Depression Detection from Social Media Comments Using Deep Learning

SYED MD. AHNAF HASAN, MD. TANVIR HOSSAIN, and ANNAJIAT ALIM RASEL,
BRAC University, Bangladesh

Depression is one of the most prevalent and serious psychological disorders. If left untreated for a long time, it can lead to serious consequences. Therefore, early intervention can be very helpful in this regard. Nowadays, people often express themselves freely on social media, which makes these platforms a very good means for detecting depression among users. Various deep learning models can be used to automate the process of detecting depression from text data as they have proven very effective in text classification tasks. In this work, we have analyzed the performances of deep learning models having RNN layers and BERT base uncased which is a pretrained transformer model on the Reddit depression dataset available on Kaggle. Results show that the BERT model outperforms the RNN-based deep learning models with accuracy, precision, recall, and f1 score values of 98.02%, 98.06%, 98.02%, and 98.02% respectively. Finally, we analyze the predictions made by the models using the LIME XAI model to gain further insight into the classification mechanism. Implementation of XAI models alongside the deep learning models can assert faith in the predictions made.

CCS Concepts: • **Computing methodologies → Natural language processing**.

Additional Key Words and Phrases: Fine Tuning, Bert Base Uncased, RNN, GRU, LSTM, LIME, XAI

1 INTRODUCTION

Depression is described as a common and serious medical condition that affects a person's emotions, cognition, and behaviors. It's affecting millions of individuals worldwide and has broad impacts on people's health, family relations, work competence, and whole human and social life. Despite its prevalence and negative consequences, depression is often undiagnosed and misdiagnosed. The traditional identification of depressive symptoms has been based primarily on basic methods, including investigations based on clinical interviews and questionnaire surveys performed in person by medical experts [19]. Although these measures have significantly contributed to defining depression as a psychiatric disorder, they are non-standardized and entail an individual assessment, which, in turn, might result in the underestimation of reliable depressive markers [18]. In recent years, an increasing number of studies have explored the use of machine learning and artificial intelligence to improve the detection and characterization of depression based on text data, particularly social media due to the vast sources of user-generated content [7]. People are more likely to share their thoughts, feelings, and experiences more freely online. Machine learning models can then analyze language patterns, sentiment expressions, and linguistic cues in the text to infer mental states and detect potential depression markers [13]. Multiple studies have indicated the performance of machine learning algorithms and models regarding depression detection from text data. Advancements have been witnessed regarding the models' performance and deployment strategies [17]. For example, Shah et al. 2020 proposed a hybrid model that incorporates BiLSTM with word embeddings and metadata features for the classification of depression in social media posts [16]. Also, Rizwan et al. 2022 have worked on the transformer language models' performance in quantifying the intensity of depression in twitter posts [12]. More studies indicate the increased interest and potential for the machine learning models to revolutionize the depression detection and intervention approaches.

Authors' Contact Information: Syed Md. Ahnaf Hasan, ahnafhasan335@gmail.com; Md. Tanvir Hossain, tanvir.eece.mist@gmail.com; Annajiat Alim Rasel, annajiat@gmail.com
, BRAC University, Dhaka, Bangladesh.

In this research work, we evaluate the performance of RNN models and the pretrained transformer model, BERT Base Uncased in identifying depressive user comments collected from Reddit. Furthermore, we utilize LIME, an XAI model to gain insights on how the predictions were made by the black box models.

