Machine Learning for Depression Risk Monitoring on Chinese Social Media: A Comprehensive Evaluation and Analysis

Zhenwen Zhang, Zepeng Li, Zhihua Guo, Jianghong Zhu, Yu Zhang, Bin Hu

Submitted to: JMIR Mental Health
on: March 11, 2024

Table of Contents

Machine Learning for Depression Risk Monitoring on Chinese Social Media: A Comprehensive Evaluation and Analysis

Zhenwen Zhang[1] MEng; Zepeng Li[1] PhD; Zhihua Guo[1] MSc; Jianghong Zhu[1] MSc; Yu Zhang[1] MSc; Bin Hu[1] PhD

[1]Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University Lanzhou CN

Corresponding Author:
Bin Hu PhD
Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University
No. 222 South Tianshui Road
Lanzhou
CN

Abstract

Background: Depression is a significant global public health issue that affects the physical and mental well-being of hundreds of millions of people worldwide. However, a substantial number of individuals with depression on social media often go undiagnosed and struggle to access timely and effective treatment, increasingly becoming a major societal health concern.

Objective: This paper aims to explore and develop an online depression risk detection method based on deep learning technology to identify individuals at risk of depression on the Chinese social media platform Sina Weibo.

Methods: We initially collected approximately 527,333 posts publicly shared over one year from 1600 individuals with depression and 1600 individuals without depression on the Sina Weibo platform. Subsequently, we developed a hierarchical Transformer network to learn semantic features for each user. This network comprises two levels of Transformer structures, one at the word level and the other at the sentence level. These Transformers are employed to extract the textual semantic features of each post, and the aggregated features of all posts for each user generate user-level semantic features. A classifier is then applied to predict the risk of depression. Finally, we conducted statistical and linguistic analyses of the content of posts from individuals with and without depression using the Chinese LIWC.

Results: We divided the original dataset into training, validation, and test sets. The training set consists of 1000 individuals with depression and 100 individuals without depression. The validation and test set each includes 600 users, with 300 individuals with depression and 300 without depression. Our method achieved an accuracy of 84.62%, precision of 84.43%, recall of 84.50%, and F1 score of 84.32% on the test set without applying sampling techniques. After applying our proposed retrieval-based sampling strategy, our method achieved an accuracy of 95.46%, precision of 95.30%, recall of 95.70%, and F1 score of 95.43%. These results strongly demonstrate the effectiveness and superiority of our proposed depression risk detection model and retrieval-based sampling technique. This provides new insights for large-scale depression detection through social media. Through language behavior analysis, it is observed that individuals with depression are more likely to use negation words (the value of "swear" is 0.001253). This may indicate the presence of negative emotions, rejection, doubt, disagreement, or aversion expressed by individuals with depression. Additionally, we also found that individuals with depression tend to use negative emotional vocabulary in their expressions (NegEmo: 0.022306, Anx: 0.003829, Anger: 0.004327, Sad: 0.005740), which may reflect their internal negative emotions and psychological state. This frequent use of negative vocabulary could be a way for individuals with depression to express negative feelings towards life, themselves, or their surrounding environment.

Conclusions: The research results indicate the feasibility and effectiveness of deep learning methods in detecting the risk of depression. This provides insights into the potential for large-scale, automated, and non-invasive prediction of depression among users of online social media.

(JMIR Preprints 11/03/2024:58259)
DOI: https://doi.org/10.2196/preprints.58259

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Machine Learning for Depression Risk Monitoring on Chinese Social Media: A Comprehensive Evaluation and Analysis

Zhenwen Zhang, MEng; Zepeng Li, PhD; Zhihua Guo, MSc; Jianghong Zhu,MSc;Yu Zhang,Msc; Bin Hu, PhD

Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou, China

Corresponding Author:

Bin Hu, PhD
Gansu Provincial Key Laboratory of Wearable Computing School of Information Science and Engineering
Lanzhou University
No. 222 South Tianshui Road
Lanzhou, 730000
China
Phone: 86 +8617352120733
Email: bh@lzu.edu.cn

Abstract

Background: Depression is a significant global public health issue that affects the physical and mental well-being of hundreds of millions of people worldwide. However, a substantial number of individuals with depression on social media often go undiagnosed and struggle to access timely and effective treatment, increasingly becoming a major societal health concern.

Objective: This paper aims to explore and develop an online depression risk detection method based on deep learning technology to identify individuals at risk of depression on the Chinese social media platform Sina Weibo.

Methods: We initially collected approximately 527,333 posts publicly shared over one year from 1600 individuals with depression and 1600 individuals without depression on the Sina Weibo platform. Subsequently, we developed a hierarchical Transformer network to learn semantic features for each user. This network comprises two levels of Transformer structures, one at the word level and the other at the sentence level. These Transformers are employed to extract the textual semantic features of each post, and the aggregated features of all posts for each user generate user-level semantic features. A classifier is then applied to predict the risk of depression. Finally, we conducted statistical and linguistic analyses of the content of posts from individuals with and without depression using the Chinese LIWC.

Results: We divided the original dataset into training, validation, and test sets. The training set consists of 1000 individuals with depression and 100 individuals without depression. The validation and test set each includes 600 users, with 300 individuals with depression and 300 without depression. Our method achieved an accuracy of 84.62%, precision of 84.43%, recall of 84.50%, and F1 score of 84.32% on the test set without applying sampling techniques. After applying our proposed retrieval-based sampling strategy, our method achieved an accuracy of 95.46%, precision

of 95.30%, recall of 95.70%, and F1 score of 95.43%. These results strongly demonstrate the effectiveness and superiority of our proposed depression risk detection model and retrieval-based sampling technique. This provides new insights for large-scale depression detection through social media. Through language behavior analysis, it is observed that individuals with depression are more likely to use negation words (the value of "swear" is 0.001253). This may indicate the presence of negative emotions, rejection, doubt, disagreement, or aversion expressed by individuals with depression. Additionally, we also found that individuals with depression tend to use negative emotional vocabulary in their expressions (NegEmo: 0.022306, Anx: 0.003829, Anger: 0.004327, Sad: 0.005740), which may reflect their internal negative emotions and psychological state. This frequent use of negative vocabulary could be a way for individuals with depression to express negative feelings towards life, themselves, or their surrounding environment.

Conclusions: The research results indicate the feasibility and effectiveness of deep learning methods in detecting the risk of depression. This provides insights into the potential for large-scale, automated, and non-invasive prediction of depression among users of online social media.

