Li *et al.*
Health Information Science and Systems (2023) 11:6
https://doi.org/10.1007/s13755-022-00197-5

Health Information Science and Systems

RESEARCH

MHA: a multimodal hierarchical attention model for depression detection in social media

Zepeng Li[1], Zhengyi An[1], Wenchuan Cheng[1], Jiawei Zhou[1], Fang Zheng[1] and Bin Hu[1,2,3*]

Abstract

As a serious mental disease, depression causes great harm to the physical and mental health of individuals, and becomes an important cause of suicide. Therefore, it is necessary to accurately identify and treat depressed patients. Compared with traditional clinical diagnosis methods, a large amount of real and different types of data on social media provides new ideas for depression detection research. In this paper, we construct a depression detection data set based on Weibo, and propose a Multimodal Hierarchical Attention (MHA) model for social media depression detection. Multimodal data is fed into the model and the attention mechanism is applied within and between modalities at the same time. Experimental results show that the proposed model achieves the best classification performance. In addition, we propose a distribution normalization method, which can optimize the data distribution and improve the accuracy of depression detection.

Keywords: Deep neural network, Social media, Depression detection, Attention mechanism, Multimodality

Introduction

According to the statistics of the World Health Organization (WHO), more than 300 million people in the world were suffering from depression in 2017, which were equivalent to 4.4% of the world population. The number of depressed patients increased by 18.04% in 10 years [1]. At present, the incidence of depression tends to be younger, and patients with depression are more prone to violent injury to others or self-harm, suicide and other bad situations than normal people, which has a serious impact on personal, family and socio-economic development. The clinical diagnosis of depression mainly depends on questionnaires or scales [2–5], but these methods have certain problems. For example, patients often conceal their true thoughts during the filling process.

With the development of the Internet, the scale of social media is expanding. On social media platforms such as Weibo, Twitter and Facebook, hundreds of millions of users share their views and life states every day, including text, pictures, videos, audios, etc., which usually contain rich emotional information. Compared with questionnaires and scales, depressed patients are more likely to express their true feelings in social media. These open and real data provide a new perspective for depression detection research.

Most depression detection studies are based on the text posted by users on social platforms to obtain users' emotional state. However, some literatures have proved that simply analyzing users' posts cannot detect their depressive tendencies very accurately [6, 7]. In addition, Chinese expressions are more diverse, so these methods often cannot obtain users' potential real emotion in Chinese depression detection. Aiming at this problem, some studies are committed to using multimodal data for depression detection [8–11]. Figure 1 is an example of pictures posted by normal user and depression user. It can be seen that there are certain differences in the pictures posted by them, so we can consider using multimodal information such as pictures for depression detection. Moreover, social media users usually express their depressive tendencies in only a small part of the data. If all data of users are considered to be of equal importance, it is possible to ignore important

*Correspondence: bh@lzu.edu.cn
[1] Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, Gansu, China
Full list of author information is available at the end of the article

Fig. 1 Example of pictures posted by normal user and depression user, where **a** represents the picture of normal user and **b** represents the picture of depression user

information that affects the classification of the model, resulting in the wrong prediction results given by the model. Therefore, it is necessary to introduce attention mechanism into depression detection, which can enable the model to automatically screen out the information that plays an important role in the prediction results [12–14]. However, most of the current depression detection studies based on multimodality only use the attention mechanism within a certain type of data. In this way, the contribution of a certain mode to the overall classification effect cannot be measured, and the model cannot be adjusted adaptively according to the proportion of information contribution.

In the task of social media user classification, there may be great differences in the average data volume of different categories of users. For example, when using social media data for depression detection, the number of posts and pictures posted by depressed users is generally less than that of normal users. The existing solution is to delete the data of each user to a certain threshold [15]. Although this method can reduce the difference between the average data volume of different categories of users to a certain extent, by comparison, for a class of users with a large average data volume, the actual data volume retained is closer to the threshold. This may induce the model to classify only according to the difference of the actual retained data volume, and it does not extract the real information in the data actually.

The lack of publicly available data sets for research is also one of the problems that the field of social media depression detection. The particularity of depression detection and the privacy policy of social media platform make researchers have to collect and use user data in range of limited privacy, which also leads to the lack of publicly available data sets. Meanwhile, the existing depression detection data set based on social media is mainly in English, while the Chinese data sets are very rare, which greatly limits the research in Chinese social media depression detection fields.

In response to the above problems, we propose a Multimodal Hierarchical Attention (MHA) model for social media depressed user detection, which can process text, pictures and auxiliary information at the same time. The proposed model uses a novel hierarchical attention mechanism, which can not only automatically screen out the information that plays an important role in the prediction results in the same modal data, but also adjust the model adaptively according to the information contribution ratio of different modal data. Since the average data volume of depressed users is much smaller than that of normal users, we propose a distribution normalization method to make the model learn more features of the

data by changing the data distribution. This method randomly deletes the pictures of the normal users according to the number of pictures of the corresponding depression users who release fewer pictures, so that the number of pictures of the two types of users is consistent. In addition, we construct a Chinese social media data set for depression detection, which contains data of 2299 users with depressive tendencies and 2307 normal users from December 24, 2020 to December 23, 2021. The experimental results show that compared with the baseline, our MHA model achieves the best performance in the Chinese social media depression detection task. Meanwhile, the proposed distribution normalization method can reduce the difference of data distribution and further improve the performance of the model.

In general, our work has the following contributions:

- We propose a multimodal hierarchical attention model, called MHA. The model can process the data of different modalities at the same time, and combine the attention mechanism to screen the information that plays an important role in the task of depression detection within and between modalities.
- Aiming at the problem that there may be great differences in the average data volume of different categories of users, we introduce a distribution normalization method, which improves the performance of the model by aligning the data distribution.
- Due to the lack of domain-specific data sets, we construct a Chinese social media data set for depression detection research. This data set would help to promote further research related to depression detection in the field of computer science and psychology.

