Identification of College Students' Depressive Tendency Through Text Big Data Analysis

Yazhou Zuo
School of Liberal Arts, Shangqiu University, Shangqiu, Henan 476000, China
Email: y575862@163.com

Keywords: text mining, long short-term memory algorithm, depressive tendency, classification

Received: January 9, 2024

Accurate identification of depressive tendencies in the early stage is beneficial to the treatment and prevention of depression. This paper presents a text-based depressive tendency recognition algorithm used to assist the judgment of depressive tendencies, which employs an intelligent algorithm to classify and identify texts and determines whether the user who publishes the text information is a patient with depressive tendency. The long short-term memory (LSTM) algorithm was used to recognize text information, and the convolution operation in the convolutional neural network (CNN) was introduced to improve the recognition performance of LSTM. After that, the convergence of the algorithm and the influence of various text vectorization on the algorithm were examined using Chinese texts on the Sina Weibo platform in the simulation experiment, and the results were compared with the support vector machine and conventional LSTM methods. It was found that the LSTM algorithm improved by using a CNN not only trained faster but also performed better in identifying depressive tendencies compared to the other two algorithms. The contribution of this paper lies in utilizing a CNN to further extract textual features, thereby enhancing the recognition performance of LSTM and providing an effective reference for accurately and quickly identifying depressive tendencies among college students.

Povzetek: Razvili so algoritem za prepoznavanje depresivnih tendenc pri študentih, ki temelji na besedilnem rudarjenju in LSTM algoritmu, izboljšanem z uporabo CNN. Rezultati kažejo, da kombinacija LSTM in CNN omogoča hitrejšo in točnejšo identifikacijo depresivnih tendenc v primerjavi s SVM in klasičnimi LSTM metodami.

1 Introduction

With the increasing pressure of modern society, the incidence of depression has gradually increased, which has become an increasingly serious public health problem worldwide [1]. College students are one of the groups with a high incidence of depression. This group usually faces the pressure of study, employment, interpersonal relationship, and other aspects, and also needs to deal with self-cognition, identity, and other issues [2]. As a serious mental illness, depression is mainly manifested by low mood, loss of interest, insomnia, and reduced sense of self-worth. However, the pathogenesis of depression is complex, and its pathophysiological process is not clear. When patients are found to have pathophysiological symptoms, the disease is often in the middle and late stage, which increases the difficulty of treatment [3]. Kolenik and Gams [4] found that the rise in mental health issues, particularly among the youth, is not a new phenomenon, and the reasons for the increase in stress, anxiety, and depression included a significant shortage of mental health professionals and regulations, as well as inequitable availability of mental healthcare opportunities. Kolenik and Gams [5] conducted an analysis of the current technological approaches and trends, and analyzed the highly interdisciplinary landscape of intersections between intelligent cognitive assistant, attitude, and behavior change, and mental health. Kolenik [6] applied the Internet of Things to psychological health assessment, aiming to monitor, understand, and identify an individual's mental health issues through their physiological, behavioral, cognitive, emotional states, as well as their environment. Kolenik, Gjoreski, and Gams [7] introduced a personal virtual assistant called permeass with novel environmental intelligence capabilities designed to assist with three mental health problems: stress, anxiety, and depression. If depressive tendency can be detected and intervened in the early stage, it can be more effective in the prevention and treatment of depression. With the improvement of computer performance, text mining technology provides new tools and methods for the identification of depressive tendency. The status of depressive tendency can be evaluated by analyzing the emotional color and semantic orientation among the text information published by college students in daily life. The related works are reviewed in Table 1.

Table 1: Related works

Author	Method	Result
Peng et al. [8]	They proposed a depression patient identification model based on a multi-kernel support vector machine (SVM)	The method was the most appropriate for identifying depressed individuals based on social

		media data.
Xing et al. [9]	They proposed a new two-level depression recognition method.	This approach yielded favorable outcomes in both gender independent and gender dependent experiments.
Tlelocoyotecatl et al. [10]	They designed a depression detection method based on identification and accumulation of symptom evidence through user posts.	The experimental results verified the effectiveness of this method.