Keywords: Depression; Social Media; Natural Language Processing; Deep Learning

Introduction

Depression is a global mental illness that can seriously affect people's physical and mental health. According to the World Health Organization, more than 300 million people worldwide suffer from depression, and 5% of them are adults. Depression is affected by various factors, such as biological, psychological, and social environments. It may affect a person's sleep and appetite and often expresses symptoms such as physical fatigue, poor concentration, and diminished interest. When depression is recurrent and reaches moderate or severe intensity, it can become a serious health disorder or even cause suicide. Although several countries and institutions have introduced medical policies and treatments for depression, most people still cannot receive timely treatments [1]. The main reason for this phenomenon lies in the inability of existing technical means to achieve early identification and large-scale detection of depression. To detect depression and assess its severity, scales such as the Patient Health Questionnaire 9-item (PHQ-9), Self-rating Depression Scale (SDS), and Hamilton Depression Rating Scale (HAM-D) have been applied to clinical depression detection [2-4]. Recently, several scholars have been exploring depression detection with physiological data, where Electroencephalogram (EEG) [5-7] and physiological images [8-9] are adopted. EEG and physiological images provide objective clinical medical evidence to help psychiatrists reveal the physiopathological types and pathogenesis of depression. Instead, the scales are susceptible to the subjects' emotional state, cooperativeness, and environment, which makes the authenticity and reliability of the test results questionable. Although these methods have enormously advanced the scientific study of depression, large-scale and efficient methods for depression detection still face many challenges. On the one hand, these methods cannot track and model patients' mental states. Long-term and short-term physiological and scale tests may fail to accurately measure subjects' mental states. On the other hand, most patients may fail to realize their condition during the early stages of depression, which may cause timely treatment [10].

With the development of Internet technology and the spread of mobile networks, WeChat, Weibo, and Twitter have become indispensable to people's daily lives and work. People increasingly rely on social media to share everyday life, express ideas, and real-time outflow emotions. In particular, users with mental health tend to have a higher tendency and dependency on social media. Some social media platforms have opened special topic forums, such as Depression and Autism SuperTopics, where users with mental disorders can pour out their emotions, share their treatment process, and seek online help. Social media users' self-reported texts contain rich information about emotions and events, such as medications, self-perceptions, and suicidal intent. Such information is dynamically evolving with the patient's mental state and treatment process and has prominent time-evolving properties, which is valuable for establishing an effective model of mental illness. Moreover, these data are diverse, frequently updated, and easily accessible, which can effectively contribute to the study of social media mental health [11-17]. Early studies explored the leverage of statistical learning methods to analyze differences between depressed and non-depressed users from Twitter in terms of emotional word usage [18-19], language style [20], and social behavior [21]. With the widespread popularity of deep learning in NLP, social media-based depression research shifted to a deep learning-based paradigm [22-27]. Specifically, user-level depression detection is treated as a long text classification task, where a user's posts are concatenated into a long text and classified by a neural network model.

This study focuses on applying deep learning methods to detect depression on Chinese social media. We utilized Sina Weibo as the data source for this study given that it is one of the most popular social media platforms in China with over 200 million active users per month. To identify the depression risk users, we first collected approximately 15,774,510 posts made by 1,600 depressed and 1,600 non-depressed users between December 2020 to December 2021 on Sina Weibo. We then developed a Hierarchical Transformer Network to study the semantic features of each user from their posts. The

HTN consists of a two-level Transformer structure that focuses on learning post-level and user-level feature representations, respectively. This model can effectively capture fine-grained semantic features at the word, sentence, and document levels, and has obvious advantages for portraying the differentiated feature representations of depressed and non-depressed users. In addition, to further improve the feature representation capability for depressed users, we propose a retrieval-based sampling strategy to select depression-related posts to train the depression risk detection model. The experimental results indicate the importance of this sampling strategy in minimizing the impact of unnecessary noisy data on model performance. Our model performance gains more than 10% improvement in all evaluation metrics after applying this sampling strategy. Our methodology provides strong support for identifying users at risk for depression through online social media data in Chinese communities, which is important for the health of all people and social harmony.

Methods

Data Collection

We focus on predicting depression risk on user-level via their social media posts. To this end, we collected a corpus includes 3,200 users, with 1,600 depressed and 1,600 non-depressed users. We gathered posts authored by each user from December 2020 to December 2021 and annotated each user as depressed or non-depressed according to linguistic patterns, rules, and psychological knowledge.

We first randomly picked several users as candidates from relevant super topics. Then, these users were annotated as depressed or non-depressed by three annotators based on predefined annotation guidelines. Finally, we obtained 1600 depressed and 1600 non-depressed users. The detailed annotation process is as follows:

- We follow the annotation guidelines [18,19,23] previously developed for the English language domain. If an user self-reported in their post that they were diagnosed with depression, then we annotate the user as depressed. However, due to the differences and characteristics inherent in languages, applying English rules comprehensively within the Chinese context is only sometimes feasible. Consequently, we have developed annotation rules better suited to the Chinese language environment. Within this context, social media users frequently employ metaphors to convey their depressive state, such as references to medication, treatment approaches, and suicidal ideation.
- We annotate users as non-depressed users if they did not explicitly express in their posts that they had suffered from depression in the past or present.

Table 1. Results of our proposed model and baseline models.

Characteristic	Depressed	Non-Depressed
Number of Users	1600	1600
Number of Posts	169,838	357,495
Number of Words	4,282,792	11,491,718

Data Preprocessing

The raw data collected from Sina Weibo often contains irrelevant or informal expressions, which may hinder the effectiveness of model training. To eliminate the impact of these factors on the model, we performed the following data preprocessing steps to clean such noise:

- We used the jieba tokenizer to segment each post into a word sequence.
- We replaced the emoticons in the posts with the corresponding emotion words.
- We removed numbers, URLs, and punctuation from posts.

- We deleted posts automatically published by Sina Weibo's robot assistant, such as birthday reminders and membership-level notifications.
- We removed duplicate posts.
- We adopted posts with posts longer than three words for training.

Ethical Considerations

All the data in this paper were obtained from Sina Weibo’s public data, which protects those who have private profiles from being subject to research studies. Hence, this analysis meets the standards to waive informed consent and similar guidelines [55]. Furthermore, we desensitized the data to protect the privacy of the users. Specifically, we removed all user information related to the identity of the users.

Problem Definition

This study aims to use deep learning and user-generated content to build a depression prediction model that automatically predicts whether each user is at risk for depression. The input to this model is each user's post and the output is the label of whether the user is depressed or not.