Related work

The current research on depression detection of social media users is mainly divided into two categories: machine learning and deep learning methods. Multimodal-based depression detection tasks basically use deep learning methods that are jointly trained by multiple models. In addition, attention mechanism can help the model automatically screen out information that plays a greater role in classification. Therefore, the depression detection method using attention mechanism is also a research hotspot in related fields.

Machine learning-based methods for depressed user detection mainly rely on feature engineering, that is, the features extracted from text are fed into machine learning models for training and predicting. Priya et al. [16] used machine learning algorithms to detect a variety of psychological problems such as anxiety, depression and stress for users who filled in the online questionnaire. Tian et al. [17] compared the identified depressed users with the general population in demographic features, circadian rhythm and emoticon. Seabrook et al. [18] analyzed the relationship between the severity of depression and the expression of emotional words. Deep learning models can extract features automatically without complex operations compared with machine learning methods. Wang et al. [19] used deep learning methods such as Bert, Roberta and XLNet based on the pre-trained language expression model for depression risk prediction, and further trained on the large-scale unlabelled data set collected from Weibo. In order to get better results, researchers usually use information other than text as features. Cao et al. [20] used FastText model as the embedding of user text, and used LSTM for sequential processing to obtain the context of user posts. He et al. [21] detailed the deep learning method for depression detection, and discussed the challenges and prospects of using deep learning technology for depression diagnosis.

Most machine learning and deep learning methods only use unimodal data, resulting in limited detection performance. Therefore, there are some works trying to use multimodal data for depression detection. The research of Koutsouleris et al. [22] showed that the combination of clinical, neurocognitive, neuroimaging and genetic information could help to improve the performance of depression detection. Cai et al. [23] provided users with positive, neutral and negative audio stimuli, and then extracted linear and nonlinear features from EEG data of three modes to distinguish patients with depression from normal people. Ceccarelli et al. [24] proposed a late multimodal fusion strategy based on feedforward neural network, which combines audio, video and text information at the same time. In view of the lack of labelled depression audio data sets, Toto et al. [25] introduced AudiBERT, which integrates the pre-trained audio and text representation model, and uses the dual attention mechanism to strengthen their representation respectively. Sardari et al. [26] used CNN autoencoder to extract features from original audio, which is better than manual feature extraction methods and other deep learning methods in this field.

In recent years, attention mechanism has been widely used in various fields. Many deep learning models combined with attention mechanism have shown excellent performance in affective computing tasks such as depression detection. Some studies have combined attention mechanism with the text information to detect depression. Song et al. [13] constructed four FAN models with good performance and high interpretability based on psychological research, and used the data on social media for depression detection. Ren et al. [14] proposed an attention network that can discover the high-level semantic

information and emotional information hidden in the text, and verified the effectiveness of the model by using the data in the social network. Mallol-Ragolta et al. [27] and Xezonaki et al. [28] used a hierarchical attention-based model to extract language features from clinical visit records to detect whether individuals have depression. There are some works that applied attention mechanism to other types of information. For example, He et al. [29] used CNN model with attention mechanism to mine potential depression patterns in facial images. Zhang et al. [30] used attention mechanism to explore the correlation between EEG signals and demographic information. In addition, some studies try to apply attention to multimodal information at the same time. Zheng et al. [31] and Niu et al. [32] used graph attention network to learn the embedded representation of different modal information to find the correlation between modalities.

Methodology

The structure of the proposed MHA model is shown in Fig. 2. It focuses on combining multimodal data and attention, using attention mechanisms both within and between modalities, discovering more important data within modalities, and assigning different importance to different modalities.

Problem definition

For user $u_i \in U, i = 1, 2, \ldots, N$, post sequence of the user is $p_i^j \in P_i, j = 1, 2, \ldots, M_i$, and the picture sequence is $c_i^l \in C_i, l = 1, 2, \ldots, L_i$, where N represents the number of users, M_i represents the number of posts of user u_i, and L_i represents the number of pictures of user u_i. The post time, dictionary features and social information of user u_i are expressed as t_i,d_i and s_i respectively. In this paper, depressed user detection is regarded as a binary classification task, that is, user label $y_i \in \{0, 1\}, i = 1, 2, \ldots, N$. The purpose of this study is to give the prediction result of users' depressive tendency $\bar{y}_i$ according to users' social media data.

Feature description

Compared with traditional machine learning methods, the most significant advantage of deep learning methods is that it can obtain higher accuracy when processing large-scale data, and there is no need to extract features manually, so they have been widely used in many fields [21, 33]. Some studies have shown that the integration of auxiliary information into deep learning methods can further improve the classification performance [34]. Therefore, this work extracts a variety of auxiliary information, including dictionary features, post time and social information, to help further improve the performance of the model.

Table 1 Social information statistics of different categories of users

	Normal group	Depression group	Ratio difference (%)
Avg_reposts	33.52	36.23	− 7.48
Avg_comments	263.23	220.98	19.12
Avg_likes	386.19	239.09	61.52
Avg_followings	358.11	285.37	25.49
Avg_followers	406.74	224.30	81.34
Avg_posts	135.09	61.27	120.48

Ratio difference = (Normal group / Depression group − 1) × 100%

At present, there are some works combined with emotion dictionary to identify depression [35]. The Chinese suicide dictionary [36] has a total of 2168 words divided into 13 categories. We select 21 antidepressants and 153 words related to depressive symptoms, and add them to the dictionary. According to the expanded dictionary, a 13-dimensional feature vector is extracted from all posts of each user.

According to our observation, the post time of depressed users is different from that of normal users. Depressed users are more likely to post at midnight or before dawn. Therefore, we regard post time as a feature of depression discrimination. Specifically, we divide the post time into six periods and count a 6-dimensional post time feature for each user.

In addition, users' social information may also help to detect depressive tendencies. Therefore, we take it as part of the auxiliary information. The social information used in this paper includes the number of Weibo posts, reposts, comments, likes, followings and followers. Table 1 shows the differences in social information between normal and depressed users. Based on users' social information, we extract a 6-dimensional feature vector for each user.

