IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 8, 2020, accepted February 3, 2020, date of publication February 13, 2020, date of current version February 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2973737

MGL-CNN: A Hierarchical Posts Representations Model for Identifying Depressed Individuals in Online Forums

GUOZHENG RAO[1,2,3], (Member, IEEE), YUE ZHANG[1,2], LI ZHANG[4], QING CONG[1,2], AND ZHIYONG FENG[1,2], (Member, IEEE)

[1]College of Intelligence and Computing, Tianjin University, Tianjin 300072, China
[2]Tianjin Key Laboratory of Cognitive Computing and Applications, Tianjin University, Tianjin 300072, China
[3]School of New Media and Communication, Tianjin University of Science and Technology, Tianjin 300072, China
[4]School of Economics and Management, Tianjin University of Science and Technology, Tianjin 300072, China

Corresponding author: Li Zhang (zhangli2006@tust.edu.cn)

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 61373165 and Grant 61672377.

ABSTRACT More users suffering from depression turn to online forums to express their problems and seek help. In such forums, there is often a large volume of posts with sensitive content, indicating that the user has a risk of suicide and self-harm. Early detection of depression using appropriate deep learning models and social media data can prevent potential self-harm. However, existing depression detection models are not powerful enough to capture critical sentiment information from the large volume of posts published by each user, which makes the performance of these models not satisfying. To address this problem, we propose a hierarchical posts representations model named Multi-Gated LeakyReLU CNN (MGL-CNN) for identifying depressed individuals in online forums. The model consists of two parts: the first one is a post-level operation, which is used to learn the representation of each post of the user, and the second one is a user-level operation, which is used to obtain the overall representation of the user's emotional state. Besides, we propose another depression detection model by changing the number of gated units in the MGL-CNN, which is named Single-Gated LeakyReLU CNN (SGL-CNN). We show how to use our models to identify depressed users through a lot of posted content. Experimental results showed that our models performed better than the previous state-of-the-art models on the Reddit Self-reported Depression Diagnosis dataset, and also performed well on the Early Detection of Depression dataset.

INDEX TERMS Depression detection, online forums, MGL-CNN, SGL-CNN, neural network architecture.

I. INTRODUCTION

Mental health problem remains a major challenge affecting modern people's health. People's mental health condition is closely related to their quality of life. It affects how they think, feel, and act. Depression, one of the most common mental disorder, is the leading cause of self-harm and suicide worldwide and affects millions of people [1]. According to the estimates from the World Health Organization [2], more than 300 million people are now suffering from depression. Moreover, from 2005 to 2015, the number of people with depression increased by at least 18%. Depression is the most preventable disorder [3]. Several studies have demonstrated that early depression detection and treatment can reduce the damage caused by the disorder [4]–[6]. However, services for the early detection and treatment of depression and other mental health disorders are extremely limited. Furthermore, many patients are reluctant to seek help from the healthcare provider. These problems prevent patients from getting timely treatment, leading to further deterioration of the condition [7], and, worst of all, suicide [8]. Therefore, it is vital to recognize people suffering from depression and provide them with psychological therapy as soon as possible [9].

An increasing number of people suffering from depression turn to online resources (Twitter, websites, Reddit, etc.) to express their psychological issues and seek help [10]–[12]. In particular, online forums that can choose to remain pseudonymous or anonymous are more popular. Using social

The associate editor coordinating the review of this manuscript and approving it for publication was Nagarajan Raghavan.

media data for early detection of depression tasks has become an effective means. Meanwhile, massive social media data makes it difficult to identify users with depression or at risk of suicide manually, which makes the development of automatic depression detection technology more critical. Early detection of depression on social media is a continuous process of data accumulation, and only when there is sufficient evidence can a certain degree of depression risk be identified. This requires collecting a large number of posts with strong temporal associations and long time spans. However, most existing approaches use data from a limited number of health centers involving privacy and confidentiality issue for depression detection [13]. There are a few methods for depression detection based on large-scale social media data. Therefore, how to build user emotional state representations and identify critical sentiment information from a large number of posted contents is very important.

Self-expression and social support can help improve the psychological state of depressed people [14]. Moreover, the words people use on social media can reveal real and significant aspects of their social and psychological worlds [15]. Natural language is related to personality, mental state, and situational fluctuation. Therefore, how to identify the linguistic style of the individuals involved is particularly important. There has been a great deal of research that is focused on the depression detection and mental health problems, starting from the analysis of text extracted from social media. For example, Mowery *et al.* conducted a large number of machine learning algorithms to classify depressive symptoms from twitter data for mental health [16]. Choudhury *et al.* collected several hundred Twitter users who have been diagnosed with depressive disorder, using a statistical classifier to estimate the risk of depression through measuring the users' emotions, language, and linguistic styles [17]. Moreno *et al.* utilized a large amount of data from Facebook, referencing to depression symptoms to ultimately determine the depression users [18]. Towards large-scale data like Reddit dataset, Yates *et al.* presented a general neural network architecture for combining posts into a representation of a user's features to assess depression and self-Harm risk [19]; Then Qing Cong *et al.* proposed a deep learning-based approach for solving the imbalanced RSDD datasets [20]. In the Natural Language Processing field, text-based depression detection can also be considered as a sentiment analysis task. In fact, in addition to depression detection, early detection techniques can also be used in many other health-related fields. For example, it might be used to identify potential pedophiles, people with suicidal tendencies, and to monitor the evolution of psychological disorders.

