(\,)$ function assigns a class based on the sign of the result.

5) Logistic Regression (LR): Applied due to its simplicity and efficiency in probabilistic prediction.[4] The LR model is suitable for binary classification tasks like distinguishing between suicidal ideation and general content in our datasets. This is achieved through the logistic function, formally represented as:

$$P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2 \cdots\cdots \beta_n X_n)}} \tag{13}$$

Here, $P(Y = 1|X)$ is the probability that the dependent variable Y is equal to 1, given the independent variables X. β_0 is the intercept and e is the base of the natural logarithm. The prediction is typically classified as 1 (suicide class) if $P(Y = 1|X)$ is greater than 0.5, and as 0 (Non-suicide class) otherwise.

[4]https://machinelearningmastery.com/logistic-regression-for-machine-learning/

6) Decision Tree (DT): Chosen for its ability to understand complex language structures and relationships [17]. DT aids in suicide risk assessment through decision paths.

7) AdaBoost: Selected to enhance the performance of decision trees. The ensemble method combines multiple weak learners to form a strong classifier and improve the accuracy of suicide risk predictions from textual data [18].

The most representative equation for AdaBoost classifier is the final model decision function. It is a weighted sum of the weak classifier's decisions characterized as:

$$H(x) = sign(\textstyle\sum_{t-1}^{T} \alpha_t \cdot h_t(x)) \tag{14}$$

In the equation (1.14), $H(x)$ represents the final decision function of the AdaBoost classifier and T is the total number of weak classifiers used. In this study, the value of T was specified as 50. α_t denotes the weight of the 't' th weak classifier in the ensemble.

As hyperparameter tuning, the parameters including batch size, epoch, optimizer, dropout rate, kernel size, and hidden units were systematically adjusted. After identifying the best combined parameters, the performance of all models was evaluated.

IV. RESULT ANALYSIS

Following precise pre-processing and data cleaning, the model training and evaluation phases proceeded. The process began by allocating 70% of the data for training, 20% for validation, and 10% for testing. This division was designed to rigorously assess model's performance and generalizability.

The procedure of result analysis initiated with Word clouds generation for each class. The word clouds provided a visual representation of word frequencies. They were created using the 'WordCloud' library for displaying words with sizes proportional to their frequency in the texts. This approach highlighted the most prominent words in larger fonts, allowing for easy identification of key terms characteristic of suicidal and non-suicidal texts. The clouds of each context for both datasets are shown in Fig. 4.

As this task primarily centered around DL techniques, performance of the DL models was examined and evaluated subsequently. Table V provides a detailed breakdown of evaluation metrices. For Reddit dataset, the proposed model exhibited the highest accuracy of 93.07%. Accrued F1-score of 0.93 indicates a balanced precision-recall relationship. The performance of the proposed model remained notable on the Twitter dataset as well, with an accuracy reaching 92.47% and mentionable precision and recall for both classes. With consistent performance against both datasets, the model establishes its strength in classification tasks.

In examining word embedding's influence on model efficacy, outcomes revealed the prominence of Skip-Gram method over alternative techniques such as CBOW and GloVe embeddings. This finding emerged as the most effective, leading the study for further analysis.

Fig. 4. Word clouds of a) Reddit b) Twitter.

TABLE V. RESULTS OBTAINED FROM DL MODELS

Data set	Models	Class	Precision	Recall	f1-score	Accuracy
Reddit	CNN	Suicide	0.92	0.91	0.92	91.30%
		Non-suicide	0.91	0.91	0.91	
	Bi-GRU	Suicide	0.93	0.92	0.92	92.12%
		Non-suicide	0.91	0.92	0.92	
	Proposed model	Suicide	0.93	0.94	0.93	**93.07%**
		Non-suicide	0.93	0.92	0.93	
Twitter	CNN	1	0.94	0.83	0.88	89.28%
		0	0.86	0.95	0.90	
	Bi-GRU	1	0.93	0.88	0.90	91.11%
		0	0.89	0.94	0.92	
	Proposed model	1	0.94	0.90	0.92	**92.47%**
		0	0.92	0.94	0.93	

On a different note, we operated two distinct optimizers: Adam and Stochastic Gradient Descent (SGD). The empirical findings indicated that the Adam optimizer continually outclassed SGD in terms of model's accuracy, securing its selection as the preferred optimizer. Table VI is showcasing the consequences.