Multimodal hierarchical attention model

In this paper, we use the data of three modalities: text, image and auxiliary information. Depressed users may not continue to show their depressive tendencies, so screening out information that clearly express depressive tendencies will be helpful to classification. This work applies the attention mechanism to pictures and auxiliary information. Through the attention mechanism, our model can pay more attention to the information that plays a positive role in identifying depressive tendency. Since the importance of different modal data may be different, we also apply the attention mechanism to the data of different modalities.

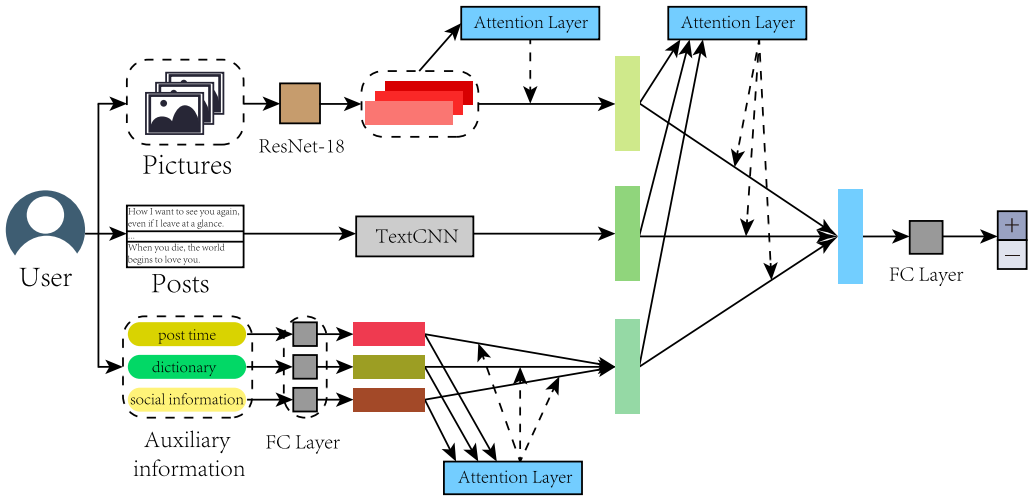

Fig. 2 The structure of MHA model

In order to screen out the features that are helpful to identify depressive tendency within and between modalities at the same time, we propose the MHA model. ResNet-18 uses residual learning for the first time, effectively alleviating the model degradation problem in deep learning [37]. It uses ImageNet data set for pre-training. Therefore, we use the pre-trained ResNet-18 to extract a feature vector for each picture c_i^l of user u_i:

$$\tilde{c}_i^l = f_{\text{ResNet}}\left(c_i^l\right) \tag{1}$$

Then the attention mechanism is applied to the extracted picture features:

$$\widetilde{C}_i = \sum_{l=1}^{L} \alpha_i^l \tilde{c}_i^l \tag{2}$$

$$\alpha_i^l = \frac{\exp\left(\tilde{c}_i^l \times W_1 + b_1\right)}{\sum_{k=1}^{L} \exp\left(\tilde{c}_i^k \times W_1 + b_1\right)} \tag{3}$$

where $\widetilde{C}_i \in \mathbb{R}^{1\times 512}$ represents the final picture feature of user u_i, α_i^l represents the attention weight of each picture feature $\tilde{c}_i^l$, $W_1 \in \mathbb{R}^{512\times 512}$ and $b_1 \in \mathbb{R}^{1\times 512}$ represent trainable full connection layer parameters.

For text data, we concatenate all posts of user u_i into a long text, and train a TextCNN model to extract text features:

$$\widetilde{P}_i = f_{\text{TextCNN}}\left(p_i^1 \mid p_i^2 \mid \ldots \mid p_i^{M_i}\right) \tag{4}$$

where $\widetilde{P}_i \in \mathbb{R}^{1\times dim}$ represents the text feature of user u_i, dim represents dimension of text feature.

For the auxiliary information t_i, d_i and s_i, we first map them to the feature space using full connection layers respectively:

$$\widetilde{t}_i = W_2 \times t_i + b_2 \tag{5}$$

$$\widetilde{d}_i = W_3 \times d_i + b_3 \tag{6}$$

$$\widetilde{s}_i = W_4 \times s_i + b_4 \tag{7}$$

where $\widetilde{t}_i$, $\widetilde{d}_i$ and $\widetilde{s}_i$ represent the feature vectors corresponding to the three auxiliary information respectively, W_2, W_3, W_4 and b_2, b_2, b_2 represent the trainable full connection layer parameters. The attention mechanism is then applied to each feature:

$$\widetilde{A}_i = \sum_{k=1}^{3} \beta_i^k a_k, a_k \in \left\{\widetilde{t}_i, \widetilde{d}_i, \widetilde{s}_i\right\} \tag{8}$$

$$\beta_i^k = \frac{\exp\left(W_5 \times a_k + b_5\right)}{\sum_{m=1}^{3} \exp\left(W_5 \times a_m + b_5\right)} \tag{9}$$

where $\widetilde{A}_i$ represents the feature vector of all auxiliary information of user u_i, β_i^k represents the attention weight corresponding to the three auxiliary information, W_5 and b_5 represent trainable full connection layer parameters.