Existing Challenges

Previous studies have employed two approaches to obtain a user-level semantic representation: sequential-based and summary-based approaches. The sequential-based approach involves concatenating users’ posts into a long text and then utilizing machine learning methods for encoding and prediction. In contrast, the summary-based approach first employs a summarization model to generate a short textual description for each user, followed by the use of machine learning methods for encoding and prediction. However, both of these methods have limitations. The sequential-based approach may inaccurately capture independent temporal user sentiment information due to the blunt concatenation of posts into a long text, and it also faces computational efficiency challenges. Additionally, the quality of the user description texts generated by the summary-based approach is difficult to control and evaluate, which can lead to poorer predictive accuracy.

In previous studies, all collected user posts have been used to train depression classification models. However, using all posts of a user to train depression detection models is not always effective. For instance, although some depressed users may post frequently, only a subset of their posts express symptoms, emotions, and thoughts related to depression. If all posts are used to train a depression classification model, it may introduce additional noise that affects the predictive accuracy of the depression model.

Proposed Prediction Model

Figure 1. The workflow of our proposed depression prediction model.

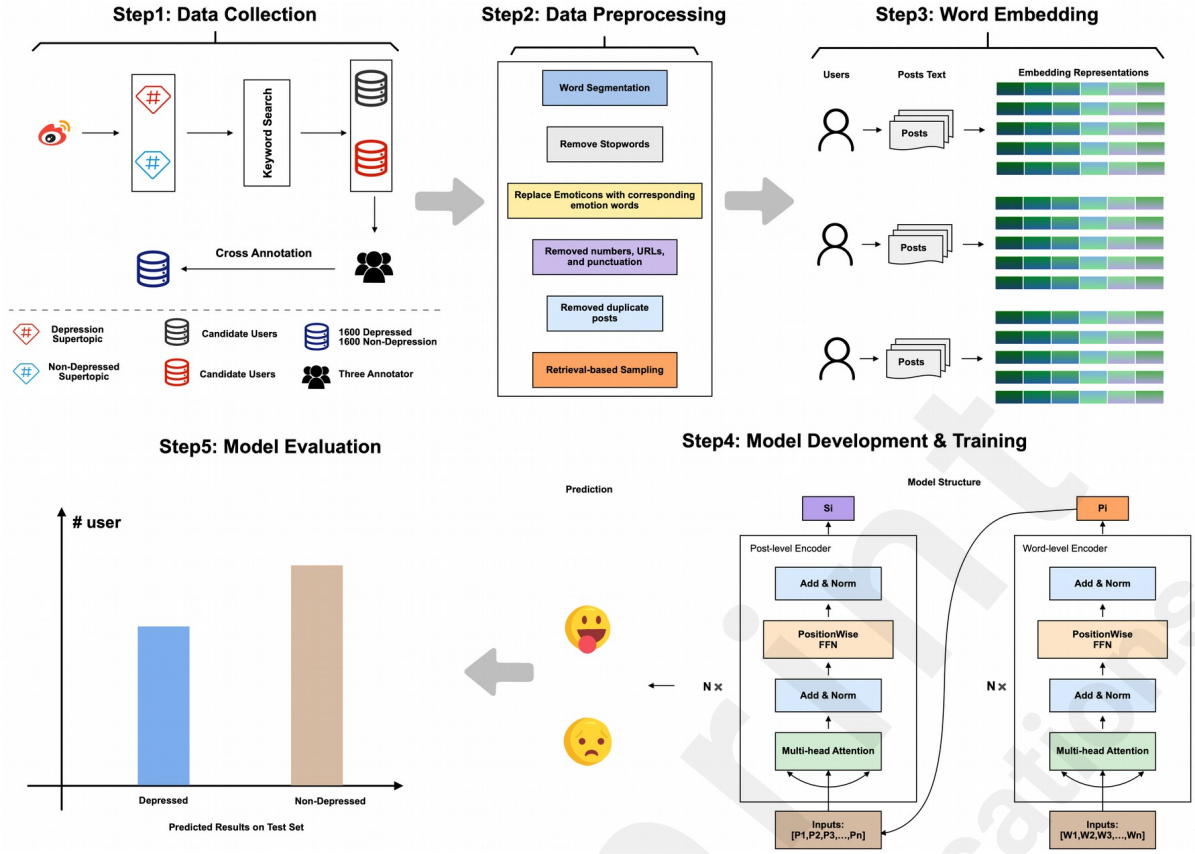

Figure 1 illustrates the workflow of our proposed depression detection model, which consists of five steps: Data Collection, Data Preprocessing, Word Embedding, Model Development & Training, and Model Evaluation. We discussed the process of data collection and preprocessing in the previous section. Therefore, we will provide detailed insights into the development and training details of the model in the following sections.

Word Embedding

Word embedding is fundamental in applying deep learning to NLP tasks, as it represents semantic information about words by mapping them to real-valued vectors in a high-dimensional vector space (e.g., 100dim, 200dim, 300dim, 768dim, etc.). One advantage of word embedding is that it can effectively represent the semantics of words in different contexts and can be further optimized. The emergence of word embedding technology has accelerated the development of NLP and facilitated the effective processing and understanding of human language.

In this paper, to obtain better word embeddings, we introduced Tencent's pre-trained word embeddings [30] (Tencent AI Lab Embedding Corpus for Chinese Words and Phrases) to initialize the embedding representations of each word in user posts. The Tencent pre-trained word embedding database was pre-trained on the Directional Skip-Gram algorithm using Wikipedia, Baidu Baike, and web text data. It includes embeddings for 12,287,936 Chinese words (200d). We first used Jieba tokenizer to tokenize each post from users. Specifically, the vocabulary of Tencent pre-trained word embedding database was adopted as an external vocabulary to guide the tokenization of user posts. Then, we looked up the embedding of each word in user posts from the Tencent's pre-trained word embeddings database and fed them into the model for further training.

Model Development and Training

As shown in Figure 1, we propose a hierarchical transformer network (HTN) to study textual semantic features from users' posts. The Transformer is an attention-based neural network architecture that has gained considerable attention in recent years, particularly in NLP and CV.

