Grenze International Journal of Engineering and Technology, June Issue

Deep Learning–based NLP Approaches for Depression Detection on Social Media Data: A Comparative Study

M D. Gulshan Firdous[1] and Thota Radha Rajesh[2]
[1-2]Vignan's Foundation for Science/Department of Computer Science and Engineering, Guntur, India
Email: firdousgull07@gmail.com, rajeshsakvara@gmail.com

***Abstract*— The detection of depression from social media text has become an important research topic in the field of Natural Language Processing (NLP). Online platforms such as Twitter and Reddit contain extensive user generated content where individuals frequently express emotions, personal struggles, and psychological states. These digital expressions provide meaningful linguistic patterns that can be analyzed to identify early signs of depression. This study presents a comparative analysis of several deep learning architectures for detecting depressive content in social media posts. The models evaluated include Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), Bidirectional LSTM (BiLSTM), Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), and Graph Convolutional Networks (GCN). A standardized preprocessing pipeline involving text cleaning, normalization, and tokenization is applied to improve data quality before feature extraction using word embedding techniques. Model performance is measured using accuracy, precision, recall, and F1-score to ensure a comprehensive evaluation. In addition, a hybrid architecture combining BiLSTM and BiGRU layers is proposed to capture both long-term contextual dependencies and short-term sequential patterns within textual data. Experimental results indicate that the proposed hybrid model out performs baseline approaches in detecting depression related language. The findings highlight the importance of contextual representation and sequential learning in building reliable AI-based systems for early mental health monitoring.**

***Index Terms*— Depression Detection, Hybrid BiLSTM–BiGRU Model, Deep Learning, Natural Language Processing, Social Media Text, Mental Health Monitoring.**

I. INTRODUCTION

Depression is a common mental health disorder that affects millions of individuals worldwide and continues to be a major public health challenge [1]. Although awareness has increased, early identification remains difficult because many people avoid seeking professional help due to stigma, social barriers, or limited medical access. As a result, researchers are investigating alternative approaches for early detection, particularly through the analysis of digital behavior and social media activity [2]. Social networking platforms such as Twitter and Reddit allow users to share personal experiences, emotions, and daily struggles in textual form [3]. These online expressions often contain linguistic indicators such as negative sentiment, repetitive thoughts, and language associated with [4]. Social withdrawal or hopelessness [5]. Traditional machine learning methods typically relied on keyword frequency, sentiment scores, and predefined lexicons. However, these approaches are limited in capturing deeper contextual meaning and semantic relationships within text [6].

Grenze ID: 01.GIJET.12.2.114_20

Deep learning techniques have significantly improved text-based mental health analysis by automatically learning complex representations from raw data [7]. Models including RNN, LSTM, GRU, CNN, and GCN are widely applied for depression detection tasks. Convolutional networks extract local textual features, whereas graph-based models analyze relational information such as word co-occurrence or user–post interactions [8]. Recurrent models are particularly effective in learning sequential dependencies. This study compares multiple deep learning architectures on a unified dataset and introduces a hybrid BiLSTM–BiGRU framework to enhance contextual and sequential pattern learning for improved depression detection.

II. Literature Review

A. NLP for Mental Health Detection

Textual data from social media platforms contain emotional cues, linguistic patterns, and psychological indicators that can assist in mental health analysis[9]. Natural Language Processing (NLP) techniques are widely applied to extract these features and identify depression-related patterns. Early approaches relied on lexicon-based methods such as LIWC, which analyze emotion-related words, personal pronouns, and cognitive expressions. However, these models depend heavily on predefined vocabularies and often fail to capture deeper contextual meaning. Word embedding models such as Word2Vec, GloVe, and fastText significantly improved text representation by mapping words into dense vector spaces using distribu- tional semantics. Word2Vec utilizes skip-gram co-occurrence learning, fastText incorporates sub-word information, and GloVe leverages global corpus statistics. These embeddings enable better semantic understanding beyond simple keyword matching. More recently, transformer-based models including BERT, RoBERTa, and DistilBERT employ bidirectional self-attention mechanisms to generate dynamic contextual embeddings[10]. Such models effectively interpret complex emotional expressions. Studies show that depressed individuals frequently use first-person pronouns, negative emotion words, and absolutist terms such as “never” or “nothing,” reflecting self-focus and hopelessness. Incorporating these lin- guistic features into deep learning frameworks enhances the accuracy of depression detection systems.