After obtaining the feature vector of the three modalities, we use the attention mechanism again to obtain the final vector representation of user u_i:

$$\widetilde{F}_i = \sum_{k=1}^{3} \gamma_i^k e_k, e_k \in \left\{\widetilde{C}_i, \widetilde{P}_i, \widetilde{A}_i\right\} \tag{10}$$

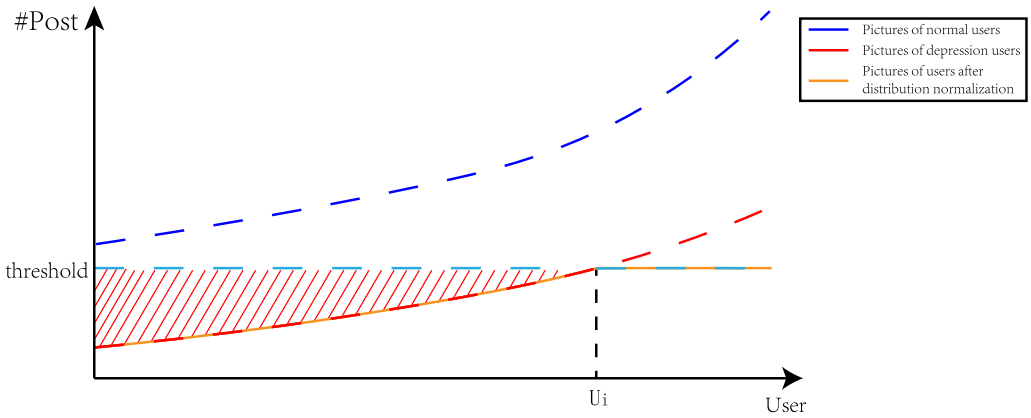

Fig. 3 Schematic diagram of distribution normalization method

$$\gamma_i^k = \frac{\exp\left(W_6 \times e_k + b_6\right)}{\sum_{m=1}^{3} \exp\left(W_6 \times e_m + b_6\right)} \tag{11}$$

where $\widetilde{F}_i$ represents the final vector representation of user u_i, γ_i^k represents the attention weight corresponding to the three modal vectors, W_6 and b_6 represent the trainable full connection layer parameters. Finally, a full connection layer is used to classify $\widetilde{F}_i$, and the depressive tendency prediction results of user u_i is obtained:

$$\tilde{y}_i = W_7 \times \widetilde{F}_i + b_7 \tag{12}$$

Distribution normalization

Due to the large difference in the number of pictures posted by social media users, if all pictures of users are used for classification, the model may directly obtain the classification results through the feature of picture number. That is, the model does not learn the feature of the data actually, which would cause the model to classify only by the difference in the number of pictures. In order to adress this problem, we set an upper threshold for the number of pictures. Specifically, for users whose number of published pictures is less than the threshold, we keep all their picture data. For users whose number of pictures exceeds the threshold, we sort the pictures according to the publishing time, and only keep the latest pictures of the user. As shown in Fig. 3, the original picture distribution of normal users and depressed users is shown as dark blue and red dotted lines respectively. After setting the upper threshold, the picture distribution of the two types of users will change to light blue and orange dotted lines respectively.

It can be seen that before the u_i point, the number of pictures of normal users is equal to the threshold, and the number of pictures of depressed users is less than the threshold, which will affect the calculation of picture attention weight by the model. In order to solve this problem, we propose a distribution normalization method, which makes the distribution of number of normal users' pictures consistent with that of depressed users. Specifically, normal users and depressed users are sorted in ascending order according to the number of pictures firstly, then one normal user and one depressed user are selected each time, and another corresponding user's pictures are randomly deleted according to the number of the user's pictures who has posted a small number of pictures. After the above operations, the distribution of number of the two groups of users' pictures tends to be consistent, as shown by the orange dotted line in Fig. 3.

After ranking the users in the normal group and depression group in ascending order according to the number of posted pictures, we find that the number of pictures of users in the normal group selected each time is greater than that of users in the corresponding depression group. Therefore, we only normalize the distribution of picture data of normal group users. Table 2 shows the statistics of the number of users' pictures in the normal group before and after using the distribution normalization method.

Table 2 Statistics on the number of pictures released by normal group users before and after distribution normalization

	Before	After
Average	131.39	49.50
Median	61	14
Sum	303,109	114,190

Data set

Social media depression detection tasks often face the problem of lack of available public data sets. An important reason is that researchers are worried about divulging users' private information, and the sensitivity of depression topics further exacerbates the scarcity of data sets. Therefore, taking Sina Weibo as the data source, we construct a Chinese social media data set for depression detection research, and desensitize all contents that may leak personal privacy.

Data source

Sina Weibo[1] is a social media platform owned by Sina. Since its launch in August 2009, the number of Weibo users has maintained explosive growth. In the fourth quarter of 2021, the number of monthly active users reached 573 million, a year-on-year increase of 10%, and the number of daily active users reached 249 million, a year-on-year increase of 11%.

Weibo has a sub-forum function called 'SuperTopic', and users can express their views in a certain sub-forum. In order to screen users with potential depressive tendencies, we obtain all text data from December 21, 2020 to December 20, 2021 in the 'Depression SuperTopic' subforum, and identify 2426 users with depression by manual annotation. We obtain all posts and corresponding image data in their personal home page. Finally, we get 167,586 original posts and 119,812 corresponding pictures without preprocessing.

In order to compare with the users in the depression group, we randomly select some users from the Weibo, and after annotating, we get 2400 users without depressive tendencies as the control group. Similarly, we obtain a total of 366,663 posts and 320,458 pictures on the personal home page of all users in the control group within the same time range.

Data annotation

For the user who may have depressive tendency obtained from 'Depression SuperTopic' and the randomly obtained control group users, we use the method of manual annotation to select the data that can be used to detect the depressive tendency. We invite three experienced psychologists to complete the annotation work. Generally speaking, for a specific user, experts need to judge whether the user is suffering from depression based on all the posts obtained from the user. In order to ensure the correct annotation, each user's posts are marked by three experts independently, and only when all experts give the same annotation results to a user, we add the user to the corresponding data set. In the screening process, we annotate the data according to the following criteria:

(1) depressive tendency present:

- One or several Weibo posts of the user in 'Depression SuperTopic' contain a clear diagnosis of depression, such as: 'I was diagnosed with major depression today', 'It has been a week since I was diagnosed with depression', etc.
- The user's posts clearly express that he/she is taking antidepressant drugs or receiving antidepressant treatment, such as: 'This time the doctor prescribed agomelatine', 'I will have my first MECT treatment next month.' et al.

