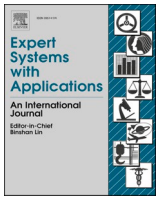
Expert Systems With Applications 249 (2024) 123834

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



An attention-based CNN-BiLSTM model for depression detection on social media text

Joel Philip Thekkekara [a,*], Sira Yongchareon [a], Veronica Liesaputra [b]

[a] School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand
[b] Department of Computer Sciences, University of Otago, Dunedin, New Zealand

ARTICLE INFO

Keywords:
Depression detection
Natural language processing
Deep learning for social media analytics

ABSTRACT

Depression has long been described as a common mental health disorder and a disease with a set of diagnostic criteria that influences the affected individuals' feelings and behavior. The prevalence of Internet use has augmented people's openness to share their experiences and struggles, including mental health disorders on social media thus researchers have tried developing classification models for depression detection using various machine learning and deep learning techniques. In this research, we propose a deep learning architecture with an attention mechanism on CNN-BiLSTM (CBA) and provide a comparative analysis to benchmark well-known deep learning models using the public dataset namely CLEF2017. We found that along with F1 score, precision and recall it is also vital to consider the Area under the curve - Receiver operating characteristic curve (AUC-ROC) and Mathews Correlation Coefficient (MCC) metrics for evaluating depression classification models since the MCC considers all the four values of a confusion matrix. Based on our experiments, the CBA model outperforms the existing state of the art model with an overall accuracy of 96.71% and scores of 0.85 and 0.77 for AUC-ROC and MCC, respectively.

1. Introduction

Depression, which is considered as a condition impacting the overall mental wellbeing of an individual, caught the attention of multiple researchers from (Migliore, Alicata, & Ayala, 1995) onwards, who desired to apply computational techniques to complement the existing clinical methodologies for detecting depression. Depression has long been described as a common mental health disorder, with an established set of diagnostic criteria (Health, Health, Excellence, Society, & Psychiatrists, 2011). According to the WHO report (Roser, 2020), 264 million individuals, which equal to 3.4 percent of the global population (2.7 % males and 4.1 % females), are projected to suffer from depression. Almost half of those at risk live in Western Pacific and South-East Asia, witnessing large manifestations in China and India. Depression is still underdiagnosed and not properly handled in many countries which eventually leads to a serious sense of self-perception and suicide, at its peak. Many affected people choose not to obtain sufficient professional support to avoid the social stigma of depression. It was found that though the manifestations of mental illnesses such as depression would lead to isolation and social withdrawal (Baron & Kenny, 1986), social media platforms are increasingly providing an effective scope for affected individuals to connect with individuals exhibiting similar manifestations, share similar experiences, and provide mutual support. Exploiting these conclusions, social media's communities can be used to contest stigma, spike need for professional help-seeking, and provide direct online help to those suffering from mental illness. A comparative study (De Choudhury, Gamon, Counts, & Horvitz, 2013), had also observed that millennial individuals tend to use online resources more. They tend to seek more health-related information due to the stigmatization of illnesses such as depression and for easily communicating their illness on social media with people suffering alike. More recently, the COVID-19 outbreak in Wuhan, China expounded in depth the impact of social media exposure (SME) and how it augmented the high prevalence of panic and mental health problems (Gao).

There has been considerable studies that has been undertaken to understand the mental workload by using EEG based analysis (Chakladar, Dey, Roy, & Dogra, 2020). However, early research on the relation between language and depression has theoretically focused primarily on the linguistic features that are manifested in 'depressed language', such as negatively valence words, e.g. as used in (Beck, Rush, Shaw, & Emery, 1979) and recurring use of first-person pronouns in (Pyszczynski, Holt, & Greenberg, 1987). It was found that there was a strong correlation

* Corresponding author.
E-mail addresses: jthekkek@aut.ac.nz (J. Philip Thekkekara), sira.yongchareon@aut.ac.nz (S. Yongchareon), veronica.liesaputra@otago.ac.nz (V. Liesaputra).

https://doi.org/10.1016/j.eswa.2024.123834
Received 25 August 2020; Received in revised form 19 January 2024; Accepted 21 March 2024
Available online 22 March 2024

between such language and depressed individuals. This led to finding the connection between linguistic output and depression (Coppersmith, Dredze, Harman, Hollingshead, & Mitchell, 2015) which analyzed the social media messages. However, the studies didn't propose any empirical observation based on an efficient computational model which took into consideration the linguistic dimension nor benchmarked the dataset.

Although Machine Learning (ML) techniques despite their promise as an approach to studying depression, the existing machine learning methods are semiautomatic as their methods require feature extraction which is manual as well as include time-consuming and labor-intensive selection methods. Thus, a classification method that is capable to directly learn from raw social media data and extract features automatically (Goodfellow, Bengio, & Courville, 2016) would be more suitable for constructing an automated social media analysis method for depression discrimination which is focused on this study. In this paper, we propose an attention-based CNN-BiLSTM to emphasize those linguistic features to help us detect depression on the CLEF2017 early risk prediction (eRisk) dataset which is a standardized corpus collection curated only for depression study (Losada, Crestani, & Parapar, 2017).

To the best of our knowledge, as of till now no studies have benchmarked deep learning models on the CLEF2017 early risk prediction (eRisk) dataset (Losada et al., 2017) and the primary focus in all depression detection related studies have revolved only on increasing the accuracy, F1-score and the precision of the model—the latter which we found, could not be used reliably in the imbalanced dataset for the depression case study, as these scores took into consideration only three values of the confusion matrix whereas the fourth value i.e. True Negative (TN) was often not considered. In the clinical field, True Negative Rate (Sensitivity and Specificity) and True Positive Rate are used as the main metrics. However, in data science, True Positive Rate (TPR) and Positive Predictive Value (PPV) known as Recall and Precision respectively (Cohen et al., 2016) are the most commonly used metrics.

