International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Detecting Depressive Tweets by Weighted Voting Ensemble Model of Attention Based Bi LSTM and BERT with Transfer Learning

Reseena Mol N. A.[1*], Dr. S. Veni[2]

Submitted: 09/12/2023 **Revised**: 19/01/2024 **Accepted**: 30/01/2024

Abstract: A prevalent mental illness in today's world is depression. Only a small portion of the millions of people who experience depression seek adequate medical care. Prolonged depression without effective medical treatment can lead to suicide, which makes early intervention a necessity. Nowadays, people share their inner feelings through social media platforms such as Twitter, Facebook, etc., so they can be used wisely for the early detection of depression among its users. Current approaches, despite integrating machine learning techniques, frequently struggle with challenges related to accuracy and effectively capturing exact patterns present in textual data. We suggest an ensemble of standard word embedding with Bi-LSTM and advanced language model BERT with transfer learning using weighted voting to seek improved performance by resolving the current limitations and scores an accuracy of 97.4%. The novelty lies in the comprehensive utilization of advanced techniques to process and analyze social media content, contributing to early detection efforts and augmenting mental health support initiatives.

Keywords: *Bi-LSTM,BERT,,Depression, transfer learning, twitter,word embedding*

1. Introduction

Major depressive disorder (MDD), otherwise known as depression, is a significant psychological illness that can have a footprint on many aspects of your existence [1]. Stress, anxiety, or other mental breakdowns can lead to depression. Humans with depression experience frequent feelings of desolation and hopelessness and lose enjoyment in all activities. It also affects numerous bodily needs, like hunger and sleep.

Depression turns productive hours into ignored hours, hurting the national market in terms of incompetent productivity. Depression thus impacts not just the person but also negatively impacts the nation's administration and business organisations [3].

As per the World Health Organisation (WHO), 3.8% of the global population is affected by depression [4]. Untreated depression leads to suicide; according to research, the mortality rate among depressed individuals is much higher than that of individuals without these disorders [2]. Additionally, studies show that the prevalence of depression has increased over time, but treatment has not kept pace [5].

People suffering from depression often hesitate to acknowledge it in public or seek medical assistance; this can be hazardous. Social media platforms can be of major use in these circumstances. People often share their true inner feelings through their posts on these platforms. Twitter is one of the most widely used social media platforms. Twitter platforms generate 200 billion tweets annually,averaging 6000 tweets per second, and they their data openly for public use[6]. We recommend using Twitter posts (tweets) as a medium to classify depressed users and non-depressed users, thereby allowing early intervention to be provided.

Through the textual comments and posts discussing various topics, individuals have the opportunity to convey their ideas,thoughts,perspectives and life experiences[7]. Analyzing the negative sentiment expressed in a person's social media posts can be extended to assess their level of depression. In this study we conduct experimental comparisons between attention based Bi-LSTM and BERT employing transfer learning. Additionally, we introduce a novel weighted voting ensemble method that outperforms individual models in terms of accuracy.

This paper is structured as follows. Related works in Section II; Methodology in Section III, Experimental Results described in Section IV and finally in V we conclude the paper.

2. Related Works

Twitter has been a leading microblogging site since its inception in 2006. Users share their feelings, opinions, and suggestions through tweets. Twitter has been utilised by several researchers for different analysis of sentiment [8,9]. Research is extensively done as these social media platforms provide huge support for developing machine learning and

[1] *Research Scholar,Department Of Computer Science,Karpagam Academy Of Higher Eduction,Coimbatore – 641021,India*
ORCID ID : 0000-0003-4133-8967.
[2]*Professor,Department Of Computer Science,Karpagam Academy Of Higher Education,Coimbatore-641021,India.*
ORCID ID : 0000-0002-2999-8463
** Corresponding Author Email: reseena.n.a@gmail.com*

deep learning paradigms. The advancement of mental disorder identification on social media has been significantly boosted by the utilization of deep neural networks (DNNs), encompassing convolutional networks (CNNs) and LSTMs[10][11]. Deep learning has shown remarkable efficacy in Natural Language Processing (NLP) tasks such as Text categorization and Sentiment analysis(SA). Numerous studies in the existing literature have concentrated on leveraging DL models for the analysis of user generated content and textual data. For instance, Khafaga et al.[12] explore into the critical role of social media in expressing emotions and stress by which detects the depression in user's messages. They introduce an innovative approach, the Multi Aspect Depression Detection with Hierarchical Attention Network (MDHAN) which binds deep learning to classify Twitter data and identify signs of depression. In comparison to established techniques such as CNN,SVM and MDL, the proposed models achieves an impressive accuracy of 96.86%. In [13]Bi-LSTM is identified as the best neural network with Word2vec embedding model. A hybrid deep learning approach utilizing CNN and Bi-LSTM was discussed in [14].