In this work, we propose two general hierarchical posts representations models for identifying depressed users on large-scale datasets, including two essential parts: the post-level operation and the user-level operation. We attempt to apply gating weight to construct the representation of the users' posts. The main contributions of our work are as follows:

- We introduce two hierarchical neural network models with gated units and convolutional networks for fulfilling depression detection task, which are named Multi-Gated LeakyReLU CNN (MGL-CNN) and Single-Gated LeakyReLU CNN (SGL-CNN). The user's dataset is divided into a certain number of posts, and we can use our models to identify the genuinely crucial sentiment features of each user's posts and suppress other unimportant information as possible.
- The proposed models can encode the relations between posts in user representation. It consists of two parts: the first one is a post-level operation, which is used to learn the representation of the user's every post, and the second one is a user-level operation, which can obtain the overall representation of the user's emotional state. The traditional convolutional neural network is weak in identifying crucial depression features. According to this situation, we add gated units to improve the performance of this task dramatically.
- Empirical results on the RSDD dataset task demonstrate that our models perform better than the state-of-the-art methods. To prove the generality of our models, we also introduce the Early Detection of Depression dataset from another online forum to estimate the risk of depression. Our methods also perform well on this dataset compared to strong existing methods, demonstrating that our framework is robust and general.

The rest of this paper is organized as follows. Section II gives an overview of related works. Section III introduces our depression detection models. Section IV describes in detail how we conduct the experiment and discusses the experimental results of our proposed models. Section V is the conclusion of our depression detection work.

II. RELATED WORK

Many studies analyze mental health-related texts in social media to better identify and understand mental health-related issues. Some of these studies use traditional machine learning methods. For instance, Schwartz *et al.* built a regression model by using Facebook data to predict multiple-granularity depression in individuals [21]. Thompson *et al.* used the clinical notes and online social media data to build a model based on Random Forest classifier [22] with bag-of-words features, detecting the risk of suicide in military personnel and veterans [23]. Furthermore, many traditional methods were used in the shared task automatic identification of content in mental health forums by the 2016 Computational Linguistics and Clinical Psychology Workshop. For instance, Malmasi *et al.* used a Random Forest meta-classification approach on top of a set of base classifiers [24]. Brew used SVM with Radial Basis Function (RBF) kernel [25].

Aside from the machine learning explorations which have achieved sound results in depression detection, many deep learning methods have had impressive successes on text classification and sentiment analysis. These methods only rely

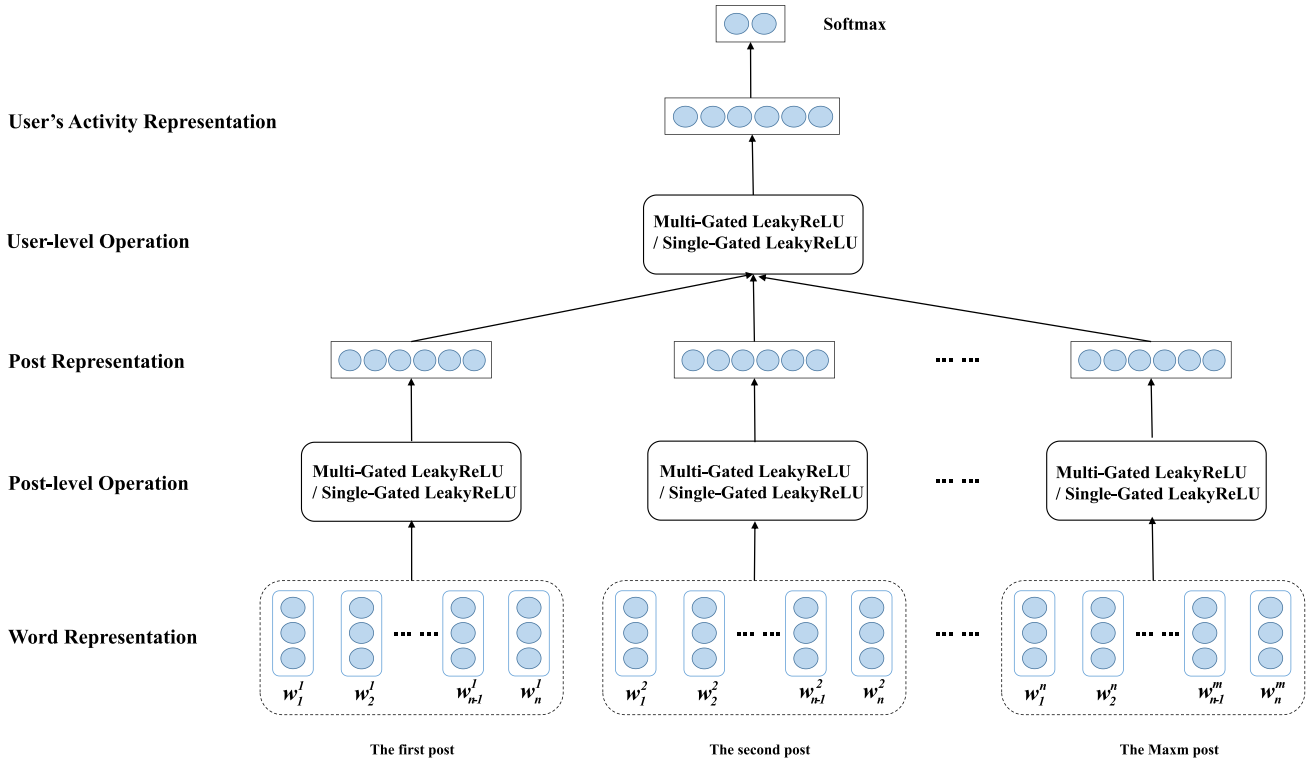


FIGURE 1. Architecture of our proposed depression detection models. Each target user has up to m posts and each post consists of n words. w_n^m stands for the n-th word in the m-th post.