Traditionally, ML systems used in non-medical domains are primarily evaluated on Recall and Precision only (Cohen et al., 2016). This is because regardless of which model is being applied to a non-clinical or non-medical problem, the value of True Negatives does not matter. However, it is imperative that the True Negatives also be involved for statistical analysis when being applied to medical situations. The absence leading to the deduction of F1, Recall and Precision score. This leads to ambiguity in clinically approved negative findings (Cohen et al., 2016). It is imperative to note that neither Recall / Sensitivity nor Precision considers the True Negatives. Thus it does not truly provide a broad picture of the actual 'accuracy'. In this study, we found it was therefore vital to consider the TN, since classifying a healthy individual as depressed was an important metric in our study, to be penalized. We thus aim to contribute to the following:

- Benchmarking for the first-time, deep learning models for depression detection on the CLEF2017 dataset, and.
- Proposing an attention-based CNN-BiLSTM model that outperforms the existing deep learning models for depression detection using the CLEF2017 dataset which takes into consideration the linguistic features, quantified using Accuracy, F1, Precision, AUC-ROC and MCC.

2. Related work and background

Multi-stage strategies have been recommended for screening an individual with depression, such as (Mitchell, Rao, & Vaze, 2011; Nease & Malouin, 2003). This was majorly owing to the comparatively high false-positive rate and low sensitivity (around 50 %) associated with non-psychiatric physician's assessments (Cepoiu et al., 2008) and short screening inventories, which were recurring problems. An additional step in the existing strategy for mental health screening could well be achieved by social media-based screening. Further, we also learned, the computational model could address the aforesaid issues which could be well complemented using artificial intelligence techniques for the task (Guntuku, Yaden, Kern, Ungar, & Eichstaedt, 2017). Utilizing user-created content (UGC) accurately may help clinicians better decide wellness levels of an individuals' psychological state. (Aldarwish & Ahmad, 2017) observed that the Social Network Sites' (SNS) utilization is growing these days, predominantly due to the extensive social media exposure. This could enable clinical clients to express their sentiments, interests, and opportunity to be offered a day by day schedule (Islam et al., 2018).

Approaches for social media data collection along with the related information about the mental health of users' have been studied. They are classified into 4 major types: a *self-reported survey* (sharing of social media data and undertaking depression survey) and *self-declared, post annotation and forum membership* (which includes keywords searching from public posts to identify users sharing their mental health diagnosis, mental illness related forums' user language, or collecting any keywords for annotation of mental illness in public posts). The approaches using public data (*self-declared, post annotation, and forum membership*) have the advantage that much larger samples can be collected cheaper and faster than surveys (Guntuku et al., 2017). Our study has therefore focused on the public depression data CLEF2017.

Traditional ML techniques that were used for depression detection are as illustrated in (AlSagri & Ykhlef, 2020). There are a few individual studies that applied traditional machine learning techniques such as logistic regression (LR), Naïve Bayes (NB), Decision Tree (DT), and Support Vector Machine (SVM) (Lai, Xu, Liu, & Zhao, 2015) for separating healthy and depressed individuals. In (Pirina & Çöltekin, 2018), classification methods were applied to various subreddits (Reddit corpora) to identify depression from social media data. After extracting features (using simple character and word bag-of-n-gram), they applied linear SVMs and showed that the F1 score was higher when Reddit data was used as opposed to the Twitter platform. In (Hasan, Rundensteiner, & Agu, 2018) the authors with a motivation to analyze the effectiveness of using the hashtags as their labels developed the Emotex and EmotexStream systems using machine learning algorithms like NB, SVM, and DT. To understand the hidden linguistic variables of mental disorders (Ríssola, Bahrainian, & Crestani, 2019), aimed to study the evolution of the mental state of a user over a while to discern the patterns of depression and the control subjects. One of the most challenging yet very important steps is the strategy adopted to label the text data. They built a training set to define features characterizing signals of depression by using negative sentiment polarity score and semantic similarity with regards to depression topic. Further, they also built their dictionary by grouping terms and concepts related to depression (78 depression-related words). This study was able to achieve an F1 score of up to 83.87 % by analyzing the temporal dimension of the individual. However, owing to the difficulty in extracting intricate meanings hidden in social media data/ posts, often the traditional classification methods may not successfully accomplish the classification well [21].

The study in (Li & Chau, 2018) was claimed as the first study for detecting depression and emotional stress on social media text by proposing a domain-specific deep learning model using the LSTM network. Deep learning, which can also be referred as representation-learning method automatically learns from raw data to provided respective representation [25]. Deep learning attempts to represent high-dimensional data which is often considered to be in an abstract level. This is achieved by continuously learning multiple levels of representation. This is further accelerated with the help of abundance of data such as videos, language and images but often not thoroughly analyzed by traditional machine learning research. Deep learning can easily process such data. Achievements and findings reported in [26] [27] (Burdisso, Errecalde, & Montes-y-Gómez, 2019) expounds this. Deep learning, per sè, has now been considered one of the powerful tools for NLP – with Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) being utilized on a wider scale. Deep learning models

have shown promising performance in various text analysis tasks, such as machine translation (Tadesse, Lin, Xu, & Yang, 2019) and sentiment analysis [28] which has attempted to address the early risk detection (ERD) problems using the SS3 framework, which stands for Sequential S3 (Smoothness, Significance, and Sanction). The approach was initially intended to be used as a document classification problem solving framework, since it is flexible. The entire framework was primarily divided into two main phases. The first phase involved dividing into multiple blocks the given input. The blocks, each of them, were then further divided into smaller chunks until the requisite words were obtained. In the second phase, they have been applied to every word confidence vectors. To generate the confidence vectors for the next level, reduction is done using a summary operator. This process continues until for the whole input a single confidence vector is generated. Finally, based on the result of this single confidence vector value the actual classification is performed. However, this framework is primarily applicable on document classification problem and depends primarily on the confidence vectors. The CBA model which our study proposes, tries to overcome this by applying attention mechanism instead of having confidence vector by adding weights to specific words of interest. High availability of computing power and data, brain-inspired deep learning architectures were explored to detect depression using linguistic characteristics. People like to express their thoughts, opinions, feelings, etc. but when they become victims of depression, the contents and context of their messages are partly/fully driven by their disorder [23]. Early research on the relation between language and depression has been mostly been in theory, by focusing on the linguistic features that are manifested in 'depressed language', such as finding the negative valence words, e.g. words here [7] and frequent use of first-person pronouns here (Pyszczynski et al., 1987). However, in (Tadesse et al., 2019) linguistic dimensions were analyzed using computational models specifically for finding 1st person singular (I, me, mine), personal pronoun (I, them, her), negations (no, not, never); Psychological processes like social processes (buddy, mate), affective processes (happy, cry, hate), cognitive processes (think, know, always), personal concerns like work, money, death were found to have a great correlation with depression. Thus, to find the connection between linguistic output and depression (Coppersmith et al., 2015), analyzed the social media messages written by individuals affected with depression