The word embedding extracted after pre-processing tweets was fed onto the hybrid learning framework to classify depressed and non-depressed tweets. They achieved an accuracy of 94.28% with a benchmark Twitter dataset [15].

In [16], from the experiments it is proven that by combining Lexicon based features with distributional representations offers better accuracy than pure classifier models and in [17] a new attribute sarcasm level is included in the proposed methodology. [18],[19]& [20] proves that incremental learning approach for valuation of words, pre trained word embeddings and mapping of emoticons to textual data play inevitable role in classification process. The same dataset was used by another classification procedure called DepressionNet [21]. They have tried to amalgamate user behaviour with user post-history using abstractive-extractive summarization. They made use of BERT and k-means clustering for extraction, followed by DistilBART and Bi-GRU with attention to abstraction. DepressionNet obtained an accuracy of 90.1% with nearly 7900 depressed and non-depressed tweets. Depression classification is also discussed in [22]. They have applied certain pre-processing like URL, hashtags, non-ASCII characters, and mention removal. The cleaned tweets were classified using different pre-trained BERT models and achieved an overall 92% accuracy.

In [23], after some text pre-processing, knowledge distillation methodology was used for feature extraction, accompanied by the ivis algorithm for feature reduction. The tweet classification was performed using different machine learning algorithms such as SVM, LR, GP, and QDA. A similar system was used in [24]. They used features like grammes, polarity, subjectivity, and POS tags. MNB and SVR were used to classify depression, obtaining accuracies of 78% and 79.7%, respectively.

In [25],[26] &[27] implemented ensemble models for classification process and [27] acquires an accuracy of 95% in the case of European tweets and proven that the motive behind the achievement of ensemble model is the excellent performance of individual models.

From the analysis of contributions mentioned above, the following conclusions were made a)Deep learning methods provides improved accuracy than machine learning methods for detecting depression.b)Hybrid models are extensively employed to perform superior to traditional single prediction models.c)Pre trained word embeddings have significant role in the efficiency of classifiers.d)Existing methodologies face problems in attaining higher accuracy and in effectively capturing precise patterns within the textual data. Hence to address the concerns revealed above, an enhanced ensemble model of Attention based Bi-LSTM and BERT with transfer learning is proposed and also compares the accuracies of Attention based Bi-LSTM and BERT with transfer learning models.

3. Methodology

From the literature study, it is clear that deep learning methods are better in terms of performance. Pre-processing is an inevitable procedure in most machine learning and deep learning algorithms. The block architecture of the depression prediction system we have come up with is illustrated in Fig. 1.

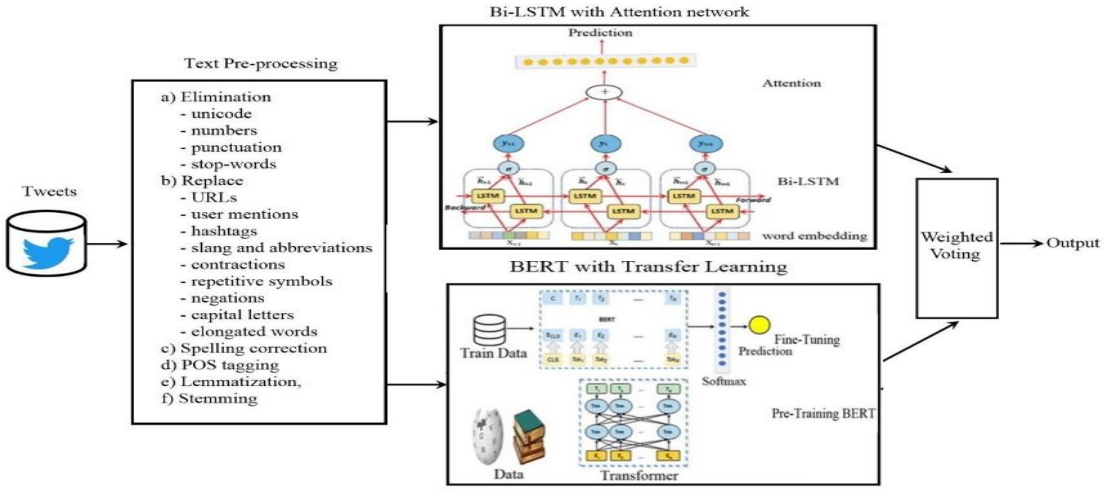

Fig. 1. Block diagram of the proposed depression prediction architecture.

