Journal of Autonomous Intelligence (2023) Volume 7 Issue 1
doi: 10.32629/jai.v7i1.1116

ORIGINAL RESEARCH ARTICLE

Enhancing conversational sentimental analysis for psychological depression prediction with Bi-LSTM

Selva Mary G.[1], John Blesswin A.[1,*], Mithra Venkatesan[2], Shubhangi Vairagar[3], Sushadevi Adagale[4], Chetana Shravage[5], Jyotsna Barpute[3]

[1] *Directorate of Learning and Development, SRM Institute of Science and Technology, Kattankulathur 603203, India*
[2] *Department of Electronics and Telecommunication, Dr. D. Y. Patil Institute of Technology, Pimpri 411018, India*
[3] *Department of Artificial Intelligence and Data Science, Dr. D. Y. Patil Institute of Technology, Pimpri 411018, India*
[4] *Department of Computer Engineering, KJEI's Trinity Academy of Engineering, Pune 411048, India*
[5] *Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri 411018, India*
*** Corresponding author:** John Blesswin A., johnblesswin@gmail.com

ABSTRACT

Human mental health (HMH) is a pervasive and impactful condition that profoundly affects an individual's cognitive, emotional, and behavioural aspects in a negative manner. Among various mental health disorders, depression is particularly prevalent, with approximately 20% of women experiencing at least one depressive episode during their lifetime. Identifying depression early on is crucial for timely intervention and support. This study examines user-generated content from major social platforms like Twitter, Facebook, and Instagram, aiming to detect potential signs of depression through behavioural symptoms such as mood changes, loss of interest, altered sleep patterns, focus difficulties, and impaired decision-making. Leveraging natural language processing and machine learning, sentiment analysis deciphers emotional context in posts and comments. A new efficient methodology utilizing Bidirectional Encoder Representations from Transformers (BERT) is proposed for efficient analysis of the posts and comments. Knowledge distillation transfers insights from a large BERT model to a smaller one, enhancing accuracy. Integrating word2vec and BERT with bidirectional long short-term memory (Bi-LSTM), the approach effectively analyses depression and anxiety indicators in social media data. Comparative assessments highlight the system's excellence, achieving a remarkable 98.5% accuracy through knowledge distillation. The proposed methodology marks a substantial stride in identifying mental health signals from social media, facilitating better early intervention and support for those facing depression and anxiety-related challenges.

Keywords: human mental health; natural language processing; sentiment analysis; social media; machine learning; sarcasm identification

ARTICLE INFO

Received: 7 August 2023
Accepted: 24 August 2023
Available online: 9 October 2023

1. Introduction

Human mental health (HMH) represents a complex neurological condition that deeply influences an individual's emotional state, cognitive functions, and actions. This places it among the most widespread and severe forms of depressive disorders. Depression, a prevalent and serious medical illness, is characterized by persistent feelings of sadness and a diminished interest in activities once found enjoyable. Furthermore, it can manifest physical symptoms due to a combination of genetic, environmental, and psychological factors. This intricate interplay gives rise to various types of depression, including persistent depressive disorder, seasonal affective disorder, and psychotic depression. In the modern digital age, online social media platforms have become integral facets

of people's lives, offering avenues for self-expression and connectivity. However, these platforms also raise concerns about their potential impact on mental well-being, including HMH. Users openly share their feelings, emotions, and sentiments on a wide array of subjects and images, making these platforms invaluable resources for gaining insights into mental health conditions. Applying sentiment analysis to social media content can assist researchers in identifying potential signs of mental health challenges. Classifying posts for indications of mental health concerns became an essential research area. Despite the benefits of sentiment analysis, accurately detecting depression presents challenges, especially when expressions may appear positive but conceal underlying negative meanings. This underscores the importance of identifying sarcasm and subtle cues in online communications. Early identification of depression is vital for effective intervention and enhancing the lives of those affected. However, the absence of a standardized procedure for depression detection complicates the prompt recognition and management of this condition, given that symptoms may overlap with those of other mental health disorders. Traditionally, diagnosing mental illnesses relies on individuals completing questionnaires designed to uncover specific patterns of emotions and social interactions[1]. Nonetheless, these methods may have limitations in terms of early detection and precision. Thus, a need arises for more advanced and data-driven approaches to enhance the assessment of mental well-being and early intervention. With timely and appropriate care and treatment, many individuals grappling with mental illness or emotional disorders have a better chance of recovery and improved wellness[2].

Recognizing the significance of early detection and intervention, researchers and healthcare practitioners are exploring novel methodologies, including predictive analytics and machine learning algorithms, to pinpoint indicators of mental health at an early stage. In this research study, the posts and comments in the social media is analysed and classified. The users are not classified in this study. However, in future, the research will also include the users and their classifications. These innovative techniques aim to offer insights into early markers of mental health problems, enabling proactive intervention and personalized assistance. Leveraging technology and data analysis, the objective is to empower healthcare providers to extend targeted and timely support to individuals facing mental health challenges, ultimately enhancing the overall mental well-being of affected individuals and contributing to improved mental health outcomes within society.

2. Literature review

Mental health problems have a profound impact on an individual's emotional well-being, cognitive functions, and social interactions. The consequences of these issues extend to society at large, prompting the need for innovative approaches to prevention and intervention. Detecting mental health problems early is vital, and the field of medical predictive analytics holds tremendous promise, potentially revolutionizing healthcare, including mental health services[3]. Researchers have effectively harnessed this technology to extract meaningful insights from data, personalize experiences, and develop intelligent automated systems[4]. Within the realm of machine learning, widely utilized algorithms such as support vector machines, random forests, and artificial neural networks have proven adept at predicting and categorizing future events[5]. These predictive algorithms play a pivotal role in identifying indicators of mental health, equipping healthcare professionals with valuable insights for accurate diagnosis and treatment.

