IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 August 2024, accepted 16 October 2024, date of publication 30 October 2024, date of current version 11 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3488081

RESEARCH ARTICLE

Integrating Bert With CNN and BiLSTM for Explainable Detection of Depression in Social Media Contents

CAO XIN AND LAILATUL QADRI ZAKARIA
Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi 43600, Malaysia

Corresponding author: Lailatul Qadri Zakaria (lailatul.qadri@ukm.edu.my)

This work was supported by the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia.

ABSTRACT Depression is a prevalent mental health condition that significantly impacts individuals' lives. Early detection of depression is crucial for timely intervention and improved outcomes. However, traditional machine learning approaches are constrained by the limited amount of annotated data and lack of model transparency. This study aims to address these challenges by leveraging social media data and advanced natural language processing techniques to develop effective and explainable models for depression detection. The study focuses on two main objectives. The first objective is to develop and evaluate fine-tuned Bidirectional Encoder Representations from Transformers (BERT), BERT with Bidirectional Long Short-Term Memory (BERT-BiLSTM), and BERT with Convolutional Neural Network (BERT-CNN) models, and compare their performance with MentalBERT, a state-of-the-art model for mental health detection. The second objective is to observe the key features used by the BERT models to make the decision-making using Transformer Interpretability Beyond Attention Visualization and Average Attention Weight methods. The study utilizes three publicly available datasets: the Depression Reddit Dataset, the Sentiment Analysis for Tweets Dataset, and the Mental Health Corpus. The results demonstrate that the proposed models, especially BERT-BiLSTM and BERT-CNN, achieve superior performance compared to MentalBERT, particularly regarding accuracy and F1-score. Notably, BERT-CNN achieved exceptional accuracy scores of 0.982, 0.961, and 1.0 on the Depression Reddit Dataset, the Mental Health Corpus, and the Sentiment Analysis for Tweets Dataset, respectively, demonstrating its robust performance across different social media contexts. The attention map visualizations provide valuable insights into the language patterns and key features associated with depression in social media posts. This study contributes to the mental health field by presenting novel and explainable models for depression detection using social media data. The proposed approaches have the potential to assist mental health professionals in early identification and intervention, ultimately improving the lives of individuals affected by depression.

INDEX TERMS Automatic depression detection, BERT, CNN, BiLSTM, MentalBERT, deep learning, Explainability AI.

I. INTRODUCTION

Mental health is an increasing condition worldwide nowadays. Roughly one in five persons suffer from a mental health issue that can significantly impact their lives in all aspects, such as school or work behavior and family and friends relationships [1]. Depression is one of the most prevalent and serious mental health conditions. Depression can lead to tiredness and poor concentration, even self-harm and suicide. Major depressive disorder instances have reportedly increased by 27.6% as a result of the COVID-19 pandemic and the public health and social measures that followed, particularly in the post-pandemic era [31]. Depression causes huge economic losses of US$ one trillion annually [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Vivek Kumar Sehgal.

With the prevalence of social media, people tend to post openly about their daily lives, personal experiences, opinions, and challenges on platforms such as Facebook, Twitter, and Reddit [3]. These platforms have provided new avenues for self-expression and social connection. Users frequently share their thinkings, feelings, and experiences through status updates, tweets, posts, and comments. Importantly, these digital traces can serve as reflections of users' mental states and personal lives to some extent [4]. People often discuss situations or events causing stress, sadness, or other emotions. Those suffering from mental health issues like depression may express symptoms such as hopelessness, guilt, irritability, negative self-talk, expressions of loneliness, decreased social interaction, and even suicidal ideation. While not definitive on their own, these social media posts can provide warning signs and risk factors indicating depressive disorders.

Early detection of depression offers significant benefits, as it enables seeking timely intervention from doctors, which can prevent further deterioration of the condition and improve outcomes [5]. Untreated depression can lead to impaired functioning, strained relationships, and decreased quality of life, underscoring the need for methods to identify signs of depression as soon as possible. One promising approach is to gather data from social media platforms where people often express their thoughts and feelings openly. This study explores the potential of BERT with deep learning (DL) for uncovering patterns of depressive language within digital footprints. The DL approach is highly utilized in several applications to perform prediction and classification [27], [29]. This would allow for the identification of at-risk individuals and the facilitation of the care they need. Analyzing social media posts presents a scalable avenue to realize the benefits of early detection and potentially improve the lives of millions suffering from this condition worldwide.

However, one key challenge in depression detection is the limited amount of data [6]. Manual annotation of large datasets for training depression classification models is expensive and time-consuming, requiring trained professionals and involves a complex annotation process [28]. This data scarcity restricts model performance. A novel classification approach using Bidirectional Encoder Representations from Transformers (BERT) is introduced to address this challenge. BERT is a pre-trained model that leverages large amounts of text data to capture contextual information and generate rich language representations [7], [26]. Unlike traditional ML models, BERT can be fine-tuned on a smaller dataset and still achieve promising results [6].

Nevertheless, developing a solution that is merely effective in identifying potential depression-related texts is insufficient in today's context [8]. The growing influence and accountability of artificial intelligence (AI) in society have raised concerns about its reliability [9]. These concerns stem from the fact that most AI solutions, particularly DL Networks, are inherently opaque or 'black-box' models [10]. The intricacy of these models, arising from their vast parameter space and complex algorithmic combinations, renders them incomprehensible to human users, meaning that the decision-making process cannot be fully understood [11]. Such models may contain biases and base their decisions on unfair, obsolete, or erroneous assumptions, which traditional approaches to evaluating model performance can overlook. As a result, there is a lack of trust in these opaque models.