3.1. Pre-processing

Tweets are not written as clear dictionary words; people tend to shorten word letters and include smileys, urls, and hashtags, which are often unwanted. So we understand the necessity of a thorough cleaning of tweets before using them further. So, we have applied 16 different Twitter text pre-processing techniques [28]. These include removing unicode strings, replacing URLs, user mentions, and hashtags, replacing slang and abbreviations, replacing contractions, removing numbers, replacing repetitions of punctuation symbols, removing punctuations, replacing negations with antonyms, handling capitalised and lowercase words, removing stop-words, replacing elongated words, spelling corrections, part of speech tagging, lemmatization, and stemming of words. Some of the example tweets before and after preprocessing along with their class are included in Table. 1.

Table 1. Tweets before and after proposed text pre-processing**.**

Class Label	**Original Text**	**Processed Text**
notdepressed	my best startup idea this week. A chrome extension that pays you for... https://t.co/hFe1x6eb9a	best start idea week crome extens day multistep
notdepressed	Wow. Someone asked Elon Musk: "what encouraging words do you have for people who want to do a startup?" His answer: If you need words of encouragement, don't do a startup. ðŸŽ¤	wow someon ask plon mu encourag word peopl want start answer need word encourag start
notdepressed	Success comes from daily impatience and decades of patience	success come daili impati decid patienc
Depressed	the real reason why you're sad? you're attached to people who have been distant with you. you're paying attention to people who ignore you. you make time for people who are "too busy" for you. you're too caring to people who are care less when it comes to you. let those people go	real reason sad attach peopl distant pay attent peopl ignor make time peopl busi dare peopl care le come let peopl go
Depressed	my biggest problem is overthinking everything	biggest problem overthink everyth
Depressed	The worst feeling is when something is killing you inside, and you have to act like you don't care.	worst feel someth kill act care

Once the pre-processing is completed, we need a powerful mechanism to correctly categorize tweets as either depressed or non-depressed. Thus, we suggest an ensemble model that combines two popular NLP-DL frameworks to obtain a better and more robust depression detection system, explained as in the following sections.

3.2. Word embedding and Bi-LSTM with attention

A popular NLP tool that represents words is word embedding. It converts corpus words into real-number vectors. The distribution of probabilities for each word appearing before or after a particular word is used to calculate these vectors. Consequently, a tweet *t* can be represented as a sequence of *n* words where each word can be mapped to a global vector. Thus, if vector $\overleftarrow{v_i}$ represents the vector of *i*-th word with a dimension of d, then:

$$t = [\overleftarrow{v_1} || \overleftarrow{v_2} || \dots || \overleftarrow{v_n}] \quad (1)$$

For acquiring word embedding, we have used the word2vec model [29]. Word2Vec creates a vector space of many dimensions so that words with similar contexts in the corpus are situated next to one another.

LSTM is a type of recurrent neural network (RNN) with numerous loops to persist information and deal with sequential data like text. To do so, LSTM replaces RNN nodes with cells. The general architecture of a cell is presented in Fig. 2. Bi-LSTM is a kind of LSTM that can process data in both directions to preserve the information.

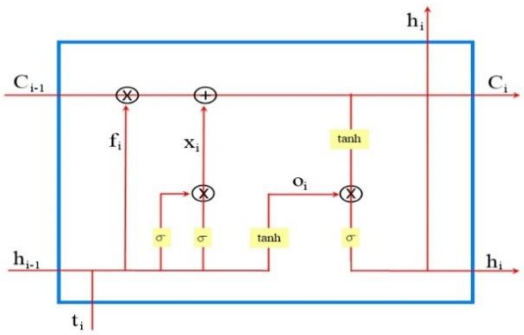

Fig. 2. Architecture of a cell in LSTM.

LSTM cell estimates the hidden state h_i at the instance i as following:

$$f_i = \sigma(W_f . [h_{i-1}, t_i] + b_f) \quad (2)$$

$$x_i = \sigma(W_x . [h_{i-1}, t_i] + b_x) \quad (3)$$

$$o_i = \sigma(W_o . [h_{i-1}, t_i] + b_o) \quad (4)$$

$$\tilde{C} = \tanh(W_c . [h_{i-1}, t_i] + b_c) \quad (5)$$

$$C_i = f_i * C_{i-1} + x_i * \tilde{C} \quad (6)$$

$$h_i = o_i * \tanh(C_i) \quad (7)$$

where σ stand for the sigmoid function, x_i is the *i*-th word vector, C_i , f_i, x_i and o_i are all gate vectors of the cell, and *W* and *b* are cell parameters. In case of Bi-LSTM, two LSTM networks processes in opposite directions. That is, a forward LSTM processes from t_1 to t_n and a backward LSTM processes from t_n to t_1.Therefore, the word feature *h* is engineered by concatenating $\overrightarrow{h}$ and $\overleftarrow{h}$, that denotes forward and backward features respectively. So *h* can be calculated as follows, where *L* is the length of a single direction LSTM.