2 RELATED WORKS

There have been several advancements in detecting depression from text data. Shah et al. in 2020 suggested a hybrid model of detecting depression based on BiLSTM, word embeddings, and metadata features of social media posts [16]. The most effective combination of Word2Vec embeddings and BiLSTM demonstrates an F1 score of 0.81. While the approach provided accurate classifications, it discloses the issue of timely depression detection and calls more studies for closing this gap in depression detection time. Samanvitha et al. in 2021 used the CLEF 2020 dataset and the following models: Naive Bayes, Logistic Regression, Random Forest, and SVM to detect depression from posts in social media [14]. The model with the highest accuracy is the Naïve Bayes classifier had the highest accuracy prerequisites, 83%, which is 2% - 4% higher compared to other models. Zahoor et al. in 2022 have reviewed sentiment analysis and category classification of restaurant reviews based on the different algorithms for Naïve Bayes, Logistic Regression, SVM, and Random Forest [20]. Thus, the results of the dataset obtained from "SWOT's Guide to KARACHI's Restaurants Cafes Dhabas HBFE & Takeouts" showed that Random Forest has the highest accuracy of 95%. Lestandy et al. in 2023 have studied the effect of word embedding dimensions on depression data classification with BiLSTM [9]. They used the Kaggle Reddit Depression Dataset and GloVe and Word2Vec embeddings with BiLSTM. Word2Vec achieved the best performance with 96.22% accuracy when trained with 500 dimensions with BiLSTM, better than GLoVe. Rizwan et al. in 2022 examined the possibility of detecting depression from social media texts using small transformer-based language models [12]. Through the application of the Twitter dataset labelled with VADER and TextBlob, depression intensity classification for models such as Electra Small Generator and Albert Base V2 was conducted. The former outperformed all models by obtaining the highest F1 score of 89%. AlSagri et al. in 2020 utilized machine learning tools for depression detection among Twitter users, with specific features produced by the network behavior and tweets extraction [3]. Accuracy of SVM is higher compared with other models, although all the metrics are lower than 80%. Data collection was assisted by the Twitter Search API and annotated manually. Borba de Souza et al. in 2022 used DAC Stacking in their study to classify depression, anxiety and their comorbidity in the SMHD Dataset [4] . They used LSTM, CNN and a Hybrid LSTM-CNN architecture and achieved impressive results yielding f-measures near 0.79 in both depression and anxiety cases. Sanga et al. in 2023 determined depression detection in a digital framework with textual analysis [15]. Attention-enabled deep learning models were trained, including an LSTM, GRU, AlBERT, DistilBERT, CNN, and Attention, on a wide range of Reddit and Kaggle datasets. The deployed models demonstrated 1.52% accuracy in the mean, with DistilBERT-BiLSTM achieving up to 94.93% accuracy and indicating AUC score of 0.9501. Akter et al. in 2023 presented the development of a novel cyberbullying detection model from a dataset obtained during the TRAC-2 Workshop [2]. The dataset consists of 25,000 comments gathered in three languages: English, Bengali, and Hindi. Several methods were used to implement the model: LSTM, BiLSTM, LSTM-Autoencoder, Word2vec, BERT, and GPT-2. Achieved results differ significantly, with proposed models' raw English reaching up to 99% accuracy, semi-noisy Bangla obtaining 95% accuracy, and noisy English 92% accuracy. Flores et al. in 2023 presented DeepScreen through the use of temporal facial landmark features from the DAIC-WOZ dataset, incorporating GRU-D and BRITS. they achieved a significant leap of 13.4% in depression screening F1 score, which was 0.85 [6]. Akter et al. in 2022 discussed the issue of a growing amount of violent content on social media. Author used Hindi, Bangla, and English

Table 1. Dataset Instances

clean_text	is_depression
i dont even deserve to live	1
sooo sick of the snow ughh	0

datasets and dealt with the problem of data scarcity in several ways, including most importantly the use of machine translation [1]. Different deep learning methods, such as LSTM, BiLSTM, BERT, and GPT-2, are then used, and the results indicate that BERT is the best performing model; BERT records an accuracy of 0.79 when used for English data, while BERT Multilingual records.72 and .69 on Bangla and Hindi datasets respectively.