TABLE VI. CHOOSING WORD-EMBEDDINGS AND OPTIMIZERS

Dataset	Methods	Skip-Gram		CBOW	GloVe
		SGD	**Adam**		
Reddit	CNN	90.85%	**91.30%**	90.52%	89.81%
	Bi-GRU	91.44%	**92.12%**	91.28%	90.75%
	Proposed Model	91.77%	**93.07%**	92.77%	92.36%
Twitter	CNN	88.25%	**89.28%**	87.34%	88.48%
	Bi-GRU	88.37%	**91.11%**	88.60%	89.17%
	Proposed Model	89.40%	**92.47%**	89.97%	90.19%

To obtain a deeper understanding of the capabilities of DL models, the accuracy-loss curve, confusion matrix and ROC curve were generated and inspected.

The ROC curves illustrated in Fig. 5 provides a visual representation of the ability of implemented classifiers to discriminate between suicide and non-suicide instances. Area under the curve (AUC) indicating the overall performance. Observing the outcomes, it is apparent that the proposed model demonstrates the highest discriminative power. The model reaches AUC values of 0.9797 and 0.9745 for Reddit and Twitter datasets, respectively. Being compared to the CNN and Bi-GRU models, performance of the proposed approach underscores the usefulness in capturing the complexities of both datasets.

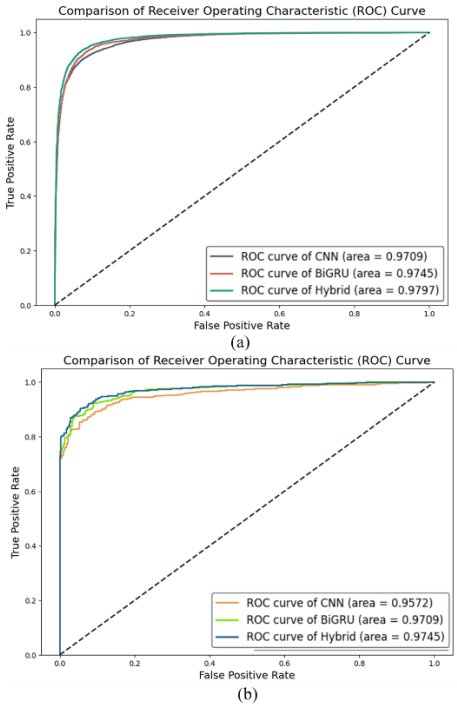

Fig. 5. ROC curves of (a) Reddit (b) Twitter.

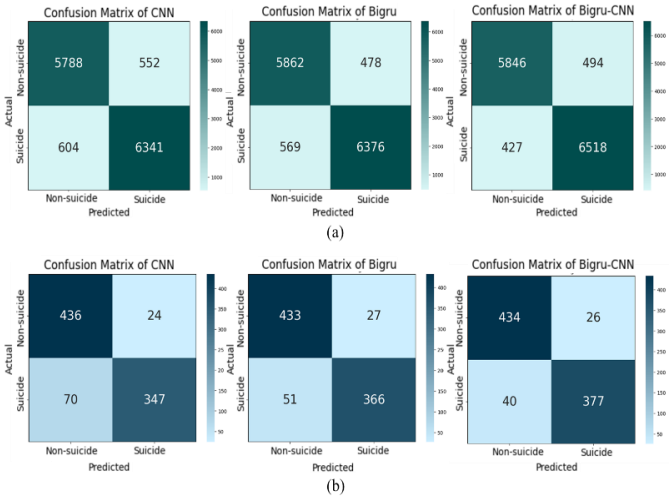

Fig. 6. Confusion matrices of (a) Reddit (b) Twitter.

We utilized confusion matrix to assess the effectiveness of our implemented models. The confusion matrix offers insights into true and false predictions and helps us understanding our model's performance. Derived metrics like precision, recall, accuracy, and F1-score gave a complete view of each model's capabilities and limitations. Fig. 6 showcases the confusion matrices for CNN, Bi-GRU, and BiGRU-CNN, starting from left.

In addition, the accuracy-loss curves in Fig. 7 and Fig. 8 guided the decision on when to stop training for avoiding overfitting. The early stopping technique was employed based on the validation loss. Here, the accuracy curve illustrated the percentage of accurate predictions in both training and validation phases. The loss curve indicated the extent to which the model's predictions differ from the actual values.