on text and are not dependent on any external features. For example, Long Short-term Memory (LSTM) [26] networks and their variants introduced the memory units and the gating mechanism to decide whether to delete or add information from memory units so that longer dependency information can be learned. Much research about sentiment analysis is based on LSTM. For instance, Gui *et al.* [27] proposed a novel cooperative multi-agent model to depression detection on Twitter. The model included a text feature extraction and an image feature extraction. The part of text feature extraction applied a gated recurrent unit and convolutional neural networks to extract the textual sentiment features. Tang *et al.* [28] developed two effective target-dependent models for sentiment classification on Twitter by using the bidirectional LSTM. In addition to LSTM, Convolutional Neural Networks (CNNs) are actively exploited for text classification in the medical domain or other NLP tasks, designed to learn to extract a hierarchy of crucial text elements. For instance, Kim first applied a simple CNN with one layer of convolution for sentence classification [29]. Yates *et al.* proposed a general neural network architecture for combining posts into a representation of a user's features to assess depression and self-Harm risk [19]. Although CNN is originally designed for Computer Vision, they are very successful in NLP tasks and easily parallelized during training for not having time dependency.

Gated Convolutional Neural Network [30] (GCNN) firstly introduced the gated units into CNN for language modeling, providing a linear path for the gradients while retaining nonlinear capabilities for reducing the gradient vanishing. This model utilizes one convolutional layer to produce gating weights and identify abstract features. Since the weights and the abstract features are convolved at the same level, the significant features identified by the gating weights is very monotonous. To better discover contextual information in text classification, Yang Liu *et al.* introduced a new CNN model (AGCNN) for sentence classification, which generated the gating weights by a variety of specialized convolution kernels to integrate the contextual information of a particular context window into the control weights [31]. And to achieve better performance with aspect-based sentiment analysis, Wei Xue *et al.* proposed a model based on gated convolutional neural networks, which can selectively output the sentiment features according to the given aspect or entity [32].

III. THE PROPOSED METHOD

We proposed two novel hierarchical depression detection models named MGL-CNN and SGL-CNN for identifying depressed individuals in online forums. Since a user's overall data consists of a list of posts, and each post consists of a list of words, the models consist of two parts: a post-level operation and a user-level operation. It first produces continuous post representations from word representations. Afterward, post representations are treated as inputs of the second part to get the user's overall emotional state representations. Users' activity representations are then used as features for

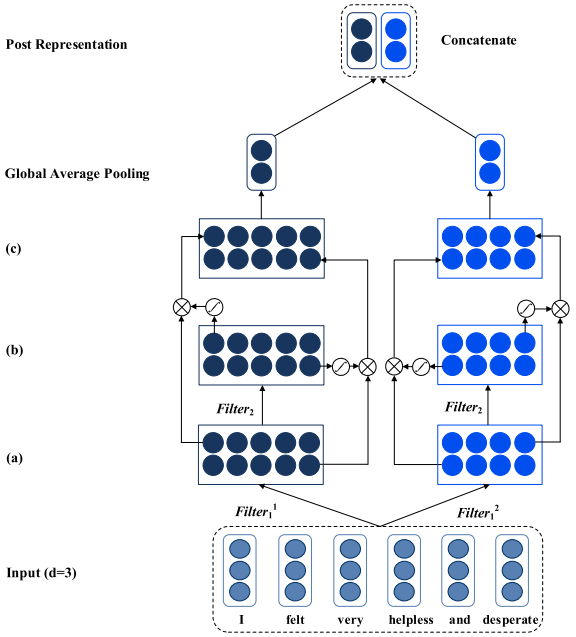


FIGURE 2. The architecture of post-level operation in SGL-CNN, denoting a user's post with the length n of 6 and the word vectors' dimension d of 3 as an example. The first convolutional layer applies $Filter_1^1$ and $Filter_1^2$ to derive feature values (a); $Filter_2$ shown here which is applied to obtain a gating weight (b); the element-wise product between (a) and (b) is used to get an enhancement feature map (c).

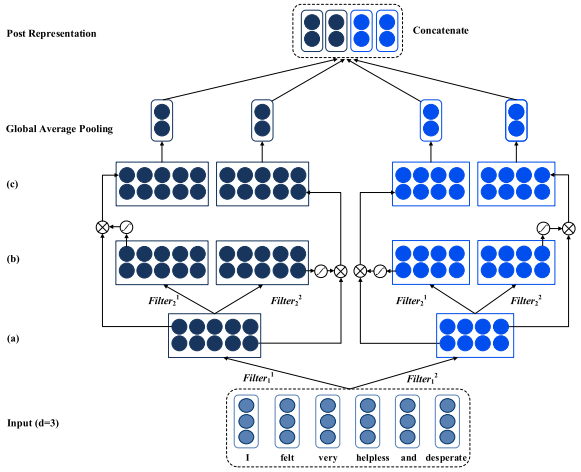


FIGURE 3. The architecture of post-level operation in MGL-CNN, denoting a user's post with the length n of 6 and the word vectors' dimension d of 3 as an example. The first convolutional layer applies $Filter_1^1$ and $Filter_1^2$ to derive feature values (a); $Filter_2^1$ and $Filter_2^2$ shown here which is applied to obtain gating weights (b); the element-wise product between (a) and (b) is used to get enhancement feature maps (c).

depression classification. The architecture of our proposed depression detection models is shown in Fig. 1. The difference between the two models is the number of gated units in their two parts.

Current natural language processing methods mostly use long short-term memory and attention mechanisms to predict the sentiment polarity of the concerned targets, which need more training time and computational cost. Our proposed models replace the recurrent connections typically used in recurrent networks with gated temporal convolutions. Meanwhile, special convolution encoders are used to convolve the inputs and obtain gating weights independently, and the hierarchical structure can reuse parameters. Therefore, the computations of our models don't have time dependency and can be easily parallelized over the individual words of every post in the user's document.