3 METHODOLOGY

The workflow of this research work is shown in Fig. 1. At first the data is split into train, validation and test sets. The splits for train, validation and test are 70%, 15% and 15% respectively. Later the tokenization is performed on the splits. Two main types of models are used in this work: RNN models and BERT Base Uncased which is a pretrained transformer model. For the RNN based models, a tokanizer object created from the tokenizer class from keras is fitted on the train split. Later the train, test and validation splits are transformed to integer sequences via the tokenzier object. Later padding and truncation operations applied on the tokenized sequences. For the BERT Base Uncased model, we use the BERT tokenizer available in the Hugging face transfomers library. The BERT tokenizer object in this case, handles tokenization, padding and truncation. The maximum length chosen to be kept after padding and truncation is 256. The newly generated train sequences are used to train the deep learning models, while the validation set is evaluated simultaneously to stop the model from training if it starts overfitting via early stopping. Consequently, the test split is evaluated using the trained models. Finally, we use LIME to interpret the local predictions of the models.

3.1 Dataset

The "Depression: Reddit Dataset (Cleaned)" dataset [8], [9] is used for training and evaluating the deep learning models. The dataset is available in the link: https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned/code. It contains 7731 labeled user comments curated from various comments in Reddit. The label ’1’ is used to denote depressed text while the label ’0’ indicates non depressed text. From Fig. 2 we can see that the dataset is nearly balanced. The dataset consists of 3900 non depressive texts and 3831 depressive texts. Two instances of each class of texts are shown in Table. 1. The dataset contains text data which are cleaned, as there are no punctuation marks, symbols, etc., furthermore the text is in lower case.

3.2 Tokenization

Deep learning models require the input which is fed to them to be numerical in nature. Therefore, the text that will be fed to the deep learning models, need to be converted to numerical format. A way to do this is tokenization. In this process, an integer value is assigned to a token most often words in the text. For the RNN models, the a tokenizer object created using Keras library is fitted on the train split. Later the train, test and validations splits are tokenized using the tokenizer object. In Fig. 3. we can see the distribution of lengths of each sequence instance. We find that the maximum sequence length among all the instances is 4239. However, most of the instances have