B. Deep Learning Models in NLP

Deep learning architectures play a crucial role in mental health prediction by analyzing complex linguistic patterns in social media text. Recurrent Neural Networks (RNNs) are among the early neural models used in NLP, as they process sequential data and capture temporal dependencies[11]. However, conventional RNN networks have inherent limitations in capturing long-term contextual dependencies in text sequences. which restrict their ability to retain long-term contextual information.To overcome these challenges, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks were introduced[12]. These models incorporate gating mechanisms to regulate information flow and preserve relevant con- text over extended sequences. LSTM utilizes input, output, and forget gates, whereas GRU employs a simplified structure with fewer parameters. Their ability to model long-term contextual and emotional patterns makes them effective for de- pression detection tasks. Bidirectional LSTM (BiLSTM) further enhances contextual understanding by processing text in both forward and backward directions.Convolutional Neural Networks (CNNs) extract local textual features using convolu- tional filters and n-gram patterns, making them suitable for short social media posts[13].. Deep Neural Networks (DNNs) rely on dense layers for feature learning but lack sequential modeling capabilities[14]. Graph Convolutional Networks (GCNs) extend analysis to relational structures such as word co-occurrence and user–post interaction networks[15].Given the strengths and limitations of these models, comparative evaluation is essential. Therefore, this study proposes a hybrid BiLSTM–BiGRU architecture to integrate contextual and sequential learning for improved depression detection.

III. Methodology

The methodology followed in this paper starts with the collection of raw social media text, such as posts, tweets or comments. These texts contain informal language, abbreviations, emojis and varying sentence lengths. The collected texts must be preprocessed that include tokenization, lower case conversion, or removal of unnecessary icons, symbols, emojis and unrelated words. The processed text is then confined to fixed lengths of words. The cleaned text is then converted to numerical form with the help of an embedding layer. This layer connects each word to a dense vector that extracts semantic and contextual information. Pre-trained word embeddings such as Word2Vec or GloVe are widely used to encode language patterns from large texts . Here, Word2Vec is used in this architecture. This layer identifies similarities between words related to depression. The embedded text is then processed through BiLSTM layer. This layer proceeds in sequence in both forward and backward directions to extract information from both past and future words. This bidirectional aspect is important in identifying depression as the meaning of a word or a phrase depend in the context that which is mentioned. Long term dependencies are retained in the memory of the model with the help of LSTM cells. Prolonged emotional states and recurring negative expressions across social media text are thus identified. The output of the BiLSTM layer is given as an input to BiGRU layer. This layer uses related short-term dependencies and unique language checks.Bidirectional nature of GRU ensures the preservation of the contextual data from both the directions. Thus, this model is able to identify explicit emotional variations with in the text. A dropout layer is then used to prevent outfitting and improve generalization. This layer randomly disables some neurons. This enables the model to learn more robust feature descriptions not merely depending on specific neurons. The architecture of the proposed model is shown in Fig. 1.