These studies have used their respective methods to identify depressive tendencies in text on social media. Some have used commonly used SVM for identification, while others have used feature vocabulary and hierarchical analysis to identify depressive tendencies. This study used deep learning algorithms for identifying depressive tendencies in text on social media and utilized the activation function in deep learning algorithms to fit the patterns of features in big data, thus making more accurate identifications of depressive tendencies. This paper briefly presents a text-based depressive tendency recognition algorithm used to assist the judgment of depressive tendency. The long short-term memory (LSTM) algorithm was used to identify text information, and the convolution operation in the convolutional neural network (CNN) was introduced to improve the recognition performance of LSTM. The convergence status of the algorithm was subsequently tested using texts on the Sina Weibo platform in simulation experiments, along with evaluating the influence of different text vectorization approaches on the algorithm. Additionally, a comparison was conducted between the CNN and LSTM combined, SVM, traditional LSTM algorithms.

The basic structure of this article is: abstract-introduction-the algorithm principle and process of depressive tendency recognition-simulation experiments-conclusions.

2 Depressive tendency recognition based on text data

The emergence of text mining technology provides new methods for the identification of depressive tendency. Patients with early depressive tendency [11] still communicate with the outside world and sometimes leave text messages on social media. These text messages usually contain the psychological and emotional activities of the publishers [12]. Data mining is used to identify the depressive tendencies of publishers by analyzing text containing psychological and emotional information, in order to take preventive measures early on.

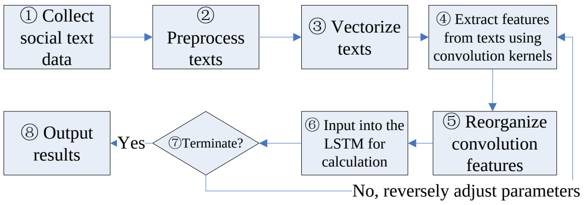
① Collect social text data
② Preprocess texts
③ Vectorize texts
④ Extract features from texts using convolution kernels
⑤ Reorganize convolution features
⑥ Input into the LSTM for calculation
⑦Terminate?
Yes
⑧ Output results
No, reversely adjust parameters

Figure 1: The training flow of the text data-based depressive tendency recognition method designed in this paper

The basic principle of using social text data for depressive tendency recognition in this paper is to first extract features from social text and then use a classifier to classify the features to obtain the recognition results of depressive tendencies [13]. The algorithms that can act as classifiers include decision tree, SVM, deep learning algorithm, etc. This paper uses a deep learning algorithm as the classifier. Among deep learning algorithms, CNN can use convolution kernels to extract local features and combine multiple local features into global features, but it cannot contact the context for sequence data. As an extension of RNN, LSTM can effectively contact the context of sequence data [14], but it is not comprehensive enough in feature extraction. Therefore, this paper combines CNN with LSTM, uses CNN to extract features, and then employs LSTM to classify features. The training process is shown in Figure 1.

① A crawler program is employed to crawl social text data from social platforms.

② The texts are preprocessed by privacy removal, word segmentation, noise reduction, etc. Privacy removal means removing the privacy information in the text data. Word segmentation means segmenting long sentences for later vectorization and recognition [15]. Noise reduction means deleting meaningless words and punctuation from text information to reduce interference.

③ The skip-gram model in Word2vec is employed to vectorize the texts, and the dimension of the text vectors is set according to the demand. Generally, the more the dimension is, the more comprehensive the semantic features can be covered by the vector, but the calculation will also increase. The vectorized texts are combined into a two-dimensional matrix, which is convenient for later feature extraction. The matrix specification is $m \times n$, where m denotes the number of words in the text and n denotes the dimension of the text vector.

④ Multiple convolution kernels in the CNN convolutional layer are used to extract the convolution feature from the two-dimensional matrix of the text vector. During the extraction process, the convolution kernel will slide in the matrix according to a certain step [16]. The extraction formula in the sliding process is:

$$Y_i = f(X_i \otimes W_i + b_i), \quad (1)$$

where Y_i is the convolution output eigenvalue of the i-th convolution kernel, X_i is the input vector of the i-th convolution kernel, W_i is the weight in the i-th kernel, and b_i is the bias of the i-th kernel.