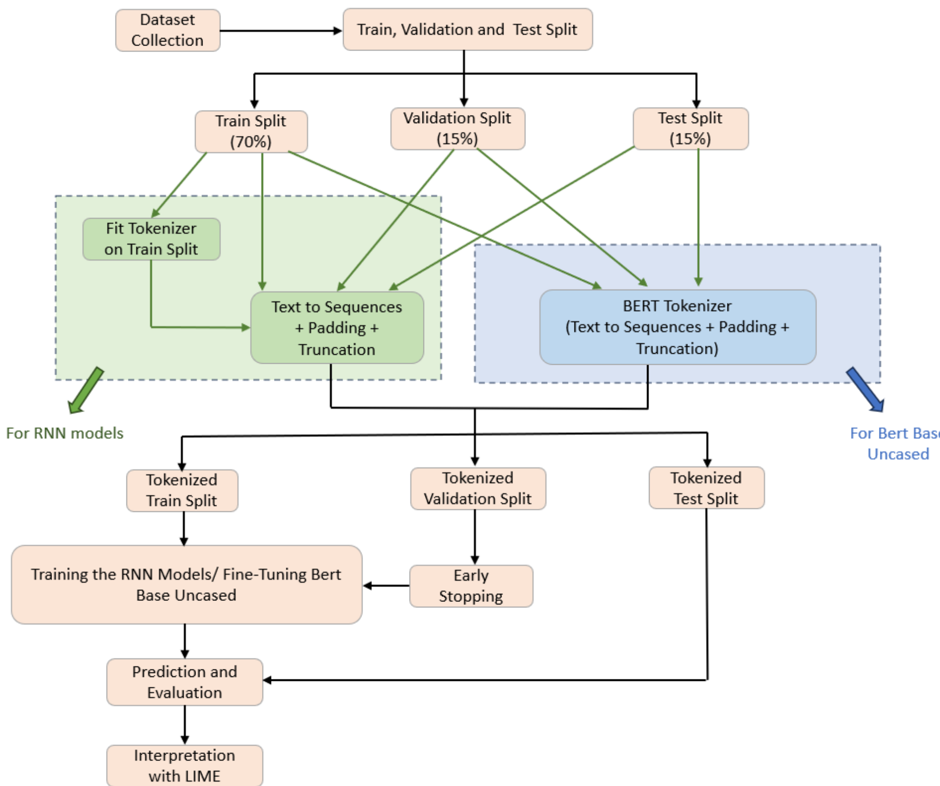

Fig. 1. Methodology

Fig. 2. Dataset Distribution

a sequence which fall in the range of 0-250. So, we choose 256 has the maximum length of the sequences. Any shorter sequence will be zero-padded and longer sequences will be truncated. For BERT Base Uncased, we use the BERT Tokenizer from Hugging face transformers library which handles tokenization, padding and truncation.

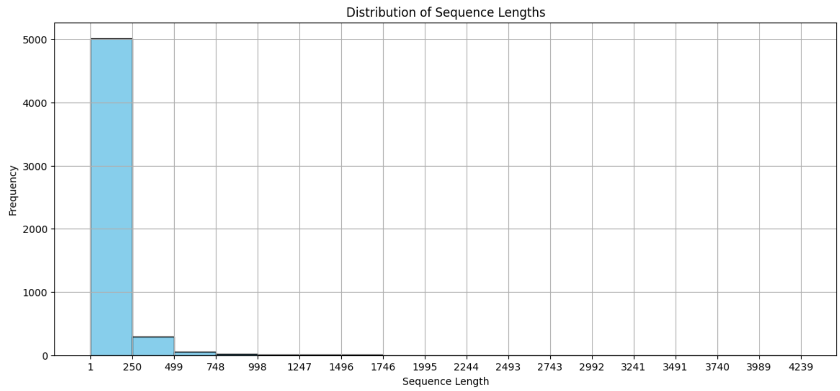

Fig. 3. Distribution of Sequence Length

3.3 GloVe

GloVe: Global Vectors for Word Representation is a technique to generate word embeddings [10]. Deep learning models work with numerical values, i.e. data they are trained on and the data they evaluate all have to be converted to numerical format. However, only converting words in a text to a single numerical value or to a one-hot encoding vector is not enough, as the meaning of a word is not effectively captured in this way. Word embeddings are vectors assigned to a word with a very high dimension. Each of these dimensions carry an aspect of a word's meaning. GloVe is an efficient technique to generate these embeddings. Word embeddings generated by GloVe can be found in various dimensions such as 25, 50, 100, 200, etc. In this project we have used the 100 dimensional word embeddings generated by GloVe.

3.4 Deep Learning Models

3.4.1 RNN Models. In this work, performance of six RNN models are evaluated. Each of the models use either unidirectional or bidirectional variations of the RNN layers. Three RNN layers are used: Simple RNN, LSTM and GRU, which are available in Keras library. Each of these layers will have 32 units. The architecture of these models in general is shown in Fig. 4. The first hidden layer will be the embedding layer. The GloVe pretrained embeddings will be used here to transformer the incoming tokens to a higher dimensional embeddding. The chosen dimesion of embeddings is 100. The next hidden layer contains either of the mentioned RNN layers, which will have 32 units. The Simple RNN layer has an incoming input and a feedback path carrying its hidden state value. It uses the hidden state value from previous time step and current time step input to generate output. It works consecutively as sequences of data like in the case of text is passed to it. The LSTM is a variation of RNN that uses separate paths for propagating long term memory and hidden state. It changes the values of the memory state based on the importance of the current input and the previous time step hidden state. It uses 3 gates: input gate, forget gate and output gates to accomplish its task. It solves the vanishing gradient problem of basic RNN. GRU is another variation of RNN that is similar to LSTM, but has one less gate than LSTM, making it faster. It does not use a separate path to propagate long term memory state, rather it uses the hidden state for that task. The final layer in the RNN models is the output layer that will have a single dense unit with sigmoid activation function to generate the probabilities of the classes.