(2) depressive tendency absent:

- None of the user's posts contain any words related to depression.
- There are depression related words in posts, but there is not enough evidence to show that the user has depressive tendency. For example, the user objectively evaluates the social news related to depression, or there are depression related keywords in shared lyrics or movie lines.

Table 3 shows the post examples of some annotated users.

Data preprocessing

Data preprocessing can eliminate redundant features and noise in the input text and improve the performance of the model. For all the posts and image data of 2426 depressed users and 2400 normal users, we propose a complete set of Chinese social media data preprocessing method. Specifically, we process all data according to the preprocessing operation listed below:

(1) In this paper, ResNet-18 is selected to extract the feature contained in users' image data. In order to conform the input structure of ResNet-18 network, we adjust the size of all images to 224×224.
(2) Due to the characteristics of Weibo, posts will be automatically generated when users share specific content. These posts should be identified and deleted. We screen out the keywords contained in these texts, and use these words to delete the corresponding posts. If the deleted post has corresponding pictures, we will delete these pictures.
(3) Delete URL links and all emoticons that appear in the post.

[1] https://weibo.com

Table 3 Examples of Weibo posts with and without suicidal ideation in 'SuperTopic'

Category	Example
Depression	Will the COVID-19 vaccine lead to the recurrence of depression? I am so afraid of a relapse of depression.
	I was diagnosed with moderate depression and mild anxiety some time ago, but I'm only 13 years old, and I'm just on my first year of junior high.
	Today is the ninth MECT. Lying on the operating table, anesthesia was injected into my body. At that moment, my body was numb, but my mind was clear. This is how I feel for the past two years. As I lay in a coffin, a sword pierced my heart.
Normal	I don't like being accosted by men, which makes me feel uncomfortable instinctively. If I am in a crowded place, I will be afraid.
	Although my parents didn't make everything I had go well, at least my values and moral cultivation are enough to be a useful person in this society, so that I can constantly expand my horizons and enjoy the freedom and beauty they yearn for.
	Taking a nap is really too easy to dream, and every time I have strange dreams.

Table 4 Statistics of normal and depression group data

Category	#Users	#Posts	#Pictures	#Length	Avg_Post	Avg_Length	Avg_Picture
Depression	2299	140,863	114,078	4,451,881	61.27	31.60	49.62
Normal	2307	311,645	303,109	10,951,214	135.09	35.14	131.39

#Users represents the number of users, #Posts represents the number of posts, #Pictures represents the number of pictures, #Length represents the total length of the post, Avg_Post represents the average number of user's posts, Avg_Length represents the average length of each user's posts, Avg_Picture represents the average number of pictures of users, Depression is the depression group data set, and Normal is the normal group data set

(4) Delete the content mentioned in the post through the '@' symbol that may cause privacy disclosure.
(5) We constructed a list of administrative divisions to match the user's current geographic location information appearing in the body of some posts and delete them.
(6) Delete all 'SuperTopic' titles appearing in posts.
(7) Delete the posts and corresponding pictures with less than 4 characters in a single post after the above operations.
(8) Delete users with too few remaining posts. We set the minimum number of posts to 3. Through the above preprocessing steps, we finally obtain 140,863 posts and corresponding 114,078 pictures of 2299 users with depressive tendencies, and for 2307 normal users, we obtain a total of 311,645 posts and corresponding 303,109 pictures. Table 4 shows the statistics of the two groups of user data.

Experiments

Experimental setup

The data set used in the experiments contains text, pictures and auxiliary information of 4606 social media users. This is a relatively balanced data set, with 2299 depressed users and 2307 normal users, respectively. We split all user data into train, validation, and test sets in a ratio of 8:1:1. Specifically, the train set contains data of 1839 depressed users and 1845 normal users, the validation set and test set both consist of data of 230 depressed users and 231 normal users.

For each user, we intercept the first 100 characters of each of his/her post and splice the intercepted content, and use 'jieba,'[2] a current relatively mature Chinese word segmentation tool, to segment the spliced text. In order to avoid the interference of some words with very low frequency to the model training, we set the size of vocabulary to 5000, and the words with very low frequency are replaced with '<UNK>'. In addition, we also add some words related to depressive symptoms and antidepressants to the word segmentation dictionary to obtain more accurate word segmentation results. The word vector corresponding to each word is initialized with a random number.

The model is implemented using PyTorch 1.10 framework, and the model is trained using a single NVIDIA A100 Tensor Core GPU. We apply grid search to determine the optimal parameters of the model, train the model for 500 epochs, and use an early-stop strategy to prevent model overfitting.

Experimental process

Firstly, the user's segmented text, all pictures and auxiliary information are sent to the proposed MHA model. The text data is sent into TextCNN model to obtain the vector representation of the text. Each picture uses the

[2] https://github.com/fxsjy/jieba.

pre-trained ResNet-18 to extract features, and combine with the attention mechanism to obtain the vector representation of the user's pictures. Three full connection layers are used separately to extract hidden layer vectors for the three types of auxiliary information: posting time, dictionary features and social information, and then an attention layer is applied to get the vector representation of auxiliary information. After obtaining the vector representation of text, picture and auxiliary information, the attention mechanism is combined again to obtain the vector representation of the user, and the final prediction result of the model is output after the full connection layer.

In addition, we feed the user's text into five widely used text classification models: TextCNN, DPCNN, FastText, Bert, and Transformer for depression detection experiments to show the superiority of the MHA model. We also try to deal with user images without using the distribution normalization method, and use the original data directly to prove the effectiveness of our proposed distribution normalization method.

Benchmark

In order to verify the effectiveness of the proposed MHA model, we select five widely used text classification models for depression detection experiments. The input of the model is the same as the user text used in the MHA model, and the output is whether the user has depressive tendency.