$$h = \overrightarrow{h_i} \oplus \overleftarrow{h_i}, h_i \in R^{2L} \quad (8)$$

Incorporating attention into Bi-LSTM can boost its overall performance. Attention mechanisms are added to any deep learning network so that the network can focus on particular parts by assigning different weights to each part of the input data. Here, the weight w_i should be added to the appraised word feature h_i. So, the sentence feature *r* can be calculated as follows:

$$k_i = \tanh(W_h h_i + b_h), k_i \in [-1,1] \quad (9)$$

$$w_i = \frac{\exp(k_i)}{\sum_{p=1}^{N} \exp(k_p)}, \sum_{i=1}^{N} w_i = 1 \quad (10)$$

$$r = \sum_{i=1}^{N} w_i h_i, \; r \in R^{2L} \quad (11)$$

Where W_h and b_h corresponds to the weight and bias in the attention layer.

3.3. BERT with Transfer Learning

BERT (Bidirectional Encoder Representations from Transformers) was introduced by Google in the year 2018 [30]. It became popular in a blink as BERT outperformed most of the previously existing language models. BERT is a potent architecture that includes multiple transformers, ELMO context embedding, and a bidirectional context learner.

BERT is simply transformer based learning model. These transformers links each output to its corresponding input and establishes attention. Specifically, BERT stacks multiple transformer encoders together to generate a feature vector. Figure. 3 presents the BERT with closer view onto the architecture of the encoder. The overall structure of BERT model is illustrated in the Fig. 4.

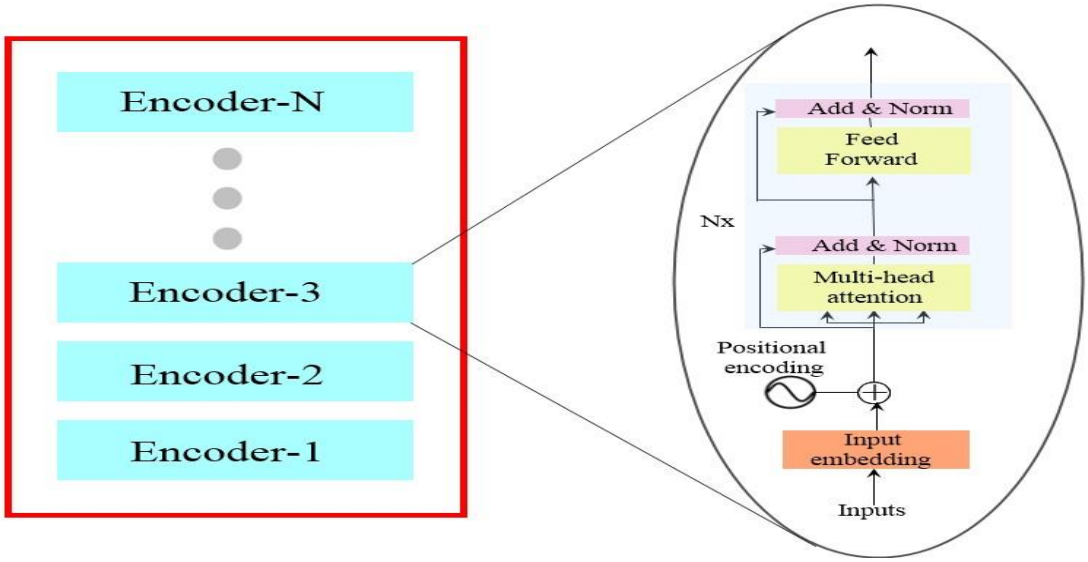

Fig. 3. Bert encoder framework.