To address the challenges associated with 'classical' AI and enhance the trustworthiness of depression detection models, this study aims to improve the explainability of the depression detection model. In this study, the first objective is to develop three BERT-based models for depression detection: fine-tuned BERT, BERT-BiLSTM, and BERT-CNN. The performance of these models is then compared with MentalBERT [12] using the same dataset. MentalBERT is a tailored version of the BERT model, specifically designed to tackle tasks associated with mental health, including depression detection. It is pre-trained on a large corpus of mental health-related text data, allowing it to capture domain-specific knowledge and language patterns associated with mental health conditions. By comparing the proposed models (fine-tuned BERT, BERT-BiLSTM, and BERT-CNN) with MentalBERT, this study aims to evaluate the effectiveness of the proposed models against a state-of-the-art benchmark tailored for mental health applications. The second objective is to understand the model's decision-making process, this study implements a novel visualization method by [13] called Transformer Interpretability Beyond Attention Visualization (TIBAV). The TIBAV method computes the relevance of different parts of the input data to the final classification decision made by a transformer-based model [14].

The main contributions of this study to the field of depression detection and natural language processing (NLP) include demonstrating the effectiveness of BERT-based models for depression detection across diverse datasets, conducting a comparative analysis with Mental-BERT, providing insights into the performance of specialized mental health models, incorporating explainability analysis to enhance the transparency and trustworthiness of the depression detection models, proposing a novel approach that combines BERT with BiLSTM and CNN architectures for improved depression detection, and contributing to the development of automated tools for early mental health detection and intervention.

The organization of the remaining sections of the study is as follows. Related work on depression detection using social media data is presented in Section II. Section III covers the methodology, including dataset introduction, model construction, and experiment setup. Section IV provides comparative results on each dataset, discussion, and model explainability visualization. Conclusion and Future Work are presented in the final Section V.

II. RELATED WORK

This section provides a comprehensive overview of the existing research on depression detection using social media data, focusing on applying ML and DL techniques and the importance of explainability in mental health contexts. Authors in [15] proposed a depression detection approach using K-nearest neighbor (KNN). The authors performed depression detection by analyzing a large collection of comments on Facebook to detect emotions that can indicate depression. The study reached ground truth results between 60-70% about various metrics levels.

Authors in [16] explored detecting depression from Twitter data by utilizing two different types of classifiers: Naïve Bayes, and a hybrid NBTree model. The authors collected tweets, preprocessed the data, and then performed sentiment analysis using TextBlob to assign polarity scores and label them as positive, negative, or neutral based on the scores. The labeled data was fed into Naive Bayes and NBTree classifiers on WEKA. Both models achieved high accuracy, with 92.34% on the 1000 tweet dataset and 97.31% on 3000 tweets. The authors show promise in using DL and social media data for depression detection and screening.

Authors in [17] constructed an SVM model utilizing multiple kernels to detect individuals experiencing depression. The authors first divided data into three categories: microblog text, user profile, and behaviors, then considered the heterogeneity between the three categories and adaptively chose the best kernel. This research highly reduced the error rate compared to Naive Bayes, Decision Trees, and KNN.

Authors in [10] utilized Random Forest for major depression disorder detection. Two ML models were proposed: a singleton model using a single random forest classifier along with decision threshold functions and a dual model that utilized two separate Random Forest classifiers, one to identify depressed subjects and another for non-depressed subjects. The research showed that the dual model surpasses the singleton model, improving early detection rates by over 10% compared to current state-of-the-art methods. The research concluded that the use of ML on social media data can aid in the early detection of depression, which is very important for timely intervention and potentially reducing the disorder's impact on public health.

Authors in [4] investigated different deep neural network architectures for detecting depression in Twitter users using posts from the CLPsych 2015 shared task dataset. They compared the CNN based model with recurrent neural network (RNN) based models. With the optimized word embeddings, their CNN models worked best for depression detection at the user level.

Authors in [18] highlighted a novel text-based multi-task BGRU (Bidirectional Gated Recurrent Unit) network approach to detect depression from clinical interviews. The paper proposed a new network with pre-trained word embeddings to analyze patient responses, predicting both depression presence and severity scores. Results show that pre-training and using sentence-level embeddings enhance accuracy. The model achieves a high F1 score of 0.84 and a low error of 3.48 on a benchmark dataset.

To develop a timely depression diagnostic system, [19] introduced a model that combines two hidden layers and a substantial bias with an RNN comprising two dense layers to implement the long-term memory (LSTM) model. The proposed approach shows impressive outcomes in identifying early signs of depression through the emotions of numerous social media users, thereby showcasing the effectiveness of RNN and LSTM architectures.

As DL models, such as BERT, continue to advance the field of depression detection, the importance of explainability cannot be overstated. In mental health applications, where the stakes are high, and decisions can have significant consequences for individuals' well-being, it is crucial to understand how these models arrive at their outputs. Explainability is essential for building trust, ensuring transparency, and facilitating the adoption of DL-based depression detection tools in clinical practice. Techniques like SHapley Additive exPlanations (SHAP) [20] and Local Interpretable Model-Agnostic Explanations (LIME) [21] can be used to estimate the importance of specific words, n-grams, or semantic features in the context of depression detection. These methods provide a more granular understanding of the model's behavior and can help identify the most informative aspects of the input text.

III. METHODOLOGY

The framework of this study is summarized in Figure 1. Generally, this study is divided into six phases: data collection, preprocessing, model construction, model training, model evaluation and comparison analysis, and finally, model explanation. Data preprocessing involves cleaning, tokenization, training, and splitting test datasets. Model construction describes the architecture of BERT and its integration with BiLSTM and CNN, respectively. Model training refers to training the three models using depression detection datasets for optimal performance. Model evaluation employs standard metrics like accuracy, precision, recall, and F1-score to assess the three models' predictive performance. Comparative Analysis compares the performance of the three models with MentalBERT to determine the more effective approach. Model explainability focuses on visualizing the model's decision-making process through attention maps to showcase explainability.

