Multimedia Tools and Applications (2022) 81:23649–23685
https://doi.org/10.1007/s11042-022-12648-y

An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM

Harnain Kour [1] · Manoj K. Gupta [1]

Received: 20 April 2021 / Revised: 2 January 2022 / Accepted: 9 February 2022 /
Published online: 18 March 2022

Abstract
Depression has become one of the most widespread mental health disorders across the globe. Depression is a state of mind which affects how we think, feel, and act. The number of suicides caused by depression has been on the rise for the last several years. This issue needs to be addressed. Considering the rapid growth of various social media platforms and their effect on society and the psychological context of a being, it's becoming a platform for depressed people to convey feelings and emotions, and to study their behavior by mining their social activity through social media posts. The key objective of our study is to explore the possibility of predicting a user's mental condition by classifying the depressive from non-depressive ones using Twitter data. Using textual content of the user's tweet, semantic context in the textual narratives is analyzed by utilizing deep learning models. The proposed model, however, is a hybrid of two deep learning architectures, Convolutional Neural Network (CNN) and bi-directional Long Short-Term Memory (biLSTM) that after optimization obtains an accuracy of 94.28% on benchmark depression dataset containing tweets. CNN-biLSTM model is compared with Recurrent Neural Network (RNN) and CNN model and also with the baseline approaches. Experimental results based on various performance metrics indicate that our model helps to improve predictive performance. To examine the problem more deeply, statistical techniques and visualization approaches were used to show the profound difference between the linguistic representation of depressive and non-depressive content.

Keywords Mental health · Twitter data · Convolutional and recurrent neural networks · Long short-term memory model

✉ Harnain Kour
18dcs008@smvdu.ac.in

Manoj K. Gupta
manoj.gupta@gmail.com

[1] Department of Computer Science and Engineering, Shri Mata Vaishno Devi University, Katra, India

1 Introduction

Depression is a serious psychiatric disorder in communities across the world and fares a proper share of the global disease count. There are above 350 million people who suffer from depression, which corresponds to more than 4.4% of the global population [18]. Additionally, two-third of patients do not seek out help. The major problem is that depression unknowingly affects the personal and social life of a person. At its extreme, it can lead to other factors like suicide, psychiatric disorders, etc. Approximately, 1 person dies every 40 s, which translates to 8,00,000 suicide deaths every year across the globe [21]. Suicides are one of the primary reasons for adolescent deaths which imply that adolescents are at great risk of depression. The research of depression and identifying it in a person is an essential task globally, as well as in the context of India, where suicide rates are alarming [69]. According to Lancet's report published in 2012, in India every hour a student commits suicide due to depression. According to the World Health Organization (W.H.O.) report, approximately 8934 students committed suicide in 2015. In the previous 5 years, approximately 39,775 students killed themselves and most of the suicides are not even reported [16]. This figure is alarming as well as calls for the need to take critical actions to address the problem and take necessary steps.

Mental health and issues related to it are very important to address at every stage of life, be it childhood, adolescence, or adulthood. The person suffering from depression usually suffers from a short-term or long-term low mood state that kills creativity or enthusiasm in day-to-day activities of life [44]. Prolonged low mood state and routine tensions can become chronic or recurrent which can lead to a serious health problem [75]. Victims of depression usually suffer from symptoms like insomnia, loneliness, loss of appetite and sleep, lack of concentration at work and personal life, and sometimes there are high chances for suicide [31, 69]. Some of the common symptoms are shown in Table 1 [13]. There are various reasons for having long-term depression in a person such as rough childhood, sexual abuse, addiction to alcohol, medical treatments, work pressure, and historical legacy of racism, colonialism, and caste [42]. Untreated depression and anxiety get worse with time and result in sleep deprivation, memory issues, heart problems, etc. Various incentives and programmes have been started for curing depression under the guidance of various countries and well-known organizations like W.H.O. Victims suffering from depression are unable to utilize these treatments because most of them are from lower and middle-class families [8, 9]. Also, developing countries do not have an effective scheme for treating depression due to a lack of funds and resources [16].

Differentiating depressed individuals from non-depressed ones is a quite challenging task as there is no practical approach available for doing so. Moreover, there are not enough resources and trained health professionals for treating depression. Most of the existed prediction techniques are not accurate due to lack of effective methods in the diagnosis of depression. Social media giants such as Twitter, Facebook, Snapchat, and Instagram, can help us for predicting depression because of the large pool of users and their activities in their respective platforms. Platforms like Twitter, generate social media data on an average of 6000 tweets per second i.e., around 200 billion tweets per year and they are providing their data for open source [7]. Figure 1 clearly shows that anxiety and depression are top-listed and major problems for young people. The young population usually spends a lot of time on social sites, and it allows researchers to collect and analyze content shared on social platforms [4, 9]. After talking to health professionals, we become familiar with the fact that there are so many unsolved problems regarding mental health and most people consider it taboo to discuss it with family or friends. However, on social media, people are willing to share their sentiments and

Table 1 Symptoms of depression

Symptom type	Examples	
Cognitive symptoms	-Lack of motivation, Lack of interest, Incurability, Pessimism, Decision making and Problem-solving issues, Memory loss, Concentration loss, Thought's clouding, Lack of Social Interest etc.	
Physical symptoms	-Pain, Appetite loss: weight, Sexual dysfunction, Hypersomnia, Disturbed sleep, insomnia, Cries, Mutism, Psychomotor retardation, Weakness, Energy loss, Fatigue etc.	
Perception of self	-Feeling alone, feeling like a burden, Feeling misunderstood, Lack of self-confidence, Lack of self-esteem etc.	
Mood and emotional symptoms	-Mental pain, Hopelessness, Helplessness, Worthlessness, feeling bad, Guilt, Sadness, Anxiety, Fear to fail, Restlessness, Irritability, Frustration, Emotional regulation, Humour, −-Anhedonia, Emptiness, Emotional blunting etc.	
Auto aggression symptoms	Desire to live, Suicidal ideation, Self-harm, Suicidal attempt etc.	
Functioning	Complex	-Financial issues, Ability to cope with life, Personal issues etc.
	Professional	-Sick leave, Loss of job, Stopping studies, Professional responsibilities etc.
	Social	-Self-exclusion, Family life, Interpersonal relationships, Social isolation, Communicating feelings etc.
	Elementary	-Autonomy, Self-care, Coping with daily tasks etc.

problems, and they expect solutions to their problems because of the vast network of social media.

Natural Language Processing (NLP) is a technique to study speech and text, and it evolved from linguistic approaches to computer algorithms with the development of computers. Initially, NLP was used to study classical tasks to well-structured grammar like language in books, but it has evolved to understand human perspective in the form of mail, web content, reviews, comments, news, social media posts, and media articles, etc. and these are more challenging to process [29, 32]. Sentiment Analysis (SA) is a technique for analyzing positive and negative characteristics of text or speech by studying sentiments. People share their thoughts, ideas, opinions, and life events by posting various textual posts and comments on their topic of interest [51]. This approach can be extended to study depression levels of a person from his/her social media content and by observing the negative sentiment scores. Social media content has progressively evolved from text to videos. However, there are various complexities in "text-related content", such as the usage of emojis, hashtags, and other languages in addition to English, which are highly valuable in understanding the sentiments of a person [20, 61]. Lots of research has proved that if the content generated by users in social networking sites is used in the correct way, it can be used to detect a person's mental state at an early stage [75]. There are many predictive algorithms and optimization techniques like Deep

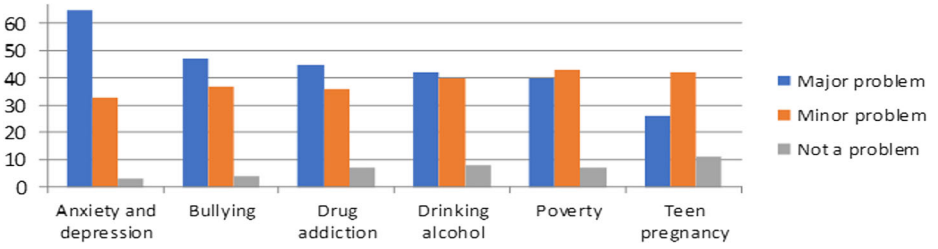

Fig. 1 Survey of the U.S. showing depression as a major problem in the young population in 2018 [36]

Learning (DL) and Machine Learning (ML) to observe patterns in the data and based on those observations to generate insights. Thus, researchers prefer to use various computational techniques to examine individuals with mental health problems, thereby predicting depression among social media users. NLP is capable of processing a number of languages in many aspects [27]. Researchers generally use classical ML algorithms such as Random Forest (RF), Support Vector Machines (SVMs), Decision Trees (DT), etc. for text analysis using binary classification methods. The efficacy of the conventional ML technique was limited as the number of correlations increased significantly due to growth in the volume of data [48]. In this research, we investigated various DL techniques and also provided a comparative report of our findings with those of conventional approaches.

Textual data classification has seen a significant boost in accuracy by deploying DL models like RNNs and CNNs in parallel with neural word embeddings [58]. CNNs were originally designed to duplicate the visual sense of human beings and animals to recognize objects and detect them in images [11] and videos, whereas RNNs are designed for sequence reading processes such as parts of speech, tagging, and language translations, etc. [68]. The uniqueness of CNNs is translational invariance i.e., extraction of unique features irrespective of angles and intensity of images in vision tasks. The addition of pooling procedures in CNN enabled it to operate on textual data as well. CNNs along with word embeddings capture syntactic and semantic information from text data. The advantage of LSTM is that it is effective at retaining vital information and is meant to avoid "vanishing gradient" problem [33]. Due to their capacity to handle arbitrary-length sequential input, it also performs well for sequence labeling tasks [60]. In this study, we have integrated DL approaches to propose our hybrid model.

In our work, we used Twitter dataset for classifying tweets, labeled as positive for depressed users and negative for non-depressed users. The remaining part of the paper is structured as follows. Section 2 discusses the related work for the analysis of depression using various techniques. Section 3 highlights the main contributions of our work. Section 4 focuses on the proposed methodology. Section 5, describes the dataset used in our work and pre-processing techniques. In Section 6, our hybrid DL based method CNN-biLSTM is explained in detail. The proposed approach is based on concept of Deep Neural Networks (DNNs) and Word Embeddings, which are still the state-of-the-art methodology to perform any learning task in NLP. Section 7 compares our proposed model with CNN and RNN based models for text classification. DL based binary classification algorithms have been adapted for the suitability of our current use case. Although, CNNs always outperform most of the tasks but are subjective in model explainability. Therefore, to add more explainability to the models we have employed an RNN based approach. Section 8 illustrates the comparative analysis, visualizations, statistical and performance analysis of the proposed work. Lastly, the conclusion of our study is given in Section 9.

2 Related work

2.1 Traditional techniques

Various ML algorithms and statistical techniques have been used for classifying text data using social media as a platform. Traditional studies show a correlation between the raised depression and social media website usage. Costello et al. [15] explained the mapping of psychological characteristics with digital records of online behavior on social platforms like

Facebook, Instagram, Twitter, etc. It was hypothesized that psychological traits can be predicted based on the language used on internet platforms and the pages liked by people. Eichstaedt et al. [19] collected the Facebook status history from 683 patients to predict their depression. ML algorithms were trained using at least 200 topics along with cross-validation techniques to avoid overfitting. Priya et al. [47] proposed a five-level prediction of depression, stress, and anxiety using five ML algorithms. Mori et al. [38] used ML algorithms on four types of Twitter information such as network, word of statistics, time, and a bag of words. The study considered 24 types of personality traits and 239 features. Tao et al. [62] predicted depressive content in social media using depression context. Data gathered from social networks was given as text to the knowledge sentiment block. The depressive contents identified from SA were used to warn and aware the family members, and social activists. Guntuku et al. [22] collected Twitter data which is around 400 million tweets in Pennsylvania, USA, and acknowledged the users whose tweets consisted of words like alone or lonely. User tweets were analyzed concerning age, time of post, daily activities of users, and gender of the user. Patterns in tweets were analyzed using ML classifiers and NLP techniques, thereby predicting the loneliness of a user.

Various popular ML techniques have attracted many researchers in the last few years. Islam et al. [27] proposed depression classification on Facebook data using ML techniques. For effective analysis, traditional ML approaches have been utilized considering different psycho-linguistic features. It has been found that DT produced better results and also the classification error rate decreased and accuracy significantly enhanced when compared with the other ML techniques. Hiraga et al. [24] utilized various conventional ML algorithms like multinomial Logistic Regression, NB, and Linear SVMs to classify mental disorders from the blogs written in Japanese. Wu et al. [73] proposed various hypothesis and their correlation based on language, time, and interaction to predict job burnout using ML algorithms like DT, Logistic Regression (LR), SVM, XGBoost, and RF on 1532 Weibo burnout users, to replace previous statistical methods based on surveys. Fatima et al. [20] used ML techniques such as SVMs, Multilayer perceptron neural network, and LR to predict postpartum depression from social media text. Features were extracted from social media platforms based on linguistics and classified as general, depression, and postpartum depression.