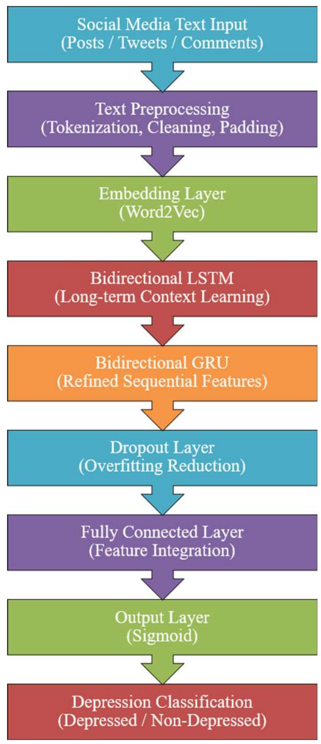


Figure 1: Architecture of the Proposed Model

Fig. 1 illustrates the architecture of the proposed The extracted features are passed through a dense layer to transform them into a suitable classification representation. A sigmoid activation function in the output layer generates a probability score indicating whether the text is depressive or non-depressive. Based on this probability threshold, the final classification is determined.The model's performance is evaluated using a confusion matrix, which summarizes classification results into true positives, true negatives, false positives, and false negatives. From these values, evaluation metrics such as accuracy, precision, recall, and F1-score are computed. This evaluation approach provides a comprehensive measure of the model's effectiveness in identifying depressive content while minimizing misclassification, thereby ensuring the reliability and robustness of the proposed system.

A. Dataset

Twitter mental health datasets are widely used for depression detection research due to the platform's open structure and frequent emotional self-disclosure. The dataset used in this study is balanced to prevent class bias, containing equal numbers of depressive and non-depressive samples. A 70–15–15 split is adopted for training, validation, and testing to ensure reliable evaluation. Word2Vec embeddings are applied to capture semantic relationships between words and improve contextual understanding. The dataset characteristics are summarized in Table 1

TABLE I: DATASET DESCRIPTION FOR DEPRESSION DETECTION

Attribute	Description
Data Source	Twitter mental health dataset
Data Type	Textual social media posts (tweets)
Language	English
Task	Binary classification (Depressed / Non-depressed)
Total Samples	400
Depressed Class	200 posts
Non-depressed Class	200 posts
Average Post Length	12–25 words
Preprocessing Applied	Lowercasing, URL removal, hashtag removal, tokenization, lemmatization, stop-word removal
Embedding Technique	Word2Vec embeddings
Training Set	70% (280 samples)
Validation Set	15% (60 samples)
Test Set	15% (60 samples)
Annotation Method	Dataset-provided labels and keyword-based self- reported indicators
Ethical Considerations	Publicly available data; user anonymity preserved; no personally identifiable information used

B. Experimental Setup

The experimental framework is developed using Python, leveraging established deep learning libraries for model con- struction and evaluation. Training is carried out with the AdamW optimization algorithm using a learning rate of2*10 and a weight decay factor of 0.01. Binary cross-entropy serves as the objective function for binary classification. The models are fine-tuned for 3 to 5 epochs with batch sizes ranging from 8 to 16. To enhance training stability and avoid overfitting, techniques such as learning-rate scheduling and early stopping are applied. The effectiveness of the models is measured us- ing standard performance metrics including accuracy, precision, recall, and F1-score, ensuring a comprehensive evaluation of depression detection capability.

IV. RESULTS AND ANALYSIS

The comparative evaluation of different models highlights their effectiveness in detecting depressive content. The evaluation metrics derived from the confusion matrix are presented in Table 2. The performance comparison of selected NLP models is summarized in Table 3.