A. DATASET

This study utilizes three publicly available depression detection datasets obtained from Kaggle: (Depression: Reddit Datase [22]), (Sentimental Analysis for Tweets [23]), and (Mental Health Corpus [24]). These datasets consist of textual data and corresponding labels. Each dataset comprises a diverse range of samples. Before training the depression detection models, the textual data from the three datasets

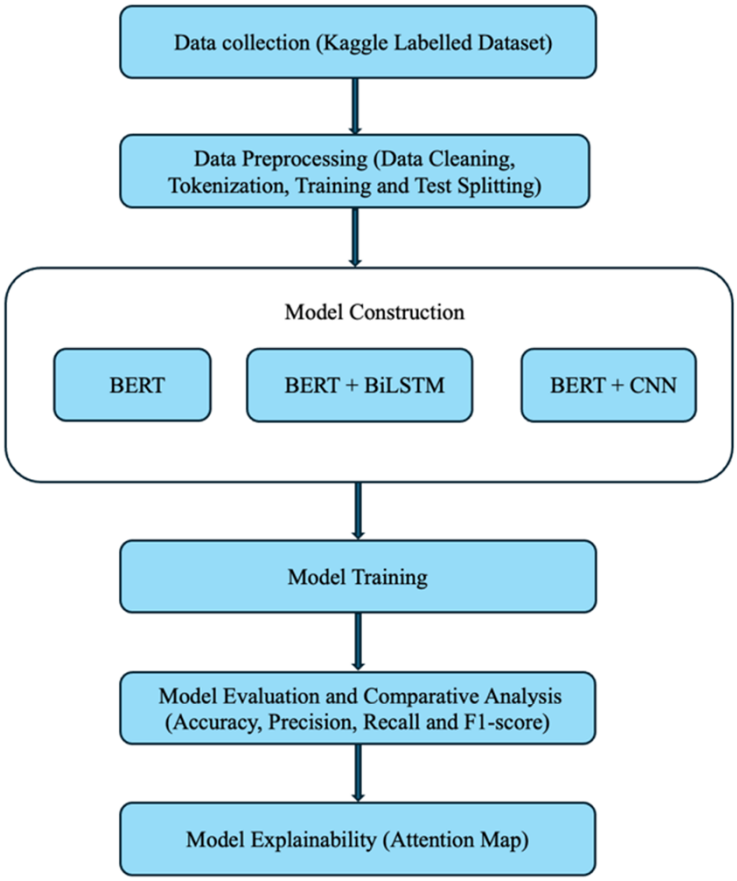

FIGURE 1. Research framework.

underwent a series of preprocessing steps to prepare it for the subsequent modeling tasks as follows:

1) **Data Cleaning**: The first step involved cleaning the raw text data. This included removing irrelevant content such as `@mentions`, hashtags, URLs, and other non-textual elements.
2) **Tokenization**: This study uses `BertTokenizer Fast` to tokenize the text following the data cleaning process. This advanced tokenizer provides efficient tokenization of text into subword units known as tokens. It converts the tokens into numerical IDs that can be used as input to the BERT-based models.
3) **Train-Test Split**: This study employs stratified splitting to divide the cleaned and tokenized datasets into training and test sets (typically 80/20). This ensures that the depression and control labels are proportionally represented in both sets, as illustrated in Table 1.

TABLE 1. Dataset description.

Dataset	Training			Test		
	Depression	Control	Total	Depression	Control	Total
Reddit Dataset	3064	3120	6184	767	780	1547
Tweet Dataset	1851	6400	8251	1600	463	2063
Mental Health Dataset	11070	11311	22381	2768	2828	5596

B. MODELS CONSTRUCTION

1) FINE-TUNED BERT

Figure 2 shows how the model was constructed. In this model, a dropout layer was introduced immediately following the pooler output of the BERT Model to adapt the BERT Model for depression detection. The pooler output serves as the input to this dropout layer, which helps to regularize the model and prevent overfitting. During the fine-tuning process, the entire model undergoes end-to-end optimization. This includes the parameters of the additional linear classifier, denoted as $W \in K \times H$, where H represents the dimension of the hidden state vectors, and K corresponds to the number of classes. Cross-entropy loss serves as the objective function minimized during fine-tuning, it measures the difference between the true and predicted class distributions. The dropout layer helps to improve the model's generalization ability. In contrast, optimizing the entire model allows it to learn task-specific patterns and make accurate predictions based on the textual input.

2) BERT-BILSTM

The BERT-BiLSTM model structure, as shown in Figure 3, integrates 2 layers of BiLSTM, each with a hidden size equal to the BERT model's hidden size, after the last hidden state

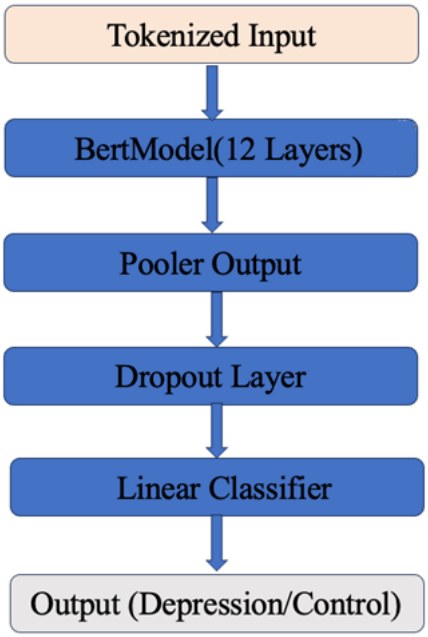

FIGURE 2. The fine-tuned BERT structure.

output of the BERT Model. The BiLSTM layer takes the last hidden state as input, allowing it to capture sequential dependencies and contextual information from both forward and backward directions. By concatenating the hidden states generated by these two LSTMs, a bidirectional representation that incorporates contextual details from both the past and the future has been obtained. The final hidden state of the BiLSTM layer is passed through a linear classifier with weights $W \in K \times 2H$, where H represents the dimension of the hidden state vectors and K corresponds to the number of classes. The factor of two in the dimension of W is due to the concatenation of the bidirectional LSTM outputs. To obtain class probabilities, a softmax activation function is applied to the output logits of the linear classifier.