2.2 Deep learning techniques

Conventional ML algorithms have performed well in predicting depression from social media-based textual content. Researchers have employed DL-based solutions to gain more insight from photos, videos, unstructured text, and emojis. Orabi et al. [43] utilized CNN and RNN to detect depression in Twitter data using root Adaptive Moment Estimation (Adam) as an optimizer, and word embedding training was done using CBOW, Skip-gram, and Random word embedding having a uniform distribution range from −0.5 to +0.5. Shrestha et al. [56] proposed an unsupervised method utilizing RNNs and anomaly detection to analyze behaviors of users on ReachOut.com (online forum). Two streamed approaches were used, one for linguistics and the other for network connection to detect depression in users. Eatedal et al. [2] utilized an RNN technique to predict depression among women in the Arab using 10,000 tweets generated by 200 users. Zogan et al. [79] proposed a new approach for identifying depressed users based on the user's online timeline tweets and user behaviors. The hybrid model comprising of CNN and Bi-GRU approaches was tested on a benchmark dataset. The semantic features were extracted which represented user behaviors. Hybrid CNN and Bi-GRU

were compared with the state-of-art techniques and found that the classification performance was improved to a greater extent. Chiu et al. [14] and Huang et al. [26] proposed a multi-model framework on the DL technique to predict depression from Instagram posts that use pictures, text, and behavior features. Tommasel et al. [64] proposed a DL technique to capture social media expression in Argentina. Time-series data generated (using markers) was fed to a neural network to forecast mental health and emotions during COVID-19. Wang et al. [70] built a dataset on Sina Weibo named Weibo User Depression Detection Data Set (WU3D) containing 10,000 depressed users and 20,000 normal users. Ten statistical features were proposed based on the social behavior, user's text, and posted pictures. Fusion F-net was proposed to train on these 10 features to detect depression. Suman [61] utilized DL models with a cloud-based smartphone application on tweets to detect depression. The sentence classifier used in this study is the RoBERTa associated with the provided tweet or Query with standard corpus tweets. The model's reliability was enhanced by the standard corpus. The patient's status of depression has been estimated and the mental health is predicted. The authors also used random noise factors and the larger set of tweet samples from which depression has been predicted.

Furthermore, Rao [48] showed that the critical sentiment information cannot be correctly captured by the traditional depression analysis models. This issue was handled by the proposed multi-gated LeakyReLU CNN model which in turn also identified the depressed characters in social media. Every user post was initially recognized and further emotional status and overall representation have been identified. The developed multi-gated has been modified into a single LeakyReLU CNN. Content posted by online users was considered in the form of the Reddit self-reported dataset. Depressed persons have been identified by this proposed model and performance results were analyzed. Sood et al. [58] used RStudio to retrieve tweets and further the sentiments were evaluated. To analyze the sentiments of the general population from Twitter, every sentiment was provided with a score. The scored tweet was based on the sentiments and for that, an innovative algorithm was developed in this study. Uddin et al. [68] researched and focused on online data in the Bangla language and analyzed depression by utilizing Long Short-Term Memory (LSTM) and deep recurrent network. Hyper-parameter tuning effects have been demonstrated. It was depicted that for a stratified dataset, the accuracy in depression detection is higher with repeated sampling. Also, an individual's depression is detected in this study with the help of proposed models and thus undesirable doings have been avoided. Pranav [46] researched and found that victims of depression used abnormal language while speaking. Processing of Twitter data for depression prognosis was done by implementing neural networks. This study stated that vagueness raised in SA can be terminated by propounding CNNs. Identification of user's depression status can be resolved by the proposed approach. This is a very fruitful prognostic tool in observing user's depression on social platforms. Rosa et al. [50] emphases on analyzing the emotional sentiments and extracting deep semantic analysis from textual data. Also, descriptive useful information of the content in natural language is extracted and a combined model training is performed using semi-supervised learning. Hybrid model DHMR was utilized to get better results. Shetty et al. [55] performed SA for posts on Twitter. KAGGLE dataset was taken as input for DL and LSTM was propounded in deep networks. Later for enhanced performance CNNs were propounded in the classification part.

2.3 Other techniques

For the earlier mental illnesses and depression identification, Alsagri et al. [5] proposed a method considered as a data-driven approach. It was concluded that depression goes to a peak based on tweets gathered from social media. The transformer-based approach has been formed based on the depression dataset. The comparative analysis of the proposed approach with the existing studies was analyzed and showed that the overall model performance got highly improved. Stephen and Prabhu [60] detected the level of depression among Twitter users using depression scores measurement through various emotions combined with sentiment scores. Different depression aspects have been underscored. The estimated scores have been correlated to major information concerning various user's depression levels. Levia et al. [33] analyzed social media users who used online platforms to reveal their mental health states. Based on the time, online messages have been analyzed in iterative order and detected the user's depression risk state earlier. The comprehensive SA combined with the ML approaches showed effective results for detecting depression's early symptoms. Zucco et al. [80] analyzed the opinion of users for performing SA. Text extraction and NLP were used to detect the opinion of users. Birjali et al. [11] analyzed user's activities from social media platforms and predicted their depression emotions. Weka tool was utilized for ML techniques for Twitter data classification. For the semantic similarities, WordNet external semantic source among the evaluated participants has been utilized. From social networks, Twitter sentiments have been extracted. Zhang [76] analyzed that depression trends and posts related to stress were highly monitored and various geographical entities have been focused by the online users. Also, this study identified that if people talk more about covid-19, depression signals significantly get increased.

2.4 Analysis table

In the Analysis table, a few studies that used ML and DL models to perform depression prediction using text data have been analyzed and discussed. Table 2 summarises the most related studies recent studies (2017–2021).

2.5 Problem formulation

The growing online social media platforms have developed a way for communication in day-to-day life. Classifying depressed individuals from non-depressed ones using lingual dialects is a challenging task. Previous works as discussed in Section 2 have talked about the problem of (1) Distinguishing depressed individuals from non-depressed individuals using social media platforms, and (2) Classifying posts that are posted by depressed people.

Furthermore, researchers have focused on unsupervised techniques, whereas in this study, a supervised technique is put forward to classify depressed users using psycho-linguistic features. Experiments and outcomes stated in previous works illustrate that both text and user detection are challenging issues. Despite the fact that Twitter provides a huge amount of data, it is a challenging task to handle this data. Some of the most common issues encountered while working with Twitter data are:

1. A huge number of images and video transactions were done parallelly with text.
2. Unstructured data with significant usage of emojis and GIFs.
3. Usage of foreign languages etc.

Table 2 Analysis of recent studies on text data using deep learning techniques and its comparison with proposed work

Ref.	Author and Year	Dataset used	Features	Techniques used	Results	Strengths	Weakness
[41]	Naseem et al. 2021	Twitter dataset	Linguistic	Word2Vec, GloVe, fastText, IWV, HyRank, BERT, DistilBERT, XLNET, and ALBERT.	Accuracy (Acc)=94.8	• Proper analysis of the dataset, thereby providing in-depth information regarding the features utilized. • Long texts support the attention layer to capture accurate syntactic features.	• Proposed model's efficiency has not been examined on other related datasets.
[14]	iChiu et al. 2021	Instagram data	Linguistic	CNN, LSTM, RNN, and RF	F1-Score (F1)=0.83	• Used DNNs to learn features within the architecture itself for enhancing dynamic feature mapping.	• Early-stage prediction of depressed users is not done. • Proposed models are evaluated on a single dataset thereby making the significance of the respective models under doubt.
[47]	Priya et al., 2020	Questionaries		DT, SVM, KNN, RF and Naïve Bayes (NB)	Acc=85 Precision (Pre)=82 Recall (Rec)=85 F1=83.6	• Efficient, elastic and simple-to-use technique for multitask classification.	• Absence of validation on generality and robustness in case of huge depression clusters.
[56]	A. Shrestha et al., 2020	ReachOut.com (Online user forum)	Linguistic and network based	Unsupervised technique based on RNN	F1=64	• Proposed unsupervised method integrates both network and psycholinguistic features, resulting in better results than baselines.	• Focused on only binary classification job via unsupervised technique. • Risky post detection needs to be improved using other network features.
[77]	W. Zheng et al., 2020	Questionnaire	Multi modal	Temporal CNN	F1=95.4% Pre=93.0% Rec=98.0%	• Graph Attention Model with a multi-modal knowledge approach	• To prove the model's efficiency, suggested model must also be

Table 2 (continued)

Ref.	Author and Year	Dataset used	Features	Techniques used	Results	Strengths	Weakness
						learns appropriate embeddings for multiple nodes. • Attention mechanism focuses on important information, thereby, improving overall results.	examined on other related datasets.
[74]	D. Xezonaki et al., 2020	DAIC-WoZ 2017 depression datasets	Linguistic	Attention-based model and Hierarchical Attention Network (HAN)	F1 = 71.5	• Words and dialogues are encoded in multiple stages via HAN thereby generating affective information.	• Explored only small datasets. • More elaborate information sources, e.g., expert knowledge bases can be considered to extract more relevant features.
[71]	JT Wolohan, 2020	Reddit dataset	Linguistic	A deep LSTM neural network with fastText	AUC = 0.93 F1 = 0.92	• Improvisation in prediction and classification performance by using hybrid of deep-LSTM and fastText.	• Analysis suffers from a data shortage and will strengthen as more data becomes available. • A comparison with the state-of-the-art is required to confirm or contradict the obtained results.
[35]	Chenhao Lin et al., 2020	Twitter dataset	Multi modal	CNN and BERT	Acc = 88.4 Pre = 90.3 Rec = 87.0 F1 = 93.6	• Used multimodal and deep neural architectures to learn features within the architecture itself.	• It is tough to validate the generality and robustness of their research on a huge depression cluster.
[39]	Murfi et al. 2019	Indonesian tweets	Topics, words and combination of topics and words as a features	Levenstein algorithm and SVM	Acc = 86.37 (Word features) Acc = 83.23 (Topic features) Acc = 86.84 (Combination features)	• Used topics in combination with words for sentiment analysis in Indonesian tweets.	• Limited to only Indonesian tweets. • Easily distracted by noisy information.

Table 2 (continued)

Ref.	Author and Year	Dataset used	Features	Techniques used	Results	Strengths	Weakness
[65, 66]	Tong et al., 2019	Twitter dataset	User profile features and linguistic features	Inverse Boosting Pruning Trees (IBPT) and Fdp--DocNADE	Acc = 91.5 F1 = 91	• Better fitness and adaptation ability, for minimizing the influence of noisy information. • It has two easily adjustable parameters.	• It requires more training time than rest of the models used in the study. • Cannot accurately process short texts.
[67]	M. Trotzek et al., 2018	CLEF 2017 Reddit dataset	User level linguistic metadata	A CNN created on different word embeddings	NA	• Addresses the early detection of depression using ML models.	• There is a need to implement and compare more DL models. • It is not shown how well these models work on data that has not been seen before.
[57]	Shuai et al., 2018	Twitter dataset	LDA topics	Ensemble model, SVM and LR	NA	• Proposed approach is designed to handle the sparse data from multiple online social networks.	• Model is not robust and stable because the employed classifier does not have a strong fitting ability on complex datasets.
[54]	Shen et al. 2017	Twitter dataset	LDA topis and domain specific keywords	Multimodal dictionary learning + LR	Acc=85 F1=85	• Learn the latent and sparse representation of features and extract the common patterns from distinct feature groups.	• Similar to the dictionary-based approach, it is efficient for low-dimensional vectors only. • It can stuck at local minima. • Detection needs to be improved so that autonomous intervention can be well-developed.

4. Labeling tweets that require professionals to label data and are very time-consuming.

For simplicity, we are limiting our problem statement to "text" and "English language". Moreover, emojis, videos, and foreign language alphabets are dropped. The unstructured noisy data is cleaned using various pre-processing techniques. The problem of labeling data is solved by using an authenticated Twitter dataset that is released for the use of psychology and computer science researchers.

The main objective of our study is to develop a hybrid DL model for depression prediction, and compare its performance to the related DL models namely, CNN and RNN. Our hypothesis is that the proposed DL model should outperform DL-based CNN and RNN models, state-of-the-art studies, and baseline models in both accuracy and robustness.

3 Core contribution of the article

DL algorithms like CNN and LSTM thrive on data, and with so many open datasets available, they give better results for classifying depressed and non-depressed tweets. RNNs are preferred as they can be fed with external pre-trained embeddings. CNN's can be utilized for text processing as they can be trained very fast. Moreover, their capability to extract local features from text is mainly prominent for NLP tasks. CNNs and RNNs can be combined together to take benefit of both architectures. We have proposed the hybrid architecture which provides the advantages of both CNNs and LSTMs. The major contribution of this study involves the following:

1. A depression detection framework has been proposed using DL utilizing textual content from social media.
2. A feature-rich CNN network is built to classify user tweets concatenated with the biLSTM network to classify social media users suffering from depression i.e., CNN-biLSTM. To the best of our knowledge, first time the work of using semantic features, statistical analysis, and DL techniques jointly with word embeddings have been utilized for depression detection.
3. The research outcomes achieved on a benchmark Twitter dataset have shown the superiority of our proposed method when compared to state-of-the-art studies.

4 Proposed methodology

The proposed architecture comprises of six modules, as shown in Fig. 2: (1) Extraction of data generated by the online users. (2) Analyzing raw data. (3) Pre-processing of raw data into clean data. (4) Feature extraction to generate machine-compatible data. (5) Depression classification differentiating depressive tweets from non-depressive tweets. (6) Parameter Evaluation to evaluate and compare the hybrid CNN-biLSTM, RNN, and CNN classification approach. In this study, depression is predicted via a hybrid CNN-biLSTM approach, using Twitter-based depression datasets. The classification error is minimized and the prediction of depression becomes more precise and accurate.

The steps given below illustrate the proposed methodology and its related flowchart is shown in Fig. 3.

