DEP-BERTCNN: A Weighted Deep Learning Model for Depression Detection in Online Forums

Anonymous ACL submission

Abstract

Depression is a prevalent mental health disorder affecting a significant portion of the global population. With the rise of social media, online forums have become a popular platform for individuals to discuss their mental health concerns, and online forums have emerged as a valuable source of data for studying depression due to the vast amounts of user-generated content they contain. In this paper, we propose a weighted deep learning model named DEP-BERTCNN for depression detection in text in online forums. DEP-BERTCNN uses a combination of a pre-trained BERT language model, an attention model and convolutional neural network to classify forum users as either depressive of non-depressive. Also in DEP-BERTCNN, we computes weights, using the TFIDF method, based on linguistic depression indicators present in users' posts to enhance the performance of the depression detection model. These weights amplify the most informative aspects of posts indicative of depression. To the best of our knowledge, our work is the first to use input embeddings weighted with depression indicators in combination with an attention model for depression detection. The DEP-BERTCNN model was trained and evaluated on the large-scale Reddit Self-reported Depression Dataset (RSDD). Our results demonstrate that the proposed model outperform several baseline methods on the RSDD dataset, demonstrating the effectiveness of combining deep learning model with linguistic indicators associated with depression symptoms.

1 Introduction

Depression is a common mental disorder that affects millions of people worldwide. According to the World Health Organization (WHO), depression is the leading cause of disability worldwide and is a significant contributor to the overall global strain of disease (Organization, 2023). Moreover, depression remains under-diagnosed and under-treated. Many individuals with depression may not seek professional help or receive adequate treatment due to stigma or lack of access to mental health services (Thornicroft et al., 2017). Therefore, detecting depression early and accurately is crucial for effective intervention and treatment. With the rise of social media and online forums, individuals with depression are increasingly seeking support and sharing their experiences online (Birnbaum et al., 2017). Online forums provide a platform for individuals to share information about their emotions and mental well-being as well as connect with others who have similar experiences. Thus, online forums present an opportunity for early detection of depression and the provision of support to those who need it. They are also a valuable data source to extract valuable background knowledge about depression such as distribution and co-occurrence of symptoms associate with depression. Thus, the development of new methods and approaches to support an accurate depression detection in online forums is necessary.

We have seen in recent years a sustained increasing interest in studying mental health challenges such as depression through social media. Many recent studies have investigated the relationship between language patterns' usage in social media and depression. Several prior works have explored the use of linguistic attributes to predict mental health conditions with some success. A review of the use of social media text-based features in detecting mental health conditions was carried out by Guntuku et al. (Guntuku et al., 2017)

Deep learning has shown promising results for natural language processing (NLP) tasks, such as text classification and language translation (LeCun et al., 2015). In recent years, deep learning models have also been applied to mental health domains, including depression detection (De Choudhury et al., 2016). However, these models often use a one-size-fits-all approach, where all features are given equal importance in the model, leading to sub-optimal performance. Therefore, incorporating domain knowledge into the deep learning model can improve the performance and interpretability of the model (Cohan et al., 2018).

In this paper, we propose a weighted deep learning model for depression detection in online forums named DEP-BERTCNN. The DEP-BERTCNN model is a deep learning model that uses a combination of a pre-trained BERT model and a convolutional neural network (CNN) models to capture the semantic meaning of words in the forum posts and classify online forum users as depressive or non-depressive. We incorporate into our model a weighting scheme, which is based on linguistic depression indicators that are associated with depression symptoms. The weighting scheme assigns more importance to the words that are associated with particular depression symptoms. The authors in these papers, (De Choudhury et al., 2014; Poulin et al., 2014; Rude et al., 2004), emphasized the importance of certain linguistic cues in the language used by depressed individuals to detect depression. Some of these works employ the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al., 2007), which is a text analysis tool that can be used to extract words in psychologically meaningful categories. All these attest to the importance of linguistic depression indicators in posts by depressed individuals to which our work aligns with. Consequently, our approach uses a weighting scheme based on the TF-IDF (Term Frequency-Inverse Document Frequency) method, which measures the importance of a word in a document based on its frequency in the document and its rarity in the corpus (Sammut and Webb, 2010), to obtain weights the depression indicators used in online forums to measure their importance. In order to further emphasize the importance of depression indicators in detecting depression in texts, we introduce an attention layer between the BERT and CNN models. The attention layer is used to obtain a user's posts' representation more weighted by the importance of tokens in the posts. We evaluate our model on the publicly available Reddit Self-reported Depression Dataset (RSDD), which is a dataset of forum posts from both individuals on Reddit online platform who claimed to have been diagnosed of depression as well as control users. DEP-BERTCNN model was compared with several baseline and state-of-the-art models. Since online forums and social media platforms have become more popular with the public to express their feelings and mental health status, it will be beneficial to have a more accurate system, such as DEP-BERTCNN, that can easily be used to detect mental health issues such as depression in users' posts. This will enable practitioners and policy makers to respond to mental health escalations more promptly. The research contributions of this work include to:

- Propose a novel weighting mechanism based on depression indicators for depression detection in online forums, which uses linguistic cues in users' posts to predict the likelihood of depression.
- Develop a deep learning model named DEP-BERTCNN, which integrates the weighting mechanism with a combination of a BERT pre-trained model, an attention layer, and a CNN model, for depression detection in online forums.
- Conduct a comparative analysis of our DEP-BERTCNN model with existing methods for the RSDD dataset.
- Providing insights into the most important words used by depressive online forum users, which can help healthcare professionals and researchers better understand the ailment and its manifestations in online forums.
- To the best of our knowledge, our work is the first to use input embeddings weighted with depression indicators in combination with an attention model for depression detection.

The rest of this paper is organized as follows: Section 2 discusses the prior related studies on depression detection in social media and online forums. Section 3 explains the proposed approach. Experimental setup, dataset and results are discussed in Section 4. Finally, section 5 covers the conclusion and possible future work.

2 Related Work

Depression detection is a challenging problem due to its complex and diverse symptoms. With the widespread use of the internet, online forums have become a valuable resource for understanding depression and its associated symptoms. In recent years, there has been a growing interest in using deep learning models to detect depression in online forums. In this section, we will review some of the existing literature related to depression detection using deep learning models on posts in online forums. The fact that the word an individual uses can be an indicator to understand the individual's psychological state has led to the development of the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker et al., 2007, 2003). By using a comprehensive dictionary, the tool enables researchers to evaluate written texts into several categories, such as social, emotional, cognitive, biological and others, based on word counts. This knowledge has influenced several works to been carried out to detect depression in social media and online forums. Thus, the data from some social media (such as Twitter and Reddit) have been harvested to distinguish the underlying online behaviors between depressed and normal users (Islam et al., 2018; Shen et al., 2017; Yazdavar et al., 2017; Zucco et al., 2017; Wang et al., 2013; Yates et al., 2017a; Stankevich et al., 2018; Coppersmith et al., 2015). These works were done using either statistical and traditional machine learning or deep learning approaches. Cohan et al. (Cohan et al., 2018) proposed a triage system for classifying posts in online mental health forums based on their severity. The authors used a combination of linguistic features and machine learning models to predict the severity of posts. Preoţiuc-Pietro et al. (Preoţiuc-Pietro et al., 2016) investigated the role of personality, age, and gender in tweeting about mental illness. The authors collected a dataset of tweets related to depression and used a combination of feature-based and topic modeling approaches to identify the language patterns associated with depression. The authors found that individuals with higher levels of neuroticism were more likely to tweet about depression.

In text classification, the words that constitute texts have different importance depending on their contextual roles in the texts. Consequently, various studies have been done on which part of texts more emphasis should be placed on as identifying relevant words in texts is paramount to improving the accuracy of a text classification task. Thus, term weighting, which is used to assign appropriate weights to terms or words in texts so that the weights represent the importance of terms or words to texts have been proposed and this have been proven to be quite effective in improving text classification accuracies. Guo et al. (Guo et al., 2019) propose a novel term weighting scheme that they combined with word embeddings. In the proposed approach several weights are assigned to different terms and the weights were applied to the word embeddings of the words to enhance the classification performance of CNNs. To represent text documents more effectively, the authors, in (Onan and Toçoğlu, 2021a), introduce a term weighted tri-gram representation model which is an inverse gravity moment-based term weighted word embedding scheme. They emphasize the word embedding vectors of relevant terms by applying the TF-IDF algorithm for sarcasm identification. In another approach (Onan and Toçoğlu, 2021b), the same authors present a weighted word embeddings combined with a clustering process to identify question topics in discussion forum posts related to a massive open online course (MOOC). In applying term weighting schemes to boost model performance Haopeng et al. Ren et al. (Ren et al., 2019) propose a model they called weighted word embedding model (WWEM). It is a variant of the Neural bag-of-words (NBOW) model for term weighting schemes and n-grams. The proposed model generates informative sentence or text representations by considering the degree of words and the word-order information in texts..