Since deep learning models require huge data, analyzing crowd emotion on social networks such as Twitter could be of great potential in a multitude of applications (De Choudhury et al., 2013; Park, Cha, & Cha, 2012; Guthier, Alharthi, Abaalkhail, & El Saddik, 2014; Resch, Summa, Zeile, & Strube, 2016). The imbalanced dataset has always been a real-world problem in medical datasets (Mazurowski et al., 2008). To deal with imbalanced data distributions (Cong et al., 2018) deployed XGBoost (an optimized distributed gradient boosting library which is a Gradient Boosting framework extension) in their Reddit Self-reported Depression Diagnosis dataset and further applied Attention – BiLSTM, to achieve precision, recall and F1-score of 0.69, 0.53, 0.60 respectively. Further, they compared this architecture with other deep learning models like LSTM, Attention-LSTM, Attention Bi-LSTM, and X-A-BiLSTM. By leveraging the characteristic of the BiLSTM network to store information in the forward and backward directions and formulating context vector by focusing on important words using attention mechanism, the authors provide a better performance over other model.

Feature extraction plays a vital role in the decision-making process of the human mind and likewise in deep learning architectures. With this aim, inspired by computer vision filed, CNN was used to extract the local features like n-gram features at different points in the sentence and pooling operation helped in dimensionality reduction. CNN is incapable of capturing the sequential correlations (temporal information) apart from other disadvantages such as data requirements leading to overfitting and underfitting, parameter tuning requirements, and computation expenses (Severyn & Moschitti, 2015). To extract the context of the given sentence, the BiLSTM network was applied to the feature labels extracted from the convolutional and pooling operation. This fusion of CNN and Bi-LSTM squeezed out more knowledge from the raw data (enhancing the semantic understanding) thus enabling more learning in the neural networks. Though not specifically implemented for the depression case study, this architecture (Liu & Guo, 2019) outperforms various existing text-based classification methods and was also successful in producing better results than BiLSTM recording accuracy of 97 %. This shows the importance of the feature extraction process towards better learning for classification tasks in deep neural networks.

This study (Dinkel, Wu, & Yu, 2019) has leveraged the advantages of deep learning models by achieving an F1 score of 0.87, surpassing all the state-of-the-art results. They revealed an association between soft sounds which were short yet interpersonal such as 'um', possibly signifying that to detect depression, researchers should focus largely on the intricate behavioral aspects of text rather than on content by the analysis of pre-trained sentence embeddings (Devlin, Chang, Lee, & Toutanova, 2018; Peters et al., 2018). Their BiLSTM model with attention pooling using word embeddings achieved an F1 score of 0.87. A new modified metric of early risk detection error (ERDE) score is proposed on an ensemble approach using CNN on word embeddings and the user-level linguistic metadata (Trotzek, Koitka, & Friedrich, 2018). The ERDE measure considers both the time taken by the model to make the decision and the correctness of the decision. A deep analysis amongst the various word embeddings with CNN was carried according to $ERDE_5$, $ERDR_{50}$, F1, Precision, Recall and the proposed metric $ERDE_o$% and $F_{latency}$ measure (defined median of posts read by the system before the positive case prediction is done) (Trotzek et al., 2018). Using the metadata features from the Linguistic Inquiry and Word Count (LIWC) alone yields promising results on the eRisk dataset used here (Trotzek et al., 2018). In summary, this research majorly highlighted on benefits of incorporating linguistic metadata features and the proposed metric into early depression detection on the CLEF2017 eRisk pilot task. Thus, we found that the hybrid model of CNN-BiLSTM with attention mechanism (to add weights on words of linguistic importance) would help us to build a more robust framework that could achieve higher performance by analyzing the text features.

We found that no benchmarking using deep learning models was done on the CLEF2017 dataset. Also, in all deep learning models of depression detection over social media, we found F1 score, precision and recall are the most common metrics apart from accuracy taken largely into consideration; however, as discussed we found in our study that these metrics did not consider all four values of the confusion matrix whereas AUC-ROC and MCC could which has been further explained in Section 4. Since we do not consider the lag time of our prediction in this study, it is inappropriate to use ERDE and $F_{latency}$ score. Thus, we benchmark deep learning models for depression detection to find the best performing model and also to define precise metrics for evaluating the CLEF2017 dataset by considering all four values of the confusion matrix. In other words, we decided to not rely only on F1, precision and accuracy scores but also to employ AUC-ROC and Mathews correlation coefficient (MCC) which has been considered to have a more reliable statistical rate as it generates the score including all the four values of confusion matrix instead of three [41].

3. Architectural model

This section describes the proposed neural network model, as illustrated in Fig. 1, for separating individuals of healthy and depressed groups.

3.1. CNN-BiLSTM attention (CBA) model

CNN-BiLSTM with attention mechanism has achieved remarkable performance in classifying positive/negative reviews (Liu & Guo, 2019). We hypothesize that the local feature selection and higher representation capability of the CNN-BiLSTM model can be conjoined with the