In the realm of machine learning, the supervised learning approach stands out as a common method, particularly within medical prediction research[6,7]. Supervised learning involves labelled training data, allowing models to learn from attributes and values associated with each data instance. This enables accurate predictions and classifications. In contrast, unsupervised learning operates without labelled data, which presents challenges in clinical applications. Despite its potential, supervised learning remains the preferred choice for predicting mental health conditions due to its precision enabled by labelled data[7–9]. The advancements in machine learning hold immense potential for the mental health field, offering deeper insights into patterns, early indicators, and personalized interventions. Incorporating machine learning into mental

health research can significantly enhance outcomes and contribute to a comprehensive understanding of mental well-being. The rise of smart sensors integrated into mobile devices and wearables opens doors to intelligent mental healthcare, facilitating cost-effective data collection on both physical and psychological states[10–12]. Past systems often lacked the ability for on-demand data collection and struggled with data overload[13]. Meanwhile, social media platforms provide avenues for open discussions on mental health, enabling data-driven insights through the analysis of user posts and reactions[14]. However, the unstructured nature of social media data, encompassing idiomatic expressions and dynamic topics, poses challenges in extracting meaningful information for accurate mental health analysis. This complexity hampers existing systems in efficiently gathering pertinent data and precisely evaluating patients' conditions. Machine learning algorithms, including decision trees, support vector machines, logistic regression, AdaBoost, and multilayer perceptron, have found utility in diagnosing imbalances in mental health[15–19]. However, the continuous monitoring of patients results in diverse types of health-related data, such as voice patterns, textual content, sensor data, and emojis, introducing an additional challenge for existing machine learning systems[20]. The limitations of these systems in effectively handling such data and comprehending the semantic nuances of text shared on social media impede their usefulness for healthcare applications[21]. An intelligent approach is imperative to accurately categorize textual data concerning mental health issues[15,22,23]. **Figure 1** provides a visual representation categorizing and classifying the application of machine learning in addressing various mental health challenges, spanning conditions like schizophrenia, anxiety, depression, bipolar disorder, post-traumatic stress disorder (PTSD), and mental health issues among children[8,24]. This comprehensive review highlights the implementation of machine learning models across diverse mental health scenarios.

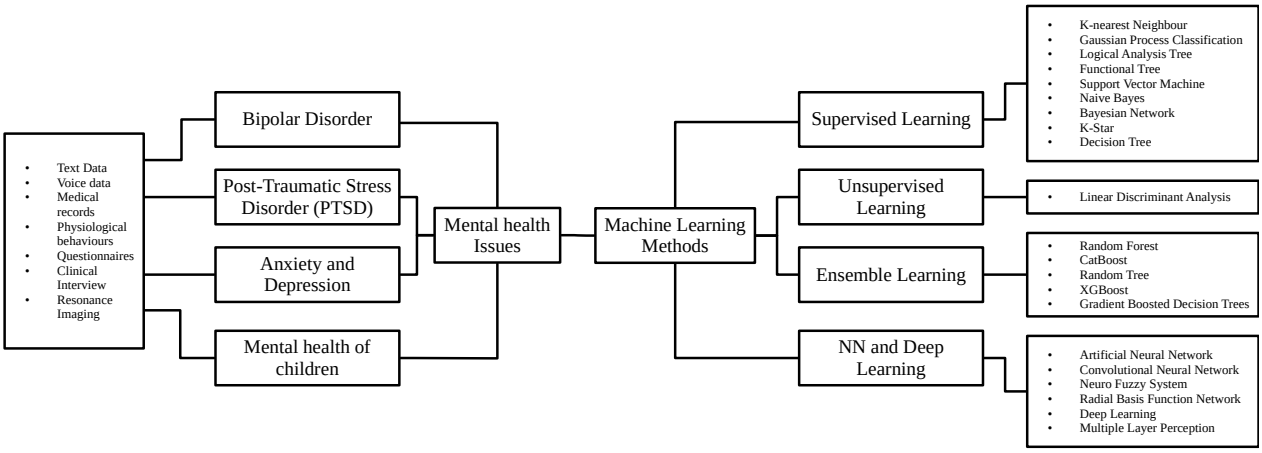


Figure 1. Systematic review and categorization of ML approaches in mental health domain.

Figure 1 illustrates the diverse machine learning strategies, further segmented into supervised learning, unsupervised learning, ensemble learning, neural networks, and deep learning techniques[25,26]. These strategies lay the groundwork for categorizing machine learning models based on their unique learning methods. Furthermore, the review encompasses an in-depth assessment of these models' performance, highlighting their efficacy within the realm of mental health. Key performance metrics like accuracy, area under the ROC curve (AUC), F1-score, sensitivity, and specificity are scrutinized to offer a comprehensive evaluation of the models' effectiveness. This all-encompassing review aims to illuminate the progress and potential of machine learning in tackling the complexities of mental health. In response to these challenges, a pioneering machine learning approach utilizing Bidirectional Encoder Representations from Transformers (BERT) is introduced to enhance the accuracy of identifying mental health issues. This innovative methodology amalgamates various data sources to achieve efficient analysis of mental health data. To effectively process depression and anxiety-related data, a bidirectional long short-term memory (Bi-LSTM) classifier is incorporated.

This research contributes significantly through:

- Introducing a fresh framework for extracting highly pertinent data related to depression and anxiety from various sources.
- Harnessing BERT as a bidirectional text representation model to capture both contextual and semantic nuances within the data. Additionally, suggesting the utilization of Bi-LSTM as a classifier to enhance the comprehension of word sequences within sentences.
- Conducting extensive experiments and evaluations, utilizing methods such as principal component analysis (PCA) and established machine learning models, and including comparative analyses for comprehensive result interpretation[13].

Following extensive hyperparameter optimizations, the model achieves a commendable accuracy rate of 98%, surpassing the performance of other benchmarked methods. This study advances the field of mental health analysis concerning social media data by introducing an innovative deep learning framework adept at managing heterogeneous data, ultimately enhancing the precision of mental health classification. The proposed approach exhibits promise in refining the monitoring and intervention of mental health issues in a timely and efficient manner.