Recent advancements in deep learning models has led to the use of the combination different deep learning approaches such as RNN, BERT and CNN for text classification. This combination of models for text classification has been demonstrated to yield better results than using BERT or CNN as standalone models (Li et al., 2019; Safaya et al., 2020). One example of the combination of of models for depression classification is DepressionNet (Zogan et al., 2021). In this model the authors proposed a social media-based depression detection framework that dealt with the posts as well as some user behaviour features. They created features using BERT-BART models and classified depression using a CNN-GRU stacked model.

The authors in (Nguyen et al., 2022), explored approaches for constraining the performance of depression detection approaches by considering the presence of depression-related symptoms in texts. They achieved this by developing nine symptom detection models that correspond to questions present in PHQ9, a clinically validated screening questionnaire this is commonly used in practical setting. This approach uses a combination of BERT and CNN models constrained by certain depression symptoms for depression detection. Our work, though similar to this, but differs as it improves on the BERT-CNN combination model by including a weighted module based on depressive linguistic patterns as well as an attention component to detect depression.

Yates et al., in their work on detecting depression and self-harm in the Reddit online forum (Yates et al., 2017b), developed the large-scale depression dataset (RSDD) and proposed two variants of a CNN model with two convolutional layers. Their models outperformed other baseline models, such as SVM and MNP, on the RSDD dataset. Our work is also based on the RSDD dataset, our goal is to demonstrate better results with our proposed model. Our proposed approached - DEP-BERTCNN is also based on deep learning architecture and uses the RSDD dataset but intends to improve on the performances reported by other approaches and methods for the same RSDD dataset. Cong et al. (Cong et al., 2018), proposed a deep learning-based model, named X-A-BiLSTM, for depression detection. The X-A-BiLSTM model consists of two components with the first used for data imbalance reduction and the other to improve classification performance. The results reported for X-A-BiLSTM is currently state-of-the-art for the depression detection using the RSDD dataset.

Our unique approach leverages on the presence of depression indicators in text to improve the performance of the depression classification model.

3 DEP-BERTCNN Methodology

Our proposed approach, named DEP-BERTCNN, is a deep learning model that uses a combination of pre-trained BERT model (Devlin et al., 2018) and convolutional neural network (CNN) models to capture the semantic meaning of words in the forum posts and classify online forum users as depressive or non-depressive. A set of posts of a single user are passed into a BERT model to obtain the corresponding embeddings. The BERT component is important as it is used to obtain the word embeddings for the corresponding users' posts. These embeddings are passed through a weighted embeddings combination module where computed depression indicators weights are combined with the posts' word embeddings, obtained from the BERT model, to generate merged weighted embeddings for the users. The BERT model is state-of-the-art language model, which can be used as a feature extractor for various textual tasks. It is deeply bidirectional, unsupervised language representation, and pre-trained using only a large text corpus. This contextual model generates a representation of each sub-word based on the other sub-words in the sentence. Sub-words in text are separated using Byte-Pair Encoding sub-word algorithm.

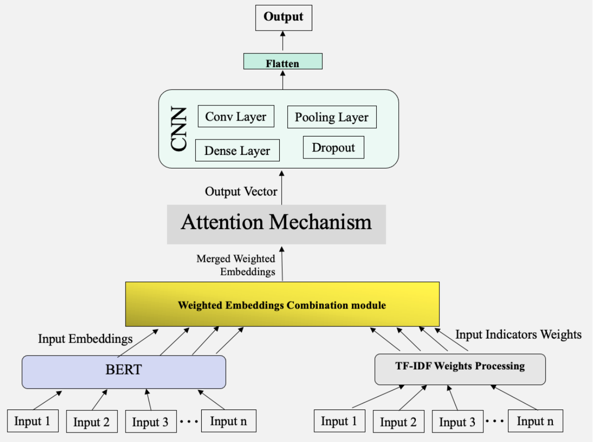

Figure 1: DEP-BERTCNN Model Architecture

As shown in figure 1, the merged weighted embeddings serve as inputs to the attention layer which creates a context vector that becomes the input for the CNN component for classification. The CNN component further extract local semantic features from the context vectors to improve the performance accuracy of the proposed model.