⑤ The convolution features extracted by the convolution kernels are reorganized. The specific operation is to combine the features extracted by different convolution kernels of the same input vector as the convolution combination feature of the input vector. In this paper, features extracted from the text segmentation vector by different convolution kernels are combined as the convolution combination features of the text segmentation vector.

⑥ The convolutional combination features of the text segmentation vector are input into the LSTM according to the word order for calculation [17].

⑦ Whether the training of the algorithm terminates is determined. If it terminates, the parameters in the algorithm are fixed, and the results are output. If it is not terminated, the weight parameters in the algorithm are adjusted reversely according to the result error, and step ④ is returned. The termination criteria include the error converging within the specified range and the training number reaching the set maximum value. If any of the two conditions is met, the training can be terminated. The error calculation uses cross entropy [18].

3 Experimental analysis

3.1 Experimental data

To assess the performance of the algorithm proposed in this paper, simulation experiments were conducted. The text data used in the simulation experiments were obtained from Sina Weibo, a Chinese microblogging platform. Using a self-developed web crawler program, text data was crawled from the 'Shudong' circle related to a user named 'Zoufan' on the Weibo platform, who committed suicide because of depression. This circle belongs to the 'Shudong' rescue project, which is a public welfare activity aimed at helping individuals with depression. As a result, this circle gathers a large number of users. The annotators have studied relevant knowledge on depression before conducting annotations and could make basic judgments on whether there is a tendency towards depression. Under the aforementioned conditions, text information was crawled from the 'Shudong' circle and initially classified using keywords related to depression. Then, further filtering and selection were performed based on the annotators' experiential knowledge. Additionally, in order to protect the privacy of crawled users, desensitization processing has been applied to the crawled texts. Part of the text data is shown in Table 2.

Table 2: Partial experimental data.

Serial number for text	Time	Content
...	...	...
425	April 7, 2019	晚上独处时，心理会猛的一声感到空虚。(English translation: When alone at night, there is a sudden feeling of emptiness in my mind.)
426	May 8, 2019	感觉对周围的一切都提不起劲，活着有些乏味。(English translation: Translation: I feel unenthusiastic about everything around me, life seems a bit dull.)
427	May 25, 2019	好事总是别人的，自己怎么老是碰上倒霉的事呢。(English translation: Good things always happen to others, why do I always encounter unlucky events?)
...	...	...

First, a web crawler program was used to crawl 5,000 microblog texts from the 'Shudong'. Then, another 5,000 microblog texts were crawled from the entire Weibo platform excluding 'Shudong'. After statistical analysis, it was found that the average length of each text was 30 words. Following the method described earlier, the microblog texts were annotated. Subsequently, 60% of the texts with depressive tendencies and 60% of the texts without depressive tendencies were chosen as training sets, while the other texts were used as test sets.

3.2 Experimental setup

The relevant parameter settings for the CNN+LSTM algorithm for depressive tendency recognition are shown in Table 2. The maximum number of iterations was 1,000 during training.

Table 3: Parameter settings for the depressive tendency recognition algorithm using CNN+LSTM.

Parameter	Setup	Parameter	Setup
Word2vec	100 dimensions	LSTM hidden layer 2	64 nodes, adopting the sigmoid activation function
Input layer	100 nodes	LSTM hidden layer 3	128 nodes, adopting the sigmoid activation function
Convolutional layer	64 convolution kernelswith a size of	Fully connected layer	Adopting the softmax function

	2×100 ; the moving step size of the convolution kernel is 1.		
LSTM hidden layer 1	32 nodes, adopting the sigmoid function	Output layer	One node

Regarding the SVM method, the sigmoid function was employed as the kernel, and the value of the penalty parameter was assigned as 1. The LSTM algorithm utilized identical parameters to those of the LSTM component in the combined CNN and LSTM approach.

3.3 Experimental results

The SVM algorithm fitted according to the training set to obtain the support vector surface, which was different from the way of gradually adjusting the parameters of the traditional LSTM and CNN+LSTM algorithms, so only the convergence curves of the traditional LSTM and CNN+LSTM algorithms during training are shown (Figure 2). It was seen from Figure 2 that the loss function of the two algorithms decreased as the number of iterations increased; the proposed CNN and LSTM combined algorithm converged faster, and the loss function value was smaller after stabilization.