3. Proposed methodology

In this section introduces our proposed framework crafted to retrieve, process, assess, and categorize mental health-related data from social networking platforms, with a specific focus on depression and anxiety. The overall architecture of our research endeavour is depicted in **Figure 2**. The information exchanged on social media platforms offers a potential wellspring of data for identifying mental health issues. However, such social media data is often devoid of structure, featuring colloquial expressions, ambiguous content, and swiftly changing subject matters. Extracting meaningful insights from this data for mental health assessment and depression detection presents substantial hurdles. As a result, our pragmatic framework encompasses several key modules, encompassing data collection, pre-processing, labelling, word substitution, and classification.

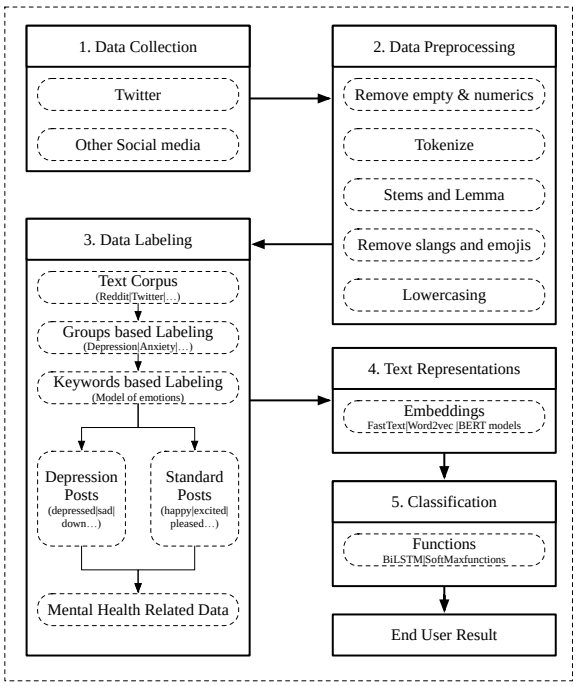


Figure 2. Schematic for the detection of mental health.

This study endeavours to devise a system employing machine learning techniques and deep learning methodologies, notably BERT and Bi-LSTM, for the purpose of mental health detection and evaluation. The

process commences with the real-time extraction of data from Twitter and Reddit through APIs, employing keyword-based queries as depicted in **Figure 2**.

Step 1: Data collection

The initial phase of our framework involves the systematic collection of data from social media platforms, focusing primarily on Twitter and Reddit. Real-time data acquisition is facilitated through APIs, utilizing keyword-based queries to retrieve relevant content. This content encompasses posts, comments, and interactions that pertain to mental health, particularly depression and anxiety. The aim is to assemble a diverse and comprehensive dataset that reflects the nuances of individuals' discussions and experiences related to these mental health conditions. The collected data serves as the foundation for subsequent stages of analysis and classification within our research work.

Step 2: Data pre-processing

Data pre-processing is a critical stage in our framework, aimed at refining the collected social media data to prepare it for subsequent analysis and classification. This stage involves several key sub-steps to ensure the data is structured, cleaned, and ready for further processing as depicted in **Figure 3**.

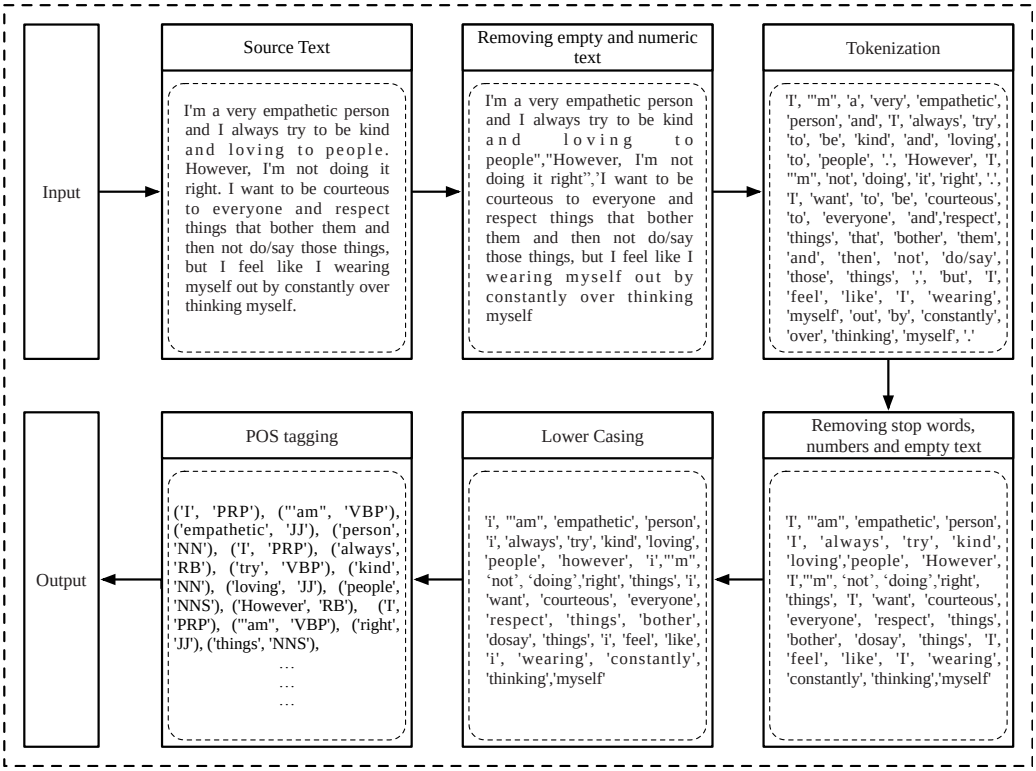


Figure 3. Pre-processing process.

(i) Removing empty and numeric entries: To ensure data quality, we begin by eliminating empty entries and numeric values from the dataset. Empty entries offer no meaningful information, while numeric values often lack contextual relevance in the context of textual data. This step streamlines the dataset by retaining only content-rich entries.

Example: Original entry: "I'm feeling down today 😔"

Empty entry: " "

Numeric entry: "12345"

(ii) Tokenization: Tokenization involves breaking down the textual content into individual tokens or words. This process enables us to analyze the data at a more granular level and facilitates subsequent text

analysis tasks. Tokenization serves as the foundation for various linguistic analyses, such as part-of-speech (POS) tagging.