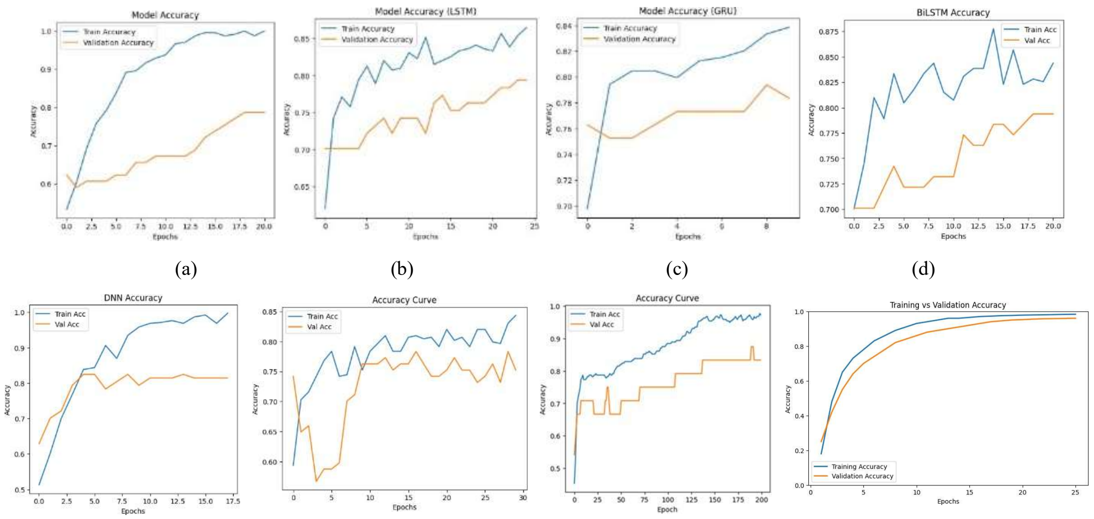
TABLE II: METRICS FOR CONFUSION MATRIX

Metric	**Formula**
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-Score	$\frac{2\times Precision\times Recall}{Precision + Recall}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Macro Average	$\frac{Metric_{class\ 0}+Metric_{class\ 1}}{2}$
Weighted Average	$\frac{(Metric_0\times Support_0) + (Metric_1\times Support_1)}{Support_0+Support_1}$

TABLE III: COMPARISON OF PERFORMANCE PARAMETERS OF SELECTED NLP MODELS

Models	Precision	Recall	F1-score	Accuracy
RNN	0.80	0.82	0.81	0.92
LSTM	0.90	0.75	0.82	0.87
GRU	0.94	0.81	0.87	0.82
BiLSTM	0.86	0.75	0.80	0.85
DNN	0.87	0.79	0.83	0.82
CNN	0.62	0.92	0.74	0.78
GCN	0.91	0.77	0.83	0.90
Proposed Model (Hybrid BiLSTM and BiGRU)	0.80	0.94	0.86	0.96

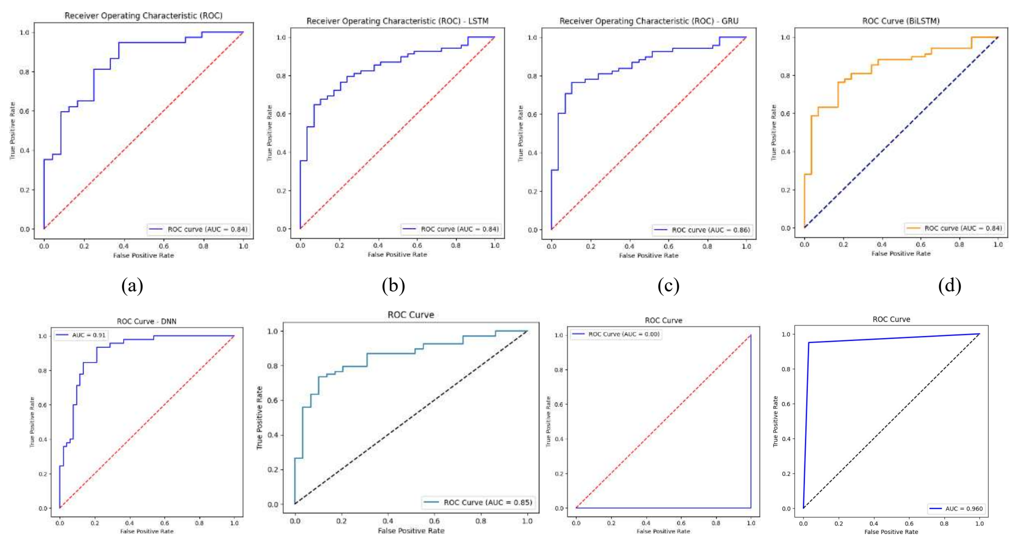
The basic RNN model achieves an accuracy of 0.92, precision of 0.80, recall of 0.82, and an F1-score of 0.81. Although it detects most depressive instances, the model exhibits lower classification reliability due to a higher number of false positives. The BiLSTM model improves contextual understanding and sequential learning, achieving a precision of 0.86, recall of 0.75, F1-score of 0.80, and accuracy of 0.85. However, relying primarily on long-term dependencies limits its ability to capture short-term emotional patterns common in social media text. The LSTM model attains a higher precision of 0.90, but its recall decreases to 0.75, resulting in an F1-score of 0.82 and an overall accuracy of 0.87, indicating reduced effectiveness in identifying a substantial portion of depressive content.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2: Accuracy and loss graphs of (a) RNN, (b) LSTM, (c) GRU, (d) BiLSTM, (e) DNN, (f) CNN, (g) GCN, and (h) Proposed Model (Hybrid BiLSTM and BiGRU).