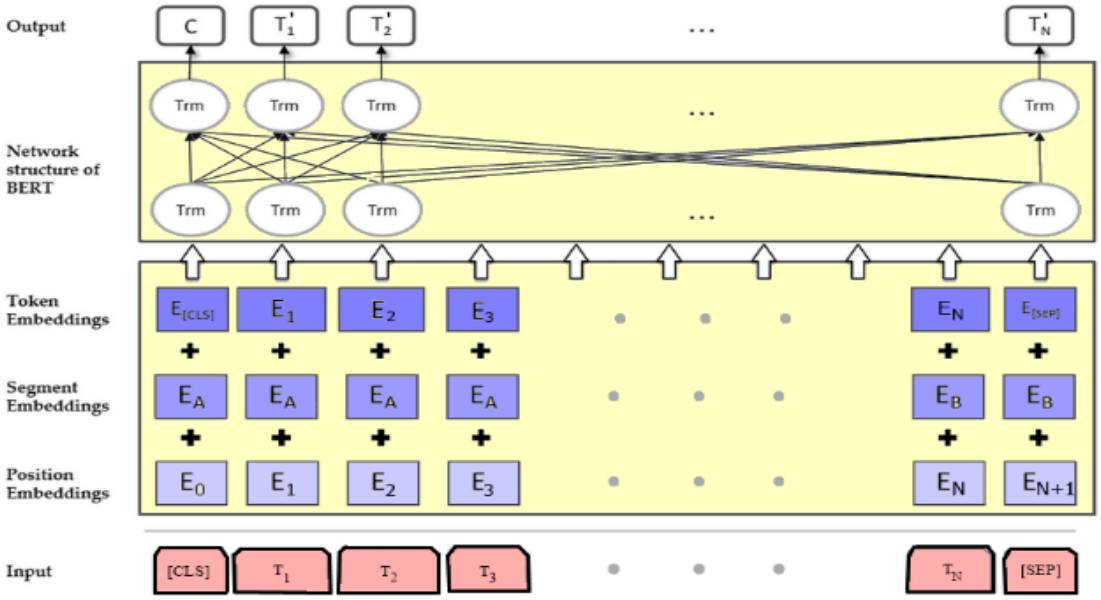

Fig. 4. Overall structure of BERT model.

With parameters that might number from 100 million to over 300 million, BERT is a large neural network architecture. Consequently, here are chances of over fitting if we try to train a BERT model from scratch using a small dataset. So, we apply fine-tuning, by which we can train the model once again on our comparatively smaller dataset.

Weighted Voting Ensemble Classifier

Ensemble classification tries to improve the overall performance of the classification by combining two or more predictive models. By using weighted voting, the classifiers are given varying weights depending on predetermined criteria, and they then cast their votes according to the weight. According to the performance accuracy of the classifier based on the training set, the weight of each classifier would be determined. The weights for each classifier would be determined using the formulas in Eq. (12).

$$W_m = \frac{A_m}{\sum A_n} \quad (12)$$

where W is the obtained weight of a model, Am is its accuracy, and An is the total combined accuracy of all classifiers utilised.

4. Experimental Results

For the successful accomplishment of experiments, we need to extract depression-related tweets from Twitter. But direct extraction of tweets from the Twitter API and successful labelling can be tiresome and may require expert involvement. Also, using a benchmark dataset is much more convenient for further study and comparisons. Therefore, we have used the Kaggle depression dataset [31]. They have provided cleaned data as well, but we have used raw data in our experiments. There are about 3496 depressed tweets and 4809 non-depressed tweets.

We have used Python to perform our experiments. We have split the entire dataset into training and test sets in the ratio 70:30. Instead of using the same train-test data for the two NLP-DL architectures, we have selected random samples. Once the pre-processing is completed, our architecture needs two different NLP-DL algorithms. For the former part, we have fitted Word2Vec to our selected training data using genism by transforming data samples into a list of lists of n-grammes. A visualisation of the feature matrix obtained is provided in Fig. 5.

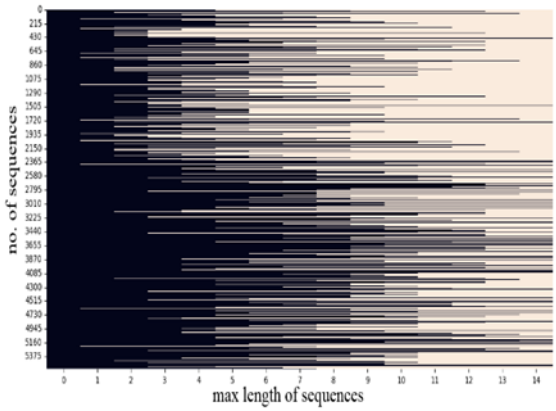

Fig. 5. Heat-map of feature matrix.

The feature matrix is passed onto the DL Bi-LSTM classifier. We have provided word sequences as input and vectors as weights to the embedding layer. An attention layer was included to estimate what parts of the real text were significant. The final dense layer would predict the probability of the text being depressed or not.