3) BERT-CNN

Figure 4 shows the structure of the BERT-CNN model, which consists of two main components: the BERT Model and the TextCNN. The input text is passed through 12 self-attention layers in the BERT Model. This study extracts the hidden states from the last 12 layers of the BERT Model, excluding the first layer, which is an embedding layer. Specifically, this study focuses on the [CLS] token embeddings. The [CLS] token in BERT is a special token added at the start of input sequences to capture the overall meaning, and it is primarily used for classification tasks. This study loops through the last 12 layers to extract the [CLS] token embeddings and concatenate them to form the input for the TextCNN. The TextCNN architecture is designed to capture important features from the input text using convolutional layers with different filter sizes. In the BERT-CNN, the TextCNN component uses 3 convolutional layers, each with three different filter sizes: 3, 4, and 5. Such

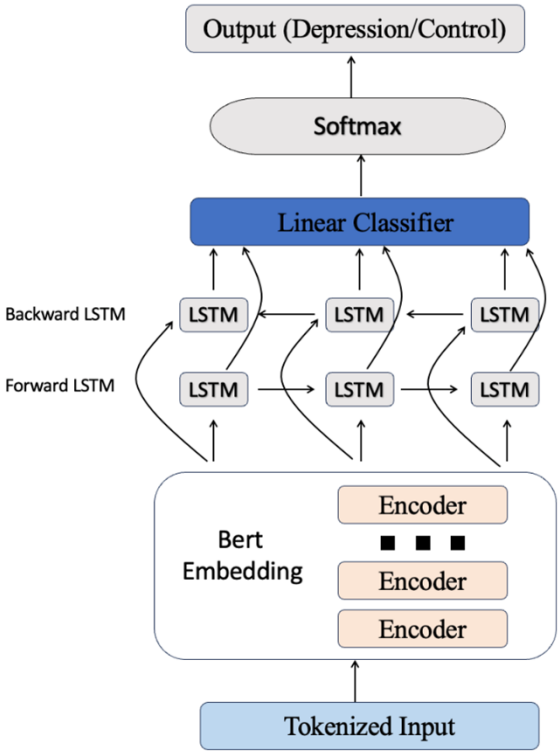

FIGURE 3. BERT-BiLSTM structure.

models may follow the application of convolutional layers, a ReLU activation function is utilized to introduce non-linearity into the model. Following the activation function, max pooling is applied to the output of each convolutional layer. The pooled features from different filter sizes are concatenated, forming a single feature vector. A dropout layer is applied to the concatenated features to mitigate overfitting and enhance the model's generalization ability. Finally, the dropout-regularized features are passed through a linear layer to produce the output logits.

4) MODELS EXPLAINABILITY

Explainability is particularly critical in mental health settings due to the sensitive nature of diagnoses and their significant impact on patients' lives. It enhances trust between patients and healthcare providers, supports informed clinical decision-making, and ensures ethical and transparent use of AI in mental health care. Moreover, explainable AI can provide valuable insights into mental health conditions' complex, multifaceted nature, potentially leading to more personalized and effective treatments. To illustrate the explainability of the depression detection model, this study employs the TIBAV method proposed by [13] and AAW method proposed by [25]. These methods go beyond traditional attention visualization techniques to provide a more comprehensive understanding of the model's decision-making process. TIBAV goes beyond traditional attention visualization by considering the entire forward pass of the model, enabling a more fine-grained understanding of the

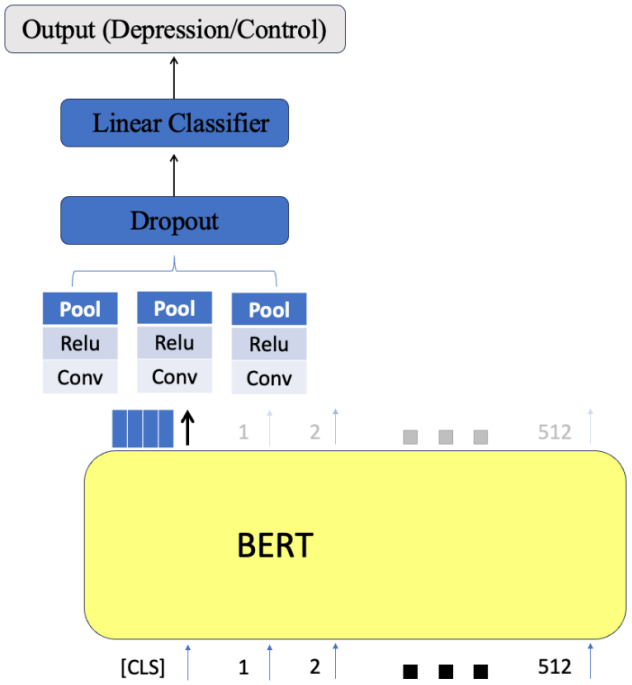

FIGURE 4. BERT-CNN structure.

importance of each input token. On the other hand, AAW simplifies the computation by averaging these attention scores across different layers and heads to understand each input token's importance.

To make the explainability method accessible and user-friendly, this study develops a comprehensive interface that combines the predictive capabilities of the depression detection model with intuitive visualizations. After the user submits the text, the depression detection model processes the input and determines whether it indicates signs of depression. This interface's most notable feature is how it explains the model's decision-making process. Instead of simply returning a binary classification (depression or control), the interface also uses the intensity of the color to display the importance for each word in the input text, obtained using the explainability method. If the model classifies the input as indicating depression, the interface highlights the words with higher importance in shades of red. The deeper the red color, the more important that word was in the model's depression prediction. Conversely, suppose the input is classified as not indicating depression (control). In that case, the interface highlights the important words in shades of blue, with the deeper blue signifying higher importance.