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 3: ROC graphs of (a) RNN, (b) LSTM, (c) GRU, (d) BiLSTM, (e) DNN, (f) CNN, (g) GCN, and (h) Proposed Model (Hybrid BiLSTM and BiGRU).

Similarly, the GRU model presented a recall of 0.81 and a precision of 0.94, resulting in an accuracy of 0.82 and an F1- score of 0.87. Although the model demonstrated strong feature classification capability, it failed to detect certain depressive indications. The CNN model achieved a high recall of 0.92, indicating strong sensitivity to depressive words and phrases. However, it recorded a low precision of 0.62, leading to an F1-score of 0.74 and an overall accuracy of 0.78, suggesting limited discrimination capability. The DNN model obtained an F1-score of 0.83, accuracy of 0.82, precision of 0.87, and recall of 0.79. While it maintained balanced performance across metrics, its overall effectiveness remained moderate. The GCN model achieved a precision of 0.91, recall of 0.77, F1-score of 0.83, and accuracy of 0.90. Although effective in extracting structural features, it failed to capture all depressive instances comprehensively.

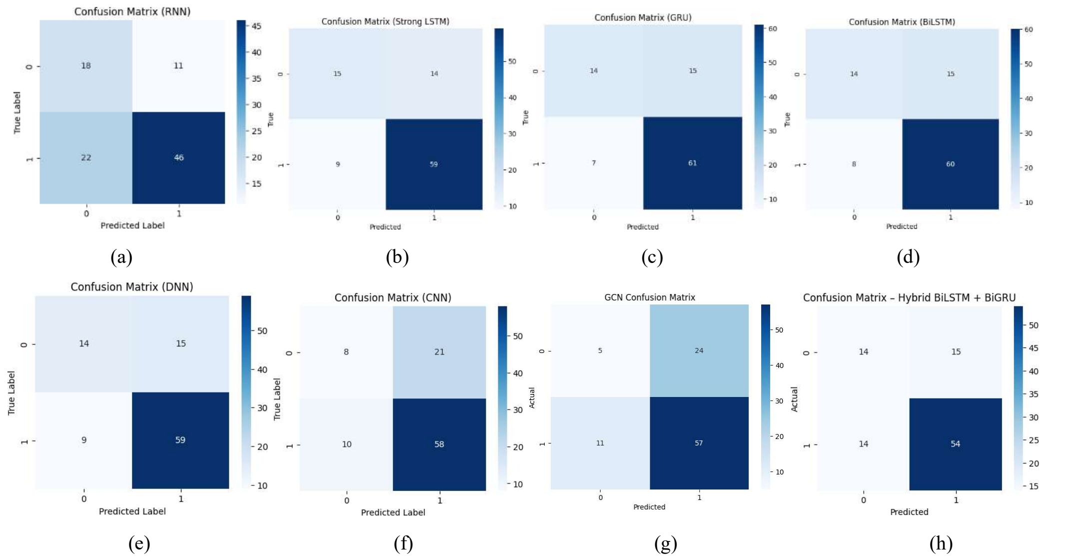


Figure 4: Confusion matrices of (a) RNN, (b) LSTM, (c) GRU, (d) BiLSTM, (e) DNN, (f) CNN, (g) GCN, and (h) Proposed Model (Hybrid BiLSTM and BiGRU).