Before being tested on the actual test set, a subset of the training set was used for system validation and the performance of the system on the test set is given in Fig. 6.

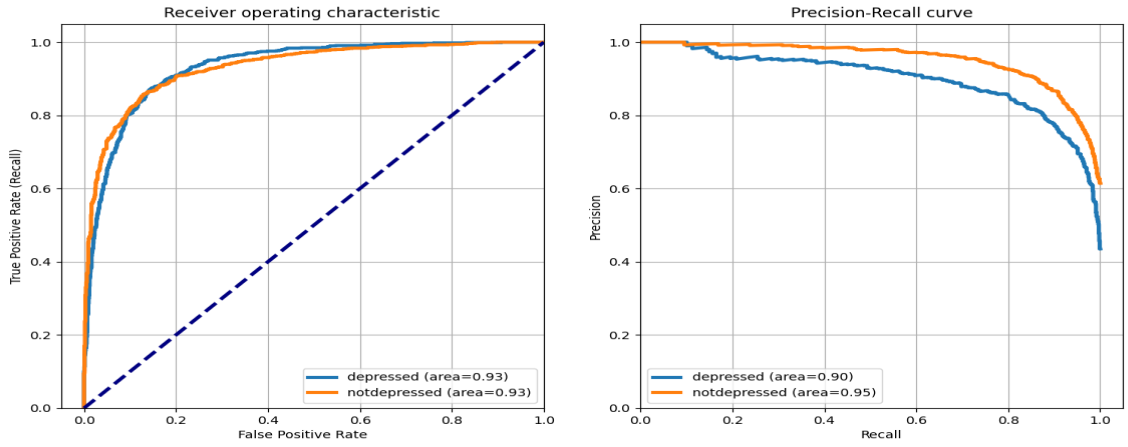

Fig. 6. ROC and Precision Recall curve of testing of Bi-LSTM + attention model.

For the second NLP-DL model, we have made use of transformers in Python. We will perform transfer learning from pre-trained DistilBERT, which is a lighter version of the BERT model. To estimate the probability of being depressed and not being depressed, we will now average the output of BERT into a single vector and add two more dense layers. Here also we have checked the accuracy and loss curves of the BERT model on our train data and additionally, the ROC and precision-recall curves are given in Fig. 7.

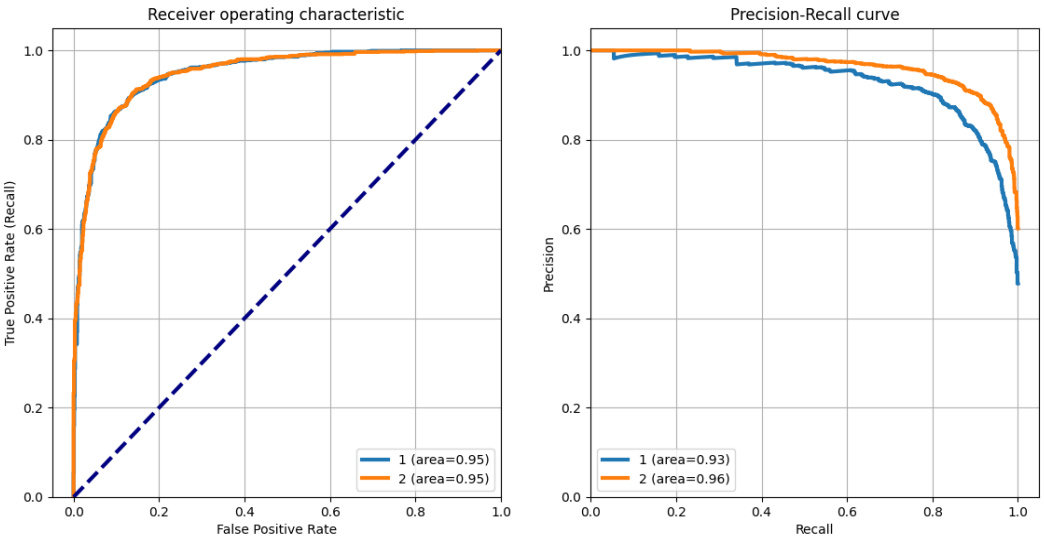

Fig. 7. ROC and Precision Recall curve of testing BERT model.

Finally, the proposed weighted averaging is performed, incorporating probabilities from both NLP and DL frameworks. A comparison table of accuracy measures and ROC-Precison Recall curve showing the performance of the proposed depression prediction system is presented in Table 2 and Figure 8.