3.1 DEP-BERTCNN Weighting Scheme

In our approach, we first obtain the word embeddings vectors for each post of a user using the BERT model. To obtain the corresponding weight for each post, we leverage on the linguistic articulation of depression texts. Thus, our weighting scheme is based on the presence of depression indicators, which are associated with depression symptoms, in the user's posts. The weighting scheme assigns more importance to the inputs that are associated with the depression symptoms. Several published works in literature have emphasized the importance of certain linguistic cues in the language used by depressed individuals to detect depression. These works employ different methods such as the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al., 2007, 2003) and Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5) (Association, 2013) to extract words related to depression. Thus, the importance of linguistic depression indicators in posts by depressed individuals has been established. Using these resources and knowledge garnered from the Patient Health Questionnaire-9 (PHQ-9) (Kroenke et al., 2001), we created indicators of depression symptoms shown in table 1.

3.1.1 Depression Indicators Weights Analysis

We carried out analysis on the occurrences of depression indicators in both depressed and non-depressed users' posts. We plotted the distribution of the computed weights for these indicators, and as shown in figures 2 and 3, there is a higher usage of depression indicators by depressed users than non-depressed users. We also discovered that the top 5 most used depression indicators were the words "bad", "die", "kill", "ruined" and "hurt".

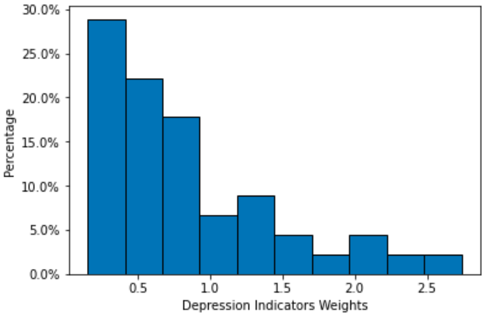

Figure 2: Histogram plot for the distribution of depression indicators weights in depressed users' posts

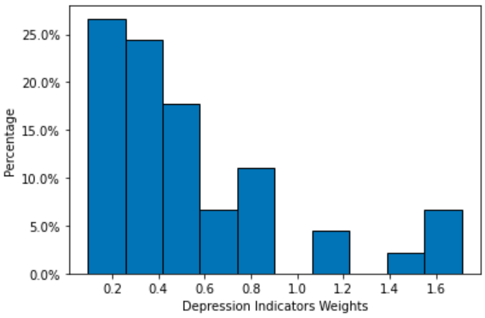

Figure 3: Histogram plot for the distribution of depression indicators weights in non-depressed users' posts

3.1.2 Depression Indicators Weights Computation

In each of a user's posts, we determine the presence of depression indicators and compute their weight scores for the posts. In this work, we use a weighting scheme based on the *TFIDF* (Term Frequency-Inverse Document Frequency) method, which measures the importance of a word in a document based on its frequency in the document and its rarity in the corpus. The weights are computed with the depression indicators as the term we are interested in for each user's posts. Thus, we compute the *TFIDF* score for every depression indicator in each of a user's posts. The formula for calculating *TFIDF* for a term in a document is given in equation 1:

$$TFIDF(t, d) = TF(t, d) * IDF(t) \qquad (1)$$

where t is the term being evaluated, d is the document containing the term, *TF(t, d)* is the term frequency of the term *t* in the document *d*. *IDF(t)* is the inverse document frequency of the term *t* calculated as:

$$IDF(t) = log(N/df(t)) \qquad (2)$$

where N is the total number of documents in the corpus and *df(t)* is the number of documents in the corpus that contain the term *t*.