C. EXPERIMENT SETUP

This study employs pre-trained BERT-base-uncased models. The BERT-base-uncased model consists of 12 transformer layers, and each word embedding possesses a dimensionality of 768. These models were trained with an initial learning rate of 2e-5 using the AdamW [32] optimization algorithm. The training process was run in batches of 16 samples each to achieve optimal performance. This process was repeated ten times (epochs) in total. To ensure the reproducibility of the experiments, setting fixed random seeds are followed. Detailed model parameters are provided below:

1) Max sequence length: 512
2) Batch size: 16
3) Number of epochs: 10
4) Dropout: 0.3
5) Learning rate: 2×10^{-5} (The initial learning rate used for training the models)
6) Warmup proportion: 0.1 (The proportion of training steps used for learning rate warmup)
7) Weight decay: 1×10^{-2} (The weight decay coefficient applied to the model's parameters during training)
8) AdamW epsilon: 1×10^{-8} (The epsilon value used in the AdamW optimizer for numerical stability)

IV. RESULTS AND DISCUSSIONS

A. COMPARATIVE RESULTS ON TWEET DATASET

Table 2 compares the results of three BERT-based models (Fine-tuned BERT, BERT-BiLSTM, and BERT-CNN) with the MentalBERT model. The fine-tuned BERT, BERT-BiLSTM, and BERT-CNN models have similar accuracy, precision, recall, and F1 scores, with minor variations. This suggests that these models are equally effective in classifying the given dataset. However, the MentalBERT model, trained on the Reddit dataset for 10 epochs, has a slightly lower accuracy, precision, recall, and F1 score than the other three models. This indicates that the MentalBERT model may not be as effective in classifying the given dataset as the other models. The BERT-CNN model has the highest accuracy, precision, recall, and F1 score among all the models. This suggests that adding the CNN layer to the BERT model might help capture more intricate patterns in the data, resulting in improved performance. The BERT-BiLSTM model has a slightly higher recall score than the Fine-tuned BERT model, while the other scores remain the same. This indicates that adding the BiLSTM layer might help capture more information from the sequence of data, resulting in improved recall.

TABLE 2. Results of four models on Reddit dataset, highest scores highlighted.

Model	Accuracy	Precision	Recall	F1
Fine-tuned Bert	0.981	**1.0**	0.979	0.981
Bert + BiLSTM	0.981	**1.0**	0.982	0.981
Bert + CNN	**0.982**	0.98	**0.990**	**0.982**
Mental Bert	0.963	0.978	0.926	0.961

B. COMPARATIVE RESULTS ON TWEET DATASET

Table 3 presents a comparative analysis of the performance of three BERT-based models (Fine-tuned BERT, BERT-BiLSTM, and BERT-CNN) and the MentalBERT model trained on the tweet dataset. The fine-tuned BERT has the highest accuracy, precision, recall, and F1 score among all

the models, indicating its strong performance in correctly identifying positive instances and overall effectiveness. BERT-CNN achieves the same performance as fine-tuned BERT in terms of accuracy, precision, recall, and F1 score, suggesting that it is also highly effective in identifying positive instances and has a similar overall performance. BERT-BiLSTM has a slightly lower accuracy, recall, and F1 score than Fine-tuned BERT and BERT-CNN but still maintains a high precision of 1.0. This indicates that it is slightly less effective in identifying positive instances than the other two models but performs well in precision. MentalBERT has the lowest accuracy, precision, and F1 score among all the models but achieves the same recall as BERT-BiLSTM. This suggests that while MentalBERT can identify positive instances to some extent correctly, it may struggle with correctly identifying negative instances, resulting in lower overall performance.

TABLE 3. Results of four models on tweet dataset, highest scores highlighted.

Model	Accuracy	Precision	Recall	F1
Fine-tuned Bert	**1.0**	**1.0**	0.998	**0.999**
Bert + BiLSTM	0.999	**1.0**	0.996	0.998
Bert + CNN	**1.0**	**1.0**	0.998	**0.999**
Mental Bert	0.998	0.996	0.996	0.996

C. COMPARATIVE RESULTS ON MENTAL HEALTH DATASET

Table 4 compares the performance of four models: Fine-tuned BERT, BERT-BiLSTM, BERT-CNN, and MentalBERT. Fine-tuned BERT achieves an accuracy of 0.935, with precision, recall, and F1 scores of 0.966, 0.956, and 0.934, respectively. This indicates that while it may not be the top-performing model, it still performs well in classifying instances from the dataset. BERT-BiLSTM outperforms Fine-tuned BERT with a higher accuracy of 0.960. It also achieves strong precision and recall, with F1 and accuracy scores of 0.960 and 0.968, respectively. This suggests that adding the BiLSTM layer aids in capturing more nuanced patterns in the data, resulting in improved performance. BERT-CNN exhibits the highest precision of 0.978 and recall of 0.969, showcasing its capability to identify positive instances while reducing FP accurately. It also achieves the highest F1 score of 0.961, showcasing its strong performance in classifying instances from the dataset. MentalBERT lags behind the other models with the lowest accuracy, precision, recall, and F1 score. This could imply that MentalBERT, specialized for mental health-related text, may be less effective on this dataset than the more general BERT-based models that have been fine-tuned.

D. DISCUSSION

Overall, the results indicate that the BERT-based models, particularly BERT-CNN and BERT-BiLSTM, are highly effective in detecting depression-related content across various datasets. These models consistently achieved high

TABLE 4. Results of four models on mental health dataset, highest scores highlighted.

Model	Accuracy	Precision	Recall	F1
Fine-tuned Bert	0.935	0.966	0.956	0.934
Bert + BiLSTM	0.960	0.968	0.969	0.960
Bert + CNN	**0.961**	**0.978**	**0.969**	**0.961**
Mental Bert	0.905	0.924	0.843	0.897

accuracy, precision, recall, and F1 scores, outperforming MentalBERT in most cases. The superior performance of BERT-based models can be attributed to several factors.