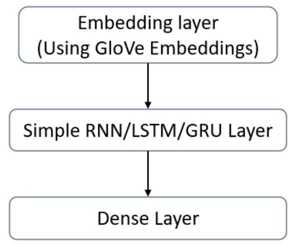

Fig. 4. Architecture of RNN models

3.4.2 Bert Base Uncased model. Transformers are attention based deep learning models with a large number of parameters which can effectively capture information from sequences by using parallel computations. Moreover, the attention mechanism can be applied multiple times on the same input to capture more context and information; this mechanism is called multi-head attention mechanism. Due to the large number of parameters of popular transformer models, it is computationally very expensive to train them from scratch. Therefore, in most cases for downstream task such ours, a pretrained transformer model is used to be fine-tuned on a dataset. BERT: Bidirectional Encoder Representations from Transformers, is a transformer based language representation model [5] which was trained on a huge corpus via self-supervised learning.The Bert Base (Uncased) model used in this work is a pretrained BERT model available in hugging face transformers library. This model has got 110 million parameters, which helps it to effectively capture information from sequence data.

3.5 LIME

LIME: Local Interpretable Model-agnostic Explanations is an explainable AI (XAI) model which is used to explain predictions made by deep learning models [11]. Deep learning models are inherently black box in nature due to their complex structure. Therefore, it is not clear how the model made a prediction or came to a conclusion for a particular instance of data. XAI models are used to explain why the deep learning model drew a particular conclusion on an instance based on the feature of the instance. LIME is used to generate local explanations for a particular instance of data by generating perturbed data points in the local neighborhood of the data point and uses a linearly interpretable model which best approximates the prediction of the black box model on that particular data.

3.6 Hyperparameters

The hyperparameters used in this research, have been summarized in Table 2. The number of the units of RNN layers will be 32. For, RNN models, the size of the word embeddings will be 100, the vocabulary size will be 10,000. The maximum length of tokenized sequences after padding and truncating will be 256. Early stopping patience will be 3; if the accuracy of validation set while training does not improve for next consecutive 3 epochs, the training will stop. The number of initial epochs to train each model is chosen to be 50, however, training would stop early if the model was found to overfit.

3.7 Performance Metrics

Performance of all the models were evaluated using four metrics: accuracy, precision, recall and F1 score. Accuracy is the ratio of the number of correctly predicted instances to the total number of

Table 2. Summary of Hyperparameters

Parameter Name	Value
Simple RNN/LSTM/GRU units	32
Word Embedding dimension (For RNN models)	100
Vocabulary Size of Tokenizer (For RNN models)	10000
Max length of sequences (After padding and truncation)	256
Early Stopping Patience	3
Number of epochs	50

instances. It works very good for balanced data but is misleading for imbalanced data. Precision is the ratio of true positives to the summation of true positives and false positives, i.e. the percent of predictions which are correct out of all the predictions made by the model for a particular class. Recall on the other hand is the ratio of true positives to the summation of true positives and false negatives which indicates the fraction of the actual labels that the model predicted correctly for a class. F1 score is a metric that is a balance between precision and recall; it is the harmonic mean of the precision and recall. Precision, recall and F1 score provide important insights to a model's performance especially for imbalanced data. Macro average, micro average or weighted average can be taken for precision, recall and F1 to combine the scores of all the classes.

4 EXPERIMENT RESULTS AND ANALYSIS

4.1 Performance Analysis of the Models

In this research, six RNN models were trained and a BERT Base Uncased pretrained transformer model was fine-tuned using the training split of the dataset. The models were simultaneously trained and evaluated on the validation split. The initial value of epochs was set to 50, however, early stopping was used via monitoring the validation set's accuracy. The patience for early stopping was selected to be 3; if the validation set's accuracy did not improve in the next 3 consecutive epochs during training, the training would be stopped to prevent overfitting. The model with the weights for which the highest validation accuracy was found was retained. After the training was complete, the models were evaluated on the test set on the four performance metrics: accuracy, precision, recall, and f1 score.