Table 2. Accuracy measures

Model	**Accuracy**	**Precision**	**Recall**	**F1-S core**
Bi-LSTM + Attention	0.9	0.89	0.889	0.897
BERT	0.93	0.92	0.927	0.925
Weighted Voting	0.974	0.96	0.969	0.966

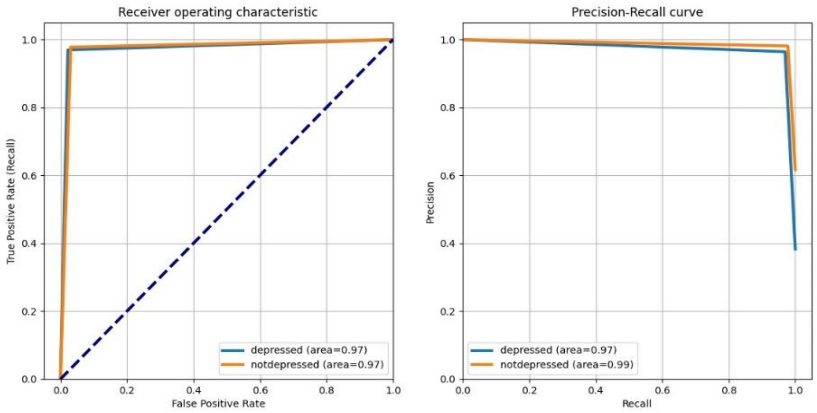

Fig. 10. ROC and Precision Recall curve of Ensemble model.

5. Conclusion

Depression is a serious mental health concern seen around the globe. Lack of treatment leads to decreased productivity or even suicide in the worst cases. Therefore, early intervention is highly recommended. People often tend to hide their mental conditions from others, which results in a lack of medical support. Social media platforms can be of greater use in these circumstances. Users often share their true feelings, knowingly or unknowingly, on their accounts, as they feel virtual space is a much safer place to open up.

In this paper, we have put forward a depression prediction system using two popular and effective NLP-DL

combinations. One is word embedding packed along Bi-LSTM with attention, and the other is language model, BERT, with transfer learning. Both model predictions were combined using weighted voting to accelerate the overall performance of the final predictions and scores an accuracy of 97.4%. Our experiment results show satisfactory performance, which can be used to provide early intervention to Twitter users in cases of depression. As technology's relentless evolution persists, our model serves as a testimony to the affirmative influence on data driven innovation can apply in enhancing the global well-being of individuals.

References

[1] M. Kerr, "Depression (major depressive disorder)," Healthline, 23-Sep-2020. [Online]. Available: https://www.healthline.com/health/ clinical-depression . [Accessed: 30-Apr-2023].

[2] L. A. Pratt, B. G. Druss, R. W. Manderscheid, and E. R. Walker, "Excess mortality due to depression and anxiety in the United States: results from a nationally representative survey," Gen. Hosp. Psychiatry, vol. 39, pp. 39–45, 2016.

[3] PricewaterhouseCoopers, "The Socio-economic impact of untreated mental illness," PwC. [Online].Available:https://www.pwc.com/m1 /en/publications/socio-economic-impact-untreated-mental-illness.html.

[4] "Depressive disorder (depression),"Who.int. [Online]. Available:https://www.who.int/news-room/fact-sheets/detail/ depression . [Accessed: 01-May-2023].

[5] R. D. Goodwin, L. C. Dierker, M. Wu, S. Galea, C. W. Hoven, and A. H. Weinberger, "Trends in U.s. depression prevalence from 2015 to 2020: The widening treatment gap," Am. J. Prev. Med., vol. 63, no. 5, pp. 726–733, 2022.

[6] Clement, J.: Twitter: number of monthly active user 2010–2019(2019),https://www.statista.com /statistics/282087/number-of-monthly-active-twitter-users/

[7] Benamara, Farah and Moriceau, Véronique and Mothe,Josiane and Ramiandrisoa, Faneva and He, Zhaolong "Automatic Detection of Depressive Users in Social Media."(2018) In: Conférence francophone en Recherche d'Informationet Applications (CORIA), 16 May 2018 - 18 May 2018 (Rennes, France).

[8] K. Sailunaz and R. Alhajj, "Emotion and sentiment analysis from Twitter text," J. Comput. Sci., vol. 36, no. 101003, p. 101003, 2019.

[9] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, "Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers-A study to show how popularity is affecting accuracy in social media," Appl. Soft Comput., vol. 97, no. 106754, p. 106754, 2020.

[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation 9, 8 (1997), 1735–1780.