With the aim of further authenticating the performance of the models, a manual dataset was created containing twenty texts. The outcome is tabularized in Table VII that portrays five texts along with assigned labels and corresponding predicted labels by the models. The left most column of the table, indicates the dataset that the models were trained on prior to manual input testing. It is apparent to note that, the proposed model demonstrated comparatively satisfactory performance in this regard as well. The proposed model that was trained on the Reddit dataset was accurate in every prediction, whereas, Twitter trained proposed model predicted one suicidal text incorrectly.

Cross -dataset testing was executed to prolong the layers of rigorous scrutiny on model's resilience. To do so, models were tested by data they have not encountered during training. For instance, models trained on the Reddit dataset underwent testing against the Twitter test split and vice versa. The corresponding values presented in Table VIII revealed the proposed model's strong adaptability against the others.

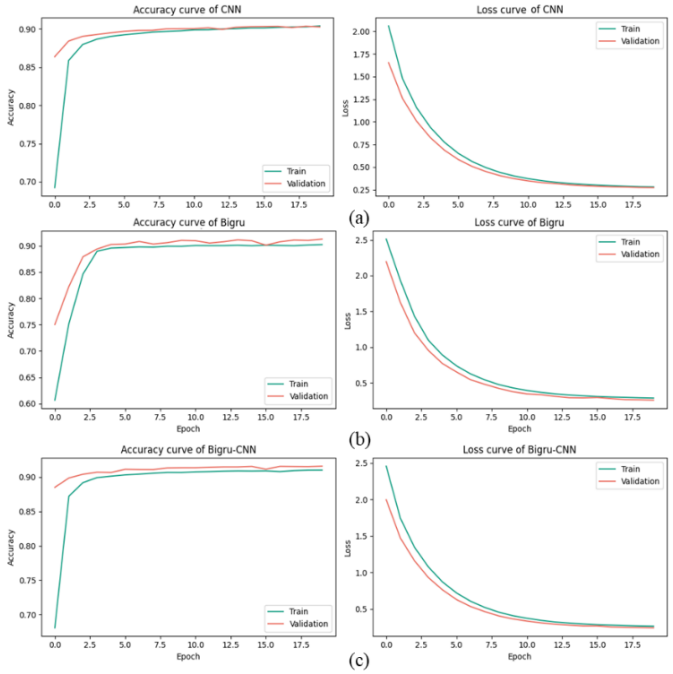

Fig. 7. Curves of (a) CNN (b) Bi-GRU (c) Proposed model of Reddit.

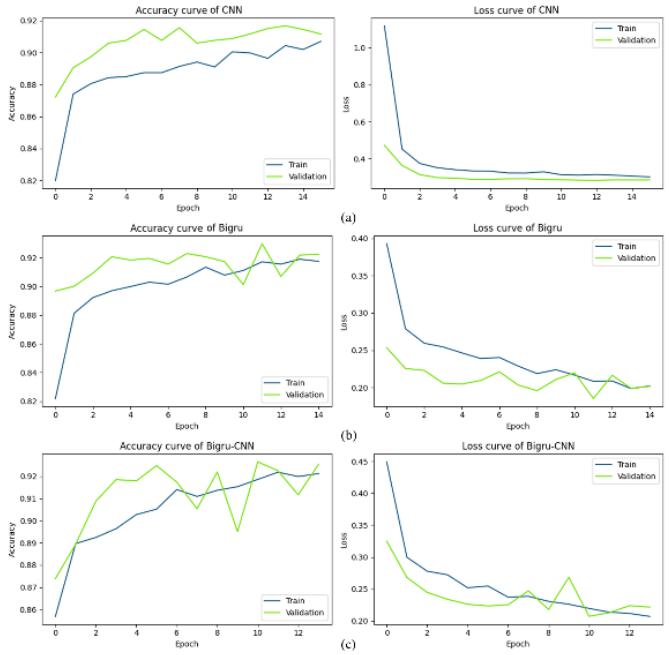

Fig. 8. Curves of (a) CNN (b) Bi-GRU (c) Proposed model of Twitter.