In this session, we mainly describe the details of the post-level operation for the two models. The structure of the user-level operation is the same as the post-level operation. The input to the models will pass through multiple layers of the convolutional neural network with gated units, making full use of limited contextual information to obtain the critical features of the post's representation.

Fig. 2 and Fig. 3 respectively display the post-level operation in SGL-CNN and MGL-CNN, which all consist of two convolutional layers and a global average pooling. The difference between them is the number of gating weights generated. In the demo model of post-level operation in MGL-CNN, the first convolutional layer uses two convolution kernels of different size to obtain an abstract feature map (no padding on the input). The second convolutional layer with the gated unit then uses convolution kernels to obtain different gating weights (padded where necessary). These gating weights are applied to do the element-wise product with the feature map generated by the first convolutional layer to get a post representation.

We have described the process by which *one* feature is extracted from *one* filter. Each word is represented by a embedding stored in a word embedding matrix $L_w \in \mathbb{R}^{d\times|V|}$ where $|V|$ is the number of words in the vocabulary and d is the dimension of word vector. Formally, let us denote a user's each post consisting of n words as $\{w_1, w_2, \ldots w_i, \ldots w_n\}$, let $x_i \in \mathbb{R}^d$ be the d-dimensional word vector corresponding to the i-th word in the post. A post embedding of length n is represented as

$$X_{1:n} = [x_1, x_2, \ldots, x_n] \tag{1}$$

In the first convolutional layer, we use CNN with multiple convolutional filters of different widths [33] to produce post's representation. The convolutional filters with different widths can be regarded as extractors to obtain multi-grained local information like N-Grams. Similarly, a convolutional filter with a width of 2 essentially captures the semantics of bigrams in a user's post. Multiple convolutional filters with different window sizes are applied to obtain multiple feature maps. Let $K \in \mathbb{R}^{s\times d}$ with the stride of 1 be a convolutional filter, which is applied to a window of s words to produce a new feature. Let $[x_i, x_{i+1}, \ldots, x_{i+s-1}]$ refers to the concatenation of word embeddings in a fixed-length window size s, which is denoted as $X_{i:i+s-1}$. A new feature a_i is generated from $X_{i:i+s-1}$ by

$$a_i = f\left(K * X_{i:i+s-1} + b\right) \tag{2}$$

where $b \in \mathbb{R}$ is a bias term, * denotes convolution operation and f is a activation function (LeakyReLU). This filter is applied to each possible window of s words in the post $\{X_{1:s}, X_{2:s+1}, \cdots, X_{n-s+1:n}\}$ to produce a feature map

$$A = [a_1, a_2, \ldots, a_{n-s+1}] \tag{3}$$

with $A \in \mathbb{R}^{(n-s+1)\times 1}$. Each feature map A from all feature maps obtained by filters with different sizes is then fed into the second convolutional layer.

The second convolutional layer consists of a convolutional layer and gated units. This layer is designed to produce different gating weights. We denote a convolution operation involving a kernel $F \in \mathbb{R}^{h\times 1}$, which is applied to contextual features A. The kernel F with window size h (padded when necessary) slides on a feature a_l to generated a gating weight

$$g_l = \begin{cases} f\left(F * A_{l-\frac{h}{2}:l+\frac{h}{2}-1} + b\right) & (h \text{ is even }) \\ f\left(F * A_{l-\frac{h-1}{2}:l+\frac{h-1}{2}} + b\right) & (h \text{ is odd }) \end{cases} \tag{4}$$

Here $g_l \in \mathbb{R}, l = 1, 2, \cdots, n\text{-}s+1$. All gating weight elements generated by the feature map A and the kernel F produce the gating weights matrix

$$G = [g_1, g_2, \cdots, g_{n-s+1}] \tag{5}$$

with $G \in \mathbb{R}^{(n-s+1)\times 1}$. Let m be the number of convolution kernels used in the second convolutional layer. We utilize the gated units of MGL-CNN to extract different gating weight matrix: $G_1, G_2, \cdots, G_m$.

Afterwards we get the output feature map O through the gating weight matrix G

$$O = A \otimes G \tag{6}$$

where $\otimes$ is the element-wise product between matrices. $O \in \mathbb{R}^{1\times(n-s+1)}$ when we use the SGL-CNN. $O \in \mathbb{R}^{m\times(n-s+1)}$ when we use the MGL-CNN.

The output O of the first convolutional layer is modulated by the gating weight G. These gating weights multiply feature map A and control what information should be propagated through the layers. To capture global information of a post, we then feed the outputs of the second convolutional layers to a global average pooling layer and concatenate all the outputs to get post representation (concatenate when in MGL-CNN model).

The obtained post representations are fed to the user-level operation to calculate the user's activity representation. We use the same method as the post-level operation. The obtained user's features are then passed to a fully connected softmax layer whose output is the probability distribution over labels. Categorical cross-entropy is used as the model's loss function. Let p^T be the target sentiment distribution for each document, p be the predicted document sentiment distribution.

$$\text{loss} = -\sum_{i\in T}\sum_{j=1}^{C} p_j^T(i) \cdot \log\left(p_j(i)\right) \tag{7}$$

where T is the training data, C is the number of categories, i is the index of the document, j is the index of class. The goal of training is to minimize the cross-entropy error between p^T and p for all training documents.

IV. EXPERIMENTS AND RESULTS

Experiments are conducted based on the Reddit Self-reported Depression Diagnosis (RSDD) dataset and the Early Detection of Depression dataset (eRisk 2017). We evaluate the performance of our proposed models by comparing them with other strong baseline models and analyze the performance of our models. The reported results are on the test set.