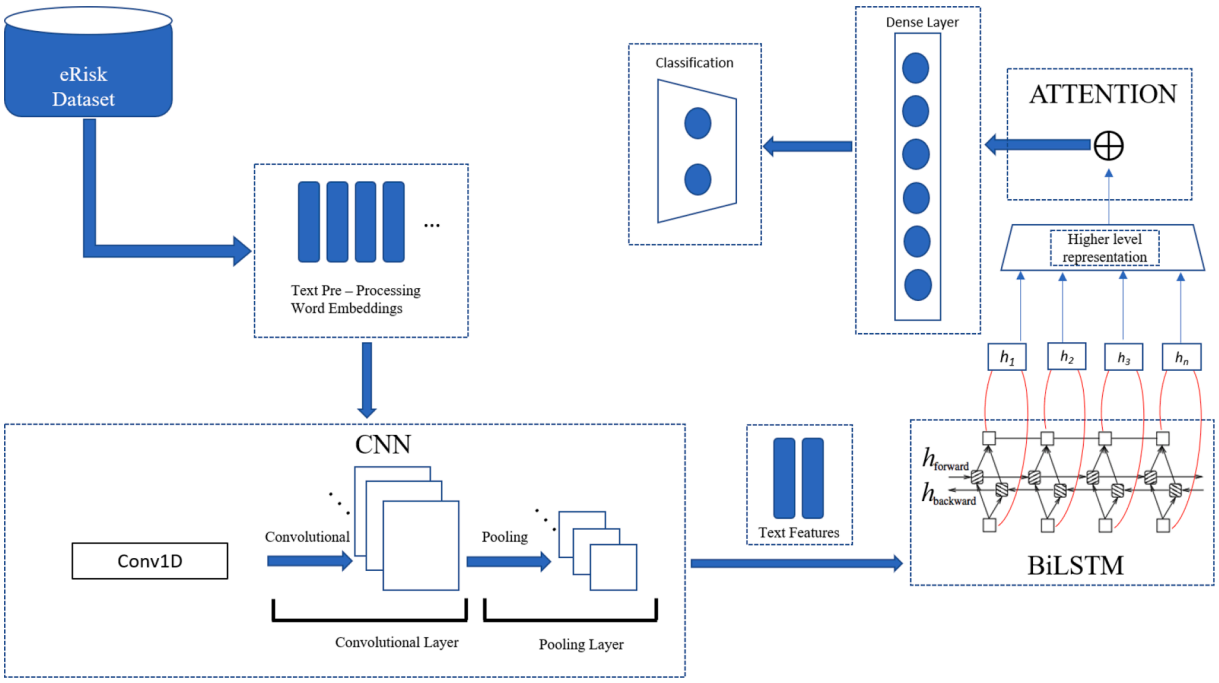


Fig. 1. CNN-BiLSTM-ATTN (CBA) Architecture for depression classification.

attention mechanism (CBA) for better understanding the linguistic dimension of our depression detection case study. CNN, as described in (Orabi, Buddhitha, Orabi, & Inkpen, 2018) learns from sliding w-grams in a given input sequence d having entries e_1, e_2, …, e_t. Early methods were largely dependent on keyword extraction to find features (Hu & Wu, 2006).

BiLSTM proved to be able to task sequential modeling better than LSTM owing to its rich ability to access both the historical and future contexts. BiLSTM used the combination of backward layer and the forward hidden layer whereas LSTM only exploited the historical context.

However, to obtain better results in text data it was crucial to focus on the significant information alone. It was found that to highlight the information from the contextual information, the attention mechanism could set different weights for words of importance (Bahdanau, Cho, & Bengio, 2014). In (Tadesse et al., 2019) Linguistic dimensions like personal pronoun (I, them, her), 1st person singular (I, me, mine), negations (no, not, never); Psychological processes like social processes (buddy, mate), affective processes (happy, cry, hate), cognitive processes (think, know, always); Personal concerns like work, money, death were found to have a great correlation with depression. The combination of BiLSTM and attention mechanism was thus assumed to further improve model performance for depression detection, helping us attain the best of both the worlds. The CBA model follows the attention model as described in the work of (Yang et al., 2016). Though Attention based CNN-LSTM model has been implemented in research studies such as script identification (Bhunia et al., 2019), our CBA model is the first model to be applied in the depression detection domain. As word sequence matter, we employ a one-dimensional convolutional layer on the sequence of length which has been pre-processed and has word embeddings of dimension 100. To obtain the feature map, we add a bias term and apply the ReLu activation function producing feature maps. This is followed by 1-max pooling with pool size 3. Bidirectional LSTM (Graves & Schmidhuber, 2005) with 128 hidden units having two parameters: dropout = 0.2 and recurrent_dropout = 0.2 is used to access both the historical and future context representations. In CBA, attention mechanism (Luong, Pham, & Manning, 2015) is employed to emphasize specifically the extracted information from the BiLSTM's backward hidden layer and forward hidden layer thereafter applies weights on words of importance on linguistic dimensions as described earlier. Finally, to classify the processed context information, a SoftMax classifier is used.

This model has the ability to capture both the sentence semantics globally and the feature of phrases locally. The attention mechanism is designed so that the network learns and captures insights from sentences for categorizing the groups. We construct a sentence representation by building representations of words and then generating a context vector. The context dependency of the importance of words is very high, i.e. the same word could be important in a context which is altogether different. To consider this sensitivity, our model includes the application of attention mechanisms as proposed in (Bahdanau et al., 2014; Lai et al., 2015). This is applied at the word level – which enables our model to pay more or less attention to specific words at an individual level in a sentence representation. Implementing attention served two benefits, i.e. resulted in better performance, as well as provided insight into words that impacted the decision making of the classification. The attention network encompasses of a word sequence encoder and a word-level attention layer. The sequence encoder involves convolutional, max-pooling, and BiLSTM by summarizing information from both directions for words to produce annotations of words and thereby incorporating the contextual information.

Given a sentence with words w_{in}, n ε[0, N], we first through an embedding matrix W_e embed the words to vectors, for i entries having input h_{in}, we have $x_i = W_e w_i$. After 1D convolution and 1-max pooling, the hidden state representation is passed to BiLSTM to produce structured representations. BiLSTM contains one forward and one backward LSTM reading sentences s_i from w_{i1} to w_{iN} and from w_{iN} to w_{i1} respectively producing two hidden states $\overrightarrow{h_{in}}$ and $\overleftarrow{h_{in}}$. h_{in} is the concatenation of these hidden encoder outputs that summarizes the contextual data of the whole sentence. Word-level attention: Every word in a sentence may not necessarily contribute alike to embody the sentence meaning. Hence, the attention mechanism will help in identifying and summarizing informative words that are crucial for understanding the meaning of a given sentence from a sentence vector.

First, we generate the hidden representation u_{in} of the encoder outputs h_{out} by computing one-layer MLP as shown from Eqs. (1) to (3).

$$u_{in} = dot\_product\,(h_{in},\ self.w) \tag{1}$$

$$u_{in} + = b \text{where } b \text{ is the biased term} \tag{2}$$

$$u_{in} = k.tanh(u_{in}) \tag{3}$$

$$\alpha_{in} = \frac{\exp(u_{in}^T u_w)}{\Sigma_t \exp(u_{in}^T u_w)} \tag{4}$$

Then, we measure the importance of the word as the similarity of u_{in} with a word-level context vector u_w and to obtain attention weights α_{in} through the softmax function as calculated in Eq. (4).

$$s_i = \sum_t \alpha_{in} h_{out} \tag{5}$$

Using Eq. (5), we then compute the sentence vector s_i as a weighted sum of the encoded sequence with the attention weights α_{in}. The final output from the attention network is given to the output layer (with SoftMax applied) for classification.