TABLE VII. PREDICTED RESULTS OF MANUAL INPUT TEXT

Trained On	Text	Actual Label	Predicted Label		
			CNN	Bi-GRU	Proposed Model
Reddit	Recommend pill to suicide	Suicide	Non-Suicide	Suicide	Suicide
	Don't tell me what to do. I am depressed	Suicide	Suicide	Suicide	Suicide
	The sun is shinning	Non-Suicide	Non-Suicide	Non-Suicide	Non-Suicide
	I want to end my life	Suicide	Non-Suicide	Non-Suicide	Suicide
	Going to commit suicide. No more toxicity.	Suicide	Suicide	Suicide	Suicide
Twitter	Recommend pill to suicide	Suicide	Non-Suicide	Non-Suicide	Non-Suicide
	Don't tell me what to do. I am depressed	Suicide	Non-Suicide	Suicide	Suicide
	The sun is shinning	Non-Suicide	Non-Suicide	Non-Suicide	Non-Suicide
	I want to end my life	Suicide	Non-Suicide	Non-Suicide	Suicide
	Going to commit suicide. No more toxicity.	Suicide	Suicide	Suicide	Suicide

As a final validation step to ensure model's reliability, K-fold cross-validation was implemented. The cross-validation determined the stability and consistency of models across multiple subsets of data.[5] This method systematically availed one-fold for testing and the rest for training and then averaged the results. The findings are structurally compiled in Table IX.

[5]https://machinelearningmastery.com/k-fold-cross-validation/

The data reveals a decline in accuracies with an increase in number of folds.

TABLE VIII. CROSS DATASET TESTING

Trained with	Tested By	Test Data Descriptions		Models Performance		
				CNN	Bi-GRU	**Proposed Model**
		Actual Class Distribution		Correctly Predicted Class		
Reddit	Twitter	0	460	246	274	319
		1	417	385	370	347
Accuracy				71.95%	73.43%	**75.94%**
Twitter	Reddit	Non-Suicide	6340	4228	4068	3994
		Suicide	6946	5209	5428	5593
Accuracy				71.04%	71.48%	**72.16%**

TABLE IX. PERFORMANCE THROUGH CROSS-VALIDATION

Folds	Reddit		Twitter	
	Model	Mean Accuracy	Model	Mean Accuracy
K=5	CNN	90.71%	CNN	88.95%
	Bi-GRU	91.68%	Bi-GRU	89.68%
	Proposed Model	**92.07%**	**Proposed Model**	**90.25%**
K=10	CNN	90.53%	CNN	88.78%
	Bi-GRU	91.57%	Bi-GRU	89.53%
	Proposed Model	**91.72%**	**Proposed Model**	**89.64%**

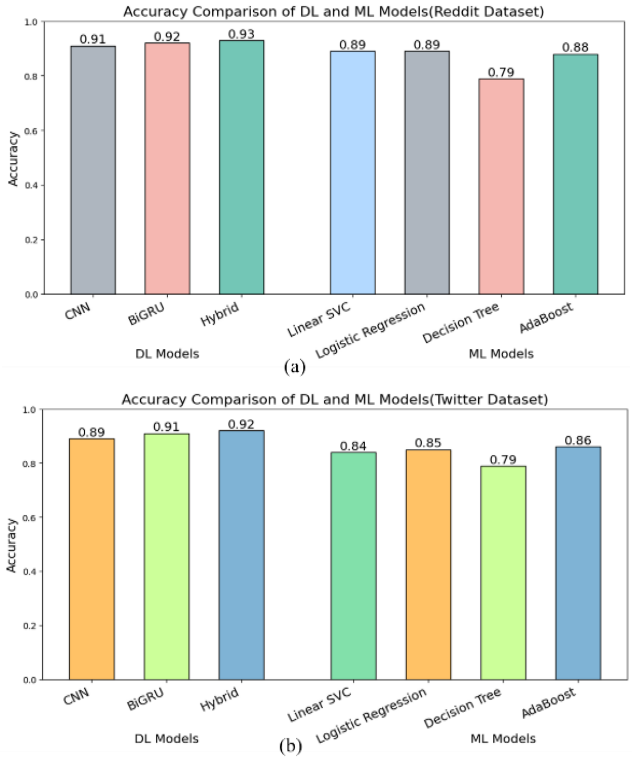

Fig. 9. Accuracy comparison among DL and ML of (a) Reddit (b) Twitter.

The research concluded with performing a comparative analysis of ML models to assess their respective effectiveness compared to Dl models. Therefore, four traditional ML models were accounted for. The performance analysis is visually portrayed with a bar chart presented in Fig. 9 which explicitly demonstrates the supreme accuracy of DL models for both datasets. It highlights the effectiveness of DL approaches over ML models in predicting suicidal contents.