A. EXPERIMENTAL DATASETS

The large-scale novel Reddit Self-reported Depression Diagnosis (RSDD) dataset [19] contains over 9,000 diagnosed users with depression, which is matched with approximately 107,000 control users who have a healthy mental state (data imbalance). On average, there are about 900 posts per user in the dataset, with 148 words per post. This dataset is created from a publicly-available online forums Reddit, which is used to train and test the model of identifying the users with depression. The RSDD dataset is magnitude larger and more high-accurate than prior work creating self-reported diagnoses datasets. The diagnosis posts, which includes false positives diagnosis such as hypotheticals, negations are all removed from the diagnosed users, and the users publishing fewer than 100 posts are also discarded. Meanwhile, in order to avoid easy identification of the diagnosed users through sensitive terms strongly associated with depression, the posts with depression terms are removed.

The Early Detection of Depression dataset (eRisk 2017) [34] can be used to develop an exploratory task on early risk detection of depression. It is a collection of posts from a set of social media users, including two categories of users: depressed users and mental health users. Both categories are unbalanced (more mental health users than depressed users). For each user, the collection contains a sequence of posts (in chronological order). The number of all users is not very high (about 486 users), but each user has a long history of writings (on average hundreds of messages from each user). Furthermore, the mean date from the first to the last submission is quite long (more than 500 days).

B. BASELINE MODELS

We compare our methods with the following baseline methods used on the Reddit Self-reported Depression Diagnosis dataset. The previous state-of-the-art model on the RSDD dataset is User model-CNN [19].

- BoW-SVM and BoW-MNB classifiers [35]. Support Vector Machines (SVM) or Multinomial Naive Bayes (MNB) combines with the post itself represented as a sparse bag of words features for depression detection tasks.
- Feature-rich-SVM and Feature-rich-MNB. The two methods use multiple features such as a sparse bag of

words features, external psycholinguistic features captured by LIWC5 [36] and emotion lexicon features [37].
- User model-CNN [19]. The depression detection model consisted of a shared architecture based on a CNN, a merge layer, model-specific loss functions, and an output layer. It was the previous state-of-the-art model on the RSDD dataset.

Besides, we introduce several popular models in natural language processing and compare our models with them.

- Long Short-Term Memory is a recurrent neural network with memory cells and three gate mechanisms, which is designed to avoid long-term dependency [26]. In our depression detection task, it takes the whole words of a post as a single sequence to obtain the post's representation and use the whole posts of a user to get the user's representation for detection.
- Bi-directional Long Short-Term Memory consists of two LSTMs, which can capture bidirectional semantic dependency and improve the abilities of memory [38].
- GRU-Attention model consists of a word- and sentence-level attention mechanisms and sequence encoders, which is based on GRU for document classification [39]. Besides, we also replace GRU with LSTM (LSTM-Attention) and Bidirectional LSTM (Bi-LSTM-Attention) to be the baselines.
- CIFG-LSTM is a variant on Long Short Term Memory, which is designed to couple the input gate and the forget gate as one uniform gate [40]. Instead of individually deciding what to forget and add, the CIFG-LSTM makes those decisions together to simplify the structure of the LSTM.
- To consider the spatial structure between words in the user's posts, we also introduce Tree-LSTM [41] to achieve the representation of words to sentences over parse tree structures rather than in a sequential way. For the depression detection task, we firstly use the Stanford CoreNLP [42] to do tokenization and split sentences on the RSDD datasets and generate dependency parses using Stanford Neural Network Dependency Parser. Then we use Tree-LSTM to obtain the post representations and LSTM to get the user's activity representations.
- We also introduce Bert [43] for the depression detection task and make some modifications. We use Bert to obtain the representation of posts that integrate the context semantics. All posts published by a user are then fed to LSTM to get the user's activity representations.

For the eRisk 2017 dataset, we choose the top methods [34] from the early detection of depression task as baselines and compare our methods against these baselines.

C. EXPERIMENTAL SETUP

The RSDD dataset consists of training, validation, and testing datasets, and each contains approximately 3,000 diagnosed users and 35,000 control users. We used the validation set to tune the hyperparameter of our models and the baselines. The eRisk 2017 dataset consists of training and testing sets. The training set of eRisk 2017 contains 83 depressed users and 403 control users, and the test set contains 52 depressed users and 349 control users.

TABLE 1. The statistical summary of the training datasets after tokenization.

Dataset	c	l	M	N	$\|V\|$
RSDD	2	148	969	38823	966881
eRisk2017	2	36	372	486	13608

TABLE 2. Optimal hyper-parameter configuration for two datasets.

Dataset	Maxm	Maxn	Embed_size	lr	Batch size
RSDD	600	100	50	0.001	64
eRisk 2017	400	30	100	0.001	128

The statistical summary of the training datasets after tokenization are shown in Table. 1. c: Number of target categories. l: Average length of a post. M: Average number of posts by users. N: Size of datasets. $|V|$: Vocabulary size.

The value of the hyperparameters of our model is shown in Table. 2. The RSDD validation set is used to select the depression detection model's hyperparameters, and the test set is used to report the results. We do not initialize the embedding layer with pre-trained embeddings such as publicly available Glove or Word2Vec. The input of the depression models is composed of original terms encoded as one-hot vectors. The input layer is then used to learn 50-dimensional and 100-dimensional embeddings of the terms (Embed_size). The learning rate (lr) is set to 0.001. For RSDD, eRisk 2017, we set the mini-batch size to 64, 128. We define *Maxm* to represent the maximum number of posts of each user and *Maxn* to represent the maximum number of tokens of each post. When one user's document which exceeds the maximum number of posts, we will shuffle the posts and randomly select posts with length *Maxm*. For example, We set our models receiving up to 600 posts ($Maxm = 600$) and a post containing up to 100 words ($Maxn = 100$) for each target user on the RSDD dataset. We have increased the maximum number of posts, but the performance on the validation data did not be improved significantly. Our proposed depression detection models are implemented in Keras.