4. Experiments and evaluation

This section gives an overall summary of the dataset used, the set up required for the various models experimented along with its primary characteristics, and the evaluation criteria used. Considering the challenges faced in the medical field to classify imbalanced medical datasets, our study specifically chose to keep the CLEF2017 dataset as it is i.e. imbalanced and analyze the performance of various deep learning models along with the proposed architecture.

4.1. Experimental setup

Table 1 describes the stacked-layer and parameters applied to each layer of the CBA model. A filter length of 5 has been used with stride as 1 in the convolutional and max-pooling operation. The input to the network consists of 188,836 train samples, 47,209 validation samples, and 59,012 test samples with 250 as sequence length which was fine-tuned as done in (Orabi et al., 2018).

The preprocessing steps include converting texts to lower case, removing digits from the text, removing bad symbols in the text, Keras tokenization (Tokenizer API) for each word. The Out of Vocabulary (OOV) words are not considered in this study. Since the models in our experiments did not have multiple different input sources nor produce multiple outputs or have models that share layers, we chose to use the Sequential API provided by Keras which is a linear method of stacking layers for all the deep learning models experimented in this paper.

The deep learning models in this architecture have used the ReLu activation function for the hidden states and SoftMax for the output layer. The loss function plays one of the most crucial roles in a deep learning model how well the output layer of a network is connected with the rest of the network. In our experiments, since binary classification has been used, we have chosen binary cross-entropy for calculating the loss function.

Table 1
Parameters Fixed In Each Layer of CBA.

Layers	Layer (type)	Output Shape	Param
1	embedding 1 (Embedding)	(250, 100)	5,000,000
2	Spatial dropout1d 1 (Spatial)	(250, 100)	0
3	Conv1d 1 (Conv1D) Kernel size = 5, Stride 1	(246, 128)	64,128
4	Max poolin 1d 1 (MaxPooling1D) Window size = 3, Stride = 1	(82, 128)	0
5	Conv1d 2 (Conv1D)	(78, 128)	82,048
6	Max pooling 1d 2 (MaxPooling1D)	(26, 128)	0
7	Conv1d 3 (Conv1D)	(22, 128)	82,048
8	Max pooling1d 3 (MaxPooling1D)	(7, 128)	0
9	bidirectional 1 (Bidirectional)	(7, 256)	263,168
10	Attention with context 1 (Attention)	(256)	25,700
11	dense 1 (Dense)	(96)	24,672
12	Dense 1 (Dense) [Output Layer]	(2)	194

4.2. Dataset and setup

4.2.1. Dataset details

Owing to the large corpus collection and vast research done in this dataset, we chose the dataset described in CLEF eRisk 2016 (Losada & Crestani, 2016) and then published as part of the conference - CLEF2017 (Losada et al., 2017) for the eRisk early signs of depression detection. It encompasses a total of 135 depressed and 752 control group posts and comments chronologically ordered. To anonymize users, usernames has been replaced with ID such as "train_subject_1". Each user has a maximum of 2000 messages collected due to the API limits and the assumption that some users might have posted very rarely as described in Table 2.

For every user in the dataset, the collection of each writer has been systematically separated into ten different chunks. The first having the oldest 10 % messages, the second containing the second oldest 10 %, so forth (Losada et al., 2017). Each message consists of title, text or both, URL, and timestamp (UTC). Since the dataset authors recorded the timestamp of different users based on a common timestamp (UTC), the authors mentioned it would be misleading to compare users based on the time that each user posted as their precise time zone is not truly known, which means they could live anywhere in the world. Therefore, all models in this paper completely discard the timestamp.

4.2.2. Training details

The performance of every model is recorded after training the data with 10-fold (iteration) cross-validation where binary cross-entropy for the loss function, Adam optimizer (JLB, 2015) with a learning rate of 0.001 have been used for 150 epochs. An attempt to increase the epochs beyond this led to overfitting. The dataset has been split by the number of individuals instead of purely by the number of texts. The binary cross-entropy for loss function is calculated as follows:

$$Loss = -\frac{1}{outputsize} \sum_{i=1}^{outputsize} y_i.\log\hat{y}_i + (1 - y_i).\log(1 - \hat{y}_i) \tag{6}$$

where $\hat{y}_i$ is the i^{th} scalar value in the model output, y_i is the corresponding target value, and *outputsize* is the number of scalar values in

Table 2
CLEF2017 Dataset's Description.

	Training Set Depressed	Training Set Control	Testing Set Depressed	Testing Set Control
No. Subjects	83	403	52	349
No. submits	30,851	264,172	18,706	17,665
Avg. num of submits/ subj.	371.7	655.5	359.7	623.7
Avg.num days from first to last submit	572.7	626.6	608.31	623.2
Avg num. words per submision	27.6	21.3	26.9	22.5

the model output.

4.3. Evaluation metric

In this paper, classification overall Accuracy, Precision, F1 score (harmonic mean of precision and sensitivity), and AUC-ROC (a measure of the separability-probability curve), and most importantly the MCC scores are taken into consideration. It was found that precision, recall, and F1 score considers only single class i.e. the assumed class of interest's positive class. Further, while considering the confusion matrix only three values are considered i.e. true positive (TP), false positive (FP), and false negative (FN), the 4th value true negative (TN) is never used in these metrics. This means that even if any value is put in the TN cell, it will not change the precision, recall, or the F1 score. To add to this, accuracy is sensitive to class imbalance. Since both classes are of interest in our study i.e. correctly predicting an individual as depressed and avoiding any individual to be erroneously considered depressed, we have treated the two classes of true and predicted class as two (binary) variables and have computed the coefficient of correlation (better prediction is achieved when the higher correlation between true and predicted values is obtained, which is a rechristened version of the phi-coefficient) i.e. MCC. MCC can be calculated using the following formula derived from the confusion matrix just as Accuracy, Precision and F1 Score:

$$Accuracy = \frac{TP + TN}{TP + TN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$F1Score = 2^{*}\frac{\text{Recall x Precision}}{\text{Recall} + \text{Precision}} \tag{9}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{10}$$

Thus, when the classifier has FP and FN value as 0, MCC = 1, which indicates a perfect positive correlation. Also, when it misclassifies i.e. TP and TN value as 0, MCC = -1 which represents perfect negative correlation. Thus, the MCC value tends to always be between −1 and 1. MCC being perfectly symmetric gives no class higher priority or importance than the other. Thus, MCC considers all values present in the confusion matrix. A higher value which is closer to one means that the model has successfully predicted both classes well albeit one class may have been sparsely or densely represented.