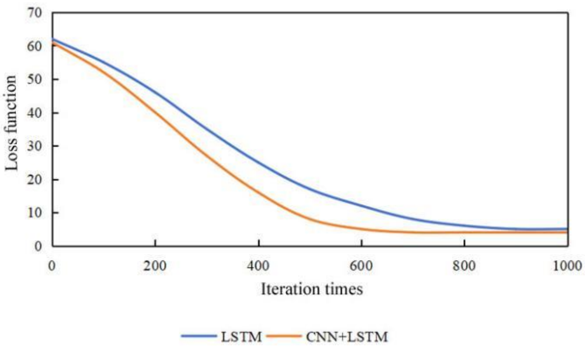

Figure 2: Training convergence curves of the traditional LSTM and CNN+LSTM algorithms.

Before comparing the performance of the CNN and LSTM combined algorithm with other algorithms, the influence of two text vectorization methods, one-hot and Word2vec, on the recognition performance of the CNN and LSTM combined algorithm was compared, as shown in Figure 3. The values of the corresponding indicators of the CNN+LSTM method using one-hot to vectorize the text were 84.48%, 84.57%, and 84.52%, respectively. The precision, recall rate, and F value of the CNN and LSTM combined algorithm using Word2vec to vectorize the text were 94.26%, 95.24%, and 94.75%, respectively. It was concluded that the CNN and LSTM combined method had higher recognition performance when Word2vec was used to vectorize the text.

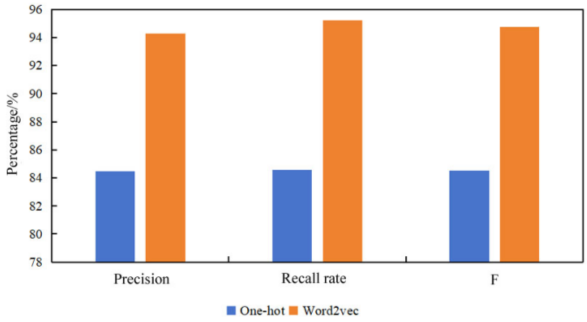

Figure 3: Performance of the CNN+LSTM algorithms under various text vectorization.

The CNN and LSTM combined algorithm was compared with the SVM and LSTM algorithms, and the recognition performance and average recognition time consumption of the three algorithms are shown in Table 4. In terms of identification performance, the CNN and LSTM combined algorithm was the best, followed by the LSTM algorithm which performed moderately well. The SVM algorithm was the worst. The average recognition time of the SVM algorithm was the longest, the average recognition time of the LSTM method was the second, and the average time of the CNN and LSTM combined method was the shortest.

Table 4: Recognition performance and average recognition time of three algorithms.

	SVM	LSTM	CNN+LSTM
Precision/%	77.36	90.12	**94.26**
Recall rate/%	76.58	89.63	**95.24**
F value/%	76.97	89.87	**94.75**
Average recognition time/s	2.23	1.59	**1.03**

4 Discussion

In daily life, people always encounter individuals or situations they do not want to deal with, which can accumulate stress and affect their mood. Generally, these negative emotions can be self-regulated, but there are also situations where it is difficult to regulate them. When negative emotions that are difficult to regulate accumulate over time, they can lead to a tendency towards depression. If not addressed in a timely manner, this psychological tendency towards depression can develop into clinical depression. Once clinical depression occurs, medication is needed for treatment. Therefore, the best approach to dealing with depression is to intervene when there is a tendency towards it and make accurate judgments about this tendency. Similarly, college students also experience stress and tend towards depression in their academic lives on campus. However, college students often have difficulty recognizing this tendency clearly and schools cannot constantly send out psychologists for diagnosis.