Unlike other deep learning models, the Transformer not only dynamically captures long-range dependencies but also exhibits faster computation speed. Inspired by this, we incorporate the Transformer into our model to better understand and encode behavior and intention from user posts. Our model consists of two levels of Transformers: a word-level Transformer and a post-level Transformer. The word-level Transformer is used to compute semantic features for each post, with word embeddings from each post as input. The sentence-level Transformer is employed to calculate aggregated semantic features for all user posts, with the input being the embeddings of all user posts. After obtaining the aggregated global feature representation, we perform classification on it to predict whether the user is depressed. Since our prediction task is a binary classification task, we use a sigmoid function for prediction. The proposed model is capable of learning fine-grained feature representations at the levels of words, sentences, and documents from user posts, which is crucial for enhancing prediction accuracy.

Model Evaluation

To effectively train and evaluate our model, we divided 1,600 users with depression and 1,600 users without depression into three sets: 1,000 for training, 300 for validation, and 300 for testing, respectively. We conducted experiments using Scikit-Learn for statistical methods and employed PyTorch for deep learning-based experiments. For the SVM and NB models, we used the RBF kernel-based SVM and MultinomialNB, respectively, during the training stage. The convolutional kernel size was set to [2, 3, 4], and the number of filters was set to 100. For other baseline models, both the hidden size and attention size were set to 256. For our proposed model, each post was padded or truncated to 512 words. The learning rate was set to 1e-3, and the batch size was selected from the range of [32, 64, 128].

Comparison Baselines

To fully assess the potential of applying deep learning to predict depression risk on social media, we adopted 11 widely used methods as baseline models. These included statistical-based methods such as SVM and NB, traditional neural network methods like CNN, LSTM, GRU, BiGRU, and BiLSTM, as well as attention-based methods such as LSTM-Attention, GRU-Attention, BiLSTM-Attention, and BiGRU-Attention.

Evaluation Metrics

We used accuracy, macro-averaged precision, macro-averaged recall, and macro-averaged F1-score to evaluate the prediction performance of our proposed model. These metrics are widely employed for assessing the performance of deep learning predictive models.

Results

Performance Comparison

Table 2. Results of our proposed model and baseline models.

Model	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Without Sampling				
SVM[a]	80.80	83.16	79.20	79.69
NB[b]	76.47	78.91	74.62	74.87
CNN[c]	79.93	80.70	80.79	79.93
LSTM[d]	71.80	73.71	69.91	69.91
GRU[e]	78.55	78.98	77.55	77.86
BiGRU[f]	67.99	67.77	67.94	67.81
BiLSTM[g]	65.92	65.88	66.06	65.81
LSTM-Attn[h]	78.55	78.34	78.15	78.23

GRU-Atten[i]	82.53	82.43	82.12	82.24
BiLSTM-Atten[j]	78.03	77.94	77.42	77.59
BiGRU-Atten[k]	80.97	80.75	80.97	80.83
HTN[l]	**84.62**	**84.43**	**84.50**	**84.32**
Random Sampling				
SVM[a]	79.24	82.37	77.38	77.78
NB[b]	76.30	79.41	74.27	74.46
CNN[c]	78.37	78.30	78.63	78.29
LSTM[d]	69.55	69.35	68.52	68.65
GRU[e]	77.68	78.03	76.69	76.98
BiGRU[f]	70.24	70.02	69.30	69.43
BiLSTM[g]	65.05	65.19	65.36	64.99
LSTM-Attn[h]	74.05	73.77	73.90	73.82
GRU-Atten[i]	80.62	80.39	80.43	80.41
BiLSTM-Atten[j]	74.39	74.20	73.72	73.87
BiGRU-Atten[k]	76.64	76.72	77.03	76.59
HTN[l]	**82.43**	**82.24**	**82.44**	**82.35**
Retrieval Sampling				
SVM[a]	92.21	93.14	91.50	92.00
NB[b]	83.56	87.44	81.78	82.43
CNN[c]	93.53	93.21	93.54	93.30
LSTM[d]	88.41	88.40	88.10	88.23
GRU[e]	92.25	92.09	92.49	92.21
BiGRU[f]	91.52	91.35	91.63	91.46
BiLSTM[g]	84.95	84.95	85.35	84.90
LSTM-Attn[h]	91.87	91.72	92.13	91.82
GRU-Atten[i]	91.27	91.16	91.34	91.15
BiLSTM-Atten[j]	91.35	91.58	90.05	91.19
BiGRU-Atten[k]	92.77	92.68	92.88	92.64
HTN[l]	**95.46**	**95.30**	**95.70**	**95.43**

[a]SVM: Support Vector Machine
[b]NB: Naive Bayes
[c]CNN: Convolutional Neural Network
[d]LSTM: Long Short-Term Memory
[e]GRU: Gated Recurrent Unit
[f]BiGRU: Bidirectional Gated Recurrent Unit
[f]BiLSTM: Bidirectional Long Short-Term Memory
[f]LSTM-Attn: Long Short-Term Memory with Attention
[f]GRU-Atten: Gated Recurrent Unit with Attention
[f]BiLSTM-Atten: Bidirectional Long Short-Term Memory with Attention
[f]BiGRU-Atten: Bidirectional Gated Recurrent Unit with Attention
[f]HTN: Hierarchical Transformer Network

Figure 2. Performance comparison of our proposed model and baseline models.