Firstly, BERT is pre-trained on a massive corpus of text, including Wikipedia and BooksCorpus, which amounts to diverse topics and linguistic structures. This extensive pre-training enables BERT to learn deep contextualized representations that can be fine-tuned for specific tasks like depression detection. In contrast, MentalBERT, though designed for mental health tasks, might not have been exposed to a wide range of contexts during its pre-training, limiting its generalizability. Additionally, integrating CNN and BiLSTM architectures with BERT in the BERT-BiLSTM and BERT-CNN models enhances their capability to extract relevant features and patterns for depression detection. CNNs excel at identifying local patterns and spatial relationships in text data, while BiLSTMs can model long-range dependencies and capture contextual information from the input text. By combining these architectures with BERT, the models can leverage the strengths of each component to improve their performance on the task. Moreover, MentalBERT is more straightforward than the BERT-based models, based on a single BERT architecture without additional components like CNNs or BiLSTMs. The increased complexity of the BERT-based models allows them to capture more complex patterns and relationships in the data, potentially leading to better performance on the task of depression detection.

However, it's important to note that while BERT-based models demonstrate superior performance, this comes at the cost of increased computational complexity. For instance, fine-tuned BERT, BERT-CNN, and BERT-BiLSTM on the mental health dataset require 50, 52, and 67 minutes to train for ten epochs. In contrast, MentalBERT, with its more straightforward structure, completes training in just 25 minutes. This significant difference in training time highlights a trade-off between model performance and computational efficiency.

We compared the proposed models, BERT-CNN and BERT-BiLSTM mode with [3], [18], and [19]. In [3], they used CNN models on the CLPsych 2015 dataset but were limited by traditional word embeddings, whereas our BERT-CNN achieved a much higher F1 score of 0.98 on the Tweet dataset. Similarly, the researcher in [3] employed a BGRU model, achieving an F1 score of 0.84, while our BERT-CNN surpassed this with 0.9610 on the Mental Health dataset, demonstrating the effectiveness of deeper contextual representations. Finally, in [19], they applied LSTM archi-

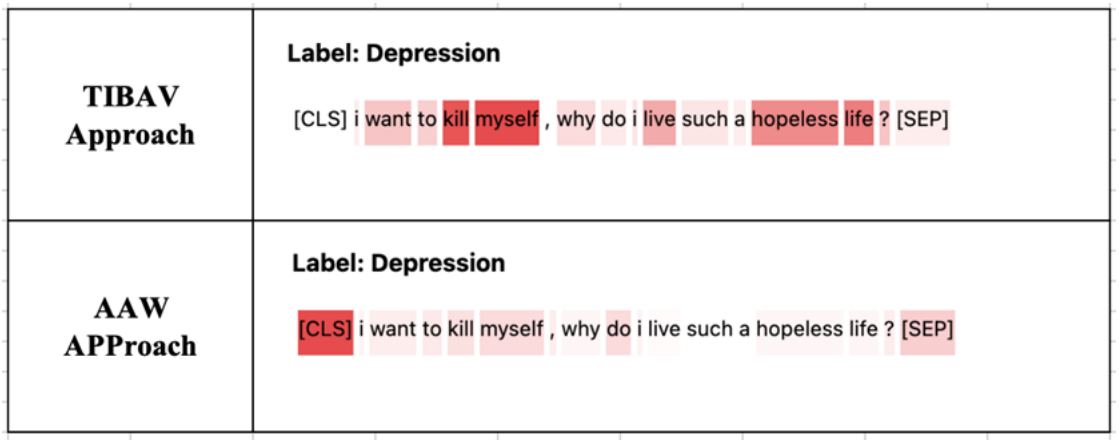

FIGURE 5. Visualization comparison between TIBAV and AAW approach.

tecture. Still, our BERT-BiLSTM model outperformed them in accuracy and F1 score, further showcasing the benefits of BERT's pre-training and enhanced feature extraction.

E. MODEL EXPLAINABILITY VISUALIZATION

This section presents a contrastive visualization analysis to demonstrate the effectiveness and explainability of the proposed models for depression detection. Two visualization techniques will be employed: TIBAV and AAW visualization. The TIBAV approach, the chosen method, provides a more comprehensive understanding of the model's decision-making process by considering the attention weights across all model layers. In contrast, the AAW visualization averages scores across all layers and heads. Furthermore, this study compares the chosen visualizations before and after training to showcase how the model's attention weights evolve and align with the relevant features and patterns associated with depression-related content. This contrastive visualization analysis allows the demonstration of the effectiveness of the models in learning to focus on the most informative aspects of the input text for accurate depression detection.

1) COMPARING TIBAV WITH AAW

The TIBAV approach and the AAW approach are shown in Figure 5. The depressive text is: ''I want to kill myself, why do I live such a hopeless life?''

The AAW approach highlights the ''[CLS]'' special token as the most important word for detecting depression. However, this is not easily understandable or interpretable for humans, as the ''[CLS]'' token does not carry any semantic meaning related to depression. On the other hand, the TIBAV approach offers a more intuitive and easily interpretable visualization. In the given example, the TIBAV visualization highlights the phrases ''end my life'' and ''hopeless life'' as the most important indicators of depression. Humans easily understand these phrases and align with the linguistic patterns and expressions commonly associated with depressive content.

The TIBAV approach assigns higher importance scores to the words and phrases most relevant to detecting depression, making the model's decision-making process more transparent and interpretable. By identifying the key phrases contributing to the model's classification, the TIBAV visualization enables a deeper understanding of the model's reasoning and the specific linguistic features it relies on to make predictions. In contrast, the AAW approach fails to provide meaningful insights into the model's decision-making process, as it assigns high importance to a special token that does not carry any semantic information related to depression. This limitation hinders the explainability of the model, making it difficult for humans to understand and trust the model's predictions.