[11] Yong Chen, Bin Zhou, Weina Zhang, Wenjie Gong, and Guangfu Sun. 2018. Sentiment Analysis Based on Deep Learning and Its Application in Screening for Perinatal Depression. In DSC. IEEE, 451–456.

[12] Khafaga, D. S., Auvdaiappan, M., Deepa, K., Abouhawwash, M., & Karim, F. K. (2023). Deep Learning for Depression Detection Using Twitter Data. INTELLIGENT AUTOMATION AND SOFT COMPUTING, 36(2), 1301-1313.

[13] M. R. Khan, S. Z. Rizvi, A. Yasin, and M. Ali, "Depression Analysis of Social Media Activists Using the Gated Architecture Bi-LSTM," 2021 Int. Conf. Cyber Warf. Secur. ICCWS 2021 - Proc., no. 2014, pp. 76–81, 2021, doi: 10.1109/ICCWS53234.2021.9703014.

[14] H. Kour and M. K. Gupta, "An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM," Multimed. Tools Appl., vol. 81, no. 17, pp. 23649–23685, 2022.

[15] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. 2017. Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution.. In IJCAI. 3838–3844.

[16] S. Munoz and C. A. Iglesias, "A text classification approach to detect psychological stress combining a lexicon-based feature framework with distributional representations," Inf. Process. Manag., vol. 59, no. 5, p. 103011, 2022, doi: 10.1016/j.ipm.2022.103011.

[17] P. KVTKN and T. Ramakrishnudu, "A novel method for detecting psychological stress at tweet level using neighborhood tweets," J. King Saud Univ. - Comput. Inf. Sci., no. xxxx, 2021, doi: 10.1016/j.jksuci.2021.08.015.

[18] J. S. L. Figuerêdo, A. L. L. M. Maia, and R. T. Calumby, "Early depression detection in social media based on deep learning and underlying emotions," Online Soc. Networks Media, vol. 31, no. August 2021, p. 100225, 2022, doi: 10.1016/j.osnem.2022.100225.

[19] S. G. Burdisso, M. Errecalde, and M. Montes-y-Gómez, "A text classification framework for simple

and effective early depression detection over social media streams," Expert Syst. Appl., vol. 133, pp. 182–197, 2019, doi: 10.1016/j.eswa.2019.05.023.

[20] K. Elshakankery and M. F. Ahmed, "HILATSA: A hybrid Incremental learning approach for Arabic tweets sentiment analysis," Egypt. Informatics J., vol. 20, no. 3, pp. 163–171, 2019, doi: 10.1016/j.eij.2019.03.002.

[21] H. Zogan, I. Razzak, S. Jameel, and G. Xu, "DepressionNet: A novel summarization boosted deep framework for depression detection on social media," arXiv [cs.LG], 2021.

[22] M. Rizwan, M. F. Mushtaq, U. Akram, A. Mehmood, I. Ashraf, and B. Sahelices, "Depression classification from tweets using small deep transfer learning language models," IEEE Access, vol. 10, pp. 129176–129189, 2022.

[23] J. Pool-Cen, H. Carlos-Martínez, G. Hernández-Chan, and O. Sánchez-Siordia, "Detection of depression-related tweets in Mexico using crosslingual schemes and knowledge distillation," Healthcare (Basel), vol. 11, no. 7, 2023.

[24] P. Arora and P. Arora, "Mining twitter data for depression detection," in 2019 International Conference on Signal Processing and Communication (ICSC), 2019.

[25] S. J. Malla and A. P.J.A., "COVID-19 outbreak: "An ensemble pre-trained deep learning model for detecting informative tweets," Appl. Soft Comput., vol. 107, p. 107495, 2021, doi: 10.1016/j.asoc.2021.107495.

[26] Prakash, K. Agarwal, S. Shekhar, T. Mutreja, and P. S. Chakraborty, "An ensemble learning approach for the detection of depression and mental illness over twitter data," Proc. 2021 8th Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2021, pp. 565–570, 2021, doi: 10.1109/INDIACom51348.2021.00100.

[27] D. Sunitha, R. K. Patra, N. V. Babu, A. Suresh, and S. C. Gupta, "Twitter sentiment analysis using ensemble based deep learning model towards COVID-19 in India and European countries," Pattern Recognit. Lett., vol. 158, pp. 164–170, 2022, doi: 10.1016/j.patrec.2022.04.027.

[28] S. Symeonidis, D. Effrosynidis, and A. Arampatzis, "A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis," Expert Syst. Appl., vol. 110, pp. 298–310, 2018.

[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional Transformers for language understanding," 2018.

[31] H. K. Cho, "Twitter Depression Dataset." 09-Aug-2021.