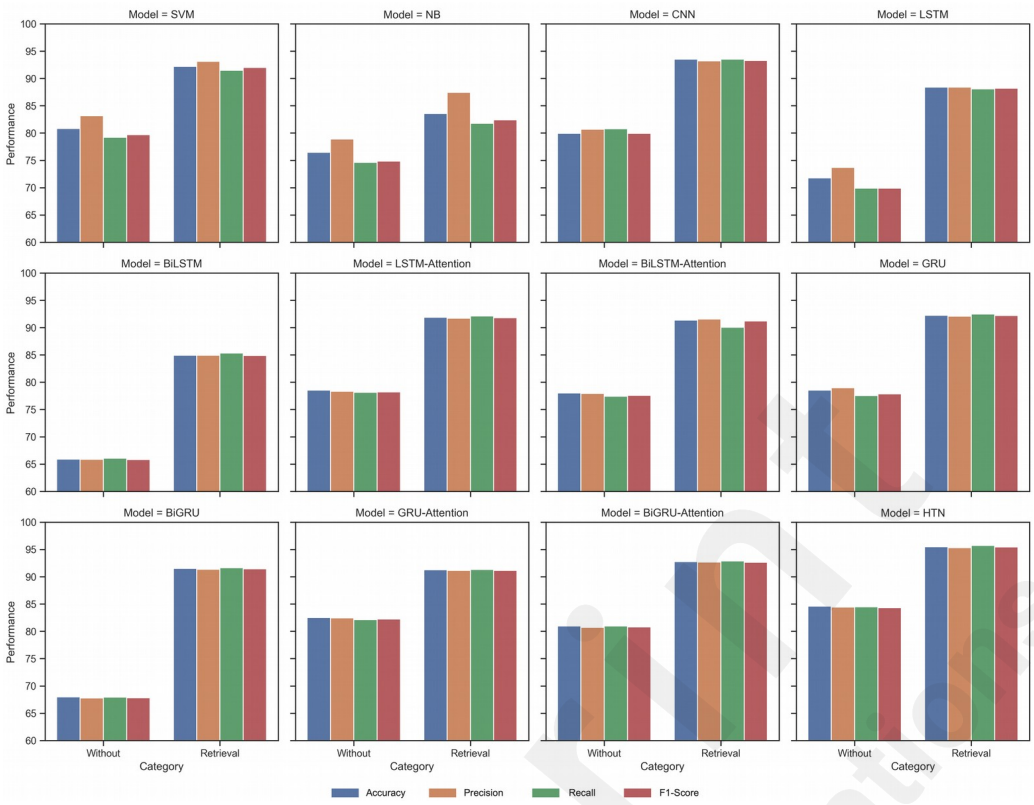

Table 2 and Figure 2 respectively present the experimental results and visualization of the baseline methods and our proposed approach on the test set. We can see that our model achieved the best results in all scenarios, with a prediction accuracy of over 95% for depression risk. As seen in Table 1, the HTN model outperforms the other baseline models, with at least a 2% improvement in the retrieval strategy and more than a 5% improvement in the other two strategies. This suggests that encoding a user's posts data with HTN is more effective than treating it as a single long text. HTN enables the model to fully consider post-interactions and intuitively fit better with human thinking. Simply treating all of a user's posts as a single long text may lead to computational and gradient challenges, limiting the model's ability to detect depression.

Unlike previous studies, this paper proposes a sampling strategy based on depression knowledge retrieval to select posts related to depression from user posts for training a depression detection model. This strategy reduces the computational overhead of model training and allows the model to focus more attention on learning about depression. We tested three different sampling scenarios and found that the retrieval-based strategy outperformed the other two by at least 10% in all evaluation metrics. In addition, we observed that the random sampling strategy performed worse than the no sampling strategy, which may be attributed to the uncertainty inherent in the random sampling process.

Compared to neural models without the attention mechanism, attention-based neural models exhibit better detection performance across all sampling strategies, with particularly significant improvements observed when using the no-sampling strategy. Models employing attention mechanisms (LSTM, GRU, BiLSTM, and BiGRU) enhance average performance by 6.09%, 6.56%, 12.06%, and 11.03%, respectively. We attribute this improvement to the fact that the attention mechanism enables the model to automatically focus more on words or phrases indicative of depression, thereby facilitating a superior semantic representation of the user.

Effect of Sampling Posts

Figure 3. Experimental results with different sampling strategies and ratios for training data.