2) BERT-BASE-UNCASED VISUALIZATION VERSUS TRAINED MODEL VISUALIZATION

The next comparison is between the BERT-based and BERT-base-uncased models without fine-tuning, which employed the TIBAV visualization approach. To illustrate the difference, the following sentence is used: ''I think I have depression because I have been feeling really down recently.'' As shown in Figure 6, the BERT-based models correctly identify the sentence as depressive and highlight keywords such as ''depression'' and ''down'' as the most important indicators. This demonstrates that the fine-tuned models have learned to focus on the relevant features and patterns associated with depressive content during training. On the other hand, the BERT-base-uncased model without fine-tuning incorrectly classifies the sentence as a control statement. This misclassification suggests that the model, without fine-tuning, needs to gain the specialized knowledge and understanding required to identify depression-related content accurately.

The contrasting visualizations of the BERT-based and BERT-base-uncased models highlight the power and importance of fine-tuning for the specific task of depression detection. By fine-tuning the pre-trained BERT model on

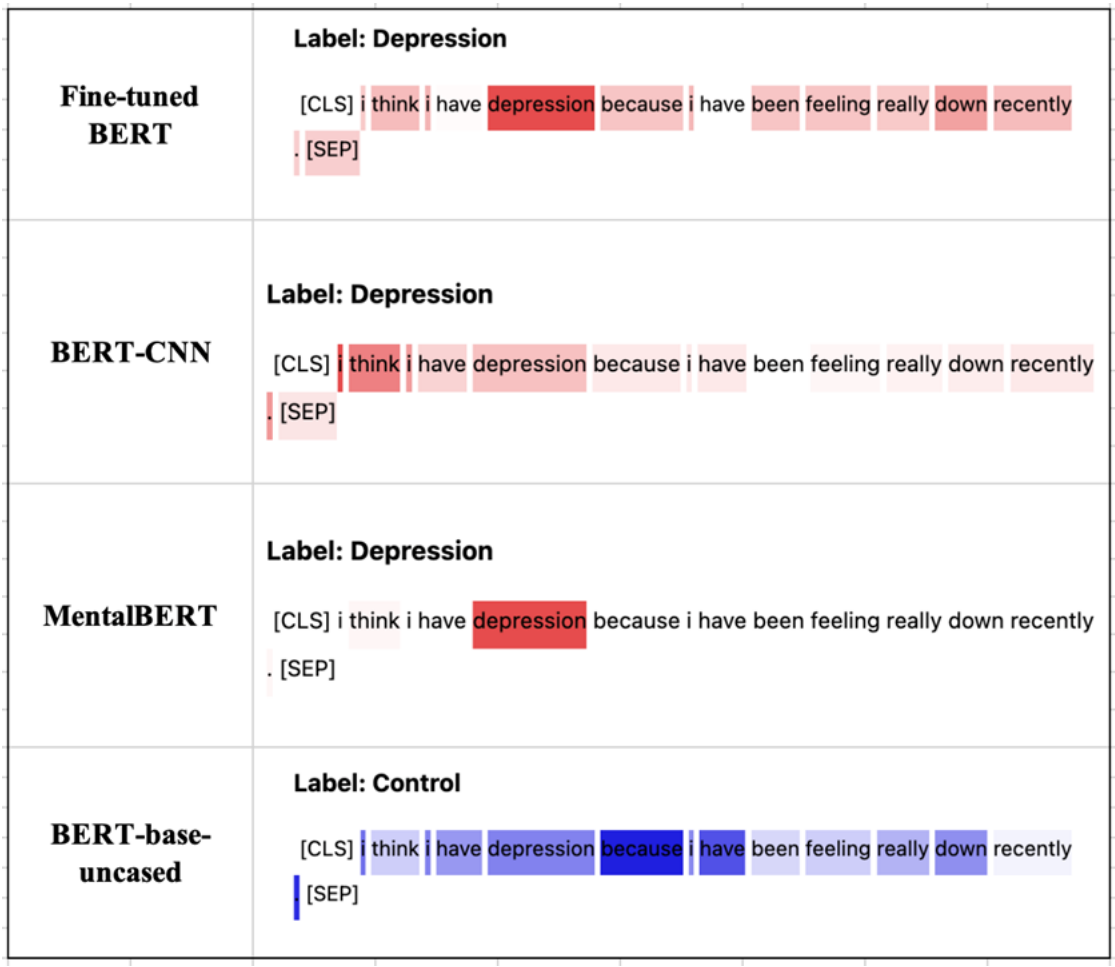

FIGURE 6. Visualization comparison between BERT-based models and BERT-base-uncased.

a labeled dataset of depressive and non-depressive text, the model learns to adapt its attention weights and focus on the most informative aspects of the input relevant to the target task.

V. CONCLUSION AND FUTURE WORK

The study's strengths include using diverse datasets to evaluate model generalizability, applying advanced DL techniques (BERT, BiLSTM, CNN) for depression detection, performing comprehensive evaluations with standard metrics, and incorporating explainability to clarify model decisions. Its main contributions to depression detection and NLP are demonstrating the effectiveness of BERT-based models across datasets, comparing with MentalBERT, enhancing transparency through explainability analysis, proposing a BERT with BiLSTM and CNN approach, and advancing automated tools for early mental health detection and intervention.

However, this study has several limitations. First, it focuses only on English data, which may reduce the generalizability to other languages and cultural contexts, as linguistic markers of depression can vary. Second, relying solely on text data may not fully capture the complexity of an individual's mental state, and incorporating multimodal data could offer a more comprehensive understanding. Third, while the AdamW optimizer is widely used in BERT-based models, it may not be the best choice for depression detection in all cases.

Future work should address these limitations by expanding the study to include data from different social media platforms and languages, exploring the integration of multimodal data (e.g., audio, visual, biometric) to enhance the accuracy and robustness of depression detection, investigating the impact of different optimizers on the models' performance and convergence. These improvements could increase the real-world applicability of the models in mental health assessment and intervention.