A single weight is computed for each post of a user since the *TFIDF* scores computed for each depression indicator in the post are summed together. Thus, let *I* be the collection of all depression indicators and *P* be corpus of all the posts for a user. For each post *p* in *P*, we compute the *TFIDF* of all indicators in *I* and sum them up to get the total *TFIDF* value which becomes the weight for *p*. This is expressed as follows:

$$w_p = \sum_{i \in I} TFIDF(i, p) \qquad (3)$$

The computed weights are normalized to have values between 0 and 1 as the sum of the computed weights for users' posts can be greater than 1. Finally, we multiply the word embeddings extracted for the post *p* using the BERT component of our model by the computed weight for *p* to obtain the weighted word embeddings for post *p*.

3.2 Weighted Posts Embeddings Combination

In this research work, we explore different embeddings combination approaches for merging the weighted users' posts embeddings to obtain a single merged weighted embedding for a user. We do this to obtain a vector representation for the user since there are several posts that constitute a label for the user. We experiment with two approaches to determine the combination approach with the best performance. These approaches are described briefly below:

#	Depression Symptom	Depression symptom Indicators
1	Depressed mood	suffering, unhappy, upset, lonely, depressed, lost, empty, sick, sad, sadness, hopeless, irritated
2	loss of Interest or pleasure	Bored, fail, dissatisfied, withdrawal, disinterested
3	Weight loss or gain	fat, thin, appetite, lost weight, overeating, hunger, hungry, undereating
4	Insomnia or hypersomnia	sleepless, sleeplessness, insomnia, awake
5	Psychomotor agitation or retardation	restless, panic, nervous, agitated, anxious, slow, posture, speed, tone
6	Fatigue	weak, tired, low energy, lethargic
7	Excessive inappropriate guilt	bad, useless, meaningless, rejected, damaged, worthless, sick, ruined, guilt
8	Decreased concentration	irritated, indecision, concentration, memory loss, losing attention
9	Thought of death or suicide	sad, afraid, helpless, suicide, die, regret, hate, dead, kill, destructive, goodbye, hurt, quit

Table 1: Depression Symptoms Indicators

3.2.1 Additive Approach

In this approach, we combine weighted embeddings of a user's posts by summing them together to get the representative embedding vector for a user as described in equation 4. The additive combination of posts' embedding vectors $\boldsymbol{ev}_{p1}, \boldsymbol{ev}_{p2}, \ldots, \boldsymbol{ev}_{pn}$ with computed posts weights $w_{p1}, w_{p2}, \ldots, w_{pn}$ is expressed mathematically as:

$$w_{p1}\boldsymbol{ev}_{p1} + w_{p2}\boldsymbol{ev}_{p2} + \ldots + w_{pn}\boldsymbol{ev}_{pn} = \sum_{i=1}^{n} w_{pi}\boldsymbol{ev}_{pi} \tag{4}$$

Here, n is the number of posts for a user. Each weight w_{pi} is a scalar and each vector $\boldsymbol{ev}_{pi}$ is of the same dimension. The result of the weighted combination is also a vector with the same dimension as $\boldsymbol{ev}_{pi}$.

3.2.2 Averaging Approach

The averaging approach takes the average of the summed weighted embeddings. That is, dividing the result of the additive approach by the n, which is the number of posts per user.

$$w_{p1}\boldsymbol{ev}_{p1} + w_{p2}\boldsymbol{ev}_{p2} + \ldots + w_{pn}\boldsymbol{ev}_{pn} = \frac{1}{n}\sum_{i=1}^{n} w_{pi}\boldsymbol{ev}_{pi} \tag{5}$$

3.3 Attention Mechanism

We add an attention layer between the weighted embeddings combination module and the CNN component of our architecture. Our objective is to provide additional focus depression-indicating components in users' posts as attention mechanism would attend to the more important components of the posts. This will allow our model to focus on the parts of the inputs that are most important for making a prediction. The merged weighted input embeddings of the users in the dataset serve as the inputs to the attention component of our architecture.

The attention mechanism was first introduced in (Bahdanau et al., 2016) for a machine-to-machine translation-related sequence-to-sequence modeling task. This mechanism was introduced to address the challenges inherent in the use of LSTM models which have difficulties in dealing with longer text sequences and focusing on specific part a text sequence. The merged weighted embeddings output from the weighted linear combination module is the input to our attention mechanism. Our attention mechanism is implemented as a multi-headed scaled dot product attention layer. The input embeddings vector is transformed into three vectors: Q(Query), K(Key), and V(Value), which are required for the attention mechanism and they are generated from different linear transformations of the input of the attention layer. Thus, the attention output Z is computed as:

$$Z = Softmax\left(\frac{Q \times K^T}{\sqrt{d_{\mathrm{dm}}}}\right) \times V \tag{6}$$

where d_{dm} represents the dimensions of Q, K and V. The attention operation assigns different weights to the transformed input vectors so that the critical input region is emphasized for better classification performance. The attention output Z then becomes the input to the CNN component.