The evaluation scores for the models on the test set are shown in Table 3. It is evident from the table, that Bert Base Uncased is the best-performing model on the test set with the highest scores across all the performance metrics. The accuracy, precision, recall, and F1 scores obtained by BERT Base are 98.02%, 98.06%, 98.02%, and 98.02% respectively. The Bidirectional GRU model and the Bidirectional LSTM models come in second and third in terms of performance scores. This shows that the bidirectionality of the gated RNN models can capture more context from the text data. The unidirectional LSTM model comes in fourth in terms of performance which is comparable to the unidirectional GRU model. The Simple RNN model is the least performing model, due to its simplicity and vanishing gradient problem, though the bidirectional implementation significantly improves its performance.

Fig. 5 shows the confusion matrix for the BERT Base Uncased model's evaluation on the test set. The pretrained transformed model is able to classify 565 instances of class 1 and 572 instances of class 0 correctly, while it classifies 20 positive instances as false negatives and 3 negative instances as false positives. The summary from the confusion matrices of all the models can be seen in Table 4. The Bidirectional GRU has competitive results in classifying the positive class compared to the Bert model, though it misclassifies 4 more negative classes than the latter model. Bidirectional

Table 3. Performance analysis of the Deep Learning models on Test Set

Model	Accuracy	Precision	Recall	F1
Simple RNN model	80.52%	83.81%	80.52%	80.06%
Bidirectional Simple RNN model	91.72%	92.31%	91.72%	91.70%
LSTM model	96.72%	96.78%	96.72%	96.72%
Bidirectional LSTM model	97.67%	97.70%	97.67%	97.67%
GRU model	96.47%	96.51%	96.47%	96.47%
Bidirectional GRU model	97.84%	97.85%	97.84%	97.84%
Bert Base Uncased	**98.02%**	**98.06%**	**98.02%**	**98.02%**

Table 4. Summary of Confusion Matrices

Model	TP	TN	FP	FN
Simple RNN model	382	552	23	203
Bidirectional Simple RNN model	503	561	14	82
LSTM model	549	569	6	36
Bidirectional LSTM model	554	571	4	32
GRU model	536	567	8	49
Bidirectional GRU model	562	568	7	23
Bert Base Uncased	**565**	**572**	**3**	**20**

LSTM is slightly better at classifying the non depressive texts than the bidirectional GRU however it performs slightly worse in predicting depressive text as it misclassifies nine more instances. The unidirectional LSTM model on the other hand is more efficient in handling depressive text and non depressive text than the unidirectional GRU model. The findings from the performance metrics and confusion metrics assert BERT base uncased as the best performing model.