A complete synopsis of the outcomes acquired from ML models, including evaluation metrices, is systematically tabulated in Table X.

TABLE X. RESULTS OBTAINED FROM ML MODELS

Data set	**Models**	**Class Label**	**Precision**	**Recall**	**f1-score**	**Accuracy**
Reddit	Linear SVC	Suicide	0.83	0.81	0.82	88.55%
		Non-suicide	0.91	0.92	0.92	
	LR	Suicide	0.84	0.80	0.82	88.55%
		Non-suicide	0.91	0.93	0.92	
	DT	Suicide	0.66	0.69	0.68	78.58%
		Non-suicide	0.85	0.83	0.84	
	AdaBoost	Suicide	0.82	0.81	0.81	87.97%
		Non-suicide	0.91	0.91	0.91	
Twitter	Linear SVC	1	0.77	0.92	0.84	83.69%
		0	0.91	0.77	0.84	
	LR	1	0.78	0.93	0.85	84.83%
		0	0.92	0.78	0.85	
	DT	1	0.77	0.80	0.78	79.36%
		0	0.82	0.79	0.80	
	AdaBoost	1	0.82	0.90	0.86	85.97%
		0	0.90	0.83	0.86	

The essence of findings from Table X indicates a decline in accuracy for most models on the Twitter dataset, however, the F1-score remained consistent for both classes. In contrast, the Reddit dataset, while displaying better accuracy, struggled to identify suicide class. When compared with the results presented in Table V, it is evident that DL models showcased higher competency over ML models.

V. DISCUSSION

Suicide has emerged itself as a global threat to humanity, demanding urgent actions and attentions to address the complex challenges it presents. As early identification of suicidal thoughts is the essential strategy to prevent suicide, our study aimed to detect suicidal tendencies by analyzing social media interactions. Pursuing that, we employed both ML and DL methods and assessed their performance. This study primarily emphasized on DL models while ML models were utilized for comparison. We conducted comprehensive pre-processing and experimented with several word embedding techniques. To verify the outcomes of the models, we utilized multi-source datasets. However, the length restriction of a

tweet often proves inadequate for expressing an individual's complex emotions, which might be responsible for the slight performance decline of the models trained on Twitter dataset. We implemented various strategy to examine the generalizability of the models. Despite the notable outcomes of our work, the models possess limitations to assess the risk variables associated with suicide psychology. Pairing our findings with the statistical analysis of the psychologists could make these models a better indicator of suicidal thoughts.

VI. Conclusion

This research aimed to develop and evaluate a detection system to identify suicidal thoughts of individuals by analyzing their social media post. The study intended to reduce suicide rate by detecting suicidal thoughts in the social media interactions of the users. To accomplish this, textual data from Reddit and Twitter was collected and both ML and DL methods was utilized with a view to assess the psychological states of the users. Through meticulous evaluation and comparison of several ML and DL methods, the study has established the commendable predictive accuracy of the proposed model. The proposed model evidently captured the discrepancies of the data in a more effective manner. It evolved into the most proficient model for suicide detection by securing accuracies of 93.07% and 92.47% respectively, on Reddit and Twitter datasets. Subsequently, a manual dataset was created, cross-dataset testing strategy was introduced and cross-validation was performed for further analysis of the generalizing ability of the models. The proposed model showcased superior results in these regards as well. Hence, the hybrid BiGRU-CNN model was proposed as an effective tool for analyzing mental health from social media content to expose suicidal thoughts.

This research concludes by highlighting the proficiencies of the proposed hybrid model and indicating a strong potential for real-world deployment. The insights attained from this study ought to lead the way for future research directions and practical applications. This research contribution is a step forward in the integration of AI into mental health services. It aimed to provide strong foundation for future advancements in NLP and AI-assisted mental health monitoring.

VII. Future Scope

Regardless of promising outcomes demonstrated by the proposed model, there always remains scope for future development. Introducing the proposed model to more diverse data from social media platforms including Facebook and Instagram could significantly enhance the model's understanding, thus accuracy. Additionally, real-world testing through direct integration with social media could provide valuable insights into its reliability and competency in practical applications. Advanced pre-trained transformer models particularly BERT, ELECTRA, and neural networks such as HANN could be employed with a view to compare them against the proposed model. Given the sensitivity of the task at hand, continuous refinement and improvement are crucial to ensure the model's resilience and real-world impact.