Example: Original sentence: “I can’t believe how amazing this is!”
Tokenized tokens: [“I”, “can’t”, “believe”, “how”, “amazing”, “this”, “is”, “!”]

(iii) Part-of-speech (POS) tagging: POS tagging involves assigning grammatical labels to each token, categorizing them based on their syntactic roles within the sentence. This step enhances our understanding of the grammatical structure and context of the text, enabling more sophisticated analyses down the line.

Example: Tokenized sentence: [“She”, “runs”, “quickly”, “in”, “the”, “morning.”]
POS tags: [“PRP”, “VBZ”, “RB”, “IN”, “DT”, “NN”, “.”]
Explanation: PRP (personal pronoun), VBZ (verb), RB (adverb), IN (preposition), DT (determiner), NN (noun), . (punctuation)

(iv) Lowercasing: To ensure consistency and reduce the complexity of the data, all text is converted to lowercase. This step mitigates the impact of capitalization variations, allowing for uniform text processing and analysis.

Example: Original text: “The SUN is shining BRIGHTLY.”
Lowercased text: “The sun is shining brightly.”

(v) Removing stop words: Stop words are common words (such as “and,” “the,” “is”) that add little semantic value to the text. Removing these stop words helps to eliminate noise and streamline the dataset, focusing our analysis on more meaningful content words.

Example: Original sentence: “I love reading books and spending time with friends.”
After stop word removal: “Love reading books spending time friends.”

By meticulously executing these data pre-processing sub-steps, we establish a clean, organized, and linguistically enriched dataset that forms the basis for subsequent stages of our framework, ultimately contributing to accurate and insightful mental health classification and evaluation.

Step 3: Text representation

In this phase of our framework, we focus on transforming the pre-processed textual data into meaningful and numerical representations that can be effectively utilized by machine learning algorithms. The choice of appropriate text representation methods plays a crucial role in capturing the semantic nuances of the data.

(i) Word2vec: Word2vec is a popular technique for text representation that captures word embeddings in a continuous vector space. It translates words into dense numerical vectors, preserving semantic relationships between words. Words with similar meanings or contexts are represented as vectors that are geometrically close in the vector space. For instance, in a word2vec representation, the vectors for “king” and “queen” would exhibit a notable similarity, reflecting their semantic association. In our proposed system, we trained the skip-gram model due to its improved performance with consistent words. The skip-gram model aims to find word representations that accurately predict adjacent words within a context (*c*). It increases the objective of the average log probability over all *X* target words and their respective contexts in a given series of words using Equation (1).

$$\frac{1}{X} = \sum_{n=1}^{X} \sum_{-c \le j \le c, j \neq 0} \log P(\omega_{x+j} | \omega_x) \tag{1}$$

where $\{\omega_1, \omega_2, \omega_3, \omega_4, \ldots, \omega_x\}$ is a given series of words.

Example: Word2vec representation: “king”: [0.5, −0.2, 0.8, ...], “queen”: [0.48, −0.18, 0.82, ...]

These vectors are mathematically constructed to capture linguistic properties, enabling semantic relationships to be mathematically inferred.

(ii) BERT (Bidirectional Encoder Representations from Transformers): BERT is a state-of-the-art transformer-based language model that produces contextualized word representations. Unlike traditional models, BERT considers the entire context of a word within a sentence, leading to highly accurate embeddings. BERT captures contextual nuances, such as word sense disambiguation, making it well-suited for tasks requiring a deep understanding of text. It is pre-trained on massive corpora, enabling it to generate context-rich embeddings that contribute to advanced language understanding. The probability prediction for a given centre word depends on the inner product of the vectors representing the input and output candidates ({ and , respectively), which are then normalized to form a probability distribution over all words in the vocabulary of size *N* using the softmax function as follows using Equation (2):

$$P(\omega_0|\omega_i) = \frac{\exp\left(v_{\omega_0}^T v_{\omega_i}\right)}{\sum_{\omega=1}^{W} \exp\left(v_{\omega}^T v_{\omega_i}\right)} \quad (2)$$

the probability of observing ω_0 given ω_i. $v_{\omega_0}^T$ represents the transposed corresponding to the word ω_0. *W* is the total number of words in the vocabulary.

However, as the size of the vocabulary increases, determining these probabilities and calculating correlated gradients for all words becomes computationally expensive[27].

Example: Original sentence: "The weather is so nice, I should go for a walk."

BERT representation: A multi-dimensional vector capturing the contextual meaning of the entire sentence.

By employing these advanced text representation techniques like word2vec and BERT, we transform the pre-processed textual data into dense, semantically meaningful numerical representations. These representations form a bridge between raw text and machine learning models, enabling accurate and effective analysis and classification of mental health-related content within our framework.

Word embedding is employed to represent words in the text corpus with real-valued vectors that capture their meanings, with words in close proximity anticipated to have similar meanings. We implement various word-embedding techniques to represent our collected text corpus (bag of words) as equivalent vectors. Word vectors are superior to older techniques, such as one-hot encoded vectors, as they maintain a semantic representation of words and require less space. We explore different word embedding techniques, such as word2vec, BERT to obtain the context of words in our text corpus (as shown in **Figure 2**). For our smart approach, we propose the use of BERT, a recent and effective method for word embedding, in our mental health problem identification.

Step 4: Classifier models

In this phase of our framework, we employ a diverse array of modern and cutting-edge classifier models to undertake the final sentiment classification on our meticulously curated dataset. Each classifier is strategically selected to harness its unique strengths and capabilities. The fusion of these models with advanced text representation techniques enhances our ability to accurately categorize mental health-related content.

Support vector machine (SVM): Support vector machines are robust classifiers known for their effectiveness in handling high-dimensional data. SVM aims to find a hyperplane that optimally separates different classes while maximizing the margin between them. In our sentiment classification task, SVM leverages the dense vectors obtained from text representation to draw decision boundaries that categorize mental health-related content based on sentiment patterns[28].

