Revista de Psicología del Deporte/Journal of Sport Psychology Vol. 30. n.°3. 2021. pp. 124-132
Universidad de Almería / Universitat Autónoma de Barcelona. ISSN 1132–239X ISSNe 1988–5636

Psychological Crisis Intervention of College Sports Majors Based on Big Data Analysis

Yajing Pang[1]

Abstract
In recent years, the psychological problems of college students have attracted extensive attention. It is of great practical significance to timely and arcuately intervene in college sports majors faced with psychological crisis (PC). However, the existing studies mainly analyze the mood and psychological state at a certain moment, but rarely track the psychological health state of different types of college students. This paper proposes a way to intervene and predict the PC of college sports majors based on big data analysis. Firstly, the massive evaluation data were collected from a psychological census database on PC of college sports majors and subjected to data mining. Besides, a PC evaluation model was established based on the decision tree (DT) algorithm. Next, the behavior big data of college sports majors in social network were fully utilized, and a PC intervention and prediction method was developed based on social network readme texts. Further, the authors extracted features from readme texts, evaluated the level of PC risk, and analyzed the longitudinal features. Finally, the proposed model was proved valid through experiments. This paper effectively applies new technologies to the data mining of the typical behaviors of college sports majors, and thereby realizes accurate PC warning. The research results are of great significance to improving the psychological health of college students.
Keywords: big data analysis; sports; college students; psychological crisis (PC) intervention

Introduction

In recent years, the psychological problems of college students have attracted extensive attention. During the transition from campus life to social life, colleges students are very likely to face various non-self-healing emotional or psychological problems, which arise from problems and perplexity about learning, employment, love, and family (Jia, 2020; Rongfeng & Chaomin, 2021; Tian et al., 2021). College sports majors are particularly vulnerable to psychological crisis (PC), under the combined pressure from cultural course learning and sports skill training (Chang & Xinyi, 2017; Ding & Yang, 2021; Luo et al., 2016). It is of great practical significance to timely and arcuately intervene in college sports majors faced with PC. In fact, PC intervention is the most important link of psychological health education for college students (Wang et al., 2017; Yue & Fangli, 2018).

One of the key indices of PC monitoring and prevention is the development of the psychological management system for college students (Cai, 2017; Liu et al., 2021; McKinley & Ruppel, 2014). Xiong (2020) surveyed the screening indices of PC, and managed psychological warning data dynamically through data mining. In this way, they monitored the psychology of high-risk groups in real time and realized accurate and effective early recognition and prewarning of students' PC. Zhang (2020) constructed a warning system of college students' PC and explained the effective application of data mining in the warning of their PC. Jiang (2018) designed a prototype system for big data analysis on college students' psychological health education, predicted the trend of their psychological behaviors, and constructed personalized emergency management mechanism for their PC, making college PC timelier and more effective. Jiang and Wang (2017) introduced fuzzy clustering algorithm to analyze college students' psychological health state, and studied the warning methods for PC, providing theoretical and practical references for colleges to implement targeted psychological health education and improve college students' psychological health. Drawing on empirical results, Fang (2021) demonstrated that a confident and optimistic attitude can effectively relieve stress and realize self-regulation and stress intervention. On this basis, they highlighted the critical importance of college students' coping strategy for stress and stressed the obviously different pressures on students of different types, levels, and disciplines.

In most colleges, PC warning is implemented by conventional measures, which have very limited effect. On the one hand, the dynamic psychological state of students cannot be acquired timely or effectively through common practices: psychological scale screening of freshmen, lectures on psychological knowledge, and setting up

[1] Hebei University of Science and Technology, Shijiazhuang 050000, China
Corresponding author: pangyj2020@163.com

psychological consultation rooms. On the other hand, despite the rapid development of the Internet technology, the information management of students in most colleges stops at simple collection, storage, and management of student information, using the most basic information technology. Few colleges have effectively mined the correlations behind the massive data.

This paper fully utilizes the massive data on college sports majors' evaluation data on PC and the big data on social network behaviors and proposes a way to intervene and predict the PC of college sports majors based on big data analysis. Section 2 collects and mines the massive evaluation data from a psychological census database on PC of college sports majors and establishes a PC evaluation model based on the decision tree (DT) algorithm. Section 3 processes the behavior big data of college sports majors in social network, develops a PC intervention and prediction method based on social network readme texts, and realized the feature extraction from readme text, the rating PC risk, and longitudinal feature analysis. Section 4 presents the experimental results and demonstrates the effectiveness of our model.