For the post-level operation of the MGL-CNN, the window sizes of the first convolutional kernels (s) are set as 2, 3, 4, 5 and 6, with 30 different kernels for each window size. In the second convolutional layer, the window sizes of kernels (h) are 1, 3, 5, 7 with 30 different kernels each window size. We use convolution kernels of size 3 in the second layer of SGL-CNN. For the user-level operation of our model, we set the same parameters as the post-level operation. We don't perform any specific tuning on the datasets. Class balancing was performed with Categorical Cross Ent in our models, which uses a softmax function and categorical cross-entropy as its loss function. All models are trained using stochastic gradient descent with the Adam optimizer [44].

TABLE 3. Performance detecting depressed users on the RSDD test set.

Method	Precision	Recall	F1
BoW-MNB	0.44	0.31	0.36
Bow-SVM	0.72	0.29	0.42
Feature-rich-MNB	0.69	0.32	0.44
Feature-rich-SVM	0.71	0.31	0.44
User model-CNN	0.59	0.45	0.51
LSTM	0.50	0.39	0.44
Bi-LSTM	0.56	0.40	0.47
Tree-LSTM	0.54	0.41	0.47
CIFG-LSTM	0.51	0.37	0.43
Bert-LSTM	0.53	0.49	0.51
GRU-Attention	0.57	0.40	0.47
LSTM-Attention	0.54	0.35	0.42
Bi-LSTM-Attention	0.62	0.39	0.48
SGL-CNN	0.51	0.56	0.53
MGL-CNN	0.63	0.48	0.54

D. RESULTS

The results in RSDD for identifying depressed users from both our methods and other baselines are shown in Table. 3. The differences between our models and baselines are statistically significant (McNemar's test,$p < 0.05$). We compare our models against several baselines using MNB and SVM classifiers with two sets of features. Although the traditional methods SVM and MNB with rich features can achieve high precision, the performance on Recall and F1 are not good compared with the state-of-the-art User model-CNN and other popular models in NLP (e.g., CNN-based and LSTM-based methods). For instance, Feature-rich-SVM and Bow-SVM give outstanding performance 0.71 and 0.72 respectively on precision but only have performance 0.31 and 0.29 respectively on recall.

Besides, our models also gain competitive results over several popular models in natural language processing. The proposed model MGL-CNN achieved precision close to Bi-LSTM- Attention but performed better on recall and F1. Moreover, we can conclude from Table. 3 that the selected seven sequence models have achieved almost the same performances on the RSDD dataset (aside from Bi-LSTM-Attention). The Bi-LSTM-Attention model achieved the best performance among them. Compared to the User model-CNN, the precision increased of 5.1%, but the recall decreased. The bidirectional architecture can look forward and backward to capture bidirectional semantic dependency and improve the abilities of the memory. Therefore, Bi-LSTM has a better performance than single directional models. And experiments on the data also indicated that the attention mechanism could help LSTM and variants achieved good results in this task.

Compared with previous work, our proposed SGL-CNN model outperforms the state-of-the-art User model-CNN in terms of Recall and F1 on depressed users (increases of 24.4% and 3.9%, respectively). Besides, our proposed MGL-CNN model outperforms the User model-CNN in terms of Precision, Recall, and F1 on depressed users (increases of 6.8%, 6.7%, and 5.9%, respectively). We can find the comprehensive result of MGL-CNN is slightly better than SGL-CNN. Our proposed models can obtain an effective improvement over the User model-CNN and perform better than other strong baseline models. We believe that the Multi-Gated (Single-Gated) LeakyReLU unit can help CNN make full use of limited contextual information to obtain the critical features of the post's representation. For our model, the first convolutional layer can capture the n-gram features of the text. The gated unit with different kernels then obtains gating weights to effectively identify language associated with negative sentiment across a user's posts and suppress the impact of unimportant information.

TABLE 4. Performance detecting depressed users on the eRisk 2017 test set.

Team	Precision	Recall	F1
UNSLA [34]	0.48	0.79	0.59
FHDO-BCSGA [34]	0.61	0.67	0.64
FHDO-BCSGB [34]	0.69	0.46	0.55
GPLC [34]	0.42	0.50	0.46
TVT-NB [45]	0.42	0.73	0.54
TVT-RF [45]	0.54	0.58	0.56
SGL-CNN	0.56	0.59	0.57
MGL-CNN	0.63	0.57	0.60

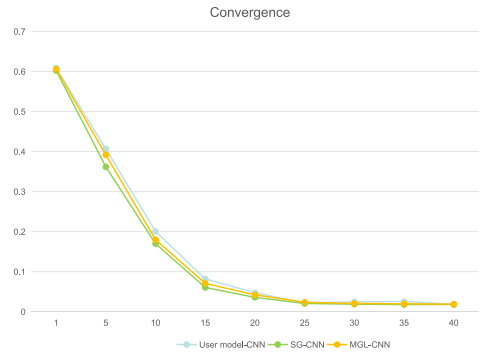


FIGURE 4. The convergence of models. The X-axis is the number of iterations and the Y-axis is the training losses.