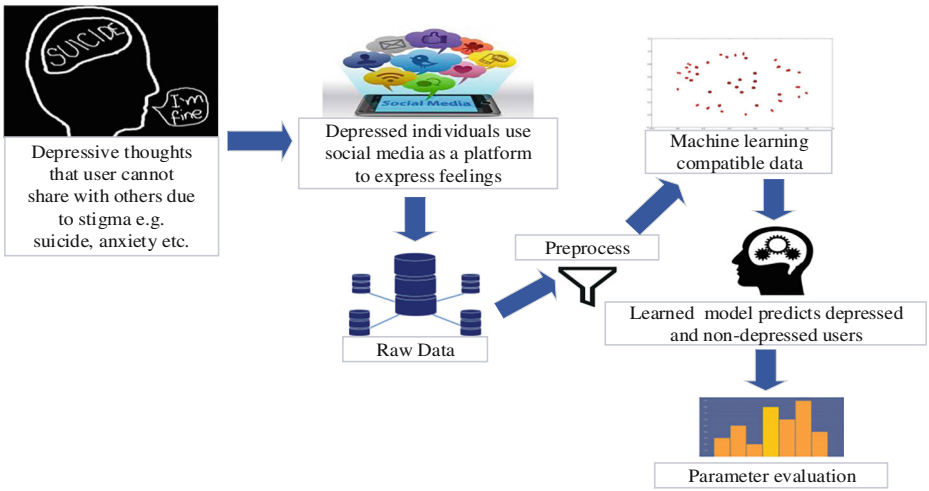

Fig. 2 The proposed system architecture

Step-1: Design a framework for SA and upload a Twitter database to predict depression. Tweets are labeled as depressed or not depressed by experts, Shen et al., [54] for mental health text analysis, and to identify sentiments of depressed or not depressed users.

Step-2: Apply pre-processing steps to the uploaded data for noise elimination. Data is processed as per the requirements resulting in considerable positive effects on the quality of feature extraction. Pre-processing techniques like data normalization, tokenization, punctuation removal, stop words removal, etc. are applied to the uploaded text data. This step provides clean and noise-free data that is further used for feature extraction.

Step-3: In this step, the feature extraction procedures are applied to the pre-processed data for extracting important and relevant features. Extracted features ascertain the relevant data dimensions to assist classification algorithms for better performance.

Step-4: To obtain better accuracy, CNN model, RNN model, and proposed hybrid classification algorithm i.e., CNN-biLSTM is used. The proposed model is evaluated against RNN, CNN, and traditional reference models on the same Twitter dataset to validate its performance. The optimized features obtained in the third step are forwarded as input to the classifiers for training and testing purposes.

Step-5: In the last step, performance parameters of the proposed depression analysis framework, such as precision, recall, F1-score, specificity, accuracy, and AUC, are calculated to validate the system.

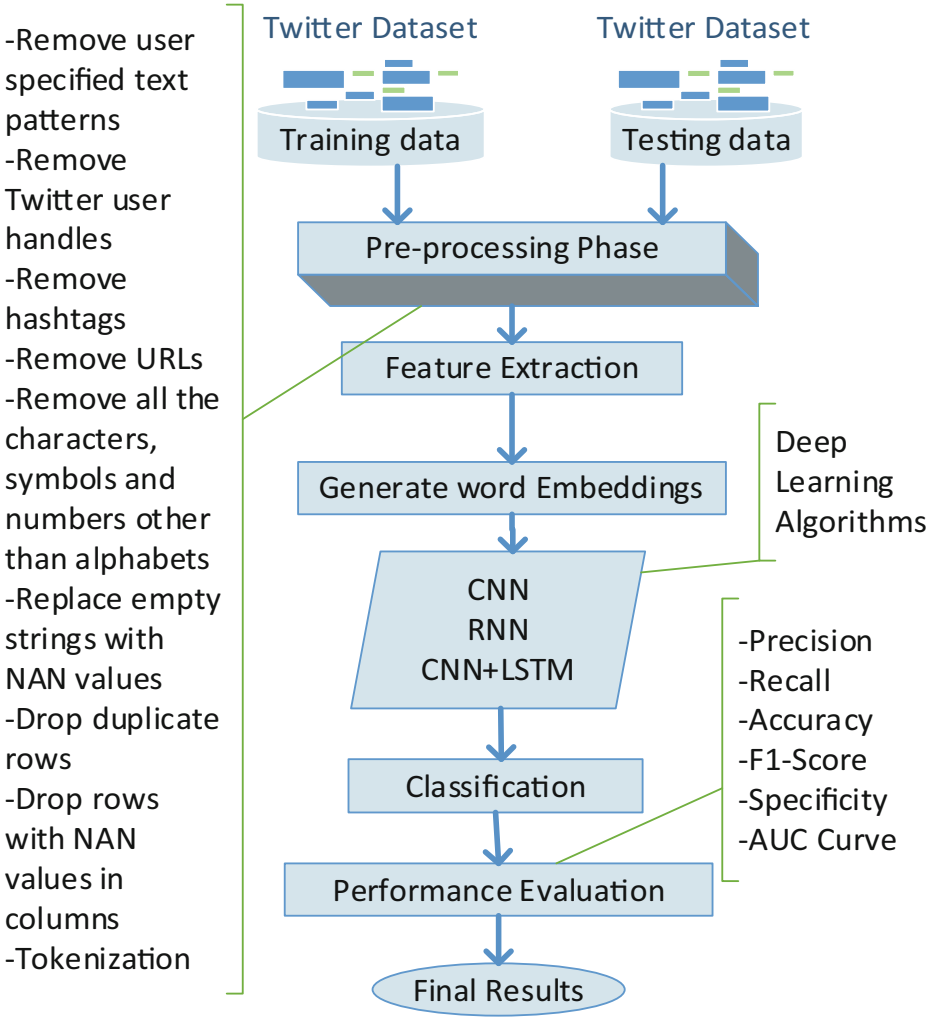

Fig. 3 Proposed flowchart

5 Dataset description and pre-processing

5.1 Dataset description

Twitter is a popular online social media platform that provides open and easy access to data. The development and validation of terms used as a vocabulary for browsing data for users with mental illness take a significant amount of time. In the past, researchers have typically followed two methods of using Twitter data:

1. Using existing datasets that are freely and publicly shared by others [34].
2. Crawling social media vocabulary, though slow, helps to get reliable data.

Table 3 Statistics of the dataset used [54]

Dataset	Depressed	Non-depressed
No. of Users	1402	300 million
No. of Tweets	2,92,564	10 billion

3. Data from resources such as Twitter impose tweet download restrictions per user per day because of the fair use policy applied to all users.

The benchmark dataset that has been considered in this study for depression prediction is taken from [54]. The information of the dataset is given in Table 3. The sample dataset is depicted in Fig. 4 below.

This dataset has been subdivided further into three complementary datasets namely, D1, D2, and D3, as discussed below:

1. **Depression Dataset (D1):** D1 consists of 2,92,564 tweets and 1402 depressed users. Every user has been labeled as depressed. It comprised of tweets obtained between the period 2009–2016. The users were labeled as "users of depression" if the anchor tweet gratified the pattern as "I am / I was / I have been identified as depressed".
2. **Non-depression Dataset (D2):** D2 consists of more than 300 million active users and 10 billion tweets. Each user is labeled as non-depressed and the tweets were gathered in December 2016.
3. **Depression-candidate Dataset (D3):** Tweets were gathered if the word "depress" was loosely used. This dataset contains 36,993 depressive candidate users and more than 35 million tweets.

For D1, D2, and D3, there are 2558, 5304 and 58,810 collected samples. Prior to anchor tweets being detected, the one-month Twitter post information of the twitter user is presented in each dataset. In the present study, D1 and D2 datasets are used and analyzed using the DL-based classification algorithms.

Depression	tweets
False	@Sephy_senpai i love it!! don't get the charge tho, it's hella chunky. the alta's so cute. only downside is that it doesn't track heartrate
True	6 years ago today I was diagnosed with depression. 6 years later I'm still fighting strong. I will win. 💪
True	@FanboysArmyPH @1DMetro @BieberParadePH\n\nI'm diagnosed with stage 957392017491 pmfpd\n(post multi fandom party depression) oh no me! 😭🔫
True	I NEED SOMEONE TO SLAP OR PUNCH OR KILL- https://t.co/Llf8m7OnUI
False	it is 3am and sam and i are yelling about the better life Blanche Dubois deserved
True	"Stop self diagnosing yourself with depression" um...I'm diagnosed bipolar which is a form of depression???????
True	i always feel so guilty abt the fact i was diagnosed with depression as if i could give the diagnosis to someone who has it worse off
True	btw i've been diagnosed as bi-polar. isn't life just so much fun. so now i have social anxiety, depression, and on top of that i'm bi-polar
False	@yifansolo ..uhm where did I blame u? Re read my tweet again.
True	#HonestyHour I have been diagnosed with psychotic depression.
True	Why are you happy then sad all of a sudden? — I was diagnosed with bipolar depression. I don't talk about it all t… https://t.co/FSgJz5u9ku
True	أستغفر الله العظيم وأتوب إليه https://t.co/qKe9zO24Lp
True	I was diagnosed in 2001 after Years of depression, irritablilty & crazy behavior. #teamfollowback
False	@takumiDisneyaka 全種類欲しい😍👍全て飲み干してペットボトルごと保存だな私だったら😉👌

Fig. 4 Sample of depressed and non-depressed tweets from a Twitter dataset

5.2 Data pre-processing

Data pre-processing [41] is an integral part of the data mining process. Real-world data is collected using different methods and is not specific to a particular domain, resulting in incomplete, unstructured, and unreliable data containing errors. Such data leads to irrelevant and erroneous predictions if analyzed directly. In our framework, various methods are used during the pre-processing phase. The first method eliminates the text patterns specified by the user. The goal of this method is to remove patterns, e.g., "user handles (@username)"; "hashtags (#hashtag)"; "URLs"; "characters, symbols, and numbers other than alphabets"; "empty strings"; "drop rows with NaN in the column"; "duplicate rows" etc. This method cleans up each tweet in the dataset and deletes all URLs in the tweet. URLs are not taken into account because they are not useful for prediction purposes and eliminating them will reduce computing complexity. The next step is to delete the date, time, digits, and hashtags. Date and time are useless for the prediction of depression, so this information is removed from the tweets. Likewise, digits are not an appropriate aspect for prediction purposes, and hashtags although may be used for prediction. It has been observed that accuracy is very low when prediction is based on hashtags. As we do not want to deviate from the trend, hashtags have also been removed. The next step is to delete emojis and remove whitespace and extra spaces in the sentence.

After this, stop words are eliminated and stemming is performed. Stopwords like are, was, at, if, etc. do not contribute to the meaning of the sentence. The NLTK [10] package consisting of a set of stopwords is used to remove stopwords from our text. Stemming [6] is a method of changing a word to its root form. Porter Stemmer is used for creating the root of a word by removing prefix or suffix (−ize, −ed, −s, −de, etc.) from the word. After cleaning up all the tweets, cleaned tweets are returned and given as input to the tokenizer, which is the next step. Tokenizing [54] raw text data is a main pre-processing step for NLP methods. Tokenizers are tools that use regular expressions to divide a given string into tokens by breaking a larger body of text into smaller lines or words. Figure 5 shows the clean depression tweets after pre-processing.

The different tokenization functions are used by importing the NLTK package. The first stage of tokenizing is to provide the datasets of cleaned positive and negative tweets as input for Tokenizer.fit_on_texts () function. It updates the internal

Depression	tweets	tidy_tweets	target
True	I was diagnosed with bipolar depression/anxiety when I was 15. I finally found meds that were working for me about 2 months before I found	i was diagnosed with bipolar depression anxiety when i was i finally found meds that were working for me about months before i found	1
False	@YukiMishima_ i don't know! that's why i'm going to the interview!	i don t know that s why i m going to the interview	0
True	@cumheremalik maybe you're right. you know something? i'm somewhat depressed, i was diagnosed with depression in summer	maybe you re right you know something i m somewhat depressed i was diagnosed with depression in summer	1
True	I was 14 when I was diagnosed with PTSD and severe depression. I'm now almost 20. So I'll tell you what I'll do. I'll take - @rregulusblack	i was when i was diagnosed with ptsd and severe depression i m now almost so i ll tell you what i ll do i ll take	1
False	exactly, i'm not the one who bought distant relatives into it either	exactly i m not the one who bought distant relatives into it either	0
True	I went to the hospital today because I had a mental breakdown lol so I was diagnosed with severe depression and anxiety	i went to the hospital today because i had a mental breakdown lol so i was diagnosed with severe depression and anxiety	1
False	I can't stop thinking about her	i can t stop thinking about her	0
True	Well the truth is out now. Yes I've been diagnosed w/major depression for awhile & yes I was hospitalized yesterday.	well the truth is out now yes i ve been diagnosed w major depression for awhile amp yes i was hospitalized yesterday	1
True	So this guy says he likes me... He hasn't seen my fucked up side yet. He doesn't know I'm diagnosed with depression, anxiety and and ED.	so this guy says he likes me he hasn t seen my fucked up side yet he doesn t know i m diagnosed with depression anxiety and and ed	1
False	@_____gunboy36 返してくるところ可愛すぎかよ		0

Fig. 5 Sample of depressed and non-depressed tweets after pre-processing

vocabulary from a list of texts and creates the vocabulary index based on the frequency of words. As a result, the word with the highest frequency has the lowest index value. Hence, this function returns the maximum number of words that is 10,275 in our framework, with an index for each word. The next stage of tokenization is the texts_to_sequences() method. It receives data of the preceding method, consisting of maximum words with an index. Its objective is to turn every word in a tweet into a sequence of integers and replace it with the corresponding integer value of the word_index dictionary. Now, the tweets get converted into integer sequences of varying lengths. After this, tweets get padded with zeroes having a length smaller than the Max_Tweet_Length which is 25 in our frame.