- *TextCNN* [38] The classical TextCNN is mainly composed of embedding layer, convolution layer and pooling layer. It is a shallow neural network. CNN was first applied to image classification and target detection in the field of computer vision, and then used for text classification and other tasks, which can achieve high classification accuracy. The word vector of TextCNN used in the experiment is initialized with random numbers.
- *FastText* [39] FastText superimposes the n-gram vector of words to obtain the representation of text. The structure of the model is very simple, which is only composed of input layer, hidden layer and output layer, but its classification accuracy is comparable to many complex depth models, and the training time is far less than that of depth models.
- *DPCNN* [40] DPCNN is a pyramid shaped deep convolution neural network. Strictly speaking, it is considered the first deep text classification convolution neural network. It can capture the long-distance dependence in the text, and the residual connection added to the model can alleviate the problems of gradient explosion and network degradation in the deep model.
- *Transformer* [41] Transformer was born in machine translation tasks and was later applied to many tasks in NLP and CV fields. Transformer is mainly composed of encoder and decoder. It only uses self-attention and introduces multi-head attention mechanism.
- *Bert* [42] Bert is a pre-trained deep language model, which is essentially a bidirectional transformer model. Bert outperforms humans in machine comprehension tasks, and shows excellent performance in many NLP tasks such as text classification.

Table 5 Performance comparison of different baseline models

Method	Accuracy (%)	F1-score (%)
DPCNN	89.80	89.43
FastText	90.24	89.80
Transformer	85.90	85.13
Bert	85.90	84.81
TextCNN	**90.46**	**90.13**

Bold values represent the best performance of all models

Results and analysis

In this section, we compare the performance of MHA model with various baseline models, and verify the effectiveness of multimodal hierarchical attention mechanism and distribution normalization method.

Performance comparison

We extract all user posts in the data set introduced in Sect. 4, and use various deep learning methods described in Sect. 5.3 to detect users' depressive tendencies. The results are shown in Table 5. The experimental results show that the five deep learning methods achieve relatively good performance. Among them, the classification accuracy of DPCNN and FastText reach 89.80% and 90.24% respectively, which shows that these classical methods can achieve high detection accuracy when dealing with text classification tasks, while Transformer and Bert only achieve 85.90%. In contrast, TextCNN shows the best performance, with classification accuracy and F1-score reaching 90.46% and 90.13% respectively. The reason may be that the performance of Bert, which is pre-trained based on large-scale corpus after fine-tuning on small data sets is not as good as that of training a new model, such as TextCNN. The structure of Transformer makes the model lose the ability to capture local features, while in depression detection task, many users only express their depression tendency in one or two words. Therefore, we use TextCNN as the text feature extractor of MHA model.

The performance of the proposed MHA model is shown in Table 6. Its accuracy and F1-score reach 92.84% and 92.78% respectively, achieving 2.38% and 2.65% improvement over the best performing baseline model

Table 6 Performance comparison of proposed models

Method	Accuracy (%)	F1-score (%)
MHA(non_att)	91.76	91.44
MHA(between_att)	92.19	92.00
MHA(inner_att)	91.54	91.79
MHA	**92.84**	**92.78**

MHA(non_att) represents the model after removing the multimodal hierarchical attention in MHA model, MHA(between_att) represents the model that only retains the attention between modalities in MHA model, and MHA(inner_att) represents the model that only retains inner modal attention in MHA model

Bold values represent the best performance of all models

Table 7 Ablation results of MHA model using different types of data

Method	Accuracy (%)	F1-score (%)
Picture + Auxiliary Information	86.33	86.39
Only Text(TextCNN)	90.46	90.13
Text + Auxiliary Information	90.89	90.91
Text + Picture	91.97	91.69
MHA (Text + Auxiliary Information+Picture)	**92.84**	**92.78**

Bold values represent the best performance of all models

Table 8 Comparison of distribution normalization results

Method	Accuracy (%)	F1-score (%)
No distribution normalization	91.97	91.90
Distribution normalization	92.84	92.78
	↑ 0.87	↑0.88

separately. We think that the improvement of model performance may be related to the use of multimodal data and attention mechanism. To verify our conjecture, we first remove the multimodal hierarchical attention in the MHA model. The experimental results show that when using only multimodal data, compared with TextCNN, MHA(non_att) achieves 1.30% and 1.31% improvement in accuracy and F1-score respectively. It shows that adding picture data and auxiliary information can remedy the deficiency of using only text data, and then improve the effect of depression detection.

Furthermore, after adding the multimodal hierarchical attention mechanism, the proposed model achieves the best performance. We also remove the attention within and between modalities in the multimodal hierarchical attention mechanism, respectively. Compared with MHA(non_att), the accuracy and F1-score of MHA(between_att) model are improved when only attention between modalities is retained. It is worth mentioning that the accuracy of MHA(inner_att) model is slightly reduced when only inner modal attention is retained, while F1-score is still improved. This may because Chinese social media users often post pictures that are not related to the text topic, and using these pictures directly for depression detection may affect the model performance. In the future work, we plan to explore the deep relationship between text and pictures to more effectively detect the depressive tendencies of social media users.

In order to explore the impact of different types of data on depression detection performance, we compare the proposed MHA model with three models using different types of data. The experimental results are shown in Table 7. It can be seen that, compared with using only text data, the addition of auxiliary information containing users' social characteristics can improve the performance of depression detection to a certain extent. This is because there are certain differences in the posting features between normal users and depressed users, which can be seen in Table 1. The model can distinguish depressed users through the characteristics contained in the auxiliary information. In addition, better performance can be achieved by using the user's picture and text data at the same time, which accuracy and F1-score reach 91.97% and 91.69% respectively. This shows that the picture data of different types of users contain significantly different features, depressed users tend to publish fewer pictures and the pictures contain more negative emotions. Compared with the other three models, the MHA model with complete data achieves the best depression detection performance, which proves the effectiveness of the data we used.

In this paper, we apply the distribution normalization method proposed in Sect. 3.4 to the image data of social media users. In order to verify the effectiveness of this method, we conduct experiments with normalized and non-normalized image data respectively. The results are shown in Table 8. It can be seen that when distribution normalization is not applied, the accuracy and F1-score of MHA model are 91.97% and 91.90% respectively. After applying distribution normalization, the accuracy and F1-score of the model are improved to 92.84% and 92.78% respectively. The above proves that the proposed distribution normalization method is helpful to better calculate the picture attention weight in the model, and can further improve the performance of the model while optimizing the data distribution.