References

[1] "One in 100 deaths is by suicide." Accessed: Nov. 09, 2023. [Online]. Available: https://www.who.int/news/item/17-06-2021-one-in-100-deaths-is-by-suicide.

[2] "Suicide." Accessed: Nov. 09, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/suicide.

[3] "NIMH » Suicide." Accessed: Nov. 09, 2023. [Online]. Available: https://www.nimh.nih.gov/health/statistics/suicide.

[4] "Accidental Deaths & Suicides in India Report 2020 : NCRB." Accessed: Nov. 09, 2023. [Online]. Available: https://www.drishtiias.com/daily-news-analysis/accidental-deaths-suicides-in-india-report-2020-ncrb.

[5] "What is Natural Language Processing? | IBM." Accessed: Nov. 09, 2023. [Online]. Available: https://www.ibm.com/topics/natural-language-processing.

[6] A. Oussous, F. Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "ASA: A framework for Arabic sentiment analysis," https://doi.org/10.1177/0165551519849516, vol. 46, no. 4, pp. 544–559, May 2019, doi: 10.1177/0165551519849516.

[7] R. Haque, N. Islam, M. Islam, and M. M. Ahsan, "A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning," Technologies 2022, Vol. 10, Page 57, vol. 10, no. 3, p. 57, Apr. 2022, doi: 10.3390/TECHNOLOGIES10030057.

[8] H. Kour and M. K. Gupta, "An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM," Multimed Tools Appl, vol. 81, no. 17, pp. 23649–23685, Jul. 2022, doi: 10.1007/S11042-022-12648-Y/FIGURES/20.

[9] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of Suicide Ideation in Social Media Forums Using Deep Learning," Algorithms 2020, Vol. 13, Page 7, vol. 13, no. 1, p. 7, Dec. 2019, doi: 10.3390/A13010007.

[10] A. Amanat et al., "Deep Learning for Depression Detection from Textual Data," Electronics 2022, Vol. 11, Page 676, vol. 11, no. 5, p. 676, Feb. 2022, doi: 10.3390/ELECTRONICS11050676.

[11] T. H. H. ; Aldhyani et al., "Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models," International Journal of Environmental Research and Public Health 2022, Vol. 19, Page 12635, vol. 19, no. 19, p. 12635, Oct. 2022, doi: 10.3390/IJERPH191912635.

[12] J. Gao, Q. Cheng, and P. L. H. Yu, "Detecting comments showing risk for suicide in YouTube," Advances in Intelligent Systems and Computing, vol. 880, pp. 385–400, 2019, doi: 10.1007/978-3-030-02686-8_30/COVER.

[13] S. Renjith, A. Abraham, S. B. Jyothi, L. Chandran, and J. Thomson, "An ensemble deep learning technique for detecting suicidal ideation from posts in social media platforms," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 10, pp. 9564–9575, Nov. 2022, doi: 10.1016/J.JKSUCI.2021.11.010.

[14] S. F. Sabbeh and H. A. Fasihuddin, "A Comparative Analysis of Word Embedding and Deep Learning for Arabic Sentiment Classification," Electronics 2023, Vol. 12, Page 1425, vol. 12, no. 6, p. 1425, Mar. 2023, doi: 10.3390/ELECTRONICS12061425.

[15] P. Ce and B. Tie, "An Analysis Method for Interpretability of CNN Text Classification Model," Future Internet 2020, Vol. 12, Page 228, vol. 12, no. 12, p. 228, Dec. 2020, doi: 10.3390/FI12120228.

[16] J. Teng, W. Kong, Y. Chang, Q. Tian, C. Shi, and L. Li, "Text Classification Method Based on BiGRU-Attention and CNN Hybrid Model," ACM International Conference Proceeding Series, pp. 614–622, Sep. 2021, doi: 10.1145/3488933.3488970.

[17] R. Li, M. Liu, D. Xu, J. Gao, F. Wu, and L. Zhu, "A Review of Machine Learning Algorithms for Text Classification," Communications in Computer and Information Science, vol. 1506 CCIS, pp. 226–234, 2022, doi: 10.1007/978-981-16-9229-1_14/FIGURES/2.

[18] N. Kalcheva, M. Todorova, and G. Marinova, "Naive Bayes Classifier, Decision Tree and Adaboost Ensemble Algorithm – Advantages and Disadvantages," 6th ERAZ Conference Proceedings (part of ERAZ conference collection), pp. 153–157, 2020, doi: 10.31410/ERAZ.2020.153.