5.3 Word embeddings

In NLP, embedding can facilitate ML applications to work with large data where a word available in highly sparse vectors can be projected down into a low dimensional embedding vector. Embeddings are dense or low dimension demonstrations of high-dimensional input vectors. Some of the recent techniques using word vectors [20, 68] learn from the given text corpus and these word embedding techniques lead to high dimensionality within the solution, typically the size of the whole corpus. Word embeddings [37] are trained in such a way that words with semantically same meanings are positioned close to each other and the vectors are created with approximately identical representations, e.g., the terms “joyful” and “miserable” have very different semantics. So, these will be represented far apart in the geometric space, thereby building more separable features out of the tokenized numerical vectors. These vectors are transformed with the help of embedded layer so that the semantic relationship between the associated word vectors is captured. The original tokenized vector does not have a relationship between different words, whereas the embedding vectors in the embedding space learn the relationship using the distance between the two vectors. Each time the training is repeated, more separable features are extracted providing more predictive power to the CNN or RNN networks. They are one of the best approaches till date to encode a sentence, paragraph, or document and can be seen as one of the breakthroughs in DL capable of solving challenging NLP issues.

We used the “Word embeddings” technique to calculate numerical vectors for every pre-processed data point. First, we converted all the sample text words into sequences for

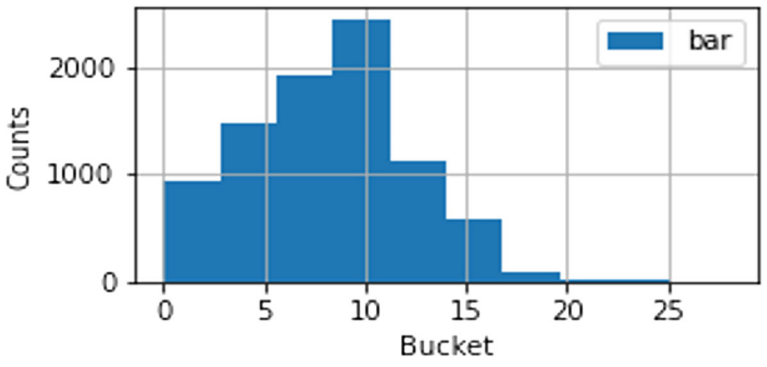

Fig. 6 Text sequence generation

generating word indexes. These indices are retrieved using the Keras text tokenizer [63]. We have ensured that the tokenizer does not assign a zero index to any word and vocabulary length is also adapted accordingly. Next, all separate words in the dataset are assigned a unique index that is used to form numeric vectors of all text samples. Initially, the length of all tweets is obtained for text sequence generation. Figure 6, illustrates a histogram of the number of tweets increasing in word length, and it is evident that the majority of tweets in the training set are fewer than 25 words long. As a result, text sequences are converted into integer sequences and zero padding is implemented. Maximum sequential length (number of words) is set to 25 because a majority of tweets in the dataset are of that length. Furthermore, 5 words are discarded as such tweets are very less in number and result in the addition of zeros to the sequence of vectors, thereby slowing model training and affecting the overall performance. The process of generating word embeddings for a tweet is illustrated in Fig. 7.

In this study, an embedding matrix with the Max_unique_words * Embedding_dim dimension is created. Embedding_dim is the length of the vector i.e., 300 in our case. This matrix is initially populated by zeroes. Every single word in the top 10,275 unique words is converted into a vector by searching inside the vector space. After this, each vector is populated as a row in the embedding matrix, and the defined vector contains 300-dimensional columns of features with a vocabulary of 10,275. We used an embedding layer of length 50 on our DNNs to produce a 25 × 50 output embedding vector for each tokenized vector. This layered neural network in our framework is trained to rebuild the linguistic context of words by taking a large body of text as input i.e., EMBEDDING_FILE. This generates a vector space, usually of several hundred dimensions, with each unique word in the corpus having a unique vector.

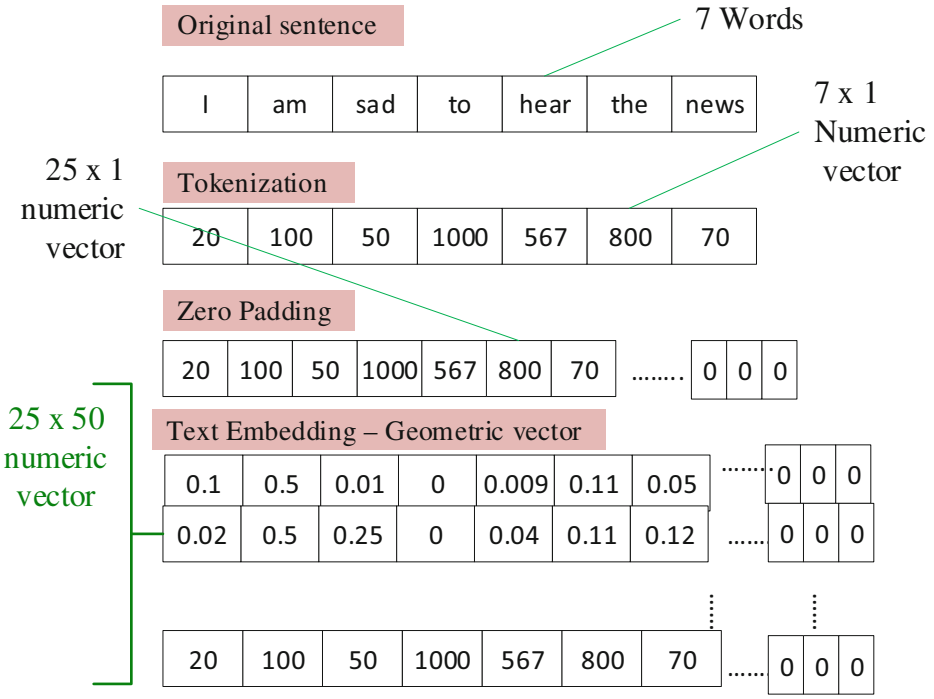

Fig. 7 Word embeddings for a single tweet

At last, we divided the positive and negative data sets for testing and training. 30% of the data is for training and remaining 70% of data is for testing. The CNN, RNN, and the proposed hybrid CNN-biLSTM models are discussed below in further sections.

6 Proposed hybrid CNN-biLSTM model

We proposed an approach by combining CNN and bidirectional-LSTM (a type of RNN), for higher classification performance to predict depressed users on Twitter. On performing multiple experiments, we found that CNN is good at extracting spatial features and performed well when contextual information with the prior sequence is not required. Whereas, RNNs are effective in extracting information when the context of adjacent elements is important to classification. In the beginning, multidimensional data is used directly as a low-level input to CNN. In the pooling and convolution operation, significant features are mined by every layer. Relative to traditional CNN, the output layer is fully connected to the hidden layer. Depression tweets are extracted in-depth and accuracy is improved by using a number of convolutional layers, pooling layers, and convolutional kernel enhancements. With over-fitting risk exposure, a complex network may occur.

As a result, the time-dependent network of recurrent neurons i.e., LSTM is incorporated with the CNN model to address the sequence problem. Similar and useless information is extracted by using the convolution kernels and the important extracted information is stored for a longer time in the state cell. Predominant results are achieved through a combination of CNN and biLSTM. The biLSTM architecture is shown in Fig. 8. Consequently, CNN-

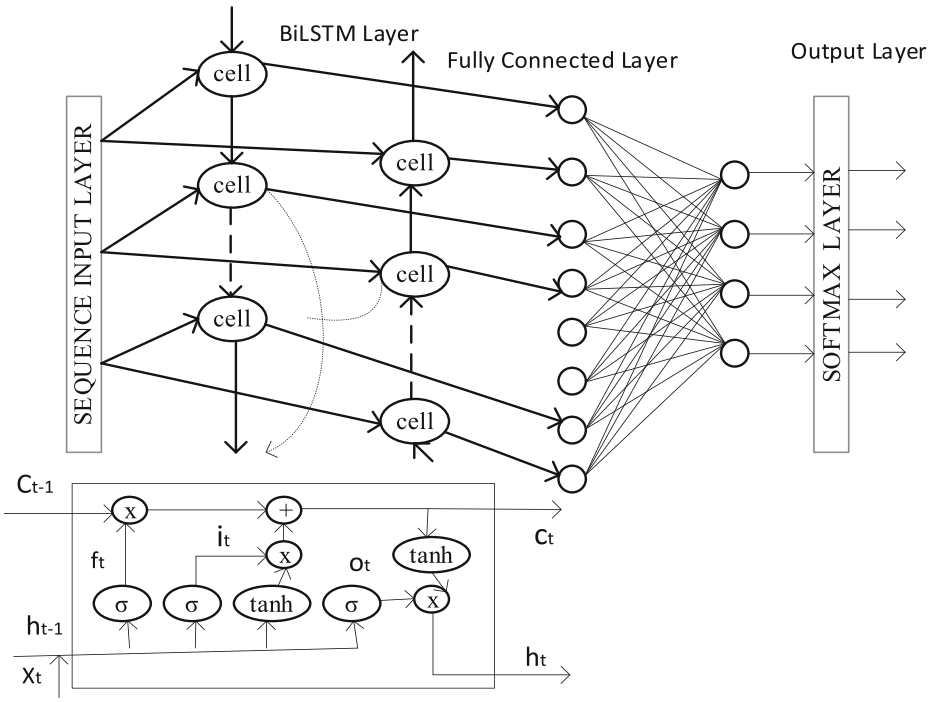

Fig. 8 Architecture of biLSTM

Table 4 Control parameters used in equations

Control parameters	Meaning
CNN (Convolution neural network)	
T	Number of tokens in the text
d	Words
$t{:}t{+}d$	Input during the t^{th} convolution
$x_t{:}t{+}d{-}1$	Embedding vectors
W_d	Learnable weights matrix
b_d	Bias
h_d	Features for filter size d = 1, 2, 3
p_d	Max pooling
LSTM (Long Short-term memory)	
t	Time
x_t	Input
h_t	Hidden state
σ	Sigmoid function
o	Convolution operator
f_t	Forget gate
i_t	Input gate
o_t	Output gate
c_t	Cell state vector

biLSTM is used where the convolutional layer facilitates the extraction of low dimensional semantic features from textual data and minimizes the number of dimensions. Moreover, biLSTM processes the text as a sequence of input. In this study, several 1-dimensional convolutional kernels are used together to achieve better performance on the input vectors. Table 4 describes the control parameters used.

Sequential input data is characterized as the average of embedding vectors of individual words as shown in Eq. (1). Unigram, Bigram, and Trigram features are extracted using the different sizes of convolutional kernels by applying them to $X1 : T$ with the use of 1D CNN. The features generated during a convolution process, in the t^{th} convolution, where a window of d words stretch from $t : t + d$ is taken as an input. The convolution process generates features for that window as follows in Eq. (2),

$$X_1 : T = [x_1, x_2, x_3, x_4, \cdots x_T] \tag{1}$$

$$h_d, t = tan\, h(W_d x_{t\ :t+d-1} + b_d\ ) \tag{2}$$

Where, $x_t : t + d - 1$ is the embedding vector of the individual words in the context window, W_d represents the parameters with the learnable weight's matrix, and b_d is the bias. Also, as different regions of the text are convolved with every filter, the generated feature map of the filter having a convolution of size d is given in Eq. (3),

$$h_d = [h_{d1}, h_{d2}, h_{d3}, h_{d4}, \cdots x_{T-d+1}] \tag{3}$$

Using different convolutional kernels with varied widths expands the scope of CNN to find the latent correlation between several adjacent words. The most important characteristic of using a convolution filter for feature extraction from textual data is to minimize the number of

trainable parameters during the feature learning process. This is achieved using a max-pooling layer following the convolutional layers [5]. The process starts with input being processed through various convolutional channels and every channel has its unique set of values. During max-pooling process, the largest value from each convolutional layer is selected and pooled to create a set of new features. Within each convolution kernel, max pooling is applied to the feature maps with convolutional size d to obtain Eq. (4). The final features of each window are extracted by concatenation of p_d for every filter size d = 1, 2, 3 and extracted the unigram, bigram, and trigram hidden features as shown in Eq. (5),

$$p_d = Maxt\left(h_{d1}, h_{d2}, h_{d3}, h_{d4}, \cdots x_{T-d+1}\right) \tag{4}$$

$$h_d = [p_1, p_2, p_3] \tag{5}$$

The most significant advantage of using a CNN-based feature extraction technique over the conventional LSTM is that the overall number of features are considerably reduced. These features are further used by a depression prediction model following the feature extraction process. The architecture of LSTM makes it overcome the "vanishing gradient" problem with sequential data using gate structures like input, output, and forget gates along with the cell states. These cell states work as an overall long-term memory for the LSTM unit and additive connections between the states. At a given time t for an LSTM cell, provided the input x_t and the intermediate state, h_t its output state is calculated as follows in eq. (6, 7, 8, 9, 10 and 11).

$$f_t = \sigma\left(W_f\, x_t + U_f\, h_{t-1} + b_f\right) \tag{6}$$

$$i_t = \sigma\Big(W_i\, x_t + U_i\, h_{t-1} + b_i \tag{7}$$

$$o_t = \sigma(W_o\, x_t + U_o\, h_{t-1} + b_o) \tag{8}$$

$$g_t = tanh\left(W_g\, x_t + U_g\, h_{t-1} + b_g\right) \tag{9}$$

$$c_t = f_t\, o\, c_{t-1} + i_t\ \ o\, g_t \tag{10}$$

$$h_t = o_t\, o\, tanh(c_t) \tag{11}$$

Where the learnable parameters are represented by W, U, and b. σ is a sigmoid function and o is the convolution operator. The LSTMs gate is represented by f_t, i_t, and o_t for forget, input, and output gates respectively. The memory state or the cell state is shown with c_t. The cell state is the only reason LSTMs are skilled to capture all the long-term dependencies in the data provided in the input sequence and can be applied to data with longer sequences. As shown in