However, with the development of the Internet, there are more and more social media platforms available for college students to use, such as Sina Weibo. Compared to communication in daily life, college students tend to express their thoughts on social platforms. However, these text messages themselves have limitations in reflecting their true thoughts and feelings. Therefore, analyzing the publicly posted text messages on social platforms can provide some insights into whether the authors have a tendency towards depression. However, on one hand, there are relatively few vocabulary words directly reflecting psychological tendencies in these publicly posted text messages; instead, it requires context or even a series of continuous posts to understand them fully. On the other hand, there is a large quantity of text information being published on social platforms and analyzing each message individually would be inefficient.

Facing big data, machine learning algorithms can mine and analyze its patterns. This article used the LSTM algorithm in deep learning to identify depressive tendencies in text-based big data. It also combined with the CNN algorithm to enhance the recognition performance of LSTM. Finally, simulation experiments were conducted using microblog text data obtained through web crawling programs. The results were compared with the SVM and traditional LSTM methods. The improved LSTM converged faster during training and had lower training error when stable; the improved LSTM algorithm, which utilizes Word2vec to vectorize text, exhibited superior performance in recognition; in comparison with the other two algorithms, the improved LSTM algorithm showed superior recognition performance. The reasons were analyzed. When identifying depressive tendencies in microblog texts, the SVM algorithm seeks a 'hyperplane' capable of dividing the high-dimensional space and subsequently employs it for classifying the microblog text. The SVM algorithm employed the kernel function’s high-dimensional mapping to effectively transform the nonlinear relationship in the text into a linear one; however, achieving a perfect fit remained challenging. As an extension of the RNN algorithm, the LSTM algorithm could effectively fit the nonlinear law using the activation function in the hidden layer. Moreover, the LSTM algorithm could use historical information to contact the context when recognizing the text, so its identification performance was significantly higher than that of the SVM algorithm. The proposed algorithm combined the CNN and LSTM algorithms. The CNN algorithm utilized the convolution kernel of the convolution layer to capture the text vector's local features, and then the arrangement of the local features were modified according to the text order. The LSTM algorithm was used to recognize depressive tendencies based on the reorganized textual features. The CNN algorithm compensates for the limitation of the LSTM algorithm in utilizing local features, so its recognition performance was superior.

5 Conclusion

This article presents a text-based depressive tendency recognition algorithm, which uses an intelligent algorithm to classify and identify text information and determine whether the user who publishes the text information is a patient with depressive tendency. The LSTM algorithm was used to identify text information, and the convolution operation in the CNN algorithm was introduced to optimize the LSTM algorithm. After that, the convergence of the algorithm and the influence of various text vectorization techniques on the algorithm are tested in the simulation experiment, and compared with the SVM and conventional LSTM methods. The main results are as follows. (1) With the increase of training times, both the traditional LSTM and CNN+LSTM methods converged, and the CNN and LSTM combined method converged faster and had a smaller loss function when it was stable. (2) Compared with one-hot, the CNN and LSTM combined method using Word2vec had superior identification performance. (3) In terms of identification performance, the CNN and LSTM combined method was the best, followed by the LSTM algorithm, while the SVM algorithm performed the worst; the average detection time of the SVM algorithm was the longest, followed by the LSTM algorithm, and the CNN and LSTM combined method was the shortest.

The limitation of this article lies in the narrow scope of depression-related information selected when using a web crawler program to collect text data. Therefore, the future research direction is to expand the range of information collected by the web crawler program to improve he universality of identification algorithms.

References

[1] Xue L, Yan Y, Fan H, Zhang L, Wang S, Chen L (2023). Future self-continuity and depression among college students: The role of presence of meaning and perceived social support. *Journal of Adolescence*, 95(Oct.), pp. 1463-1477. https://doi.org/10.1002/jad.12219

[2] He X, Lopez E (2023). Level of insomnia, symptoms of depression, and anxiety among college students with mobile phone addiction: basis for guidance and counseling program enhancement. *Applied Science*, 13(10), pp. 1641-1648.