Logistic regression: Logistic regression is a widely used linear classifier that estimates the probability of a given sample belonging to a particular class. It is particularly suitable for binary classification tasks. In our

framework, logistic regression exploits the dense vector representations of text to compute the likelihood of mental health-related content falling into specific sentiment categories[14].

Random forest: Random forest is an ensemble learning technique that constructs a multitude of decision trees during training and combines their outputs to make predictions. Each decision tree contributes to the final classification, enhancing accuracy and reducing overfitting. In our sentiment classification, random forest effectively aggregates the predictions of decision trees to assign sentiment labels to mental health-related content[15].

AdaBoost: AdaBoost is another ensemble learning method that combines multiple weak classifiers to create a strong classifier. It iteratively adjusts the weights of misclassified samples to prioritize difficult-to-classify instances. In our framework, AdaBoost utilizes the power of boosting to iteratively improve its classification performance, ultimately yielding accurate sentiment labels for mental health-related content[29].

Long short-term memory (LSTM) with convolutional neural network (CNN): To delve into the realm of deep learning, we employ LSTM models combined with CNN for sentence classification. This architecture has exhibited promising outcomes in various natural language processing tasks. CNN layers extract pertinent features from the text, while LSTM units capture long-range word relationships, thereby facilitating precise sentiment classification[30].

Bidirectional LSTM (Bi-LSTM): In an effort to further enhance sequence classification, we introduce a bidirectional LSTM model. Comprising two LSTM units operating in both left and right directions, this model captures past and future context information simultaneously. The Bi-LSTM architecture effectively preserves long-term word relationships and incorporates comprehensive contextual information, making it an ideal candidate for accurate sentiment classification of mental health-related content[31,32].

The fusion of these diverse classification models with our advanced text representation techniques equips our framework with the capability to accurately categorize and evaluate sentiment in mental health-related social media content. Each classifier brings a distinct perspective and strength, contributing to the overall efficacy of our sentiment analysis approach. Bidirectional long short-term memory (Bi-LSTM) significantly enhances sentiment classification by capturing comprehensive contextual information from both past and future words in a sequence. This bidirectional processing enables Bi-LSTM to recognize long-range dependencies and intricate sentiment nuances, resulting in improved accuracy in sentiment analysis. Studies have demonstrated that Bi-LSTM effectively handles complex sentence structures and reduces ambiguity by considering the entire context, leading to better sentiment classification outcomes. This approach is particularly valuable for tasks such as mental health sentiment analysis on social media content, where subtle expressions and contextual cues play a vital role in understanding users' emotional states.

This proposed approach has demonstrated exceptional accuracy in sequence classification tasks, including traffic event analysis with social networking data[18]. The LSTM architecture contains the main component called the memory cell $\{C_t\}$ that is being updated by using the input gate $\{i_t\}$ and forget gate $\{f_t\}$ as shown in Equations (4) and (6)–(8).

$$f_t = \sigma\left(\omega_f . (h_{t-1}, x_t) + b_f\right) \quad (3)$$

$$i_t = \sigma(\omega_i . (h_{t-1}, x_t) + b_i) \quad (4)$$

$$C_{t-2} = (\tanh(\omega_c . h_{t-1}, x_t) + b_c) \quad (5)$$

$$C_t = f_t \times C_{t-1} + i_t \times C_{t-2} \quad (6)$$

$$o_t = \sigma(\omega_o . (h_{t-1}, x_t) + b_o) \quad (7)$$

$$h_t = o_t \times \tanh(C_t) \quad (8)$$

where, f_t—forget gate (Equation (3)), represents the forget gate at time *t*, determines the amount of information to discard from the cell state. σ is the sigmoid activation function, outputting values between 0 and 1. A value

close to 0 means "forget" while a value close to 1 means "keep." ω_f are the weights associated with the forget gate. h_{t-1} is the hidden state from the previous time step. x_t is the input at the current time step. b_f is the bias term for the forget gate. i_t—input gate (Equation (4)), represents the input gate at time *t*, determines how much of the new information should be stored in the cell state. ω_i are the weights associated with the input gate. b_i is the bias term for the input gate. C_{t-2}—candidate cell state (Equation (5)), represents the candidate values that could be added to the cell state. ω_c are the weights for creating the candidate cell state. b_c is the bias term for the candidate cell state. C_t—cell state (Equation (6)), represents the cell state at time *t*. o_t—output gate (Equation (7)). ω_o are the weights associated with the output gate. b_o is the bias term for the output gate. h_t—hidden state (Equation (8)), represents the hidden state (or output) at time *t*.

The first part involves the utilization of state-of-the-art machine/deep learning techniques for depression/anxiety identification, accompanied by dimensionality reduction to enhance accuracy. We aim to optimize the performance of our framework by efficiently analysing the data collected from Reddit and Twitter, enabling precise identification of mental health problems based on social media posts. The second crucial aspect of our approach is to apply fine-tuning and model optimization, specifically knowledge distillation. Knowledge distillation involves training a compressed model to learn from a larger pretrained BERT model in a sequential manner, allowing it to become smarter and lighter[19].

4. Experimentation and result analysis

In this section, we present the experimental results of our proposed scheme. We utilized data collected from Twitter and Reddit using their respective APIs, Tweepy and PRAW. Given its comprehensive nature and focus on mental health, the social media in mental health detection (SMHD) dataset from Reddit serves as an invaluable tool for this research to develop algorithms that can detect signs of mental health issues in textual data. By utilizing this benchmark dataset, researchers can ensure that their models are trained on a representative sample of real-world data. To gather data on depression and anxiety, we employed a keyword-based approach based on the circumplex model of affect, which involved using predetermined terms to classify text into emotion categories such as happy, angry, and sad. This allowed us to collect significant data on depression and anxiety from social media platforms. **Table 1** provides a summary of the data sets collected through our proposed framework. We merged the data sets and prepared training and test sets for our experiments.

Table 1. Description of the data sets collected from Reddit and Twitter.