4.4. Results and discussions

In this work, we have thus benchmarked the various deep learning models for depression detection from the eRisk pilot task as illustrated in Table 3, which has not been done before.

To compare the performances of the deep learning models, we performed the paired *t*-test statistical analysis (Demšar, 2006). The significance value, p, was set to 0.05. It was observed that the performances of LSTM and BiLSTM (p = 0.13), CNN-BiLSTM and CNN-LSTM (p = 0.066), CNN and CNN-BiLSTM (p = 0.129) have no significant difference statistically. All other paired combinations including the proposed CBA when paired with each of the models had a statistically significant difference i.e. $p < 0.05$. We further found that CNN based models generated better results than others, which however wasn't effective for CNN-BiGRU model possibly because LSTMs compared to GRU's can remember longer sequences, in theory. It is also assumed to outperform the GRU models in long-distance relation modelling tasks (Kaiser & Sutskever, 2015; Yin, Kann, Yu, & Schütze, 2017).

We further used the Wilcoxon signed-rank test (Huang & Ling, 2005) a non-parametric alternative to the paired *t*-test which is used to compare two related samples. This test yielded the best result for CBA when compared against every other model, with every negative rank producing the CBA model greater than other models and CBA also scored the highest mean rank.

Table 3
Benchmarking Clef2017 (Our implementations).

Model	Accuracy	F1 Macro	AUC-ROC	Precision	MCC
LSTM	94.26	0.81	0.76	0.9	0.65
BiLSTM	94.3	0.81	0.76	0.9	0.65
CNN	95.41	0.86	0.82	0.87	0.74
CNN-LSTM	95.39	0.86	0.81	0.91	0.73
CNN-BiLSTM	95.47	0.86	0.82	0.90	0.72
BiLSTM-Attention	93.75	0.82	0.79	0.84	0.64
CNN BiGRU	93.3	0.8	0.75	0.88	0.62
CNN-BiLSTM Attention (CBA)	**96.71**	**0.89**	**0.85**	**0.93**	**0.77**

Further, to test and rank all the implemented models we performed the Friedman test (Demšar, 2006). Finding the significance value is less than 0.05, we rejected the null hypothesis and concluded that at least one of all the models for classification has a different effect. After analyzing the mean rankings, we found CBA to have mean rank value 8.00 which was the highest amongst all models. We, therefore, inferred that CBA was statistically, the most effective model. Since the overall accuracy experimentally recorded in Table 3 verified the statistical results of Friedman in Table 4, paired *t*-test in Table 5 and Wilcoxon tests in Table 6 we could have concluded that the CBA architecture performed well, in comparison. We found that accuracy alone could not be depended upon for an imbalanced dataset (Chicco & Jurman, 2020). Thus, we have further tried to analyze the models largely on the AUC-ROC and MCC metrics as described in Section 4, to avoid discrepancies in the evaluation. As it specifically takes into consideration all the four values of the confusion matrix and does not bias results based on any class disproportionately. The AUC-ROC expounds the ability of a model to distinguish between the classes (Huang & Ling, 2005) whilst the MCC, which is not affected by the imbalanced nature of the dataset, evaluates all the values of the confusion matrix especially the TN values which have been seldom considered otherwise. As described in Table 3, AUC-ROC and MCC scores of the various models we experimented, the higher the score inclined towards one we found the better is the model. We noticed that overall accuracy did not guarantee that the model may have had fared well in all values of the confusion matrix. We found that though models had accuracy above 90 % it could not necessarily be considered a high performing model, as accuracy do not reflect the actual performance of the model when the data was imbalanced and such models, in reality, fared poorer, especially in the computational medical field such as depression detection.

Thus, keeping the metrics discussed in Section 4 in mind we found

Table 4
Friedman test statistical analysis.

Models	Mean Rank
CNN	4.88
LSTM	3.70
BiLSTM	4.00
Attention	1.08
CNN-BiLSTM	5.54
CNN-GRU	2.80
CNN-LSTM	6.00
CBA	8.00

Table 5
Paired*t* test statistical Analysis.

MODELS	CNN	LSTM	BiLSTM	Attention	CNN-BiLSTM	CNN-GRU	CNN-LSTM	CBA
CNN		0.001	0.001	0	0.129	0	0.001	0
LSTM	0.001		0.13	0.003	0	0	0	0
BiLSTM	0.001	0.13		0.003	0.001	0	0	0
Attention	0	0.003	0.003		0	0.005	0	0
CNN-BiLSTM	0.129	0	0.001	0		0	0.066	0.001
CNN-GRU	0	0	0	0.005	0		0	0
CNN-LSTM	0.001	0	0	0	0.066	0		0.001
CBA	0	0	0	0	0.001	0	0.001	

Table 6
Wilcoxon signed ranks test statistical Analysis.