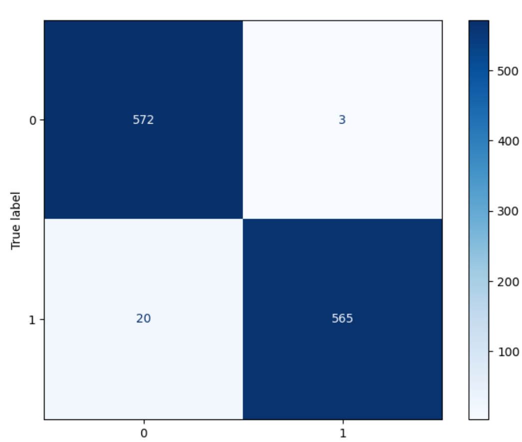

Fig. 5. Confusion Matrix of Bert Base Uncased model

4.2 LIME Explanation

LIME was used to gain an insight on the prediction mechanism of the black box model used for detecting depression from the text data. Fig. 6 shows LIME explanation of 952^{nd} instance in the test split which is a depressive text and was correctly classified by the BERT Base Uncased model. The LIME interpretation shows the probability for each of the classes for the text instance. Since the probability of the 'Depressed' class is higher than the 'Not Depressed' class, this instance is assigned the 'Depressed' class. The interpretation also shows which words in the text contribute to how much in predicting each of the classes. It assigns weights to each of the tokens in the sentence for this purpose. The interpretation provides clarity to the prediction mechanism of the deep learning model, as which words are emphasized for predicting which class are indicated by assigning weights and corresponding color. The LIME interpretation of the prediction can be used to validate the prediction made by the model and can hence improve faith in the model's estimation.

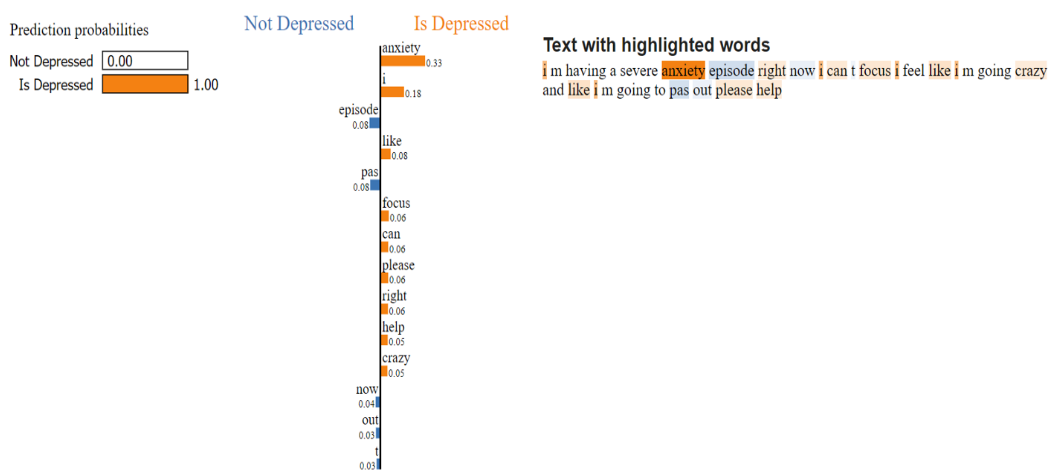

Fig. 6. LIME Interpretation of a depressive text classified by BERT Base Uncased.

5 CONCLUSION

Detecting depression among individuals can be a difficult task, however, the use of deep learning in this regard can make this task simpler with automation and high efficiency. Text data from social media can be effectively used with deep learning techniques to identify depression among users. In this research, we have trained six RNN based deep learning models and fine tuned the BERT Base Uncased model, and analyzed their performance on the Reddit Depression dataset from Kaggle. The BERT Base Uncased model was observed to obtain the highest scores across all the performance metrics for this task. Finally, we used LIME to interpret the predictions, to gain an insight into the black box models' prediction mechanism. LIME interpretation shows the words that were focused on for each class while making the estimation by the deep learning model. This work encourages the use of pretrained transformer models in downstream tasks such as detecting depressions from text as well as the usage of XAI models in adding transparency to the predictions made. This work can further be improved in many ways such as, by using a larger and more diverse dataset, analyzing the performances of more pretrained transformers and other XAI models may also be used to compare the interpretation of the predictions.

ACKNOWLEDGMENTS

We would like to express our gratitude to Annajiat Alim Rasel sir for his supervision and guidance on this project.

REFERENCES

[1] Mst Shapna Akter, Hossain Shahriar, Nova Ahmed, and Alfredo Cuzzocrea. 2022. Deep learning approach for classifying the aggressive comments on social media: Machine translated data vs real life data. In *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 5646–5655.