Sources	Health issue	No. of posts	Description
Twitter	Depression	35,000	Contents, posts and comments
	Anxiety	45,000	Contents, posts and comments
Reddit	Depression	65,000	Discussions, posts and comments
	Anxiety	60,000	Discussions, posts and comments

In evaluating our model's performance, we employed well-established performance metrics, including precision, recall, and accuracy. To comprehensively assess the classification effectiveness, we utilized a confusion matrix, a fundamental tool that presents the count of accurate and erroneous predictions[23]. Leveraging the confusion matrix, we derived the following performance metrics using Equation (9):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP}$$

where, *TP*—true positive, *FP*—false positive, *TN*—true negative, *FN*—false negative.

Accuracy: Accuracy gauges the overall correctness of our model's predictions. It is computed as the ratio of correctly classified instances to the total number of instances in the dataset.

Precision: Precision measures the proportion of correctly predicted positive instances among the instances predicted as positive. It quantifies the model's ability to avoid false positives.

Recall (sensitivity): Recall calculates the ratio of correctly predicted positive instances to the total actual positive instances. It assesses the model's ability to capture all positive instances without omission.

These metrics, derived from the confusion matrix, provide a comprehensive evaluation of our model's performance, facilitating a deeper understanding of its strengths and areas for improvement in the context of sentiment classification for mental health-related content.

Result analysis

In our conducted experiment, we thoroughly evaluated the efficacy of our novel Bi-LSTM and BERT-based knowledge distillation approach in comparison to a range of state-of-the-art machine learning algorithms. The assessment encompassed two distinct datasets comprising text posts related to depression and anxiety, sourced from prominent social media platforms Reddit and Twitter. Our comparative analysis involved benchmarking our proposed method against RF, LG, SVM, CNN, AdaBoost, and LSTM algorithms, all tasked with sentiment analysis for identifying depression and anxiety-related sentiments within our gathered data. The outcomes of this extensive evaluation are diligently presented in **Figures 4** and **5**. For RF, our experimental setup entailed utilizing 150 iterations and 100 estimators. In the case of SVM, we harnessed a training parameter ridge estimator coupled with a radial basis function (kernel = rbf). Notably, our exploration included the application of three prevalent word embedding techniques—TF-IDF, word2vec, and fastText—on SVM, NB, RF, and AdaBoost models, as meticulously delineated in **Figure 4**. This strategic approach aimed at unraveling the most effective feature extraction methods for the nuanced task of depression and anxiety detection within our dataset.

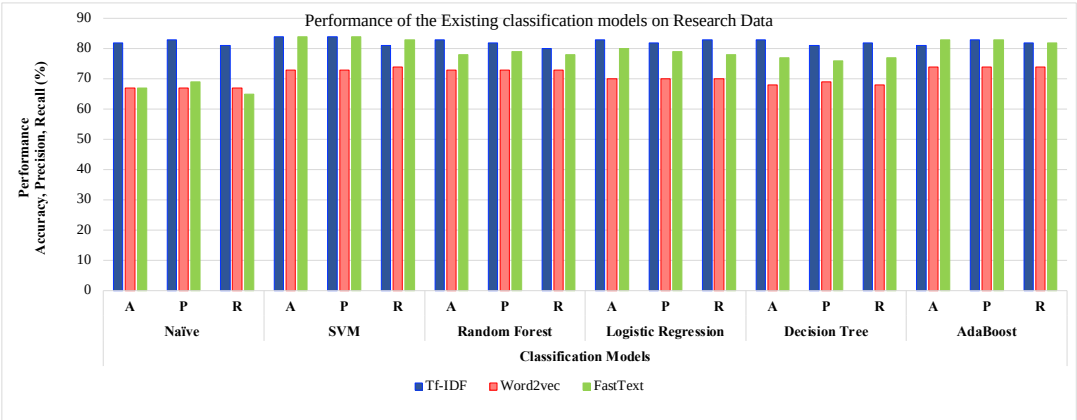


Figure 4. Performances of classical machine learning algorithms with our collected data. (A, accuracy; P, precision; and R, recall in %).

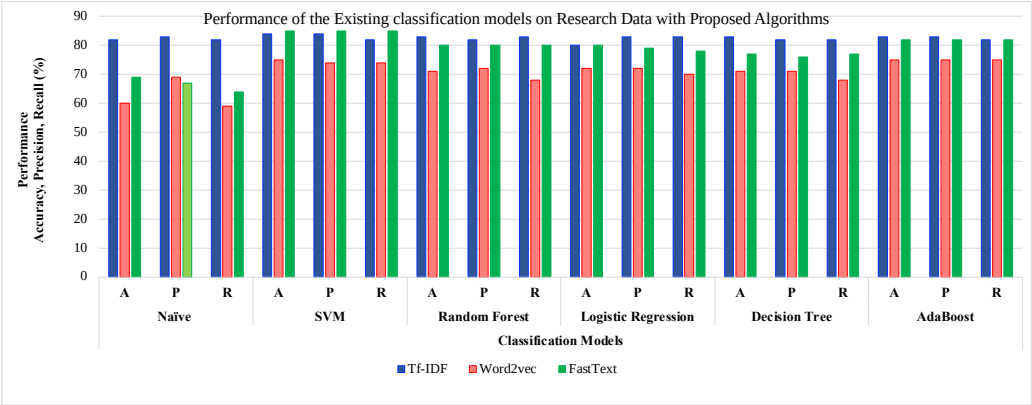
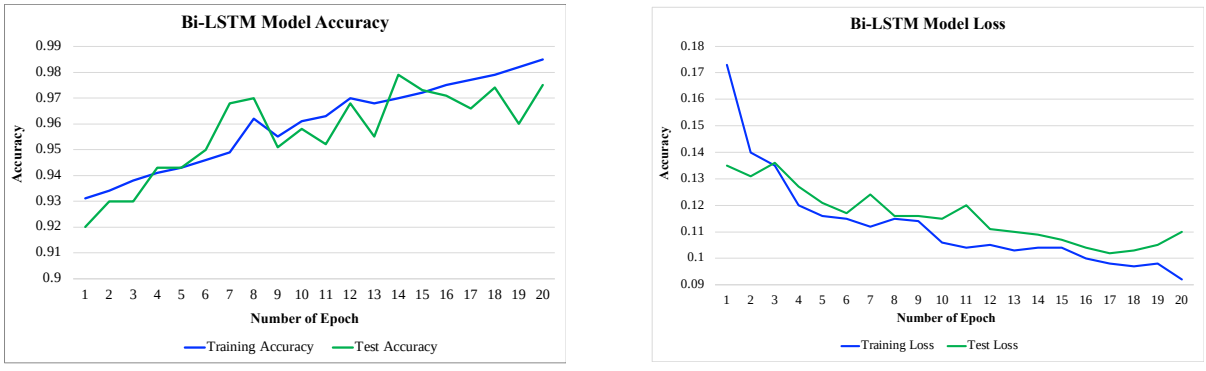