		N	Mean Rank	Sum of Ranks	
CNN - CBA	Negative Ranks	25[a]	13.00	325.00	a. CNN < CBA
	Positive Ranks	0[b]	0.00	0.00	b. CNN > CBA
	Ties	0[c]			c. CNN = CBA
	Total	25			
LSTM - CBA	Negative Ranks	25[d]	13.00	325.00	d. LSTM < CBA
	Positive Ranks	0[e]	0.00	0.00	e. LSTM > CBA
	Ties	0[f]			f. LSTM = CBA
	Total	25			
BiLSTM - CBA	Negative Ranks	25 [g]	13.00	325.00	g. BiLSTM < CBA
	Positive Ranks	0 [h]	0.00	0.00	h. BiLSTM > CBA
	Ties	0[i]			i. BiLSTM = CBA
	Total	25			
Attention - CBA	Negative Ranks	25[j]	13.00	325.00	j. Attention < CBA
	Positive Ranks	0 [k]	0.00	0.00	k. Attention > CBA
	Ties	0 [l]			l. Attention = CBA
	Total	25			
CNN-BiLSTM - CBA	Negative Ranks	25 [m]	13.00	325.00	m. CNN-BiLSTM < CBA
	Positive Ranks	0[n]	0.00	0.00	n. CNN-BiLSTM > CBA
	Ties	0[o]			o. CNN-BiLSTM = CBA
	Total	25			
CNN-GRU - CBA	Negative Ranks	25[p]	13.00	325.00	p. CNN-GRU < CBA
	Positive Ranks	0[q]	0.00	0.00	q. CNN-GRU > CBA
	Ties	0[r]			r. CNN-GRU = CBA
	Total	25			
CNN-LSTM - CBA	Negative Ranks	25 [s]	13.00	325.00	s. CNN-LSTM < CBA
	Positive Ranks	0 [t]	0.00	0.00	t. CNN-LSTM > CBA
	Ties	0[u]			u. CNN-LSTM = CBA
	Total	25			

that the best results amongst all our experimented models were reported for the CBA architecture which yielded the highest AUC-ROC and MCC score as well as the highest accuracy score. We found that though CNN layers have played an important role, the major roles were contributed by the BiLSTM and the Attention layers together towards the improvement of the overall scores and gaining insights about the contribution of a word for the classification decision. For example, a negative valence sentence having personal pronouns - "I don't know.. I would think the person was trying to befriend if someone stuck an arrow in me and set me on fire" was found to carry higher ranking for words such as "I", "me", "fire" etc. which confirms our assumption that linguistic dimension plays an important role i.e. important features which can be used to distinguish between depression and non-depression. Thus, personal pronouns reveal early signs of depression which the CBA model identified and was represented by the heat map (word-level attention) and reveals that such word representations scored a higher attention probability. Our work involves having the theory that linguistic dimensions like personal pronoun (I, them, her), 1st person singular (I, me, mine), negations (no, not, never); have a great correlation with depression (Tadesse et al., 2019). The combination of CNN-BiLSTM was then used to implement this theory and further improve model performance by adding attention weights on such linguistic aspects for depression detection. The heatmap in Fig. 2, visualizes the linguistic theory and our practical experiments to be true that personal pronouns have a higher probability score. The attention map is illustrated in Fig. 2 which is generated using the attention probabilities of the trained attention

Fig. 2. Heat map for visualizing the word-level attention attained by the CBA model.

network when the above sentence was given as a test sample. We further also realized that it is indeed crucial to have MCC metrics while evaluating a model, especially when there may be any class disproportionately (imbalanced) represented and not just solely focusing upon the accuracy of a model particularly in binary classification to achieve unbiased results of the model. The CBA model, however, lacks deeper exploitation of the attention mechanism for understanding the words which lead to social and affective processes other than just personal pronouns/ singular in the textual and visual context as well as include multi-modal frameworks (Gui et al., 2019) which is in our pipeline for future work which would help models to classify depressed individuals better.

Our future work would include having performance evaluation with a balanced dataset as well as include more deep learning models and incorporating advanced pre-trained word embeddings using GloVe and BERT to extract features used extensively in the studies such as (Zogan et al., 2021) and (Kour & Gupta, 2022). We also plan to better realize depression in individuals by harnessing short-formed 'chatting language' which includes but not only limited to acronyms and abbreviations which is very different from the English text. Delving deeper into emojis and emoticon dictionary learning is planned as part of our future work.

5. Conclusion

Our work benchmarked CLEF2017 dataset (a reliable corpus of social media text data) and performed a comparative evaluation on some of the commonly used deep learning models (LSTM, BiLSTM, CNN, CNN-LSTM, CNN-BiLSTM, BiLSTM Attention, CNN BiGRU) for depression detection and found our proposed CBA model outperforms all these models. Our deliberation to choose imbalanced data for our experiments and further using AUC-ROC and MCC to eliminate the high prevalence of biased results which depended primarily on the metrics such as precision, accuracy, and F1 score were largely studied. We intend to include the recent CLEF 2018 (Losada, Crestani, & Parapar, 2018; Wang, Huang, & Chen, 2018) and CLEF 2019 (Losada, Crestani, & Parapar, 2019) dataset and shall be studied in future studies. We aim to learn and experiment the non-verbal cues from social media text data in the future to understand the subtle nuances of mental illness. Though the current proposed model works excellently on tasks which explicitly mentions about depression, in the future, we intend to analyze and experiment off-topic depression datasets, which would not necessarily have any explicit mentioning about depression.

CRediT authorship contribution statement

Joel Philip Thekkekara: Writing – original draft. **Sira Yongchareon:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

Aldarwish, M. M., & Ahmad, H. F. (2017). Predicting depression levels using social media posts. *Paper presented at the 2017 IEEE 13th international Symposium on Autonomous decentralized system (ISADS)*.

AlSagri, H. S., & Ykhlef, M. (2020). Machine Learning-based Approach for Depression Detection in Twitter Using Content and Activity Features. *arXiv preprint arXiv: 2003.04763*.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173.

Beck, A., Rush, A., Shaw, B., & Emery, G. (1979). *Cognitive therapy of depression*. Guilford Press. New York.

Bhunia, A. K., Konwer, A., Bhunia, A. K., Bhowmick, A., Roy, P. P., & Pal, U. (2019). Script identification in natural scene image and video frames using an attention based convolutional-LSTM network. *Pattern Recognition, 85*, 172–184.

Burdisso, S. G., Errecalde, M., & Montes-y-Gómez, M. (2019). A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications, 133*, 182–197. https://doi.org/10.1016/j.eswa.2019.05.023

Cepoiu, M., McCusker, J., Cole, M. G., Sewitch, M., Belzile, E., & Ciampi, A. (2008). Recognition of depression by non-psychiatric physicians—a systematic literature review and meta-analysis. *Journal of General Internal Medicine, 23*(1), 25–36.

Chakladar, D. D., Dey, S., Roy, P. P., & Dogra, D. P. (2020). EEG-based mental workload estimation using deep BLSTM-LSTM network and evolutionary algorithm. *Biomedical Signal Processing and Control, 60*, Article 101989.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1), 6.