This paper applies information technology and computer technology to the popular field of psychology. The knowledge of different disciplines was fully integrated to generate practical and social values. Through further improvement and development, the proposed strategy could be applied to such fields as clinical care, college management, and psychotherapy.

Evaluation Data-Based PC Intervention and Prediction

The PC of college sports majors could be affected by complicated factors. Figure 1 summarizes the possible consequences of college sports majors' PC under various potential factors. Figure 1 shows the possible consequences of the PC of college sports majors. The stressors mainly refer to the pressures from the following sources: autonomy and independence, social and interpersonal relationships, learning, career, major and sudden changes, heterosexual relationship, family, and economy. This paper firstly collects and mines the massive evaluation data from a psychological census database, and then constructs a PC evaluation model based on the DT, a popular algorithm in data mining. The psychology of the students was rated against the SCL-90 scale. The flow of the proposed model is illustrated in Figure 2.

The evaluation model outputs whether the object has a PC. If the object has a PC, the output will be 1; otherwise, the output will be 0. A total of 6 characteristic attributes were selected to judge whether a sports college student is under PC: personality, mental state, physical health, family economy, cultural course learning state, and sprots skills mastery.

The characteristic attribute with the highest information gain ratio was taken as the current node of the DT. This selection process was implemented recursively to set up a complete DT. The information gain ratio can be calculated in the following steps:

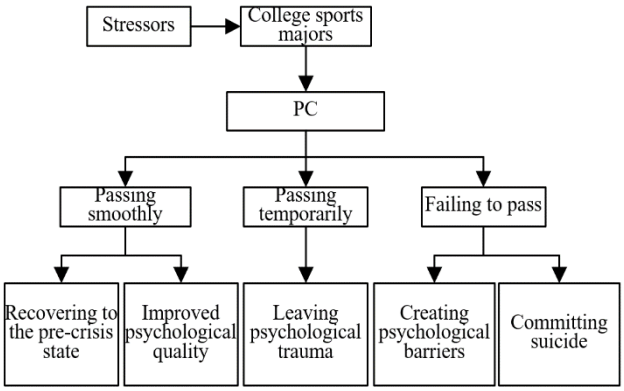

Figure 1. Possible consequences of college sports majors' PC

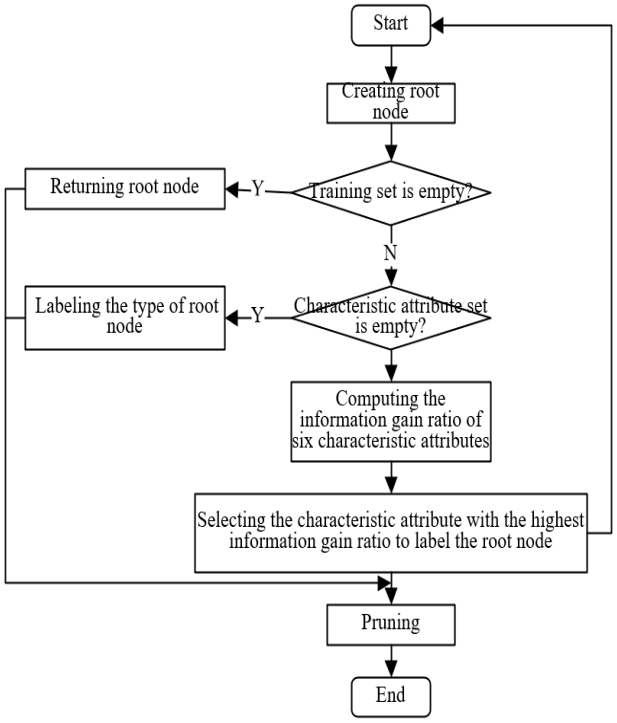

Figure 2. Flow of DT-based PC evaluation model

Step 1. Compute class information entropy

Let R be the training set of PC evaluation data of college sports majors, which contains r data samples. It is assumed that the r samples belong to n classes: D_i (i=1, 2, …, n). Let r_i be the number of samples in D_i; $\delta_i=r_i/R_i$ be the probability of a random sample belonging to D_i, where