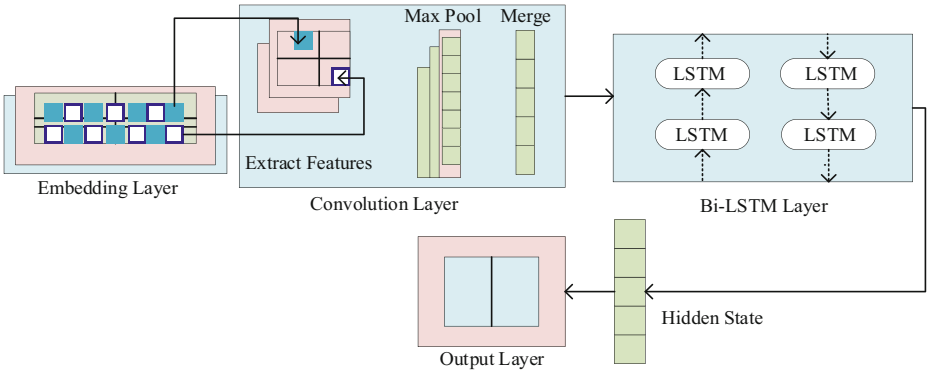

Fig. 9 Architecture of CNN-biLSTM for the depression prediction

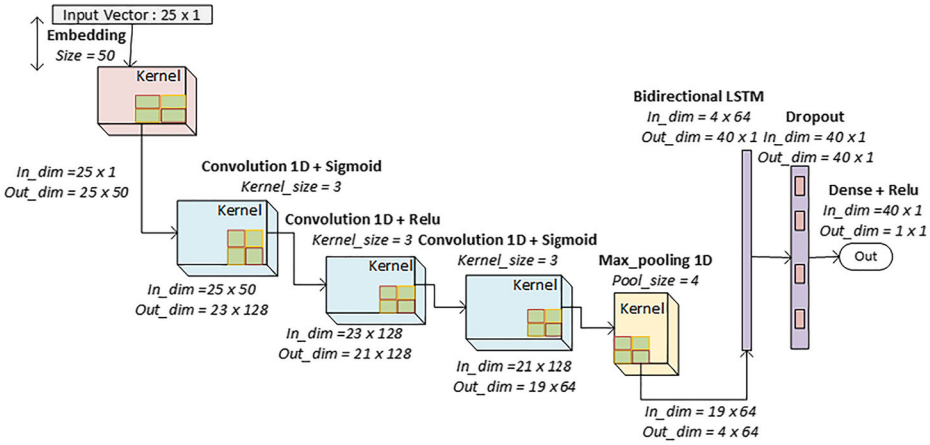

Fig. 10 Layers and parameters used in CNN-biLSTM neural network

Figs. 9 and 10, the CNN part of the network comprises of three convolution layers with a variable number of filters. The first two layers have 128 filters having a kernel size of 3 × 3 with sigmoid and Rectified linear activation unit (Relu) as activation functions. The third convolution layer has 64 filters with a kernel size of 3 × 3 and a sigmoid activation function.

Table 5 Parameters used in our hybrid model

Layer (type)	Output shape	Parameters
Embedding	(none, 25, 50)	5,13,800
Conv1D	(none, 23, 128)	19,328
Conv1D	(none, 21, 128)	49,280
Conv1D	(none, 19, 64)	24,640
MaxPooling1D	(none, 4, 64)	0
Bidirectional	(none, 40)	13,600
Dropout	(none, 40)	0
Dense	(none, 1)	41
Total parameters		6,20,689
Trainable parameters		6,20,689
Non-trainable parameters		0

This is trailed by a Max-pool layer with a 4 × 4 kernel size. Finally, we used a biLSTM with slightly different hidden computations. Since computations are carried out in both forward and backward directions, this helps us in dealing with the drawback of RNN i.e., only information from previous computations is used as the next step. We used a "dropout layer" with a "keep probability" of 0.1 to prevent overfitting on the training data. We used Relu activation for the output layer and the model is trained on "binary_crossentropy" loss and "root mean squared propagation (RMSprop)" optimizer. Table 5 shows parameters used in our model and pseudocode of our proposed work is illustrated in Algorithm 1.

Input : (X_train) training features, (Y_train) targets
Hyperparameters: embedding_size, total_filters, kernel_size, pool_size, strides, dropout_rate, and lstm_units, kernel_type, loss_function, optimiser, total_epochs, batch_size Randomly Initialize the model()

"CNN model is prepared with biLSTM for text predictions and trained for 50 epochs"
--------Embedding(input)
-------Initialise Sequential model()
----------**// Add Embedding layer as input layer**
---------model = model.add(Embedding_layer(vocabulary, embedding_size))
---------**// 1st Convolution layer**
---------model = model.add (Sequential_Layer (Convolution1D [total_filters, kernel_size, layer_name = "Conv_1D_1"], -activation_fucntion = 'sigmoid'))
---------**// 2nd Convolutional layer**
---------model = model.add (Sequential_Layer (Convolution1D [total_filters, kernel_size, layer_name = "Conv_1D_2"], activation_function = 'Relu'))
---------**// 3rd Convolutional layer**
---------model = model.add (Sequential_Layer (Convolution1D [total_filters, kernel_size, layer_name = "Conv_1D_3"], activation_function = 'sigmoid'))
---------**// Max pooling layer**
---------model = model.add (max_pool_layer(pool_size, strides))
---------**// biLSTM layer**
model = model.add (biLSTM_layer (lstm_units, activation_function, recurrent_activation, dropout, return_sequences))
---------**//Model dropout**
---------model = model.add (Dropout(dropout_rate))
---------**// Dense layer**
---------model = model.add (Dense_layer(kernel_type, total_units, activation_function = 'Relu')
-------**// Compilation**
-------model.compile (loss_function, optimizer='RMSprop')
model.fit (X_train, Y_train, total_epochs, batch_size

7 Comparative methods

7.1 CNN model

CNN [3] has revolutionized research and innovation using ML. This can be seen through the applications of DL techniques and their performance on the ImageNet [17] dataset for object detection problems in Computer Vision. Not only Computer Vision, but research has been successful in the text analysis domain with the use of CNNs. CNNs are excellent feature extractors, extracting domain-specific characteristics during each epoch, e.g., retrieving high-level characteristics, such as edges, and low-level characteristics such as objects in images. Using appropriate filters, convolution layers, dimensionality reduction operation, and pooling operations like Max-pooling and Average-pooling, time dependencies and specific features can also be captured. The role of a convolution layer is to tune the input features by reducing the spatial size of the convolution matrix into a form that is easier to process with no loss of critical classification information. The matrix involved in the realization of convolution

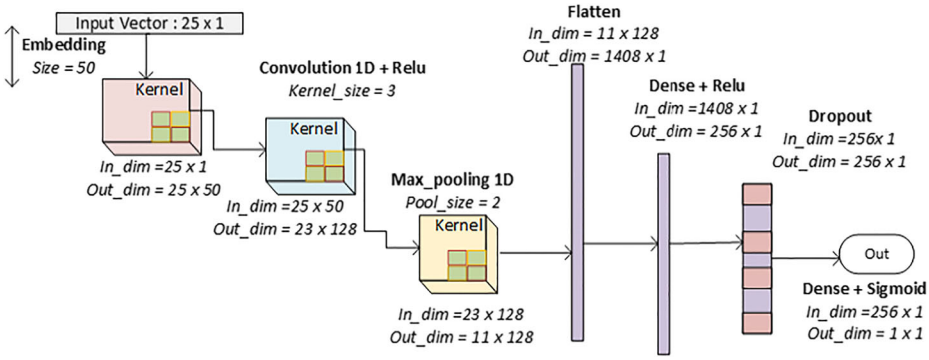

Fig. 11 Layers and parameters used in CNN neural network

information is referred to as Kernel or Filter. The kernel moves with a specific stride and each time a matrix multiplication operation is performed between the kernel and a portion of the input vector that superimposes itself on the kernel. This not only reduces the computing power needed to process the data but is also useful in extracting key features which are position invariant and contribute to effective model training.

In this study, convolutional layers, pooling layers, drop-out layers, and dense layers have been used. The neural network mostly used convolutional layers for training. The architecture of CNN to classify the text is shown below in Fig. 11.

The network parameters are taken up by the fully connected layers. The input features are extracted by the convolution layer and the resulting convolution matrix is obtained from pooling. The problem of over-fitting is resolved by the dropout layer and, as a result, during training, co-adaptation is prevented by random dropping units. We have presented the application of DNN with CNN + Maxpool to improve classification performance compared to RNN. For the convolution layer, we used a 3 × 3 kernel size with a stride of 1 tailed by a Max-pooling layer with a 2 × 2 kernel size. Finally, the convoluted features are fed into a dense layer. Since the network parameters are close to ~1 M, we used a "dropout" layer with a "keep probability" of 0.2 to avoid overfitting the training data. Sigmoid activation has been used for the output layer and the model is trained on the "BinaryCrossentropy" loss function and "RMSprop" optimizer. A l long sentence as the $d\,X\,l$ matrix is called the S sentence matrix, and by using linear filters, CNN carried out the convolution on the given input. The W weight matrix is indicated by a filter of size r and length d. W represents parameters as $d\,X\,r$. $S \in \mathbb{R}^{dXl}$is the input matrix, feature map vector with a convolutional operator $O = [O_0, O_1, \ldots., O_{s-h}] \in \mathbb{R}^{s-r+1}$ with filter W applied repeatedly to sub-matrices S as shown in eq. (12),

$$O_i = W.S_{i:i+h-1} \tag{12}$$

where, $i = 0, 1, 2 \ldots s - r$, (.) = dot product operation and $S_{i:j}$ = S's sub matrix from i to j rows. Each feature map O was served to the pooling layer for generating possible features. The highly significant feature v, captured by the max-pooling layer selects the highest value of feature map as shown in eq. (13),

$$v = \max_{0 \le i \le s-r} O_i \tag{13}$$

7.2 Recurrent neural networks – RNN model

RNN and LSTM have been highly popular in the area of NLP and speech recognition. Applications such as language modeling, sentiment classifying, contextual modeling, named entity recognition, and neural machine translation at the character level had produced excellent experimental results with RNN-based sequential modeling [23, 59]. In 2016, the Neural Machine Translation System [72] of Google utilized a deep LSTM network for multi-lingual translation. Intuitively, LSTM has been found useful for learning the context among adjacent words improving the classification performance where class ambiguity exists. RNN is a group of neural networks that support sequential data modeling. Such networks are derived from the feed-forward networks which exhibit the same behavior in relation to the function of the human brain.

RNNs have a memory that captures sequential information using the unrolled units that have hidden states and, thus, weights and biases are shared over time as shown in Fig. 12. For every element in a sequence, the output of RNN utilizes the previous calculations; therefore, RNNs are defined as recurring. Also, the future calculations are based on this captured information. As a result, RNNs are successful in NLP issues such as automatic translations where all inputs and outputs are highly dependent on each other. RNNs use information in the long sequences, but practically, they limit themselves to limited steps and capture short-term dependencies. Generally, in RNN, the initial layer is the encoder, which converts text into a sequence of token indices. Post encoder layer, data is routed through the embedded layer, and sequences of word indices are converted into a sequence of trainable vectors. Words of the same significance show same training vectors. Through iteration of elements, RNN processes the input sequence of "one step-of-time" to the input of another "step-of-time". Final processing is done using the dense layer after RNN has converted the single vector sequence.

In our study, we used a network of SimpleRNN [49] + Dense layers with "Relu" [1] activation to solve the problem of depression prediction classification on the Twitter dataset.

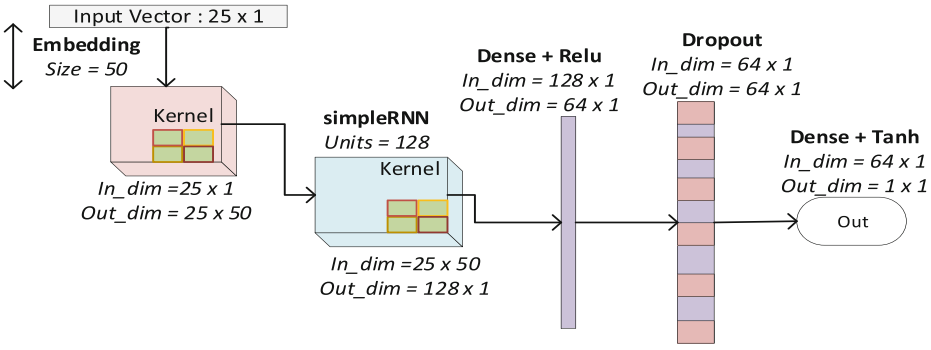

Fig. 12 Layers and parameters used in RNN neural network

SimpleRNN is a fully connected RNN in which the output of the prior time step is delivered to the next step. We used a "dropout" with a "keep probability" of 0.4 and a "tanh" activation function for the output layer. The model is trained on "binary_crossentropy" loss and "Adam" [30] optimizer with a learning rate of 0.001. RNN requires 3-dimensional data as input provided by the embedding layer. The design of our model is illustrated in Fig. 12. RNNs have proven to be very promising in improving baseline performance, although they do have certain limitations. As the calculation of RNNs is slow, they are not useful in accessing the information over the long term. Furthermore, RNNs cannot account for future inputs for the present state. As a result, other DL approaches were explored to address the issue. The other types of RNNs are LSTMs [25], biLSTMs [12] and are the most frequently used RNN type. They just have a different way of calculating the hidden state but these are far better at catching the long-term dependencies that RNNs cannot.

8 Results and discussion

This section shows experimental setups and evaluation measures considered while conducting experiments. Moreover, it discusses results achieved by conducting experiments for the proposed CNN-biLSTM approach. It also includes a performance comparison between the proposed method and other state-of-the-art studies. Moreover, this section presents statistical analysis and visualizations for lingual dialects used by depressed and non-depressed users.