The results on the Early Detection of Depression dataset for our models and the current best-performing methods are shown in Table. 4. The absolute values of the metrics from baselines illustrate that the early detection of depression task is difficult. In terms of F1, performance is low. The highest F1 is 0.64. Some methods, e.g., FHDO-BCSGB, opted for optimizing precision but had a low recall, while other methods, e.g., UNSLA, chose for optimizing recall but had low precision. This might be related to the scale and creation of the dataset. We can find that our proposed models (SGL-CNN and MGL-CNN) achieve performance close to several state-of-the-art methods in terms of Precision, Recall, and F1 on depressed users. Besides, our models are not aimed at improving one indicator like the baseline models, but perform well on all three indicators(Precision, Recall, and F1). Comparison between the result of our models and that

of the latest methods suggested that our proposed general neural network architecture can also be applied to the early detection of depression in different online forums. Besides, as shown in Fig. 4, the comparison is based on the changing of training loss within the 40 epochs. We can find that they have almost the same convergence speed, and the results of our models are even slightly better than the state-of-the-art User model-CNN model, indicating that although our models are more complex, the convergence speed does not decrease.

V. CONCLUSION

In this work, we proposed two hierarchical posts representations models for identifying depressed individuals, which was more accurate and efficient than general early depression detection models. The proposed models can effectively represent the user's overall emotional state through their posts. We applied our models on the large-scale Reddit Self-reported Depression Diagnosis dataset and found that it substantially outperformed strong existing methods in terms of Precision, Recall, and F1. However, the absolute values of the metrics illustrate that depression detection on large-scale datasets in social media is still a challenging task and worthy of further exploration. And for demonstrating that our models focus on learning representations of the user's posts from different online forums, we also applied our models on the Early Detection of Depression dataset. We found that it also achieved performance close to strong previously-proposed methods.

Our work is significant from several perspectives: we provide strong models to identify depressed users on social media and a method for large-scale public mental health studies about depression, and do a more in-depth study of the close connection between social media and mental health; we demonstrate the possibility of sensitive applications in combining clinical care with users' online activities, where doctors could be notified and help in time if the activities of user suggest they have symptoms of depression. For future work, we will explore the application of MGL-CNN and SGL-CNN to general document-level sentiment analysis.

ACKNOWLEDGMENT

The authors would like to thank all authors' contributions to this work. Besides, they would like to thank the anonymous reviews for their valuable comments.

REFERENCES

[1] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Multimodal assistive technologies for depression diagnosis and monitoring," *J. Multimodal User Inter.*, vol. 7, no. 3, pp. 217–228, Nov. 2013.

[2] World Health Organization. (2016). *Out of the Shadows: Making Mental Health a Global Development Priority*. [Online]. Available: https://www.who.int/mental_health/advocacy/WB_event_2016/en/

[3] R. F. Muñoz, P. J. Mrazek, and R. J. Haggerty, "Institute of medicine report on prevention of mental disorders: Summary and commentary.," *Amer. Psychologist*, vol. 51, no. 11, pp. 1116–1122, Sep. 2005.

[4] H. Aron, "Depression: The benefits of early and appropriate treatment," *Amer. J. Managed Care*, vol. 13, no. 4, p. 92, 2007.

[5] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in Reddit social media forum," *IEEE Access*, vol. 7, pp. 44883–44893, 2019.

[6] J. A. McGillivray and M. P. McCabe, "Early detection of depression and associated risk factors in adults with mild/moderate intellectual disability," *Res. Developmental Disabilities*, vol. 28, no. 1, pp. 59–70, Jan. 2007.

[7] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu, "Depression detection via harvesting social media: A multimodal dictionary learning solution," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 3838–3844.

[8] M. J. Friedrich, "Depression is the leading cause of disability around the world," *JAMA*, vol. 317, no. 15, p. 1517, Apr. 2017.

[9] A. J. Ferrari, R. E. Norman, G. Freedman, A. J. Baxter, J. E. Pirkis, M. G. Harris, A. Page, E. Carnahan, L. Degenhardt, T. Vos, and H. A. Whiteford, "The burden attributable to mental and substance use disorders as risk factors for suicide: Findings from the global burden of disease study 2010," *PLoS ONE*, vol. 9, no. 4, Apr. 2014, Art. no. e91936.

[10] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in *Proc. Workshop Comput. Linguistics Clin. Psychol., From Linguistic Signal Clin. Reality*, 2014, pp. 51–60.

[11] A. H. Yazdavar, H. S. Al-Olimat, M. Ebrahimi, G. Bajaj, T. Banerjee, K. Thirunarayan, J. Pathak, and A. Sheth, "Semi-supervised approach to monitoring clinical depressive symptoms in social media," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, 2017, pp. 1191–1198.

[12] C. Zucco, B. Calabrese, and M. Cannataro, "Sentiment analysis and affective computing for depression monitoring," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2017, pp. 1988–1995.

[13] G. Coppersmith, K. Ngo, R. Leary, and A. Wood, "Exploratory analysis of social media prior to a suicide attempt," in *Proc. 3rd Workshop Comput. Lingusitics Clin. Psychol.*, 2016, pp. 106–117.

[14] M. De Choudhury and E. Kiciman, "The language of social support in social media and its effect on suicidal ideation risk," in *Proc. Int. AAAI Conf. Web Social Media*, Mar. 2017, pp. 32–41.

[15] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annu. Rev. Psychol.*, vol. 54, no. 1, pp. 547–577, Feb. 2003.

[16] D. L. Mowery, A. Park, C. Bryan, and M. Conway, "Towards automatically classifying depressive symptoms from Twitter data for population health," in *Proc. Workshop Comput. Model. People's Opinions, Personality, Emotions Social Media (PEOPLES)*, 2016, pp. 182–191.

[17] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Proc. 7th Int. AAAI Conf. weblogs social media*, 2013, pp. 128–137.

[18] M. A. Moreno, L. A. Jelenchick, K. G. Egan, E. Cox, H. Young, K. E. Gannon, and T. Becker, "Feeling bad on Facebook: Depression disclosures by college students on a social networking site," *Depress. Anxiety*, vol. 28, no. 6, pp. 447–455, Jun. 2011.