[2] Mst Shapna Akter, Hossain Shahriar, Alfredo Cuzzocrea, Fan Wu, and Juanjose Rodriguez-Cardenas. 2023. A Trustable LSTM-Autoencoder Network for Cyberbullying Detection on Social Media Using Synthetic Data. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 5418–5427.

[3] Hatoon S AlSagri and Mourad Ykhlef. 2020. Machine learning-based approach for depression detection in twitter using content and activity features. *IEICE Transactions on Information and Systems* 103, 8 (2020), 1825–1832.

[4] Vanessa Borba de Souza, Jeferson Campos Nobre, and Karin Becker. 2022. DAC stacking: A deep learning ensemble to classify anxiety, depression, and their comorbidity from Reddit texts. *IEEE Journal of Biomedical and Health Informatics* 26, 7 (2022), 3303–3311.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Ricardo Flores, Avantika Shrestha, and Elke A Rundensteiner. 2023. DeepScreen: Boosting Depression Screening Performance with an Auxiliary Task. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 5213–5222.

[7] Manju Lata Joshi and Nehal Kanoongo. 2022. Depression detection using emotional artificial intelligence and machine learning: A closer review. *Materials Today: Proceedings* 58 (2022), 217–226.

[8] Manuel Kanahuati-Ceballos and Leonardo J Valdivia. 2024. Detection of depressive comments on social media using RNN, LSTM, and random forest: comparison and optimization. *Social Network Analysis and Mining* 14, 1 (2024), 1–16.

[9] Merinda Lestandy et al. 2023. Exploring the Impact of Word Embedding Dimensions on Depression Data Classification Using BiLSTM Model. *Procedia Computer Science* 227 (2023), 298–306.

[10] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[12] Muhammad Rizwan, Muhammad Faheem Mushtaq, Urooj Akram, Arif Mehmood, Imran Ashraf, and Benjamín Sahelices. 2022. Depression classification from tweets using small deep transfer learning language models. *IEEE Access* 10 (2022), 129176–129189.

[13] Ramin Safa, Peyman Bayat, and Leila Moghtader. 2022. Automatic detection of depression symptoms in twitter using multimodal analysis. *The Journal of Supercomputing* 78, 4 (2022), 4709–4744.

[14] S Samanvitha, AR Bindiya, Shreya Sudhanva, and BS Mahanand. 2021. Naïve Bayes Classifier for depression detection using text data. In *2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*. IEEE, 418–421.

[15] Prabhav Sanga, Jaskaran Singh, and Prakhar Priyadarshi. 2023. Unmasking Depression via Attention-modulated Text Analysis using Deep Learning. In *2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI)*. IEEE, 335–340.

[16] Faisal Muhammad Shah, Farzad Ahmed, Sajib Kumar Saha Joy, Sifat Ahmed, Samir Sadek, Rimon Shil, and Md Hasanul Kabir. 2020. Early depression detection from social network using deep learning techniques. In *2020 IEEE region 10 symposium (TENSYMP)*. IEEE, 823–826.

[17] Matthew Squires, Xiaohui Tao, Soman Elangovan, Raj Gururajan, Xujuan Zhou, U Rajendra Acharya, and Yuefeng Li. 2023. Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment. *Brain Informatics* 10, 1 (2023), 10.

[18] Nicola Tanner. 2023. *An integrated mindfulness and physical exercise intervention: Evaluating the effects on school staff.* Ph. D. Dissertation. Loughborough University.

[19] Md Zia Uddin, Kim Kristoffer Dysthe, Asbjørn Følstad, and Petter Bae Brandtzaeg. 2022. Deep learning for prediction of depressive symptoms in a large textual dataset. *Neural Computing and Applications* 34, 1 (2022), 721–744.

[20] Kanwal Zahoor, Narmeen Zakaria Bawany, and Tehreem Qamar. [n. d.]. Evaluating text classification with explainable artificial intelligence. *Int J Artif Intell ISSN* 2252, 8938 ([n. d.]), 8938.