8.1 Experimental setup and evaluation metrics

All DL techniques have been implemented using Anaconda Navigator in Python 3.7 and Keras (an open-source library based on TensorFlow). For calculating performance on the implemented algorithms, we used information retrieval metrics extracted from the confusion matrix of the classifier [27, 79]. A confusion matrix is a way to summarize a classification model consisting of multiple sub-metrics. The sub-metrics derived from confusion metrics include precision, recall, accuracy, F1-score, sensitivity, and AUC curves. It analyzes the model performance for each class in a classification problem. Sometimes confusion metrics are referred to as error matrices because of the tabulated representation showing correct and incorrect predictions. The description of metrics is given in Tables 6 and 7.

The performance of classification model is evaluated by plotting the distribution of prediction ratios for various classes on test data for known actual values, in the form of a confusion matrix. The confusion matrix for RNN, CNN, and CNN-biLSTM are shown in

Table 6 Terms used to compute confusion matrix

Terms	Explanation
P (Actual Positive)	The actual positive case, which is depressed in our task.
N (Actual Negative)	The actual negative case, which is not depressed in our task.
TN (True Negative)	The actual case is not depressed, and the predictions are not depressed as well.
FN (False Negative)	The actual case is not depressed, but the predictions are depressed.
FP (False Positive)	The actual case is depressed, but the predictions are not depressed.
TP (True Positive)	The actual case is depressed, and the predictions are depressed as well.

Table 7 Description of evaluation metrics

Evaluation metrics	Description	Equations
Accuracy	It is defined as the sentiments classified correctly with respect to the entire available classified sentiments.	$Accuracy = \frac{TP+TN}{TP+FN+TN+FP}$
Precision rate	Precision numerically presents how good the classifier is in predicting true values out of overall predicted values. High precision means the number of incorrectly Positively classified samples (False positive) is less and vice versa.	$Precision = \frac{TP}{TP+FP}$
Recall rate	This term is used to measure the accuracy of the classifier on the actual ground truth values. The higher the value of recall implies that the incorrectly negatively classified examples (False negatives) are less and vice versa.	$Recall = \frac{TP}{TP+FN}$
F-measure	It is a better performance metric that can be used which is a combination of both precision and recall is F1-Score, i.e., a harmonic means of both the matrices.	$F{-}measure = 2 * \frac{Precision * recall}{Precision + recall}$
Error rate	It is defined as the efficiency measure of communication channels and is derived by subtracting a hundred from the accuracy rate.	$100 - accuracy = Error\ rate$
AUC Curve	It can be defined as the exact integral of the curve that shows the variations in classification.	$AUC = \frac{1}{2}\left(\left(\frac{TP}{TP+FN}\right) + \left(\frac{TN}{TN+FP}\right)\right.$

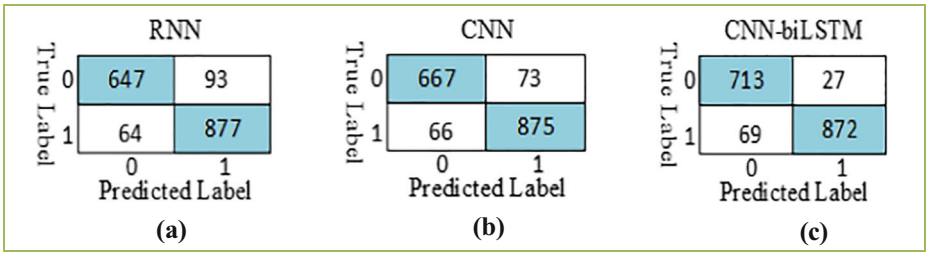

Fig. 13 Confusion matrix for **a** RNN **b** CNN and **c** CNN-biLSTM

Table 8 Experimental results

Performance metrics	RNN	CNN	CNN-biLSTM
Accuracy	90.66	91.73	94.28
Precision	90.41	92.29	96.99
Recall	93.19	92.98	92.66
FI Score	91.78	92.64	94.78
Specificity	87.43	90.13	96.35

Fig. 13a–c. CNN-biLSTM obtained a significantly larger number of true positive and true negative values, demonstrating it as an efficient classifier for our dataset.

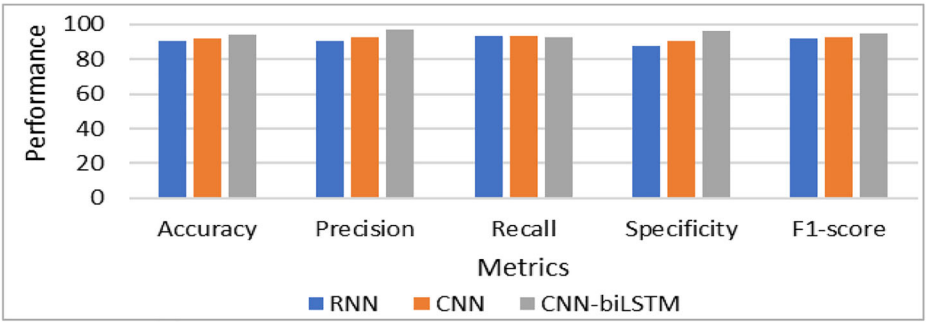

Fig. 14 Comparison of RNN, CNN and CNN-biLSTM

8.2 Comparative analysis

The proposed hybrid CNN-biLSTM model is compared with the CNN and RNN models to evaluate depression prediction using accuracy, precision, recall, F1-score, and specificity in relevance to the test set of the used dataset. Table 8 depicts that CNN-biLSTM shows higher accuracy, precision, F1-score, and specificity i.e., 94.28, 96.99, 94.78, and 96.35 respectively as compared to RNN and CNN models. The proposed model aims to enhance the precision score while maintaining a relatively stable recall value, thereby, ensuring no incorrect

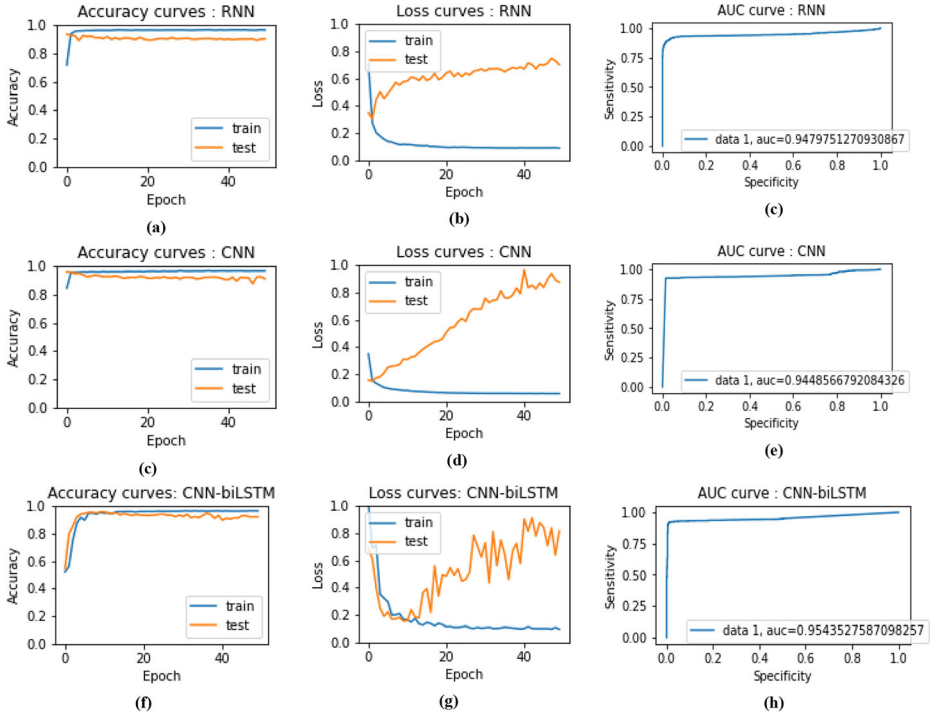

Fig. 15 Comparison of **a** Accuracy graph for RNN, **b** Loss graph for RNN, **c** AUC graph for RNN, **d** Accuracy graph for CNN, **e** Loss graph for CNN, **f** AUC graph for CNN, **g** Accuracy graph for CNN-biLSTM, **h** Loss graph for CNN-biLSTM, and **i** AUC graph for CNN-biLSTM

depression predictions are made. The performance metrics obtained in Table 8 using different models is graphically shown in Fig. 14.

From Fig. 14 it is evident that RNN model has the lowest prediction performance compared to CNN and CNN-biLSTM. It is because of the inability of RNN to deal with "vanishing gradient" and "exploding gradient" problems. Although LSTM can tackle both problems, but it can only learn information before the current word but not after it. The semantics of a word in a sentence is related not only to previous historical information but also to subsequent information that comes after it. Instead of using LSTM, this study employs biLSTM incorporating bidirectional information as well as overcoming the problem of "vanishing gradient" and "exploding gradient". The best results are obtained when CNN's deep feature extraction capabilities are combined with the biLSTM model.

To analyze the model performance more clearly, a comparison of model graphs per epoch for accuracy, loss, and AUC is shown in Fig. 15. Accuracy in the case of training and testing data for most models follows a general Hilton curve and stabilizes around 0.90, as shown in

Table 9 Classification comparison with state-of-the-art studies using F1-scores and accuracy measures

State of art studies \ Performance metrics	F1-score	Accuracy
Naseem et al. [41]	–	0.94
Zhou et al. [78]	0.84	–
Nadeem et al. [40]	0.86	0.81
Chenhao Lin et al. [35]	0.936	0.884
J. Samuel et al. [52]	–	0.91
Shuai et al. [57]	–	0.904
Jamil et al. [28]	0.73	0.75
Shen et al. [54]	0.85	0.85
Tong et al. [65]	0.91	0.915
Tong et al. [66]	0.90	0.89
Proposed method	**0.9478**	**0.9428**

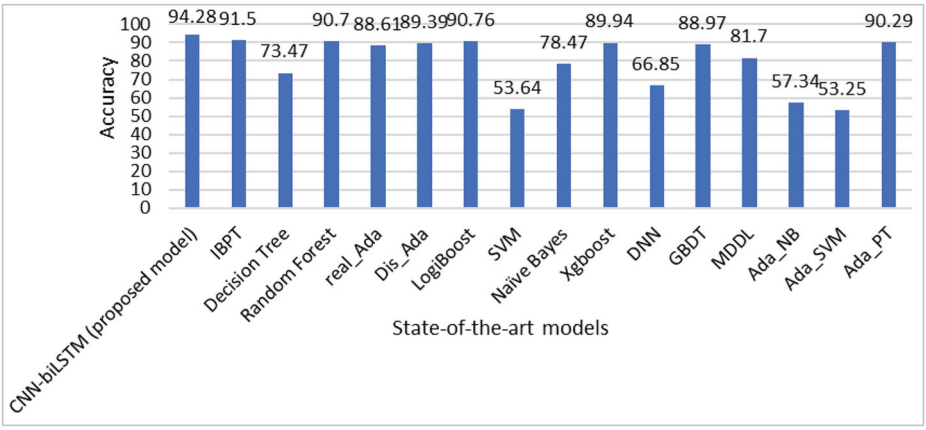

Fig. 16 Comparison of accuracy of CNN-biLSTM model with state-of-art models [65] on same dataset

the graphs. However, the loss function, particularly for test data, indicates an unstable noise output for both RNN and CNN. Moreover, in the case of proposed hybrid CNN-biLSTM model, the propagated noise is reduced. In the case of RNN and CNN, the overall AUC score is around 0.94. The AUC value of the CNN-biLSTM is slightly higher, i.e., 0.95432, indicating improved performance. As a result of the proposed hybrid CNN-biLSTM model, the accuracy of the depression analysis prediction improves while the model loss decreases.

Table 9 compares F1-scores and accuracies of state-of-the-art studies with the proposed hybrid CNN-biLSTM model. In comparison to other existing studies for depression prediction based on Twitter data, it is notable that our proposed hybrid model increases not only accuracy but also the overall F1-score. Figure 16 compares the accuracies of various algorithms (as implemented in Ref. 65 on the same Twitter dataset) to our proposed model. The proposed hybrid CNN-biLSTM model outperforms traditional approaches with an accuracy of 94.28. This implies that the hybrid deep learning models could be investigated in the future for depression analytics.

It is concluded that CNN effectively extracts local features from different locations in a sentence but does not capture the contextual features of a word token. Convolution, pooling, and fully connected layers allow CNN to adapt and learn important features using backpropagation algorithm. The convolution operation is used for weight sharing across neighborhood positions, allowing kernels to extract local information within a given space. Moreover, CNN learns relevant feature patterns using a pooling operation. The Max-pooling layer extracts important information from input feature maps and outputs the most significant value in each map while discarding others, thereby shrinking the number of input features. In comparison to LSTM, RNN has the shortcoming of being unable to handle the "vanishing gradient" and "exploding gradient" problems, as well as extracting specific context information from a long sentence. On the other hand, LSTM efficiently tackles the "vanishing gradient" and "exploding gradient" problems along with contextual feature extraction. However, the problem with LSTM is that it is unidirectional, indicating that it does not consider the effect of next word in a sentence on the current context. Furthermore, the bidirectional nature of biLSTM model concentrates on the important contextual features of a sentence, and the embedding layer extracts not only word-level but also character-level embedding vectors.

Thus, the proposed CNN-biLSTM model efficiently addresses the shortcoming of CNN and RNN by extracting both local features and contextual information from the features obtained from convolutional layer. This validates our hypothesis that integrating CNN and biLSTM improves localized feature extraction while also leveraging biLSTM's multi-directional enhanced RNN functionalities.