[19] A. Yates, A. Cohan, and N. Goharian, "Depression and self-harm risk assessment in online forums," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1–14.

[20] Q. Cong, Z. Feng, F. Li, Y. Xiang, G. Rao, and C. Tao, "X-A-BiLSTM: A deep learning approach for depression detection in imbalanced data," in *Proc. IEEE Int. Conf. Bioinformatics Biomed. (BIBM)*, Dec. 2018, pp. 1624–1627.

[21] H. A. Schwartz, J. Eichstaedt, M. L. Kern, G. Park, M. Sap, D. Stillwell, M. Kosinski, and L. Ungar, "Towards assessing changes in degree of depression through Facebook," in *Proc. Workshop Comput. Linguistics Clin. Psychol., From Linguistic Signal Clin. Reality*, 2014, pp. 118–125.

[22] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[23] P. Thompson, C. Bryan, and C. Poulin, "Predicting military and veteran suicide risk: Cultural aspects," in *Proc. Workshop Comput. Linguistics Clin. Psychol., From Linguistic Signal Clin. Reality*, 2014, pp. 1–6.

[24] S. Malmasi, M. Zampieri, and M. Dras, "Predicting post severity in mental health forums," in *Proc. 3rd Workshop Comput. Lingusitics Clin. Psychol.*, 2016, pp. 133–137.

[25] C. Brew, "Classifying ReachOut posts with a radial basis function SVM," in *Proc. 3rd Workshop Comput. Lingusitics Clinical Psychol.*, 2016, pp. 138–142.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] T. Gui, L. Zhu, Q. Zhang, M. Peng, X. Zhou, K. Ding, and Z. Chen, "Cooperative multimodal approach to depression detection in Twitter," *AAAI*, vol. 33, pp. 110–117, Sep. 2019.

[28] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," 2015, *arXiv:1512.01100*. [Online]. Available: https://arxiv.org/abs/1512.01100

[29] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*. [Online]. Available: http://arxiv.org/abs/1408.5882

[30] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," 2017, *arXiv:1612.08083*. [Online]. Available: https://arxiv.org/abs/1612.08083

[31] Y. Liu, L. Ji, R. Huang, T. Ming, C. Gao, and J. Zhang, "An attention-gated convolutional neural network for sentence classification," *IDA*, vol. 23, no. 5, pp. 1091–1107, Oct. 2019.

[32] W. Xue and T. Li, "Aspect based sentiment analysis with gated convolutional networks," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 1–10.

[33] D. Tang, Q. Bing, and T. Liu, "Learning semantic representations of users and products for document level sentiment classification," in *Proc. Meeting Assoc. Comput. Linguistics Int. Joint Conf. Natural Lang. Process.*, 2015, pp. 1014–1023.

[34] D. E. Losada, F. Crestani, and J. Parapar, "eRISK 2017: CLEF lab on early risk prediction on the Internet: Experimental foundations," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Cham, Switzerland: Springer, 2017.

[35] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2012, pp. 90–94.

[36] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," Univ. Texas Libraries, Austin, TX, USA, Tech. Rep. LIWC2015, 2015.

[37] J. Staiano and M. Guerini, "DepecheMood: A lexicon for emotion analysis from crowd-annotated news," 2014, *arXiv:1405.1605*. [Online]. Available: http://arxiv.org/abs/1405.1605

[38] A. Graves, N. Jaitly, and A.-R. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2013, pp. 273–278.

[39] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 1480–1489.

[40] J. Chung, C. Gulcehre, K. H. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*. [Online]. Available: https://arxiv.org/abs/1412.3555

[41] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process.*, vol. 1, 2015, pp. 1–11.

[42] D. Chen and C. Manning, "A fast and accurate dependency parser using neural networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 740–750.

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[44] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

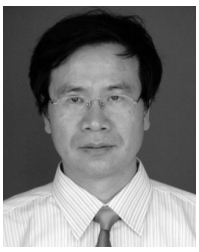
[45] M. P. Villegas, D. G. Funez, M. J. G. Ucelay, L. C. Cagnina, and M. L. Errecalde, "LIDIC-UNSL's participation at ERISK 2017: Pilot task on early detection of depression," in *Proc. CLEF*, 2017, p. 1.

GUOZHENG RAO (Member, IEEE) received the Ph.D. degree in computer applications technology from Tianjin University, in 2009. He is currently an Associate Professor with the College of Intelligence and Computing, Tianjin University. He also works with the Tianjin Key Laboratory of Cognitive Computing and Applications and School of New Media and Communication, Tianjin University. His research interests include sentiment analysis and knowledge engineering. He is a member the ACM and China Computer Federation (CCF).

YUE ZHANG received the B.Sc. degree in software engineering from Liaoning University. She is currently pursuing the master's degree with Tianjin University. Her research is in the field of sentiment analysis.

LI ZHANG received the Ph.D. degree in management science and engineering from Tianjin University, in 2006. She is currently an Associate Professor with the School of Economics and Management, Tianjin University of Science and Technology. Her research interests include sentiment analysis and supply chain management.

QING CONG received the B.Sc. degree in law and M.Eng. degree in computer technology from Tianjin University, where he is currently pursuing the Ph.D. degree. His major research involves sentiment analysis.



ZHIYONG FENG (Member, IEEE) received the Ph.D. degree in mechanical Engineering from Tianjin University, in 1996. He is currently a Professor of computer software and the Associate Dean of the School of Computer Software, Tianjin University. His major researches involve software engineering, knowledge engineering, and distributed software. He is a Distinguished Member of the China Computer Federation (CCF), and serves as the Deputy Director of Service Computing Committee of CCF, the Executive Member of Education Committee of CCF, and a member of the Software Engineering Committee of CCF, and the Chairman of the ACM China Tianjin Chapter.