3.4 CNN Component

The inclusion of the CNN model aligns with our main objective to focus our model's classification capacity on the main parts of a users' posts that indicate depression. The CNN model will further compress the text embeddings into an inherent latent representation containing the depression-indicating features, which improves the strength of the classification model. Also, recent published works (Guo et al., 2019) have demonstrated the effectiveness of CNN models for NLP text classification tasks.

Our model's CNN component comprises convolutional, pooling, dropout and dense layers. We use convolution to further extract low-level local features form our context vectors. The CNN component takes the context vectors produced by the attention component as its input to classify users as 'depressive' or 'non-depressive'. The CNN component consists of multiple layers of interconnected neurons. The first layer is the convolutional layer, which applies a set of filters to the input. the subsequent layers consist of more convolutional, dropout and pooling layers, which reduces the size of the output from the previous layer. The final layer of the CNN component is a fully connected layer that carry out classification based on the features extracted by the earlier layers.

4 Experiments and Results

In this section we discuss the dataset used for the experiments, the experimental setup for this work, and the results obtained, which are compared to other reported results for the same dataset.

4.1 Dataset

The Reddit Self-Reported Depression (RSDD) dataset is a collection of posts from Reddit that were self-reported by individuals who identified as having depression (Yates et al., 2017b). The RSDD dataset contains over two million posts from Reddit users who self-identified as having depression, along with associated metadata such as timestamp. The posts were collected between January 2006 and October 2016 using a combination of keyword filtering and manual selection and cover a wide range of topics related to depression, including personal experiences, symptoms, coping mechanisms, and treatment options. The dataset contains 9,210 diagnosed users and 107,274 control or non-depressed users which are divided approximately equally in train, validation and text datasets. On average each user in the dataset has 969 posts. The mean post length is 148 tokens. The dataset is in the JSON lines format. Each line is an array representing one user and several posts of a user combine to have a label. The label field contains the user's label ('control' or 'depression'), and the posts field contains (timestamp, untokenized post) pairs.

We carried out preprocessing tasks on the dataset by removing all tags, urls, emojis from the posts as they often do not convey useful information to the model. Also, all special characters were removed and the texts were converted to lowercase

4.2 Ethical Concerns

As social media data are often sensitive, privacy concerns and the risk to the individuals in the data are issues to be considered. Thus, the authors of the RSDD dataset followed all ethical considerations in creating the dataset as it contains only publicly available Reddit posts. They ensured that annotators were shown only anonymized posts and made to agree to make no attempts to deanonymize or contact the Reddit users. Consequently, we also follow these ethical guidelines.

4.3 Experimental Setup

In our experiments, we use the BERT based model with an embedding dimension of 100 and input length to be 150 since the mean post length in the the dataset is 148 tokens. As for the CNN component, there are four hidden layers (3 Conv1D layers & 1 dense layer) with dropout and pooling layers. The CNN component also comprises the relu activation function, a dense output layer and an embedding layer serving as the input layer of the model. The hidden layers have filters as 128 with kernel size of 5. We trained the model over 20 epochs. The model was trained using stochastic gradient descent with the Adam optimizer (Kingma and Ba, 2014) and a learning rate of 1e-5.

We evaluated different variants of our proposed model to emphasize the importance of the weighted linear combination module and the attention model in our proposed approach.

- **BERTCNN:** This model variant serves as the baseline model for our approach. Here, the model does not include component which creates posts weights using depression indicators, and the attention components. The output of the BERT model is merged as input to the CNN model.

- **BERTCNN+W:** This model variant includes only the weighted combination component but not the attention component. The weighted module is used to weight the output of the BERT model before being merged to serve as input to the CNN model.

- **BERTCNN+A:** This model variant includes only the attention component but not the weighted combination component. The output of the BERT model is merged and serve as input to the attention layer whose output becomes the input to the CNN model.