8.3 Statistical analysis

In this study, a t-test is applied to determine the significant difference among two groups i.e., depressed and non-depressed tweets. t-test [45] is a parametric test that determines whether two sets are different from one another or not. The aim of the test is to determine whether there is a noteworthy difference among the average length of string for both depression and non-depression tweets. The t-test statistics is given by Eq. (14),

tidy_tweets	len_tidy_tweets
do as diagnos things depress confirm wrong as adhd believe	10
hate friend med far wk diagnos liver failure depression mild paranoia chronic fatigu	13
bulli years torn apart bulli diagnos depress anxieti	8
diagnos depress before	3
self help tip for deal depress mental illness diagnos bipolar medicin	11

Fig. 17 Tweet length for non-depressed users

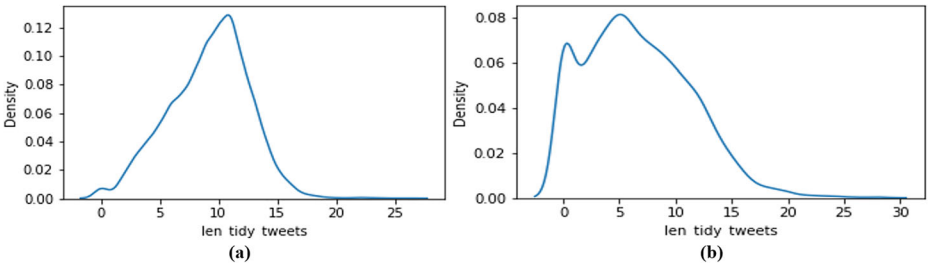

Fig. 18 Distribution plot for **a** depressed users and **b** non-depressed users

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{14}$$

where, t represents t-value to be calculated, $\overline{x_1}$ and $\overline{x_2}$ represents the means of two groups (depressed and non-depressed) whose sample distributions are compared, $\overline{x_1} - \overline{x_2}$ represents the difference in sample means, $s_1$and s_2 represents standard error of the two distributions, and n_1 and n_2 represent a number of observations in each group respectively. For a t-test, the degree of freedom (df) is the least of the two (n_1–1, n_2 – 1). In order to perform statistical analysis, the average length is calculated for both depressed and non-depressed tweets. The length of tweets is calculated and shown in Fig. 17. Density distribution plots are plotted for the length of the string as shown in Fig. 18a, b. After this, the mean of two distributions has been calculated i.e., μ_1= 9.17 for depressed strings and μ_2 = 6.63 for non-depressed strings. The null hypothesis assumes that the mean of two population sample distributions is equal (H_0: $\mu_1 = \mu_2$) and to test

Table 10 Result of the t-test

t-test		
Test statistic	DF	Sig. (2 tailed)
34.749	1	0.000

Fig. 19 Word cloud of **a** depressed users and **b** non-depressed users

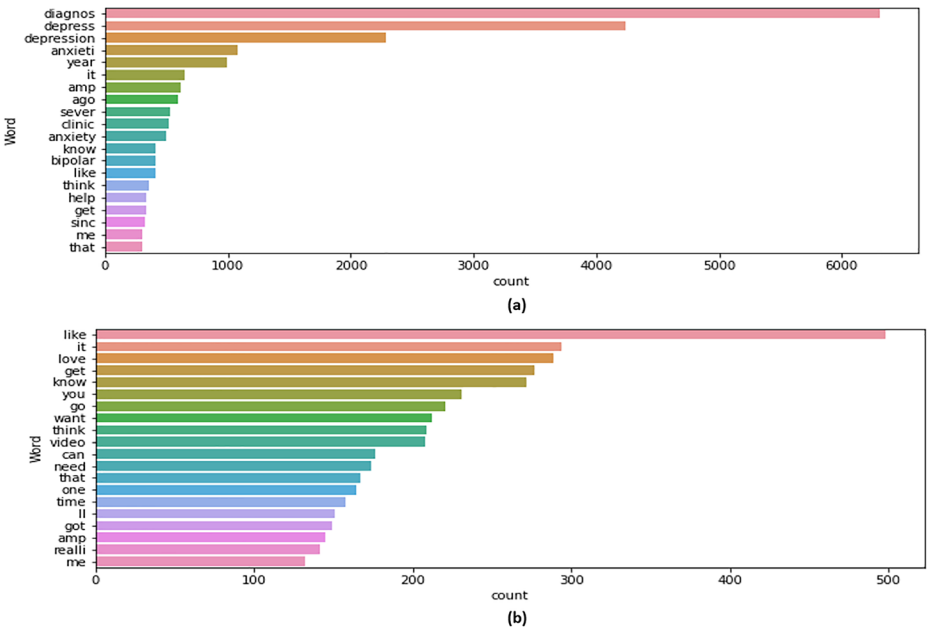

Fig. 20 Words frequency distribution plots of **a** depressed users and **b** non-depressed users

the alternate hypothesis against the null hypothesis we use a t-test. The alternate hypothesis emphasizes the significant difference between the means of both distributions.

P value is a critical value that is an important threshold and acts as a test statistic for the test results. The p value is a level that permits us to conclude when to discard the null hypothesis i.e., $H_0 : \mu_1 = \mu_2$ or to inaugurate that the two groups are dissimilar. The p value obtained from t-test = 0.000. However, we compare p value with the critical value (α). In our study, the critical value is chosen as $\alpha = 0.02$.

1. If p value > α (Critical value): t-test fails to reject the null hypothesis of the test and establishes that the distributions have the same mean.
2. If p value < α (Critical value): t-test rejects the null hypothesis of the test and establishes that the means of the sample distributions are different.

In this study, we performed the two-tailed test. The calculated value for t-test statistics is 34.749. The tabulated value for t-test statistics at $\alpha = 0.02$ and degree of freedom = 1 is 31.821. The t-test statistics is shown in Table 10. Inference from the t-test is that the null hypothesis is rejected and the distribution for the tweet length for both depressed ones and non-depressed ones are different at 0.02 level of significance.

Figure 19a, shows the word cloud [53] for depressed users depicting that the words depress and depressed, are frequently announced by depressed users on a social media platform. This demonstrates how frequently depressed people express and post their sentiments on social media platforms. Besides that, other words related to mental health such as anxiety, mental disorder, bipolar, and PTSD have seemed to appear in more than 20% of depressed users. Figure 19b shows the word cloud for non-depressed users depicting positive and life-enriching attitudes, as well as the sensitization of self-love, as shown by terms such as love and good.

The word frequency plot for the first 20 words is shown in Fig. 20a, b representing words like severe, depressed, help, anxiety, bipolar, etc., are often used by depressed users. The top words used by the non-depressed users, on the other hand, are love, like, know, think, etc. Moreover, we can conclude from plot analysis that these individuals are more likely to have a depressive self-analysis, which they actively reported on various social media platforms.

9 Conclusion

Depression is one of the most common mental disorders permeating worldwide. It is important to educate ourselves about depression on an individual, communal, and global scale. Addressing the issue and helping individuals suffering from depression should be given utmost priority. For classification-based problems, NB, DT, and RF are generally used in the text-based SA. In this study, we presented an innovative methodology to predict depression using Twitter raw dataset comprising of three subtypes (D1, D2, and D3). A real-world dataset has been used for categorizing non-depressed and depressed users in the proposed model. The proposed hybrid model i.e., CNN-biLSTM is characterized by introducing interplay between CNN and biLSTM network. Our proposed model uses biLSTM that can process longer text sequences and tackle the "vanishing gradient" and "exploding gradient" problems, unlike RNN. Moreover, our approach extract features using convolution layers and enhanced recurrent network architectures. On comparing CNN-biLSTM model with "state-of-the-art" studies, it is evaluated that the former model shows better performance in terms of various evaluation metrics. It is concluded through experimental studies that the CNN-biLSTM model is the one that achieved the best accuracy of 94.28%, precision of 96.99%, F1-score of 94.78%, specificity of 96.35%, and AUC score of 95.43%.

This work has a lot of potential to be studied further in the future; for instance, we can increase the model's accuracy by exploring different combinations of neural network layers and activation functions. The pre-trained language techniques such as Deep contextualized word representations (ELMo) and Bidirectional Encoder Representations from Transformers (BERT) can be used in the future, and train them on a large corpus of depression-related tweets. It can be challenging to use such pre-trained language models due to the restriction imposed on sequence length of a sentence. Nevertheless, studying these models on this task helps to unearth their pros and cons. Our future work aims to detect other mental illnesses in conjunction with depression

to capture complex mental issues prevading into an individual's life. Apart from Twitter raw data, various other ML methods can be evaluated on different other social media networks.

Abbreviations *LSTM,* Long Short-Term Memory network; *WHO,* World Health Organization; *NLP,* Natural Language Processing; *RNN,* Recurrent Neural Network; *CNN,* Convolutional Neural Network; *RMSProp,* root mean squared propagation; *ML,* machine learning; *SVM,* Support Vector Machine; *DT,* Decision Tree; *RF,* Random Forest; *NB,* Naïve Bayes; *KNN,* k-nearest neighbour; *LR,* Logistic Regression; *GloVe,* Global Vectors for Word Representation; *BERT,* Bidirectional Encoder Representations from Transformers; *Bi-LSTM,* bidirectional LSTM; *Acc,* Accuracy; *F1,* F1-score; *Pre,* Precision; *Rec,* Recall; *TN,* True Negative; *P,* Actual Positive; *N,* Actual Negative; *Adam,* root Adaptive Moment Estimation; *FP,* False Positive; *FN,* False Negative; *TP,* True Positive; *ELMo,* Deep contextualized word representations

Acknowledgements The authors are thankful to Dr. Abdul Majid, Head of Department, Psychiatry, SKIMS Medical College, Jammu & Kashmir (India) for providing consistent guidance and help.

Funding The authors are thankful to Shri Mata Vaishno Devi University for funding the research grant under TEQIP-III (Technical Education Quality Improvement Program-III).

Data availability The Twitter data used in this article for research of depression is released by (Shen et al., 2017). Derived data supporting the findings of this study are available from the corresponding author [18dcs008@smvdu.ac.in] on request.

Declarations

Conflict of interests The authors declare that they have no conflict of interest.

References

1. Agarap AF (2018) Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375. https://arxiv.org/abs/1803.08375
2. Alabdulkreem E (2021) Prediction of depressed Arab women using their tweets. Journal of Decision Systems:1–16. https://doi.org/10.1080/12460125.2020.1859745
3. Albawi S, Mohammed TA, Al-Zawi S (2017) Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET). IEEE, pp 1–6. https://doi.org/10.1109/ICEngTechnol.2017.8308186
4. Almeida H, Briand A, Meurs MJ (2017) Detecting Early Risk of Depression from Social Media User-generated Content. In CLEF (Working Notes). http://ceur-ws.org/Vol-1866/paper_127.pdf
5. AlSagri HS, Ykhlef M (2020) Machine learning-based approach for depression detection in Twitter using content and activity features. IEICE Transactions on Information and Systems 103(8):1825–1832 https://www.jstage.jst.go.jp/article/transinf/E103.D/8/E103.D_2020EDP7023/_pdf
6. Alshaer HN, Otair MA, Abualigah L, Alshinwan M, Khasawneh AM (2021) Feature selection method using improved CHI Square on Arabic text classifiers: analysis and application. Multimedia Tools and Applications 80(7):10,373–10,390. https://doi.org/10.1007/s11042-020-10,074-6
7. Arora P, Arora P (2019) Mining twitter data for depression detection. In: 2019 International Conference on Signal Processing and Communication (ICSC). IEEE, pp 186–189. https://doi.org/10.1109/ICSC45622.2019.8938353
8. Beard C, Millner AJ, Forgeard MJ, Fried EI, Hsu KJ, Treadway MT, … Björgvinsson T (2016) Network analysis of depression and anxiety symptom relationships in a psychiatric sample. Psychological Medicine 46(16):3359–3369. https://doi.org/10.1017/S0033291716002300
9. Biradar A, Totad SG (2018) Detecting Depression in Social Media Posts Using Machine Learning. In: International Conference on Recent Trends in Image Processing and Pattern Recognition. Springer, Singapore, pp 716–725. https://doi.org/10.1007/978-981-13-9187-3_64
10. Bird S, Loper E (2004) NLTK: the natural language toolkit. In: Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics, p 31