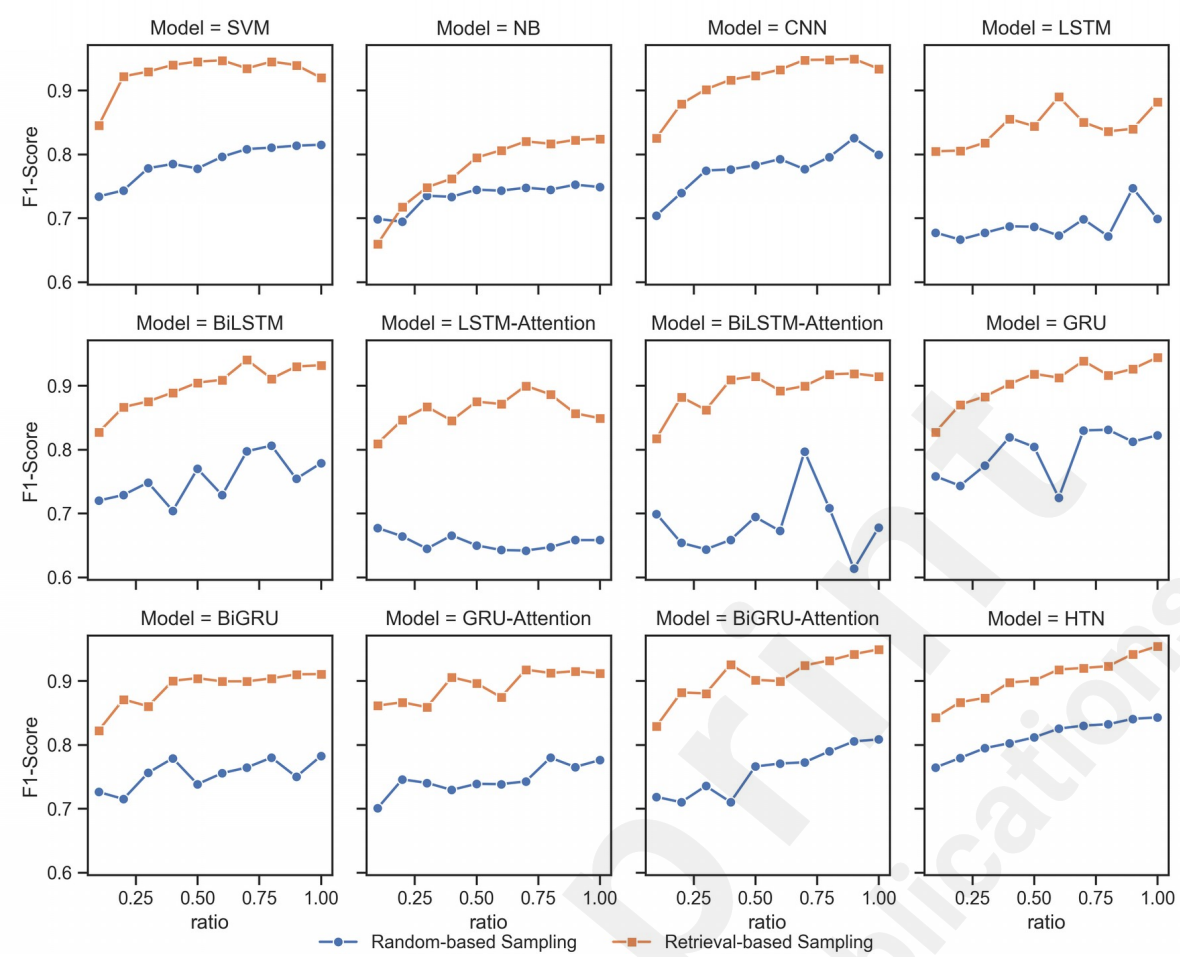

Figure 3 illustrates the F1 scores of each model across various sampling strategies and sampling ratios. It is evident that the deployment of effective sampling strategies can substantially improve the depression detection capabilities of the models. By employing a retrieval-based sampling strategy to select posts pertinent to depression, not only is the computational complexity of the model reduced, but the model also gains a better focus on acquiring knowledge related to depression from user posts. We observed that the retrieval-based sampling strategy consistently demonstrated a stable upward trend as the sampling rate increased incrementally, in contrast to the random sampling strategy, which exhibited more pronounced fluctuations. We attribute this primarily to the fact that the retrieval-based sampling strategy ensures the selection of posts related to depression in every sampling iteration. Conversely, the post selection process in the random sampling strategy is a probabilistic one that does not guarantee the relevance of a user's post to depression in each selection.

Group Behavior Analysis

Table 3. Statistical Results on post's characteristics between depressed and non-depressed users.

Characteristic	Depressed	Non-Depressed
Words/Post[a]	27	38
Posts/User[b]	132	266
Posts/User/Week[c]	2.74	5.55
FirstPerson-("□□")/Post[d]	0.03	0.16
FirstPerson-("□")/Post[e]	0.49	0.13
DepressionMentioned/Post[f]	0.72	0.15
DrugsMentioned/Post[g]	0.32	0.05

[a]Words/Post: Average number of words per post

[b]Posts/User: Average number of posts per user
[c]Posts/User/Week: Average number of posts per user per week
[d]FirstPerson("□□")/Post: Average number of mentions of first person "□□" per post
[e]FirstPerson("□")/Post: Average number of mentions of first person "□" per post
[f]DepressionMentioned/Post: Average number of mentions of keywords "□□□□□□" per post
[g]DrugsMentioned/Pos: Average number of mentions of depression-related drug names per post

Table 3 shows a comparison of seven behavioral characteristics between users with depression and those without. We note that non-depressed users are more socially engaged, posting more frequently. Additionally, users with depression tend to use the first-person pronoun "□" (“I”) more frequently than those without, who prefer the first-person pronoun "□□" (“we”). This indicates that users with depression may be more self-focused and have less interaction with others, whereas non-depressed users are more group-oriented and engage in more interactive behaviors. Furthermore, users with depression are more likely to focus on depression-related topics on social media, such as discussing their condition, treatment processes, and medication, while non-depressed users mention and discuss these topics less frequently.

Table 4. Statistical Results on using of modal particle between depressed and non-depressed users.

Characteristic	Depressed	Non-Depressed
Total number		
□	228,566	684,919
□	11,885	19,611
□	26,538	18,540
□	6,460	12,092
□	134,764	263,502
□	18,926	27,109
Avg Number/Post		
□	1.31	1.87
□	0.07	0.05
□	0.19	0.08
□	0.04	0.03
□	0.77	0.72
□	0.11	0.74

Table 4 indicates the comparative results of the use of modal particles between users with depression and those without depression. It can be observed that the usage of "□" ("de") is more frequent in both depressed and non-depressed users, while "□" ("ne") is used the least. The main reason is that "□" is commonly used as a modifier in almost all sentences, whereas "□" and "□" are often used in contexts expressing questions or uncertainties. It's worth noting that "□" ("ba") is used more frequently in the language expressions of users with depression, while "□" ("a") is used more frequently in the language expressions of non-depressed users. These two words are typically used at the end of sentences, where "□" ("ba") is often used to modify completed events, while "□" ("a") is typically used to modify events that are about to happen. In the expressions of users with depression, "□" ("ba") is more often expressed as "□□" ("okay"), "□□" ("all right"), "□□□□" ("just like this"), "□□□" ("go die"), etc. On the other hand, "□" ("a") is often combined in expressions of non-depressed users as "□□□□" ("really happy"), "□□□□□□" ("so that's how it is"), and "□□□□□□" ("you're really good to me").

Table 5. Statistical Results on using of punctuations between depressed and non-depressed users.

Characteristic	Depressed	Non-Depressed
Total number		

□	308,817	938,400
□	94,066	150,969
□	50,171	298,079
□	23,471	33,107
~	3,854	18,688
.....	27,067	59,779
Avg Number/Post		
□	1.77	2.56
□	0.54	0.41
□	0.29	0.81
□	0.13	0.16
~	0.72	0.05
.....	0.16	0.09

Table 5 shows the comparative results of punctuation use between users with depression and those without. We found that users with depression tend to use periods more frequently than non-depressed users, whereas non-depressed users favor commas more than those with depression. We speculate that this pattern may be due to the fact that users with depression often experience low mood and slowed thinking, which could make their expressions appear more cautious and negative. A period can be interpreted as a conclusion or a clear break between ideas, potentially reflecting the psychological desire of these patients to terminate or avoid further communication. In contrast, non-depressed users typically exhibit active and divergent thinking patterns. They frequently use commas to separate components of sentences and to express incomplete thought processes. We also noted that non-depressed users are more inclined to use exclamation marks (“!”). This is consistent with the experimental results concerning the interjection “□” (“a”) presented in Table 4. Furthermore, we observed that users with depression tend to use the tilde “~” and ellipses more frequently. These symbols are commonly employed in the Chinese internet context to convey a sense of helplessness or resignation.

Table 6. LIWC feature comparison results between depressed and non-depressed users.