Cohen, J. F., Korevaar, D. A., Altman, D. G., Bruns, D. E., Gatsonis, C. A., Hooft, L., … De Vet, H. C. (2016). STARD 2015 guidelines for reporting diagnostic accuracy studies: Explanation and elaboration. *BMJOpen, 6*(11).

Cong, Q., Feng, Z., Li, F., Xiang, Y., Rao, G., & Tao, C. (2018). XA-BiLSTM: A deep Learning approach for depression detection in imbalanced data. *Paper presented at the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). CLPsych 2015 shared task: Depression and PTSD on twitter. *Paper presented at the Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *Paper presented at the Seventh international AAAI conference on weblogs and social media*.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*(Jan), 1–30.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv: 1810.04805*.

Dinkel, H., Wu, M., & Yu, K. (2019). Text-based Depression Detection: What Triggers An Alert. *arXiv preprint arXiv:1904.05154*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks, 18*(5–6), 602–610.

Gui, T., Zhu, L., Zhang, Q., Peng, M., Zhou, X., Ding, K., & Chen, Z. (2019). Cooperative multimodal approach to depression detection in twitter. *Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence*.

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences, 18*, 43–49.

Guthier, B., Alharthi, R., Abaalkhail, R., & El Saddik, A. (2014). Detection and visualization of emotions in an affect-aware city. *Paper presented at the Proceedings of the 1st International Workshop on Emerging Multimedia Applications and Services for Smart Cities*.

Hasan, M., Rundensteiner, E., & Agu, E. (2018). Automatic emotion detection in text streams by analyzing twitter data. *International Journal of Data Science and Analytics, 7*(1), 35–51. https://doi.org/10.1007/s41060-018-0096-z

Health, N. C. C. f. M., Health, N. I. f., Excellence, C., Society, B. P., & Psychiatrists, R. C. o. (2011). *Common mental health disorders: identification and pathways to care* (Vol. 123): RCPsych Publications.

Hu, X., & Wu, B. (2006). Automatic keyword extraction using linguistic features. *Paper presented at the Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06)*.

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering, 17*(3), 299–310.

Islam, M. R., Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H., & Ulhaq, A. (2018). Depression detection from social network data using machine learning techniques. *HealthInformation Science and Systems, 6*(1), 8.

JLB, D. P. K. (2015). *Adam: A method for stochastic optimization.* Paper presented at the 3rd international conference for learning representations, San Diego.

Kaiser, Ł., & Sutskever, I. (2015). Neural gpus learn algorithms. *arXiv preprint arXiv: 1511.08228*.

Kour, H., & Gupta, M. K. (2022). An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM. *Multimedia Tools and Applications, 81*(17), 23649–23685.

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. *Paper presented at the Twenty-ninth AAAI conference on artificial.* intelligence.

Li, W., & Chau, M. (2018). Applying deep Learning in depression detection. *Paper presented at the PACIS*.

Liu, G., & Guo, J. (2019). Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing, 337*, 325–338. https://doi.org/10.1016/j.neucom.2019.01.078

Losada, D. E., & Crestani, F. (2016). A test collection for research on depression and language use. *Paper presented at the International Conference of the Cross-Language Evaluation Forum for European Languages*.

Losada, D. E., Crestani, F., & Parapar, J. (2017). eRISK 2017: CLEF lab on early risk prediction on the internet: Experimental foundations. *Paper presented at the International Conference of the Cross-Language Evaluation Forum for European Languages*.

Losada, D. E., Crestani, F., & Parapar, J. (2018). Overview of eRisk 2018: Early risk prediction on the internet (extended lab overview). *Paper presented at the Proceedings of the 9th International Conference of the CLEF Association*.

Losada, D. E., Crestani, F., & Parapar, J. (2019). Overview of eRisk at CLEF 2019 Early Risk Prediction on the Internet (extended overview).

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., & Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks, 21*(2–3), 427–436.

Migliore, M., Alicata, F., & Ayala, G. (1995). A model for long-term potentiation and depression. *Journal of Computational Neuroscience, 2*(4), 335–343.

Mitchell, A. J., Rao, S., & Vaze, A. (2011). International comparison of clinicians' ability to identify depression in primary care: Meta-analysis and meta-regression of predictors. *The British Journal of General Practice, 61*(583), e72–e80.

Nease, D. E., & Malouin, J. M. (2003). Depression screening: A practical strategy. *Journal of Family Practice, 52*(2), 118–126.

Orabi, A. H., Buddhitha, P., Orabi, M. H., & Inkpen, D. (2018). Deep learning for depression detection of twitter users. *Paper presented at the Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*.

Park, M., Cha, C., & Cha, M. (2012). Depressive moods of users portrayed in twitter. *Paper presented at the Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Pirina, I., & Çöltekin, Ç. (2018). Identifying depression on reddit: The effect of training data. *Paper presented at the Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*.

Pyszczynski, T., Holt, K., & Greenberg, J. (1987). Depression, self-focused attention, and expectancies for positive and negative future life events for self and others. *Journal of Personality and Social Psychology, 52*(5), 994.

Resch, B., Summa, A., Zeile, P., & Strube, M. (2016). Citizen-centric urban planning through extracting emotion information from twitter in an interdisciplinary space-time-linguistics algorithm. *Urban Planning, 1*(2), 114–127.

Ríssola, E. A., Bahrainian, S. A., & Crestani, F. (2019). Anticipating depression based on online social media behaviour. *Paper presented at the International Conference on Flexible Query Answering Systems*.

Roser, H. R. a. M. (2020). Mental Health. Retrieved from https://ourworldindata.org/mental-health.

Severyn, A., & Moschitti, A. (2015). Twitter sentiment analysis with deep convolutional neural networks. *Paper presented at the Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in reddit social media forum. *IEEE Access, 7*, 44883–44893. https://doi.org/10.1109/access.2019.2909180

Trotzek, M., Koitka, S., & Friedrich, C. M. (2018). Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*.

Wang, Y.-T., Huang, H.-H., & Chen, H.-H. (2018). A neural network approach to Early risk detection of depression and anorexia on social media text. *Paper presented at the CLEF (Working Notes)*.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. *Paper presented at the Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*.

Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of cnn and rnn for natural language processing. *arXiv preprint arXiv:1702.01923*.

Zogan, H., Razzak, I., Jameel, S., & Xu, G. (2021 July). Depressionnet: Learning multi-modalities with user post summarization for depression detection on social media. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval* (pp. 133–142).