[3] He L, Chan JCW, Wang Z (2021). Automatic depression recognition using CNN with attention mechanism from videos. *Neurocomputing*, 422, pp. 165-175. https://doi.org/10.1016/j.neucom.2020.10.015

[4] Kolenik T, Gams M (2021). Persuasive Technology for Mental Health: One Step Closer to (Mental Health Care) Equality?. *IEEE Technology and Society Magazine*, 40(1), pp. 80-86. https://doi.org/10.1109/MTS.2021.3056288

[5] Kolenik T, Gams M (2021). Intelligent Cognitive Assistants for Attitude and Behavior Change Support in Mental Health: State-of-the-Art Technical Review. *Electronics*, 10(11), 1-34. https://doi.org/10.3390/electronics10111250

[6] Kolenik, T. (2022). Methods in Digital Mental Health: Smartphone-Based Assessment and Intervention for Stress, Anxiety, and Depression. *Integrating Artificial Intelligence and IoT for*

Advanced Health Informatics, pp. 105-128. https://doi.org/10.1007/978-3-030-91181-2_7

[7] Kolenik T, Gjoreski M, Gams M (2020). PerMEASS-Personal Mental Health Virtual Assistant with Novel Ambient Intelligence Integration. *AAI4H@ECAI*, pp. 8-12.

[8] Peng Z, Hu Q, Dang J (2017). Multi-kernel SVM based depression recognition using social media data. *International Journal of Machine Learning and Cybernetics*, 10(1), pp. 43-57. https://doi.org/10.1007/s13042-017-0697-1

[9] Xing Y, Liu Z, Li G, Ding Z, Hu B (2022). 2-level hierarchical depression recognition method based on task-stimulated and integrated speech features. *Biomedical Signal Processing and Control*, 72, pp. 1-10. https://doi.org/10.1016/j.bspc.2021.103287

[10] Tlelo-Coyotecatl I, Escalante H J, Montes y Gómez M (2022). Depression Recognition in Social Media based on Symptoms' Detection. *Procesamiento del Lenguaje Natural*, 68, pp. 25-37.

[11] Akbari H, Sadiq M T, Rehman A U, Ghazvini M, Naqvi RA, Payan M, Bagheri H, Bagheri H (2021). Depression recognition based on the reconstruction of phase space of EEG signals and geometrical features. *Applied Acoustics*, 179, pp. 1-17. https://doi.org/10.1016/j.apacoust.2021.108078

[12] Cai H, Qu Z, Li Z, Zhang Y, Hu X, Hu B (2020). Feature-level Fusion Approaches Based on Multimodal EEG Data for Depression Recognition. *Information Fusion*, 59, pp. 127-138. https://doi.org/10.1016/j.inffus.2020.01.008

[13] Yildirim-Celik H, Eroglu S, Oguz K, Karakoc-Tugrul G, Erdogan Y, Isman-Haznedaroglu D, Eker C, Gonul AS (2022). Emotional context effect on recognition of varying facial emotion expression intensities in depression. *Journal of Affective Disorders*, 308, pp. 141-146. https://doi.org/10.1016/j.jad.2022.04.070

[14] Li X, Zhang X, Zhu J, Mao W, Sun S, Wang Z, Chen X, Hu B (2019). Depression recognition using machine learning methods with different feature generation strategies. *Artificial Intelligence in Medicine*, 99(Aug.), pp. 101696.1-101696.15. https://doi.org/10.1016/j.artmed.2019.07.004

[15] Jeyalakshmi C, Murugeswari B, Karthick M (2017). Recognition of emotions in Berlin speech: A HTK based approach for speaker and text independent emotion recognition. *Pakistan Journal of Biotechnology*, 14(1), pp. 63-69.

[16] Singh P, Srivastava R, Rana K, Kumar V (2021). A multimodal hierarchical approach to speech emotion recognition from audio and text. *Knowledge-Based Systems*, 229, pp. 1-17. https://doi.org/10.1016/j.knosys.2021.107316

[17] Batbaatar E, Li M, Ryu K H (2019). Semantic-Emotion Neural Network for Emotion Recognition from Text. *IEEE Access*, pp. 1-13. https://doi.org/10.1109/ACCESS.2019.2934529

[18] Alswaidan N, Menai M (2020). Hybrid Feature Model for Emotion Recognition in Arabic Text. *IEEE Access*, 4, pp. 1-12. https://doi.org/10.1109/ACCESS.2020.2975906