- **DEP-BERTCNN:** This is our complete proposed model which includes both the weighted combination and the attention mechanism.

To address the imbalance inherent in the dataset, we adopt an under-sampling method by randomly selecting as many non-depressive users as depressive users available in the RSDD dataset. Thereafter we selected 3000 each of depressed and non-depressed users for the train set and 2000 each of depressed and non-depressed users for the validation set as well as for test set.

4.4 Results

We report the results of our experiments using the two weighted posts embeddings combination approaches with the different variants of our model as shown in table 2. The results show that the additive approach performs better than the averaging approach.

Models	Additive Approach			Averaging Approach		
	Precision	Recall	F1	Precision	Recall	F1
BERTCNN	0.50	0.49	0.49	0.47	0.45	0.46
BERTCNN+W	0.60	0.54	0.57	0.56	0.53	0.54
BERTCNN+A	0.58	0.52	0.55	0.52	0.50	0.51
DEP-BERTCNN	**0.72**	**0.61**	**0.66**	0.60	0.54	0.57

Table 2: Results additive and averaging weighted posts embeddings combination approach

Using our best weighted posts embeddings combination approach, the additive approach, we compare our results with other works depression detection evaluated on the RSDD dataset as shown in table 3. The results show that our model outperform the other models in terms of precision, recall and F1 scores.

Models	Precision	Recall	F1
Bow-MNB	0.44	0.32	0.36
Bow-SVM	0.72	0.29	0.41
Feature-rich-MNB	0.69	0.32	0.44
Feature-rich-SVM	0.71	0.31	0.43
User model-CNN	0.59	0.45	0.51
LSTM	0.50	0.39	0.44
Attention-LSTM	0.54	0.35	0.42
Attention-BiLSTM	0.62	0.39	0.48
X-A-BiLSTM	0.69	0.53	0.60
BERTCNN	0.50	0.49	0.49
BERTCNN+W	0.60	0.54	0.56
BERTCNN+A	0.58	0.52	0.55
DEP-BERTCNN	**0.72**	**0.61**	**0.66**

Table 3: Results from using our additive combination of weighted embeddings approach compared with other models

5 Conclusion and Future Works

In this paper we proposed a deep learning model named DEP-BERTCNN for detecting depression in online forums using the large scale Reddit Self-Harm Depression Dataset (RSDD). DEP-BERTCNN combines a pre-trianed BERT and CNN models with a weighting scheme based on depression indicators and an attention layer for depression classification. The weighting scheme is used to compute weights from the depression indicators present in users' posts using the TFIDF method. These pre-computed weights are then used to produce a weighted vector representation for each user's posts. To further emphasize the importance of the linguistic depressive indicators in users' posts we introduced an attention component that takes the merged weighted vectors as inputs. The outputs of the attention component are fed as inputs to the CNN component of our model for further extraction of semantic context from users' posts and subsequent classification as either depressive or non-depressive. Our work is unique in the sense that it is the first to use input embeddings weighted with depression indicators in combination with an attention model for depression detection. Our results outperform the current state-of-the-art for depression detection using the RSDD dataset.

One future direction for this work is to develop a model to extract depression symptoms and their severity from users' posts in online forums and social media platforms. Such a model can then be used to obtain valuable depression symptoms statistics from online posts. We claim that the methods described in this paper can straightforwardly be reused for this task.

6 Limitations

As the availability of depression-labeled datasets is a major challenge in using deep learning models for depression detection, our experiments in this work is limited to one dataset. Our experimental results are, therefore, compared to other works using the same RSDD dataset for depression detection. This work is restricted to only using the depression indicators for weights computation though it is possible for the analysis to go further to finding out the importance of specific depression indicators for model prediction.

References

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders*, 5th edition. American Psychiatric Association.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Michael L Birnbaum, Asra F Rizvi, Jamie Confino, Christoph U Correll, and John M Kane. 2017. Role of social media and the internet in pathways to care for adolescents and young adults with psychotic disorders and non-psychotic mood disorders. *Early intervention in psychiatry*, 11(4):290–295.

Arman Cohan, Quynh Vu Do, and Nathaniel Collier. 2018. Triaging content severity in online mental health forums. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1490–1500. Association for Computational Linguistics.