The proposed hybrid BiLSTM–BiGRU model achieves a precision of 0.80, recall of 0.94, F1-score of 0.86, and the highest accuracy of 0.96, significantly outperforming all selected baseline models. The high F1-score indicates an opti- mal balance between precision and recall, while the superior recall demonstrates the model's capability to minimize false negatives. By integrating BiLSTM and BiGRU layers, the proposed architecture effectively captures both long-term con- textual dependencies and short-term emotional patterns. Consequently, the hybrid model delivers more robust and reliable performance for depression detection from social media text compared to standalone deep learning approaches.

V. Conclusion

Selected existing NLP models are evaluated along with the proposed hybrid BiLSTM and BiGRU model and their performance parameters are compared. The proposed model achieved an accuracy of 0.96, a recall of 0.94. It also dis- played an unprecedented high score of F-1 of 0.86. The combination of two bidirectional recurrent layers of two high performing models have considerably reduced both false positives and false negatives successfully. Long term parameters depending on context is efficiently dealt by BiLSTM. Short term language parameters are handled effectively by BiGRU. Thus, the hybrid model with the application of multiple recurring learning techniques equips to become a more robust and reliable architecture to detect depression in the texts from social media. It also promises a strong potential not only in the applications of mental health monitoring but also in various fields like customer sentiment and opinion mining, fraud and anomaly detection, disease prediction, education analytics, workplace well-being and burnout detection, Cybersecurity and insider threat detection and public opinion and social trend analysis. Future work may focus on expanding the dataset, incorporating multilingual data, and integrating attention mechanisms to further improve performance and interpretability of the proposed model.

References

[1] "Depression: let's talk," World Health Organization, Mar. 30, 2017. [Online]. Available: https://www.who.int/news/item/30-03-2017–depression-let-s-talk-says-who-as-depression-tops-list-of-causes-of- ill-health

[2] N. Wang, R. Kamil, S. A. R. Al-Haddad, and N. Ibrahim, "Early warning of adolescent depression: A joint detec- tion model combining social media speech behaviour and psychological scales," J. Mind Behav., vol. 46, no. 2, pp. 188–212, 2025.

[3] S. F. Waterloo, S. E. Baumgartner, J. Peter, and P. M. Valkenburg, "Norms of online expressions of emotion: Comparing Facebook, Twitter, Instagram, and WhatsApp," New Media Soc., vol. 20, no. 5, pp. 1813–1831, 2018.
[4] D. Gu et al., "Cognitive change in online mental health communities," Inf. Process. Manage., vol. 60, no. 2, p. 103192, 2023.
[5] A. Caputo, "Emotional and symbolic components of hikikomori experience," Mediterr. J. Clin. Psychol., vol. 8, no. 1, 2020.
[6] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art," Multimed. Tools Appl., vol. 82, no. 3, pp. 3713–3744, 2023.
[7] I. H. Sarker, "Deep learning: A comprehensive overview," SN Comput. Sci., vol. 2, no. 6, p. 420, 2021.
[8] R. Chiong et al., "Depression detection using social media texts," Comput. Biol. Med., vol. 135, p. 104499, 2021.
[9] S. Lyu et al., "Detecting depression via text analysis," Front. Psychiatry, vol. 14, p. 1121583, 2023.
[10] O. Karakaya and Z. H. Kilimci, "FastText+BiLSTM for classification," PeerJ Comput. Sci., vol. 10, p. e1831, 2024.
[11] A. Khan and A. Sarfaraz, "RNN-LSTM-GRU based language transformation," Soft Comput., vol. 23, no. 24, pp. 13007–13024, 2019.
[12] M. M. A. Busst et al., "Ensemble BiLSTM approach," IEEE Access, vol. 12, pp. 3528–3539, 2024.
[13] T. Zhang et al., "Text feature extraction using CNN," in Commun. Comput. Inf. Sci., vol. 727, 2017, pp. 472–485.
[14] W. Ma et al., "Mental health research with graph neural networks," in Proc. IEEE ICDEW, 2025, pp. 197–210.
[15] Z. Rahimi and M. M. Homayounpour, "Impact of preprocessing on word embedding quality," Lang. Resour. Eval., vol. 57, no. 1, pp. 257–291, 2023.