Characteristic	Depressed	Non-Depressed
Negate	0.014860	0.001409
Swear	0.001253	0.000733
Interjunction	0.097806	0.084063
PastM	0.005300	0.004043
PresentM	0.011282	0.009824
FutureM	0.006698	0.006788
ProgM	0.027634	0.019808
Social	0.064925	0.053611
Family	0.005307	0.004541
Friend	0.002057	0.001615
Humans	0.017434	0.015656
Affect	0.070012	0.047391
PosEmo	0.013584	0.028104
NegEmo	0.022306	0.010997
Anx	0.003829	0.001747
Anger	0.004327	0.002616
Sad	0.005740	0.002321
CogMech	0.195500	0.151659
Insight	0.020294	0.016495

Cause	0.012366	0.010585
Certain	0.015263	0.012144
Bio	0.037529	0.031510
Body	0.013333	0.008986
Health	0.012634	0.005975
Home	0.004320	0.003871
Money	0.004426	0.007977
Death	0.004234	0.001913
Psychology	0.018474	0.015496

Negate: negative word
Swear: obscene language
Interjunction: modal particle
PastM: past
PresentM: present
FutureM: future
ProgM: continuation
Social: social word
Family: family word
Friend: friend word
Humans: human word
Affect: emotion process word
PosEmo: positive emotion word
NegEmo: negative emotion word
Anx: anxiety word
Anger: anger word
Sad: sad word
CogMech: cognitive process word
Insight: insight word
Cause: cause word
Certain: certain word
Bio: biology process word
Body: body word
Health: health word
Home: home word
Money: money word
Death: death word
Psychology: psychology word

We used the Chinese LIWC dictionary to analyze the differences in language use between users with depression and non-depressed users, and Table 6 presents the comparative results. The results in Table 6 show that users with depression are more likely to use negative vocabulary, such as “Swear,” “Affect,” “PosEmo,” “NegEmo,” “Anx,” “Anger,” “Sad,” etc., than non-depressed users. Depressed users seemed to prefer discussing past and present events (PastM, PresentM), whereas non-depressed users appeared to focus more on possible future events (FutureM). We speculated that this difference might be because many depressed users were more heavily influenced by their family of origin and were more inclined to reflect on the impact of past events on them in their posts. Furthermore, we also noticed that depressed users were comparatively more negative than non-depressed users when discussing topics related to Social, Family, Friend, and Home. Additionally, we found that words such as “Bio,” “Body,” “Health,” “Death,” and “Psychology” were used more frequently in the posts of depressed users. The primary reason for this is that posts by depressed users may express their

intentions related to suicide or self-harm, or they may involve sharing cases and discussions about the condition among fellow patients, encompassing the diagnosis process, physical condition, and medication.

Discussion

Principal Results

This paper explores the automatic prediction of depression risk among users on online social media using deep learning methods and develops and validates the model on a large-scale dataset of online social media users. The research findings indicate that the model we developed exhibits significant advantages in predicting depression risk, confirming the effectiveness and advanced capabilities of using deep learning for depression risk prediction. The paper has several implications:

Firstly, with the rapid development of social media technology, an increasing number of young people are using social media to share their emotions and document their lives. Social media has become a crucial platform for them to express emotions, seek support, and establish social connections. However, mental health issues among young people are becoming more pronounced, making it a key societal concern. Social media serves as an important tool for them to communicate their feelings and connect with others. Nevertheless, it also presents a challenge in effectively utilizing social media data to identify and support individuals who may be facing mental health issues. An increasing number of individuals with mental health problems, especially those with depression, do not actively seek help from professionals. This results in a lack of timely treatment and support, causing them to miss optimal intervention opportunities. Furthermore, there is an increasing shortage of clinical psychologists to meet the growing mental health needs of the population. Hence, exploring automated depression risk identification technologies based on artificial intelligence, particularly deep learning, has become an crucial and essential research topic in addressing the current societal challenges.

Moreover, this study developed a hierarchical Transformer network and proposed a retrieval-enhanced post-sampling technique to enhance the performance of depression risk detection. Experimental results indicate that our developed approach outperforms all baseline methods, achieving a prediction accuracy and F1 score of 84% across three independent experiments. With the application of the retrieval sampling technique, the performance of almost all methods reaches approximately 90%. Compared to methods without sampling, there is a performance improvement of over 10% across the four metrics. This strongly demonstrates the effectiveness and advanced capabilities of our approach in predicting depression risk.

Finally, linguistic analysis revealed that depressed users exhibit more conservative and reserved social behavior on social media compared to non-depressed users. Not only do they make fewer posts, but their posts are also shorter in length. This may reflect their negativity in social interactions and a relative avoidance of social engagement. The reduced social engagement could be a result of the loneliness, frustration, or lack of motivation commonly felt by depressed individuals. Additionally, depressed users express more negative emotions in their posts. Through linguistic sentiment analysis, we found that posts by depressed users contain more negative sentiment words, a difference that is more pronounced compared to non-depressed users. This further highlights the psychological distress and negative emotional experiences that depressed individuals may encounter on social media. These characteristics offer insights into the behavioral characteristics of depressed users, providing direction for developing more accurate and personalized depression risk prediction models.

Limitations

This study has several limitations. Firstly, due to noticeable individual differences among users on

different social media platforms, the research model and findings in this paper may not accurately assess the risk of depression in online users, nor do they account for the diversity among individuals. Secondly, our focus was narrowed to Sina Weibo users, and they may not entirely represent the Chinese population or all Chinese social media users. Therefore, the research results may not be generalizable to users on other social media platforms or other populations with different medical conditions.

Conclusions

In this study, we investigate the use of deep learning techniques to predict the risk of depression based on social media data. We collected posts from 3,200 online users over a one-year period in order to develop and validate a depression risk detection model. The proposed hierarchical Transformer network demonstrated exceptional performance on the collected data, yielding predictive accuracy of over 95% across four commonly employed evaluation metrics. Furthermore, we introduced a retrieval-based post sampling technique, which significantly improved our model's ability to detect the risk of depression. This research provides technical support for the automatic identification of users at risk of depression on Chinese online social media, thereby effectively supporting online platforms in engaging in societal risk management.

Acknowledgements

This work was supported in part by the Sci-Tech Innovation 2030-Major Project of Brain science and brain-inspired intelligence technology (2021ZD0202003); in part by the National Natural Science Foundation of China under Grant 62227807 and 62072219; in part by and the Fundamental Research Funds for the Central Universities (lzujbky-2023-10), and in part by the Supercomputing Center of Lanzhou University.