Qing Cong, Zhiyong Feng, Fang Li, Yang Xiang, Guozheng Rao, and Cui Tao. 2018. X-a-bilstm: a deep learning approach for depression detection in imbalanced data. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1624–1627.

M. M. Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. Clpsych 2015 shared task evaluation. `http://www.cs.jhu.edu/~mdredze/datasets/clpsych_shared_task_2015/`.

Munmun De Choudhury, Scott Counts, Eric J. Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work amp; Social Computing*, CSCW '14, page 626–638, New York, NY, USA. Association for Computing Machinery.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 2098–2110, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Sharath Chandra Guntuku, David Bryce Yaden, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Bao Guo, Chunxia Zhang, Junmin Liu, and Xiaoyi Ma. 2019. Improving text classification with weighted word embeddings via a multi-channel textcnn model. *Neurocomputing*, 363:366–374.

Md Rafiul Islam, A R M Kamal, Nasrin Sultana, Rashedul Islam, Mohammad Ali Moni, et al. 2018. Detecting depression using k-nearest neighbors (knn) classification technique. In *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, pages 1–4.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Kurt Kroenke, Robert L. Spitzer, and Janet B.W. Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

Wenting Li, Shangbing Gao, Hong Zhou, Zihe Huang, Kewen Zhang, and Wei Li. 2019. The automatic text classification method based on bert and feature union. In *2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 774–777.

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. In *Annual Meeting of the Association for Computational Linguistics*.

Aytug Onan and Mansur Alp Toçoğlu. 2021a. A term weighted neural language model and stacked bidirectional lstm based framework for sarcasm identification. *IEEE Access*, 9:7701–7722.

Aytuğ Onan and Mansur Alp Toçoğlu. 2021b. Weighted word embeddings and clustering-based identification of question topics in mooc discussion forum posts. *Computer Applications in Engineering Education*, 29(4):675–689.

World Health Organization. 2023. Depression. `https://www.who.int/health-topics/depression#tab=tab_1` [Accessed: (May 3, 2023)].

J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1):547–577.

James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Linguistic inquiry and word count: Liwc [computer software]. Austin, TX: liwc.net.

Chris Poulin, Brian Shiner, Paul Thompson, Linas Vepstas, Yinong Young-Xu, Benjamin Goertzel, Bradley Watts, Laura Flashman, and Thomas McAllister. 2014. Predicting the risk of suicide by analyzing the text of clinical notes. *PloS one*, 9(1):e85733.

Daniel Preoţiuc-Pietro, Johannes C. Eichstaedt, Gregory Park, Maarten Sap, Laura K. Smith, Vahe A. Tobolsky, and H. Andrew Schwartz. 2016. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, pages 114–120. AAAI Press.

Haopeng Ren, ZeQuan Zeng, Yi Cai, Qing Du, Qing Li, and Haoran Xie. 2019. A weighted word embedding model for text classification. In *Database Systems for Advanced Applications*, pages 419–434, Cham. Springer International Publishing.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Claude Sammut and Geoffrey I. Webb, editors. 2010. *TF–IDF*, pages 986–987. Springer US, Boston, MA.

Guangxuan Shen, Fanglin Wang, Jian Zeng, Jun Zhu, and Deng Cai. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 3838–3844.

M. Stankevich, V. Isakov, D. Devyatkin, and I. Smirnov. 2018. Feature engineering for depression detection in social media.

Graham Thornicroft, Somnath Chatterji, Sara Evans-Lacko, Michael J. Gruber, Nancy Sampson, Sergio Aguilar-Gaxiola, Ali Al-Hamzawi, et al. 2017. Undertreatment of people with major depressive disorder in 21 countries. *The British Journal of Psychiatry*, 210(2):119–124.

X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, and Z. Bao. 2013. A depression detection model based on sentiment analysis in microblog social network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 201–213.

A. Yates, A. Cohan, and N. Goharian. 2017a. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017b. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Amir Hossein Yazdavar, Mohammad Dehghani, Mohammadreza Khodabakhsh, Sara Rahmatizadeh, and Mehrdad Shahbazi. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1191–1198.

Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. 2021. 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*, SIGIR '21, page 133–142, New York, NY, USA. Association for Computing Machinery.

Carlo Zucco, Barbara Calabrese, and Mario Cannataro. 2017. Sentiment analysis and affective computing for depression monitoring. In *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*, pages 1988–1995.