Case study

In order to demonstrate the effectiveness of the MHA model more intuitively, we select some representative

Table 9 Examples of case studies of different models

User	Text	Label	Baseline	MHA (non_att)	MHA
1	Do people want others to know that they are ill? Why do I want everyone to know that I suffer from depression.				
	Today's makeup is so beautiful. It's unreasonable not to take more photos. So how can such a beautiful girl get depression?				
	In the afternoon, I poured out all my Quetiapine Fumarate, Escitalopram Oxalate, Sertraline and even Melatonin for swallowing.	1	1	1	1
2	I also want to fall in love, but I have to go to work when I wake up and sleep when I get off work.				
	Today's wind seems to speak. It says in my ear: I want you to die.				
	It is just a beautiful girl who has grown up for another year.	0	1	0	0
3	Why have a headache and vomit? When will it be good? I'm so tired.				
	Since you can't control your thoughts, release them freely and don't force or embarrass yourself.				
	Just shopping it when you're unhappy.	1	0	0	1
4	Alas... Recently, my mind has become more and more gloomy.				
	I haven't taken a good picture of the clouds this summer.				
	I love this book very much.	1	0	0	0

users in the data set, and use the baseline model, the MHA(non_att) model and the MHA model to detect their depressive tendencies. The results are shown in Table 8. Among them, user 1, user 3 and user 4 are annotated as depressed users, and user 2 is a normal group user.

As can be seen from Table 9, the posts of user 1 contain a large number of contents with obvious depressive tendency, such as depressive symptoms and antidepressants. Therefore, TextCNN and other baseline models that only use text information can also correctly classify the user. For user 2, his/her posts contain some words related to depression, so TextCNN mistakenly classifies the user as having depressive tendency. However, after combining pictures and auxiliary information, the MHA(non_att) and MHA model extract the real emotional features of the user and give accurate prediction results, which shows that the addition of multimodal information helps to improve the performance of depression detection task.

User 3 is annotated as depressed user, but only MHA correctly identifies the user's implied depressive tendency. This may because only a small part of all the data of the user expresses relatively obvious depressive tendency. For TextCNN and MHA(non_att), they assign equal importance to all data, which causes those that express depressive tendencies to be ignored. The multimodal hierarchical attention used by the MHA model can focus on data that has more impact on the classification results, and thus can identify those depressed users that are difficult to be detected in social media. For User 4, we comprehensively judge that the user has depression tendency according to all his/her posts. The depression tendency contained in the data is too vague, and even humans need to make careful judgment. Therefore, all models fail to identify the user's true emotion correctly. We find that almost all the data of the user doesn't express his/her depressive tendency, thus making it difficult for the model to predict. In the future work, we will try to build the knowledge graph of social media depression users to further improve the performance of depression detection by introducing external knowledge.

Conclusion

In this paper, we propose a multimodal hierarchical attention model called MHA for the detection of depressive tendency in social media. The model can process multimodal data at the same time, and use the attention mechanism within and between modalities to screen the information that plays an important role in depression detection. We also construct a Chinese social media depression detection data set. Experiments show that the proposed MHA model can accurately identify users with depressive tendencies on social media. Moreover, we propose a distribution normalization method, which can align user data with different distribution and improve the performance of depression detection.

Acknowledgements

This research was supported by Supercomputing Center of Lanzhou University.

Author contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by ZA, WC, JZ, FZ and BH; The first draft of the manuscript was written by ZL and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported in part by the National Key Research and Development Program of China (Grant No. 2019YFA0706200), in part by the National Natural Science Foundation of China (Grant No. 61802158), and in part by the Fundamental Research Funds for the Central Universities (lzujbky-2021-66, lzujbky-2018-125).

Data availability
Not applicable.

Code availability
Not applicable.

Declarations

Conflict of interest
The authors have no relevant financial or non-financial interests to disclose.

Human and animal rights
Not applicable.

Ethical approval
Not applicable.

Consent to participate
Not applicable.

Consent for publication
All authors have checked the manuscript and have agreed to the submission.

Author details
[1]Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, Gansu, China. [2]School of Medical Technology, Beijing Institute of Technology, Beijing 100081, Beijing, China. [3]CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200000, Shanghai, China.