Conflicts of Interest

None declared

Abbreviations

NLP: Natural Language Processing
CV: Computer Vision
SVM: Support Vector Machine
NB: Naive Bayes
CNN: Convolutional Neural Network
LSTM: Long Short-Term Memory
GRU: Gated Recurrent Unit
BiGRU: Bidirectional Gated Recurrent Unit
BiLSTM: Bidirectional Long Short-Term Memory
LSTM-Attn: Long Short-Term Memory with Attention
GRU-Atten: Gated Recurrent Unit with Attention
BiLSTM-Atten: Bidirectional Long Short-Term Memory with Attention
BiGRU-Atten: Bidirectional Gated Recurrent Unit with Attention
HTN: Hierarchical Transformer Network

References

1. Evans-Lacko S, Aguilar-Gaxiola S, Al-Hamzawi A, Alonso J, Benjet C, Bruffaerts R, Chiu WT, Florescu S, de Girolamo G, Gureje O, Haro JM, He Y, Hu C, Karam EG, Kawakami N, Lee S, Lund C, Kovess-Masfety V, Levinson D, Navarro-Mateu F, Pennell BE, Sampson NA, Scott KM, Tachimori H, Ten Have M, Viana MC, Williams DR, Wojtyniak BJ, Zarkov

Z, Kessler RC, Chatterji S, Thornicroft G. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the WHO World Mental Health (WMH) surveys. Psychol Med. 2018 Jul;48(9):1560-1571. PMID: 29173244

2. Cosco TD, Lachance CC, Blodgett JM, Stubbs B, Co M, Veronese N, Wu YT, Prina AM. Latent structure of the Centre for Epidemiologic Studies Depression Scale (CES-D) in older adult populations: a systematic review. Aging Ment Health. 2020 May;24(5):700-704. PMID: 30661386.
3. Richter P, Werner J, Heerlein A, Kraus A, Sauer H. On the validity of the Beck Depression Inventory. A review. Psychopathology. 1998;31(3):160-8. PMID: 9636945.
4. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001 Sep;16(9):606-13. PMID: 11556941
5. Zimmerman M, Martinez JH, Young D, Chelminski I, Dalrymple K. Severity classification on the Hamilton Depression Rating Scale. J Affect Disord. 2013 Sep 5;150(2):384-8. PMID: 23759278
6. Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., ... & Patras, I. Deap: A database for emotion analysis; using physiological signals. IEEE transactions on affective computing. 2011; 3:18-31.
7. Liao, S. C., Wu, C. T., Huang, H. C., Cheng, W. T., & Liu, Y. H. Major depression detection from EEG signals using kernel eigen-filter-bank common spatial patterns. Sensors. 2017; 17:1385.
8. Cai, H., Han, J., Chen, Y., Sha, X., Wang, Z., Hu, B., ... & Gutknecht, J. A pervasive approach to EEG-based depression detection. Complexity. 2018; 1-13.
9. Feinstein, A. Multiple sclerosis and depression. Multiple Sclerosis Journal. 2011; 17(11), 1276-1281.
10. Zheng Y, Chen X, Li D, Liu Y, Tan X, Liang Y, Zhang H, Qiu S, Shen D. Treatment-naïve first episode depression classification based on high-order brain functional network. J Affect Disord. 2019 Sep 1;256:33-41. PMID: 31158714
11. Rodrigues, S., Bokhour, B., Mueller, N., Dell, N., Osei-Bonsu, P. E., Zhao, S., ... & Elwy, A. R. Impact of stigma on veteran treatment seeking for depression. American Journal of Psychiatric Rehabilitation. 2014; 17:128-146.
12. Fox AB, Smith BN, Vogt D. How and when does mental illness stigma impact treatment seeking? Longitudinal examination of relationships between anticipated and internalized stigma, symptom severity, and mental health service use. Psychiatry Res. 2018 Oct; 268:15-20. PMID: 29986172.
13. Zhu J, Li Z, Zhang X, Zhang Z, Hu B. Public Attitudes Toward Anxiety Disorder on Sina Weibo: Content Analysis. J Med Internet Res. 2023 Apr 4;25:e45777. doi: 10.2196/45777. PMID: 37014691
14. Eysenbach G, Powell J, Englesakis M, Rizo C, Stern A. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. BMJ. 2004 May 15;328(7449):1166. PMID: 15142921
15. De Choudhury, M., & De, S. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In Proceedings of the international AAAI conference on web and social media. 2014; (8)71-80.
16. Naslund JA, Aschbrenner KA, Marsch LA, Bartels SJ. The future of mental health care: peer-to-peer support and social media. Epidemiol Psychiatr Sci. 2016 Apr;25(2):113-22. PMID: 26744309
17. Pruksachatkun, Y., Pendse, S. R., & Sharma, A. Moments of change: Analyzing peer-based cognitive support in online mental health forums. In Proceedings of the 2019 CHI conference on human factors in computing systems. 2019; 1-13.

18. Lin, H., Jia, J., Nie, L., Shen, G., & Chua, T. S. What Does Social Media Say about Your Stress?. In IJCAI. 2016, July; 3775-3781
19. Squarcina L, Villa FM, Nobile M, Grisan E, Brambilla P. Deep learning for the prediction of treatment response in depression. J Affect Disord. 2021 Feb 15;281:618-622. doi: 10.1016/j.jad.2020.11.104. Epub 2020 Nov 17. PMID: 33248809.
20. Akbari, M., Hu, X., Liqiang, N., & Chua, T. S. From tweets to wellness: Wellness event detection from twitter streams. In Proceedings of the AAAI Conference on Artificial Intelligence. 2016, February; Vol. 30, No. 1.
21. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. CLPsych 2015 shared task: Depression and PTSD on Twitter. Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality. 2015; pp. 31-39.
22. Park, M., Cha, C., & Cha, M. Depressive moods of users portrayed in Twitter. Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining. SIGKDD. 2012; pp. 1-8.
23. Coppersmith, G., Dredze, M., & Harman, C. Quantifying mental health signals in Twitter. Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality. 2014, June; pp. 51-60.
24. Xu R, Zhang Q. Understanding Online Health Groups for Depression: Social Network and Linguistic Perspectives. J Med Internet Res. 2016 Mar 10;18(3):e63. PMID: 26966078
25. Reece, A. G., & Danforth, C. M. Instagram photos reveal predictive markers of depression. EPJ Data Science. 2017; 6(1), 15.
26. Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., ... & Zhu, W. Depression detection via harvesting social media: A multimodal dictionary learning solution. In IJCAI. 2017, August; pp. 3838-3844.
27. Depression and self-harm risk assessment in online forums. arXiv preprint arXiv:1709.01848.
28. Yang, L., Jiang, D., Han, W., & Sahli, H. DCNN and DNN based multi-modal depression recognition. 2017 Seventh International Conference on Affective Computing and Intelligent Interaction. 2017, October; pp. 484-489. IEEE.
29. Gui, T., Zhu, L., Zhang, Q., Peng, M., Zhou, X., Ding, K., & Chen, Z. Cooperative multimodal approach to depression detection in twitter. Proceedings of the AAAI conference on artificial intelligence. 2019, July; Vol. 33, No. 01, pp. 110-117.
30. Song, Y., Shi, S., Li, J., & Zhang, H. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018, June; Volume 2 pp. 175-180.