Figure 5. Performances of classical machine learning algorithms with proposed algorithms.

Upon analysis, we observed that AdaBoost exhibited a commendable accuracy of 83% when employing 150 estimators and a learning rate of 0.8. Concurrently, SVM demonstrated a strong performance with an accuracy of 84%, employing a training parameter ridge estimator along with a radial basis function (kernel = rbf). Remarkably, SVM showcased a noteworthy performance edge over AdaBoost, manifested through clear label separation within the processed and vectorized text corpora. Intriguingly, however, the utilization of the word2vec algorithm led to a decline in the performance of several classical machine learning algorithms as observed in **Figure 4**. This was attributed to the inherent characteristic of word2vec to disregard unseen words during its training phase.

In our investigation, we incorporated principal component analysis (PCA) as a means to effectively mitigate the dimensionality of the text data. Notably, this approach yielded notable performance enhancements across diverse vectorization techniques, encompassing BERT and word2vec. The utility of dimensionality reduction lies in its capacity to streamline predictive models, culminating in heightened efficiency when extrapolating predictions onto novel data instances.



(a) **(b)**

Figure 6. Comparison of training and test data for proposed Bi-LSTM model, **(a)** training accuracy vs. test accuracy; **(b)** training loss vs. test loss.

As shown in **Figure 6**, applying PCA led to increased accuracy for most classifiers, except for random forest (RF), Naive Bayes (NB), and logistic regression (LR) when using word2vec vectors. This is because word2vec cannot handle new words, leading to decreased accuracy in these models. Moreover, BERT outperforms word2vec for both anxiety and depression prediction tasks because it can distinguish and capture two different semantic meanings by generating two different vectors for the same word in a given text corpus. This ability to contextualize words effectively contributes to its superior performance in sentiment analysis tasks. **Figure 6** provides a comparison of accuracy and loss during the training and testing phases of the proposed Bi-LSTM model. The model accuracy (**Figure 6a**) and model loss (**Figure 6b**) are plotted for 20 epochs using BERT-based text representation. The training accuracy is calculated after applying the model to

the training data, while the test accuracy represents the accuracy on the test data. Throughout the 20 epochs, both training and test accuracy follow a similar trend, indicating that the model is well-regularized and has a representative data batch. The small difference between training and test accuracy suggests that the model has been properly configured and has a good generalization performance.

In the realm of mental health analysis on social media, a temporal examination of user posts offers a distinct advantage. Analysing tweets or posts over a prolonged period can shed light on the progression or regression of an individual's mental state. This longitudinal approach has the potential to highlight patterns, trends, and periodic fluctuations in sentiment, which might remain obscured in cross-sectional analyses. Such insights can be invaluable for early detection of mental health issues, understanding the impact of external events on an individual's mental state, and even predicting potential crises. Furthermore, a time-based analysis can also cater to the dynamic nature of mental health, where symptoms can ebb and flow. Recognizing these patterns might pave the way for timely interventions, targeted support, and more personalized mental health care strategies. In light of these potential benefits, our future endeavours in this research domain will prioritize the incorporation of a temporal perspective. We aim to analyse social media posts across different time frames to gain a more holistic and nuanced understanding of users' mental well-being.

5. Conclusion

In this research study, we have successfully developed a robust framework for detecting mental health problems, specifically depression and anxiety, using advanced deep learning techniques like BERT, Bi-LSTM, and knowledge distillation based on social media content from platforms like Reddit and Twitter. Our proposed framework significantly enhances the accuracy of smart healthcare systems in identifying mental health-related issues at an early stage. The key features of our framework include data collection from social networks using APIs, a pre-processing module to convert unstructured data into meaningful information, and a text labelling technique based on keywords and the circumplex model to extract relevant features related to mental health. We employed the latest text embedding technique, BERT, to transform words into vectors that capture the semantic meaning of the text, thus improving the accuracy of the classification task using attention mechanisms. Additionally, we introduced response-based knowledge distillation to build a smaller and smarter model for depression/anxiety detection and classification, based on the neural responses of the teacher model (BERT). We collected and prepared our own data set from Twitter and Reddit, focusing on relevant textual data to create an intelligent model for smart healthcare systems. Through an extensive experiment, we demonstrated that the proposed BERT-Bi-LSTM model outperforms other machine learning classification models. The model's superiority lies in its ability to combine the strengths of both BERT and Bi-LSTM, effectively capturing the syntactic and contextual information of each word. Moreover, the application of response-based knowledge transfer using BERT and fine-tuning for depression/anxiety detection yielded remarkably high accuracy.