Received: 9 July 2022 Accepted: 6 September 2022
Published: 18 January 2023

References

1. Organization WH, et al. Depression and other common mental disorders: global health estimates. Technical report: World Health Organization; 2017.
2. Beck AT, Steer RA, Brown GK. Beck Depression Inventory (BDI-II) vol. 10, (1996)
3. Niu L, Jia C, Ma Z, Wang G, Yu Z, Zhou L. Validating the geriatric depression scale with proxy-based data: a case-control psychological autopsy study in rural China. J Affect Disord. 2018;241:533–8.
4. Brandt WA, Loew T, von Heymann F, Stadtmüller G, Tischinger M, Strom F, Molfenter J, Georgi A, Tritt K. How does the icd-10 symptom rating (ISR) with four items assess depression compared to the BDI-II? A validation study J Affect Disord. 2015;173:143–5.
5. Maske UE, Hapke U, Riedel-Heller SG, Busch MA, Kessler RC. Respondents' report of a clinician-diagnosed depression in health surveys: comparison with DSM-IV mental disorders in the general adult population in germany. BMC Psychiatry. 2017;17(1):1–10.
6. Harris JR. No two alike: human nature and human individuality. Twin Res Hum Genet. 2006;9(5):703–4.
7. Sisask M, Värnik A, Kolves K, Konstabel K, Wasserman D. Subjective psychological well-being (WHO-5) in assessment of the severity of suicide attempt. Nord J Psychiatry. 2008;62(6):431–5.
8. Shen G, Jia J, Nie L, Feng F, Zhang C, Hu T, Chua T-S, Zhu W. Depression detection via harvesting social media: a multimodal dictionary learning solution. In: IJCAI; 2017. p. 3838–44.
9. Gui T, Zhu L, Zhang Q, Peng M, Zhou X, Ding K, Chen Z. Cooperative multimodal approach to depression detection in twitter. In: Proceedings of the AAAI conference on artificial intelligence, vol. 33; 2019. p. 110–117.
10. Cao L, Zhang H, Feng L. Building and using personal knowledge graph to improve suicidal ideation detection on social media. IEEE Trans Multimed. 2020.
11. Zogan H, Razzak I, Jameel S, Xu G. 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; 2021. p. 133–142.
12. Park J, Moon N. Multimodal depression detection system based on attention mechanism using ai speaker. In: Proceedings of the Korean Society of broadcast engineers conference. The Korean Institute of Broadcast and Media Engineers; 2021. p. 28–31.
13. Song H, You J, Chung J-W, Park JC. Feature attention network: interpretable depression detection from social media. In: Proceedings of the 32nd Pacific Asia conference on language, information and computation; 2018.
14. Ren L, Lin H, Xu B, Zhang S, Yang L, Sun S, et al. Depression detection on reddit with an emotion-based attention network: algorithm development and validation. JMIR Med Inform. 2021;9(7):28754.
15. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep learning-based text classification: a comprehensive review. ACM Comput Surv. 2021;54(3):1–40.
16. Priya A, Garg S, Tigga NP. Predicting anxiety, depression and stress in modern life using machine learning algorithms. Procedia Comput Sci. 2020;167:1258–67.
17. Tian X, Batterham P, Song S, Yao X, Yu G. Characterizing depression issues on sina weibo. Int J Environ Res Public Health. 2018;15(4):764.
18. Seabrook EM, Kern ML, Fulcher BD, Rickard NS. Predicting depression from language-based emotion dynamics: longitudinal analysis of Facebook and twitter status updates. J Med Internet Res. 2018;20(5):9267.
19. Wang X, Chen S, Li T, Li W, Zhou Y, Zheng J, Chen Q, Yan J, Tang B, et al. Depression risk prediction for chinese microblogs via deep-learning methods: Content analysis. JMIR Med Inform. 2020;8(7):17958.
20. Cao L, Zhang H, Feng L, Wei Z, Wang X, Li N, He X. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP); 2019. p. 1718–1728.
21. He L, Niu M, Tiwari P, Marttinen P, Su R, Jiang J, Guo C, Wang H, Ding S, Wang Z, et al. Deep learning for depression recognition with audiovisual cues: a review. Information Fusion. 2022;80:56–86.
22. Koutsouleris N, Dwyer DB, Degenhardt F, Maj C, Urquijo-Castro MF, Sanfelici R, Popovic D, Oeztuerk O, Haas SS, Weiske J, et al. Multimodal machine learning workflows for prediction of psychosis in patients with clinical high-risk syndromes and recent-onset depression. JAMA Psychiat. 2021;78(2):195–209.
23. Cai H, Qu Z, Li Z, Zhang Y, Hu X, Hu B. Feature-level fusion approaches based on multimodal EEG data for depression recognition. Inf Fusion. 2020;59:127–38.
24. Ceccarelli F, Mahmoud M. Multimodal temporal machine learning for bipolar disorder and depression recognition. Pattern Anal Appl 1–12 (2021)
25. Toto E, Tlachac M, Rundensteiner EA. Audibert: A deep transfer learning multimodal classification framework for depression screening. In: Proceedings of the 30th ACM international conference on information & knowledge management; 2021. p. 4145–4154.
26. Sardari S, Nakisa B, Rastgoo MN, Eklund P. Audio based depression detection using convolutional autoencoder. Expert Syst Appl. 2022;189:116076.
27. Mallol-Ragolta A, Zhao Z, Stappen L, Cummins N, Schuller B. A hierarchical attention network-based approach for depression detection from transcribed clinical interviews. 2019.
28. Xezonaki D, Paraskevopoulos G, Potamianos A, Narayanan S. Affective conditioning on hierarchical attention networks applied to depression detection from transcribed clinical interviews. In: INTERSPEECH; 2020. p. 4556–4560.
29. He L, Chan JC-W, Wang Z. Automatic depression recognition using CNN with attention mechanism from videos. Neurocomputing. 2021;422:165–75.

30. In: 2020 42nd annual international conference of the IEEE Engineering in Medicine \& Biology Society (EMBC). IEEE; 2020. p. 128--133.
31. Zheng W, Yan L, Gou C, Wang F-Y, Graph attention model embedded with multi-modal knowledge for depression detection. In: 2020 IEEE international conference on multimedia and expo (ICME). IEEE; 2020. p. 1--6.
32. Niu M, Chen K, Chen Q, Yang L. Hcag: A hierarchical context-aware graph attention model for depression detection. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2021. p. 4235–4239.
33. Orabi AH, Buddhitha P, Orabi MH, Inkpen D. Deep learning for depression detection of twitter users. In: Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic; 2018. p. 88–97.
34. Nobles AL, Glenn JJ, Kowsari K, Teachman BA, Barnes LE. Identification of imminent suicide risk among young adults using text messages. In: Proceedings of the 2018 CHI conference on human factors in computing systems; 2018. p. 1–11.
35. Safa R, Bayat P, Moghtader L. Automatic detection of depression symptoms in twitter using multimodal analysis. J Supercomput. 2021;78:4709.
36. Lv M, Li A, Liu T, Zhu T. Creating a chinese suicide dictionary for identifying suicide risk on social media. PeerJ. 2015;3:1455.
37. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.
38. Kim Y. Convolutional neural networks for sentence classification. Eprint Arxiv (2014)
39. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
40. Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization. In: Proceedings of the 55th annual meeting of the Association for Computational Linguistics (vol. 1: Long Papers); 2017. p. 562–570.
41. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Advances in neural information processing systems, vol. 30; 2017.
42. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)