REFERENCES

[1] (2021). *National Institute of Mental Health*. Accessed: Apr. 23, 2024. [Online]. Available: https://www.nimh.nih.gov/health/statistics/mental-illness

[2] (2019). *World Health Organization*. Accessed: May 5, 2024. [Online]. Available: https://www.who.int/teams/mental-health-and-substance-use/promotion-prevention/mental-health-in-the-workplace

[3] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in Reddit social media forum," *IEEE Access*, vol. 7, pp. 44883–44893, 2019.
[4] A. Husseini Orabi, P. Buddhitha, M. Husseini Orabi, and D. Inkpen, "Deep learning for depression detection of Twitter users," in *Proc. 5th Workshop Comput. Linguistics Clin. Psychol., From Keyboard to Clinic*, 2018, pp. 88–97.
[5] S. G. Burdisso, M. Errecalde, and M. Montes-y-Gómez, "A text classification framework for simple and effective early depression detection over social media streams," *Exp. Syst. Appl.*, vol. 133, pp. 182–197, Nov. 2019.
[6] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" 2019, *arXiv:1905.05583*.
[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[8] M. Choras, M. Pawlicki, D. Puchalski, and R. Kozik, "Machine learning—The results are not the only thing that matters! What about security, explainability and fairness?" in *Computational Science—ICCS*, V. V. Krzhizhanovskaya, G. Závodszky, M. H. Lees, J. J. Dongarra, P. M. A. Sloot, S. Brissos, and J. Teixeira, Eds., Cham, Switzerland: Springer, 2020, pp. 615–628.
[9] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," 2020, *arXiv:2006.11371*.
[10] F. Cacheda, D. Fernandez, F. J. Novoa, and V. Carneiro, "Early detection of depression: Social network analysis and random forest techniques," *J. Med. Internet Res.*, vol. 21, no. 6, Jun. 2019, Art. no. e12554.
[11] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.
[12] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly available pretrained language models for mental healthcare," 2021, *arXiv:2110.15621*.
[13] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 782–791.
[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.
[15] M. R. Islam, A. R. M. Kamal, N. Sultana, R. Islam, M. A. Moni, and A. Ulhaq, "Detecting depression using K-nearest neighbors (KNN) classification technique," in *Proc. Int. Conf. Comput., Commun., Chem., Mater. Electron. Eng. (ICME)*, Feb. 2018, pp. 1–4.
[16] K. A. Govindasamy and N. Palanichamy, "Depression detection using machine learning techniques on Twitter data," in *Proc. 5th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, Madurai, India, May 2021, pp. 960–966.
[17] Z. Peng, Q. Hu, and J. Dang, "Multi-kernel SVM based depression recognition using social media data," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 1, pp. 43–57, Jan. 2019.
[18] H. Dinkel, M. Wu, and K. Yu, "Text-based depression detection on sparse data," 2019, *arXiv:1904.05154*.
[19] A. Amanat, M. Rizwan, A. R. Javed, M. Abdelhaq, R. Alsaqour, S. Pandya, and M. Uddin, "Deep learning for depression detection from textual data," *Electronics*, vol. 11, no. 5, p. 676, Feb. 2022.
[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017, *arXiv:1705.07874*.
[21] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, 2016, pp. 1135–1144.
[22] (2022). *Depression: Reddit Dataset (Cleaned)*. Accessed: May 3, 2024. [Online]. Available: https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned
[23] (2021). *Sentimental Analysis for Tweets*. Accessed: May 3, 2024. [Online]. Available: https://www.kaggle.com/datasets/gargmanas/sentimental-analysis-for-tweets
[24] (2023). *Mental Health Corpus*. Accessed: May 3, 2024. [Online]. Available: https://www.kaggle.com/datasets/reihanenamdari/mental-health-corpus
[25] A. A. Falaki and R. Gras, "Attention visualizer package: Revealing word importance for deeper insight into encoder-only transformer models," 2023, *arXiv:2308.14850*.
[26] R. H. Chassab, L. Q. Zakaria, and S. Tiun, "An adjusted BERT architecture for the automatic essay scoring task," in *Proc. 5th Int. Multi-Conf. Artif. Intell. Technol.*, 2021, pp. 40–41.
[27] R. H. Chassab, L. Q. Zakaria, and S. Tiun, "An optimized LSTM-based augmented language model (FLSTM-ALM) using fox algorithm for automatic essay scoring prediction," *IEEE Access*, vol. 12, pp. 48713–48724, 2024, doi: 10.1109/ACCESS.2024.3381619.
[28] Y. Y. Chang and N. Omar, "Data annotation architecture for automatic depression detection," *Asia–Pacific J. Inf. Technol. Multimedia*, vol. 12, no. 1, pp. 39–56, 2023.
[29] S. K. Hamed, M. J. A. Aziz, and M. R. Yaakub, "Enhanced feature representation for multimodal fake news detection using localized fine-tuning of improved BERT and VGG-19 models," *Arabian J. Sci. Eng.*, pp. 1–17, Aug. 2024, doi: 10.1007/s13369-024-09354-2.
[30] N. N. W. N. Hashim, N. A. Basri, M. A.-E. A. Ezzi, and N. M. H. N. Hashim, "Comparison of classifiers using robust features for depression detection on bahasa Malaysia speech," *IAES Int. J. Artif. Intell. (IJ-AI)*, vol. 11, no. 1, p. 238, Mar. 2022, doi: 10.11591/ijai.v11.i1.pp238-253.
[31] D. F. Santomauro et al., "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic," *Lancet*, vol. 398, no. 10312, pp. 1700–1712.
[32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

CAO XIN received the bachelor's degree in computer science and technology from Jiangnan University, China, in 2019. He is currently pursuing the master's degree in computer science (software technology) with the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Malaysia. His research during the program focuses on analyzing social media data for mental health detection. His research interests include data science, artificial intelligence, social media analytics, deep learning, and natural language processing.

LAILATUL QADRI ZAKARIA is currently a Senior Lecturer with the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM). She is a member of Asian Natural Language Processing Lab (ASLAN). She actively contributes to academic research and teaching in these fields with UKM. Her work focuses on developing advanced algorithms and models for text analysis and processing, particularly in Asian languages. Her research interests include natural language processing, text analytics, and semantic web technology.