Author contributions

Conceptualization, SV; methodology, SA; software, CS; validation, SA, CS and JB; formal analysis, MV; investigation, SMG and JBA; data curation, SV; writing—original draft preparation, SMG and JBA; writing—review and editing, MV; visualization, CS, SA and JB. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Shrestha K. Machine learning for depression diagnosis using Twitter data. *International Journal of Computer Engineering in Research Trends* 2018; 5(2): 56–61. doi: 10.22362/ijcert/2018/v5/i2/v5i208
2. Tadesse MM, Lin H, Xu B, Yang L. Detection of suicide ideation in social media forums using deep learning. *Algorithms* 2020; 13(1). doi: 10.3390/a13010007
3. Seppälä J, De Vita I, Jämsä T, et al. Smartphone and wearable sensors-based m-health approach for psychiatric disorders and symptoms—A systematic review and link to m-RESIST project. *JMIR Mental Health* 2018; 6(2): 1–21. doi: 10.2196/mental.9819
4. Drissi N, Ouhbi S, Idrissi MAJ, et al. On the use of sensors in mental healthcare. *Intelligent Environments* 2019; 26: 307–316. doi: 10.3233/AISE190058
5. Dang NC, Moreno-García MN, De la Prieta F. Sentiment analysis based on deep learning: A comparative study. *Electronics* 2020; 9(3): 483. doi: 10.3390/electronics9030483
6. Fiallos A, Jimenes K. Using reddit data for multi-label text classification of twitter users interests. In: Proceedings of the 2019 Sixth International Conference on eDemocracy & eGovernment (ICEDEG); 24–26 April 2019; Quito, Ecuador. pp. 324–327.
7. Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision. Available online: https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf (accessed on 28 September 2023).
8. Mary GS, Blesswin AJ, Kumar SM. Self-authentication model to prevent cheating issues in grayscale visual secret sharing schemes. *Wireless Personal Communications* 2022; 125: 1695–1714. doi: 10.1007/s11277-022-09628-8
9. Sau A, Bhakta I. Screening of anxiety and depression among seafarers using machine learning technology. *Informatics in Medicine Unlocked* 2019; 16: 100149. doi: 10.1016/j.imu.2018.12.004
10. Neha S, Nivya, Shekar PHC, et al. Emotion recognition and depression detection using deep learning. *International Research Journal of Engineering and Technology (IRJET)* 2020; 7(8): 3031–3036.
11. Ali F, El-Sappagh S, Islam SMR, et al. An intelligent healthcare monitoring framework using wearable sensors and social networking data. *Future Generation Computer Systems* 2021; 114: 23–43. doi: 10.1016/j.future.2020.07.047
12. Yadav A, Vishwakarma DK. Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review* 2020; 53(6): 4335–4385. doi: 10.1007/s10462-019-09794-5
13. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019); 2–7 June 2019; Minneapolis, Minnesota, America. pp. 4171–4186.
14. Kowsari K, Meimandi KJ, Heidarysafa M, et al. Text classification algorithms: A survey. *Information* 2019; 10(4): 150. doi: 10.3390/info10040150
15. Garcia-Ceja E, Riegler M, Nordgreen T, et al. Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing* 2018; 51: 1–26. doi: 10.1016/j.pmcj.2018.09.003
16. Garcia-Ceja E, Galván-Tejada CE, Brena R. Multi-view stacking for activity recognition with sound and accelerometer data. *Information Fusion* 2018; 40: 45–56. doi: 10.1016/j.inffus.2017.06.004
17. Zogan H, Razzak I, Wang X, et al. Explainable depression detection with multi-modalities using a hybrid deep learning model on social media. *World Wide Web* 2022; 25(1): 281–304. doi: 10.1007/s11280-021-00992-2
18. Kumar A, Sharma A, Arora A. Anxious depression prediction in real-time social data. In: Proceedings of the International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019; 4 March 2019; Uttarakhand, India. pp. 1–7.
19. Kim J, Lee J, Park E, Han J. A deep learning model for detecting mental illness from user content on social media. *Scientific Reports* 2020; 10(1): 11846. doi: 10.1038/s41598-020-68764-y
20. Dansana D, Adhikari JD, Mohapatra M, Sahoo S. An approach to analyse and forecast social media data using machine learning and data analysis. In: Proceedings of the 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA); 13–14 March 2020; Gunupur, India. pp. 1–5.
21. Wang B, Wang A, Chen H, et al. Evaluating word embedding models: Methods and experimental results. *APSIPA Transactions on Signal and Information Processing* 2019; 8(E19). doi: 10.1017/ATSIP.2019.12
22. Qiu XP, Sun TX, Xu YG, et al. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences* 2020; 63(10): 1872–1897. doi: 10.1007/s11431-020-1647-3
23. Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine* 2018; 13(3): 55–75. doi: 10.1109/MCI.2018.2840738
24. Vaswani A, Shazeer S, Parmar N, et al. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 4–9 December 2017; Long Beach, CA, USA. pp. 1–11.
25. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation* 1997; 9(8): 1735–1780. doi: 10.1162/neco.1997.9.8.1735
26. Tadesse MM, Lin H, Xu B, Yang L. Detection of depression-related posts in reddit social media forum. *IEEE Access* 2019; 7: 44883–44893. doi: 10.1109/ACCESS.2019.2909180

27. Burdisso SG, Errecalde M, Montes-y-Gómez M. A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications* 2019; 133: 182–197. doi: 10.1016/j.eswa.2019.05.023
28. Moraes R, Valiati JF, Neto WPG. Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications* 2013; 40(2): 621–633. doi: 10.1016/j.eswa.2012.07.059
29. Munikar M, Shakya S, Shrestha A. Fine-grained sentiment classification using BERT. In: Proceedings of the 2019 International Conference on Artificial Intelligence for Transforming Business and Society (AITB); 5 November 2019; Kathmandu, Nepal. pp. 2–5.
30. Jang B, Kim M, Harerimana G, et al. Bi-LSTM model to increase accuracy in text classification: Combining word2vec CNN and attention mechanism. *Applied Sciences* 2020; 10(17): 5841. doi: 10.3390/app10175841
31. Blesswin AJ, Mary GS, Kumar SM. Multiple secret image communication using visual cryptography. *Wireless Personal Commns* 2022; 122(4): 3085–3103. doi: 10.1007/s11277-021-09041-7
32. Zeberga K, Attique M, Shah B, et al. A Novel text mining approach for mental health prediction using Bi-LSTM and BERT model. *Computational Intelligence and Neuroscience* 2022; 2022: 7893775. doi: 10.1155/2022/7893775