11. Birjali M, Beni-Hssane A, Erritali M (2016) A method proposed for estimating depressed feeling tendencies of social media users utilizing their data. In: International Conference on Hybrid Intelligent Systems. Springer, Cham, pp 413–420. https://doi.org/10.1007/978-3-319-52,941-7_41
12. Brahma S (2018) Improved sentence modeling using suffix bidirectional lstm. arXiv preprint arXiv: 1805.07340. https://arxiv.org/abs/1805.07340
13. Chevance A, Ravaud P, Tomlinson A, Le Berre C, Teufer B, Touboul S, … Tran VT (2020) Identifying outcomes for depression that matter to patients, informal caregivers, and health-care professionals: qualitative content analysis of a large international online survey. The Lancet Psychiatry 7(8):692–702. https://doi.org/10.1016/S2215-0366(20)30191-7
14. Chiu CY, Lane HY, Koh JL, Chen AL (2021) Multimodal depression detection on instagram considering time interval of posts. Journal of Intelligent Information Systems 56(1):25–47. https://doi.org/10.1007/s10844-020-00599-5
15. Costello C, Srivastava S, Rejaie R, Zalewski M (2021) Predicting Mental Health From Followed Accounts on Twitter. Collabra: Psychology 7(1). https://doi.org/10.1525/collabra.18731
16. Deaths and suicides in India (2015) National Crime Records Bureau. Ministry of Home Affairs. Government of India. http://www.isbtonline.com/current-affairs-details.php?id=6084&National-Suicide-Report,-A-student-commits-suicide-every-hour-in-India:-NCRB
17. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L (2009) ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, pp 248–255. https://doi.org/10.1109/CVPR.2009.5206848
18. Depression, W.H.O (2017) Other common mental disorders: global health estimates. World Health Organization, Geneva, pp 1–24 https://www.who.int/mental_health/management/depression/prevalence_global_health_estimates/en/
19. Eichstaedt JC, Smith RJ, Merchant RM, Ungar LH, Crutchley P, Preoţiuc-Pietro D, … Schwartz HA (2018) Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences 115(44):11,203–11,208. https://doi.org/10.1073/pnas.1802331115
20. Fatima I, Abbasi BUD, Khan S, Al-Saeed M, Ahmad HF, Mumtaz R (2019) Prediction of postpartum depression using machine learning techniques from social media text. Expert Systems 36(4):e12409. https://doi.org/10.1111/exsy.12409
21. Gilbert P (2007) Psychotherapy and counselling for depression. Sage
22. Guntuku SC, Schneider R, Pelullo A, Young J, Wong V, Ungar L, … Merchant R (2019) Studying expressions of loneliness in individuals using twitter: an observational study. BMJ Open 9(11):e030355 https://bmjopen.bmj.com/content/9/11/e030355.abstract
23. Hameed Z, Garcia-Zapirain B (2020) Sentiment classification using a single-layered BiLSTM model. IEEE Access 8:73,992–74,001. https://doi.org/10.1109/ACCESS.2020.2988550
24. Hiraga M (2017) Predicting depression for japanese blog text. In Proceedings of ACL 2017, Student Research Workshop (pp. 107–113). https://www.aclweb.org/anthology/P17-3018.pdf
25. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Computation 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
26. Huang, Y. C., Chiang, C. F., & Chen, A. L. (2019). Predicting Depression Tendency based on Image, Text and Behavior Data from Instagram. In DATA (pp. 32–40). https://www.scitepress.org/Papers/2019/78336/78336.pdf
27. Islam MR, Kabir MA, Ahmed A, Kamal ARM, Wang H, Ulhaq A (2018) Depression detection from social network data using machine learning techniques. Health Information Science and Systems 6(1):1–12. https://doi.org/10.1007/s13755-018-0046-0
28. Jamil Z (2017) Monitoring tweets for depression to detect at-risk users (Doctoral dissertation, Université d'Ottawa/University of Ottawa)
29. Kim K, Moon J, Oh U (2020) Analysis and Recognition of Depressive Emotion through NLP and Machine Learning. The Journal of the Convergence on Culture Technology 6(2):449–454. https://doi.org/10.17703/JCCT.2020.6.2.449
30. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. https://arxiv.org/abs/1412.6980
31. Kohrt BA, Speckman RA, Kunz RD, Baldwin JL, Upadhaya N, Acharya NR, … Worthman CM (2009) Culture in psychiatric epidemiology: using ethnography and multiple mediator models to assess the relationship of caste with depression and anxiety in Nepal. Annals of Human Biology 36(3):261–280. https://doi.org/10.1080/03014460902839194
32. Kumar R, Nagar SK, Shrivastava A (n.d.) Depression Detection Using Stacked Autoencoder From Facial Features and NLP. 10.24113/ojssports.v7i1.115

33. Leiva V, Freire A (2017) Towards suicide prevention: early detection of depression on social media. In: International Conference on Internet Science. Springer, Cham, pp 428–436. https://doi.org/10.1007/978-3-319-70,284-1_34
34. Li, Y., Mihalcea, R., & Wilson, S. R. (2018). Text-based detection and understanding of changes in mental health. In International Conference on Social Informatics (pp. 176–188). Springer Cham.. Springer, . doi: https://doi.org/10.1007/978-3-030-01159-8_17
35. Lin C, Hu P, Su H, Li S, Mei J, Zhou J, Leung H (2020) Sensemood: Depression detection on social media. In: Proceedings of the 2020 International Conference on Multimedia Retrieval, pp 407–411. https://doi.org/10.1145/3372278.3391932
36. Major Depressive Disorder among teens https://www.pewsocialtrends.org/2019/02/20/most-u-s-teens-see-anxiety-and-depression-as-a-major-problem-among-their-peers/. Accessed 10 March 2021
37. Mandelbaum A, Shalev A (2016) Word embeddings and their use in sentence classification tasks. arXiv preprint arXiv:1610.08229. https://arxiv.org/abs/1610.08229
38. Mori K, Haruno M (2021) Differential ability of network and natural language information on social media to predict interpersonal and mental health traits. Journal of Personality 89(2):228–243. https://doi.org/10.1111/jopy.12578
39. Murfi H, Siagian FL, Satria Y (2019) Topic features for machine learning-based sentiment analysis in Indonesian tweets. International Journal of Intelligent Computing and Cybernetics
40. Nadeem M (2016) Identifying depression on Twitter. arXiv preprint arXiv:1607.07384
41. Naseem U, Razzak I, Khushi M, Eklund PW, Kim J (2021) Covidsenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis. IEEE Transactions on Computational Social Systems
42. Oquendo MA, Ellis SP, Greenwald S, Malone KM, Weissman MM, Mann JJ (2001) Ethnic and sex differences in suicide rates relative to major depression in the United States. American Journal of Psychiatry 158(10):1652–1658. https://doi.org/10.1176/appi.ajp.158.10.1652
43. Orabi AH, Buddhitha P, Orabi MH, Inkpen D (2018, June) Deep learning for depression detection of twitter users. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pp 88–97 https://www.aclweb.org/anthology/W18-0609.pdf
44. Paffenbarger RS Jr, Lee IM, Leung R (1994) Physical activity and personal characteristics associated with depression and suicide in American college men. Acta Psychiatrica Scandinavica 89:16–22. https://doi.org/10.1111/j.1600-0447.1994.tb05796.x
45. Park CW, Seo DR (2018) Sentiment analysis of Twitter corpus related to artificial intelligence assistants. In: 2018 5th International Conference on Industrial Engineering and Applications (ICIEA), Singapore, pp 495–498. https://doi.org/10.1109/IEA.2018.8387151
46. Pranav KR (2018) Neural Network Based System to Detect Depression in Twitter Users via Sentiment Analysis. https://doi.org/10.1136/bmjopen-2019-030355
47. Priya A, Garg S, Tigga NP (2020) Predicting anxiety, depression and stress in modern life using machine learning algorithms. Procedia Computer Science 167:1258–1267. https://doi.org/10.1016/j.procs.2020.03.442
48. Rao G, Zhang Y, Zhang L, Cong Q, Feng Z (2020) MGL-CNN: A hierarchical posts representations model for identifying depressed individuals in online forums. IEEE Access 8:32,395–32,403 https://ieeexplore.ieee.org/abstract/document/8998086
49. Recurrent neural networks and simpleRNN layer https://keras.io/api/layers/recurrent_layers/simple_rnn/. Acessed 2 April 2021
50. Rosa RL, Schwartz GM, Ruggiero WV, Rodríguez DZ (2018) A knowledge-based recommendation system that includes sentiment analysis and deep learning. IEEE Transactions on Industrial Informatics 15(4):2124–2135. https://doi.org/10.1109/TII.2018.2867174
51. Rustagi A, Manchanda C, Sharma N, Kaushik I (2021) Depression anatomy using combinational deep neural network. In: International conference on innovative computing and communications. Springer, Singapore, pp 19–33. https://doi.org/10.1007/978-981-15-5148-2_3
52. Samuel J, Ali GGMN, Rahman MM, Esawi E, Samuel Y (2020) COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification. Information 11:314
53. Seo J, Yoo K, Choi S et al (2019) The latent learning model to derive semantic relations of words from unstructured text data in social media. Multimedia Tools and Applications 78(28):649–28,663. https://doi.org/10.1007/s11042-018-6211-2
54. Shen G, Jia J, Nie L, Feng F, Zhang C, Hu T, Zhu W (2017) Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. In IJCAI (pp. 3838–3844). http://hcsi.cs.tsinghua.edu.cn/Paper/Paper17/IJCAI17-SHENGUANGYAO.pdf
55. Shetty NP, Muniyal B, Anand A, Kumar S, Prabhu S (2020) Predicting depression using deep learning and ensemble algorithms on raw twitter data. International Journal of Electrical and Computer Engineering 10(4):3751. https://doi.org/10.11591/ijece.v10i4.pp3751-3756

56. Shrestha A, Serra E, Spezzano F (2020) Multi-modal social and psycho-linguistic embedding via recurrent neural networks to identify depressed users in online forums. Network Modeling Analysis in Health Informatics and Bioinformatics 9(1):1–11. https://doi.org/10.1007/s13721-020-0226-0
57. Shuai HH, Shen CY, Yang DN, Lan YFC, Lee WC, Philip SY, Chen MS (2018) A comprehensive study on social network mental disorders detection via online social media mining. IEEE Transactions on Knowledge and Data Engineering 30(7):1212–1225
58. Sood A, Hooda M, Dhir S, Bhatia M (2018) An initiative to identify depression using sentiment analysis: a machine learning approach. Indian J Science Technol 11(4):1–6. https://doi.org/10.17485/ijst/2018/v11i4/119594
59. Soutner D, Müller L (2013) Application of LSTM neural networks in language modelling. In: International Conference on Text, Speech and Dialogue. Springer, Berlin, Heidelberg, pp 105–112 https://link.springer.com/chapter/10.1007/978-3-642-40585-3_14
60. Stephen JJ, Prabu P (2019) Detecting the magnitude of depression in Twitter users using sentiment analysis. International Journal of Electrical and Computer Engineering 9(4):3247. https://doi.org/10.11591/ijece.v9i4.pp3247-3255
61. Suman SK, Shalu H, Agrawal LA, Agrawal A, Kadiwala J (2020). A novel sentiment analysis engine for preliminary depression status estimation on social media. arXiv preprint https://arxiv.org/pdf/2011.14280.pdf
62. Tao X, Zhou X, Zhang J, Yong J (2016) Sentiment analysis for depression detection on social networks. In: International Conference on Advanced Data Mining and Applications. Springer, Cham, pp 807–810. https://doi.org/10.1007/978-3-319-49,586-6_59
63. Tensorflow and text preprocessing https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/Tokenizer. Acessed 20 March 2021
64. Tommasel A, Diaz-Pace A, Rodriguez J M, Godoy D (2021) Capturing social media expressions during the COVID-19 pandemic in Argentina and forecasting mental health and emotions. arXiv preprint arXiv:2101.04540. https://arxiv.org/abs/2101.04540
65. Tong L, Zhang Q, Sadka A, Li L, Zhou H (2019) Inverse boosting pruning trees for depression detection on Twitter. arXiv preprint arXiv:1906.00398
66. Tong L, Liu Z, Jiang Z, Zhou F, Chen L, Lyu J, Zhang X et al (2019) Cost-sensitive Boosting Pruning Trees for depression detection on Twitter. arXiv preprint arXiv:1906.00398
67. Trotzek M, Koitka S, Friedrich CM (2018) Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Transactions on Knowledge and Data Engineering 32(3):588–601. https://doi.org/10.1109/TKDE.2018.2885515
68. Uddin, A. H., Bapery, D., & Arif, A. S. M. (2019). Depression Analysis from Social Media Data in Bangla Language using Long Short Term Memory (LSTM) Recurrent Neural Network Technique. In 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2) (pp. 1–4). IEEE. doi: https://doi.org/10.1109/IC4ME247184.2019.9036528
69. W. H. Organization "Suicide data," World Health Organization 2019. https://www.who.int/teams/mental-health-and-substance-use/data-research/suicide-data
70. Wang Y, Wang Z, Li C, Zhang Y, Wang H (2020) A Multitask Deep Learning Approach for User Depression Detection on Sina Weibo. arXiv preprint arXiv:2008.11708. https://arxiv.org/abs/2008.11708
71. Wolohan JT (2020) Estimating the effect of COVID-19 on mental health: Linguistic indicators of depression during a global pandemic. In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. (2020) https://www.aclweb.org/anthology/2020.nlpcovid19-acl.12/
72. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Dean J (2016) Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144. https://arxiv.org/abs/1609.08144
73. Wu J, Ma J, Wang Y, Wang J (2021) Understanding and Predicting the Burst of Burnout via Social Media. In: Proceedings of the ACM on Human-Computer Interaction, 4(CSCW3), pp 1–27. https://doi.org/10.1145/3434174
74. Xezonaki D, Paraskevopoulos G, Potamianos A, Narayanan S (2020). Affective Conditioning on Hierarchical Networks applied to Depression Detection from Transcribed Clinical Interviews. arXiv preprint arXiv:2006.08336. https://arxiv.org/abs/2006.08336
75. Zafar A, Chitnis S (2020) Survey of depression detection using social networking sites via data mining. In: 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, pp 88–93. https://doi.org/10.1109/Confluence47617.2020.9058189
76. Zhang Y, Lyu H, Liu Y, Zhang X, Wang Y, Luo J (2020) Monitoring Depression Trend on Twitter during the COVID-19 Pandemic. arXiv preprint arXiv:2007.00228. https://arxiv.org/abs/2007.00228

77. Zheng W, Yan L, Gou C, Wang FY (2020) Graph Attention Model Embedded With Multi-Modal Knowledge For Depression Detection. In: 2020 IEEE International Conference on Multimedia and Expo (ICME). IEEE, pp 1–6. https://doi.org/10.1109/ICME46284.2020.9102872
78. Zhou TH, Hu GL, Wang L (2019) Psychological disorder identifying method based on emotion perception over social networks. International Journal of Environmental Research and Public Health 16(6):953. https://doi.org/10.3390/ijerph16060953
79. Zogan H, Wang X, Jameel S, Xu G (2020) Depression detection with multi-modalities using a hybrid deep learning model on social media. arXiv preprint arXiv:2007.02847. https://arxiv.org/ftp/arxiv/papers/2003/2003.04763.pdf
80. Zucco C, Calabrese B, Cannataro M (2017) Sentiment analysis and affective computing for depression monitoring. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, pp 1988–1995. https://doi.org/10.1109/BIBM.2017.8217966