$R=R_1+R_2+\ldots+R_n$. For a given PC evaluation sample set of college sports majors, the information volume required for classification, i.e., the class information entropy, can be calculated by:

$$CAIE\left(R_1, R_2, ..., R_n\right) = -\sum_{i=1}^{n} \delta_i log_2 \delta_i \quad (1)$$

Step 2: Compute conditional information entropy

Suppose characteristic attribute P has u different values $\{p_1, p_2, \ldots, p_u\}$. Based on P, R can be divided into d subsets: $\{R_1, R_2, \ldots, R_u\}$. It is assumed that R_j contains all the evaluation data samples in R, which have the same value p_j, and the number of class D_i samples in subset R_j is r_{ij}. Then, the conditional information entropy for dividing R into d subsets by characteristic attribute P can be computed by:

$$COIE\left(P\right) = -\sum_{j=1}^{u} \frac{r_{1j} + r_{2j} ... + r_{nj}}{r} CAIE\left(R_{1j} + R_{2j} ... + R_{nj}\right) \quad (2)$$

$$CAIE\left(R_{1j} + R_{2j} ... + R_{nj}\right) = -\sum_{i=1}^{n} \delta_{ij} log_2 \delta_{ij} \quad (3)$$

The weight of subset j can be expressed as:

$$\frac{r_{1j} + r_{2j} ... + r_{nj}}{r} \quad (4)$$

Formula (4) shows the weight of subset j equals the ratio of the number of evaluation samples in the subset with characteristic attribute $P=p_j$ to that number in R. The smaller the $COIE(P)$, the more reasonable the subset division. The probability δ_{ij} for a sample in R_j belonging to class D_i can be calculated by:

$$\delta_{ij} = \frac{r_{ij}}{\left|r_j\right|} \quad (5)$$

Step 3: Compute information gain

The information gain of the DT branch for characteristic attribute P can be calculated by:

$$IG\left(P\right) = CAIE\left(R_1, R_2, ..., R_n\right) - COIE\left(P\right) \quad (6)$$

Step 4: Compute split information entropy

Suppose characteristic attribute P has u different values. Then, the evaluation samples in R were classified with characteristic attribute P as the benchmark. The split information entropy of P can be calculated by:

$$SIE\left(P\right) = -\sum_{j=1}^{u} \delta_j log_2 \delta_j \quad (7)$$

R_j is the subset of value j in characteristic attribute P. Then, δ_j can be calculated by:

$$\delta_j = \frac{R_j}{r} \quad (8)$$

Step 5: Compute information gain ratio

The information gain ratio of characteristic attribute P can be calculated by:

$$IG - DA\left(P\right) = \frac{IG\left(P\right)}{SIE\left(P\right)} \quad (9)$$

PC Intervention Prediction Based on Social Network Readme Texts

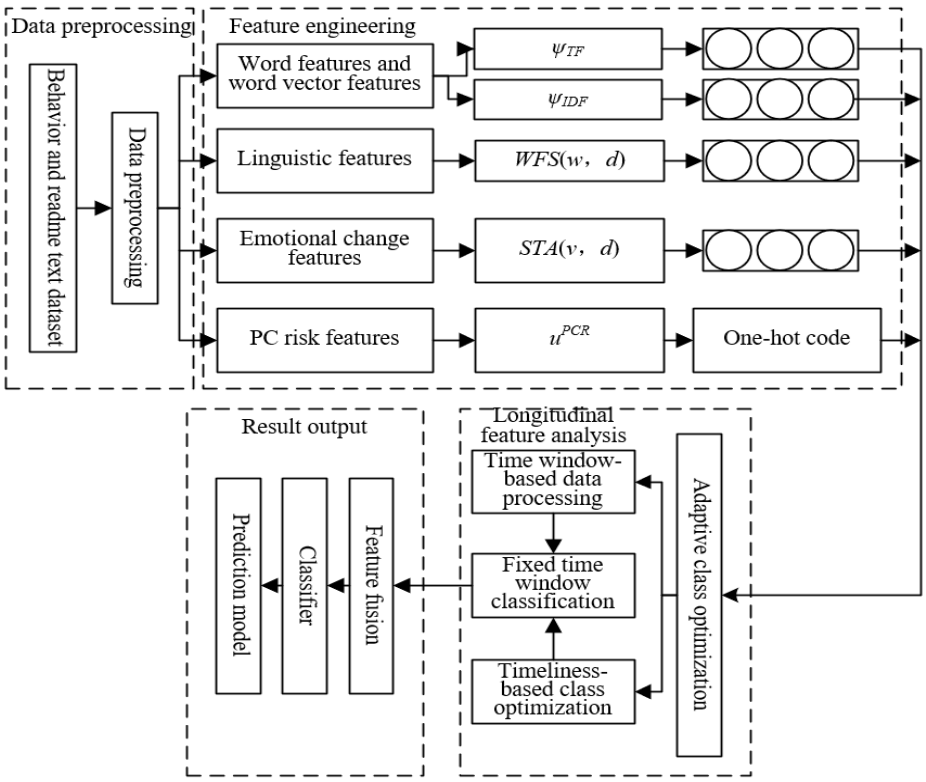

Figure 3. Structure of PC intervention prediction model

The big data on social network behaviors provide important samples for judging whether college sports majors face PC. In social network, the readme texts of college sports majors cover complex contents. Multi-feature transformation is needed to convert the texts into the original data for PC analysis. Figure 3 shows the structure of our PC intervention prediction model based on social network readme texts.

Feature extraction from readme texts

From the posts of college sports majors on social network, it is possible to acquire social speeches containing various information within a period. The text contents of social speeches can be directly represented in units of word. Here, each word is taken as a dimension to build a bag-of-word model. Then, the frequency of words and inverse documents in the model were counted to realize the mapping of a social speech text to a high-dimensional vector space.

Let $Q=\{q_1, q_2, \ldots, q_N\}$ be the set of all the words q in the corpus, where N is the number of different words in the corpus; $W=\{w_1, w_2, \ldots, w_M\}$ be the set of all the readme sentences in the corpus, where M is the number of readme sentences in the corpus; $W(q_i)$ be the set of readme sentences containing word q_i in set W ($W(q_i) \in R$); H be the number of readme sentences in $W(q_i)$.

For a target sentence w_j (length: $|w_j|$) belonging to set W, the frequency of word q_i in that sentence is denoted as $E(q_i, w_j)$. Then, the word frequency of q_i in w_j can be calculated by:

$$\psi_{TF\text{-}ij} = \frac{E(q_i, w_j)}{|w_j|} \tag{10}$$

The inverse document frequency (IDF) corresponding to q_i can be calculated by:

$$\psi_{IDF\text{-}i} = log\left(\frac{M}{1+H}\right) \tag{11}$$

The word-IDF frequency, which measures the importance of a word, can be calculated by:

$$(\psi_{TF\text{-}IDF})_{ij} = \psi_{TF-ij} \cdot \psi_{IDF-i} \tag{12}$$

In the vector space, w_j can be described as an N-dimensional vector:

$$u(w_j) = \left[(\psi_{TF-IDF})_{1j}, (\psi_{TF-IDF})_{2j}, \ldots, (\psi_{TF-IDF})_{Nj}\right] \tag{13}$$

For the contextual information of social speech texts, the relevant features can be extracted with the word2vec model for a single-layer neural network and trained by continuous bag-of-words model or skip-gram model.

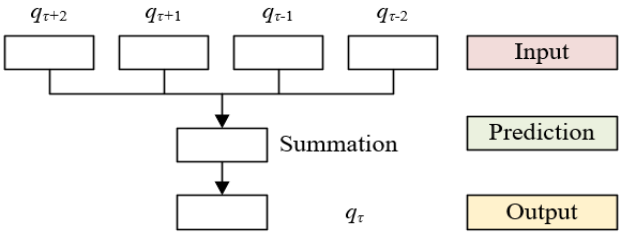

Figure 4. Structure of continuous bag-of-words model

The continuous bag-of-words model can predict the vector of a word based on its context. Figure 4 shows the structure of continuous bag-of-words model. The prediction problem can be modeled by conditional probability: $CP(q_\tau | q_{\tau-d}: q_{\tau+d})$. Suppose a given readme sentence $w=\{q_1, q_2, \ldots, q_\varphi\}$ contains φ words, with q_τ being the target word and d being the size of contextual window. Then, the following objective function can be established to maximize the log likelihood function in the above formula:

$$OF = \frac{1}{\varphi}\sum_{\tau=1}^{\varphi} logCP(q_\tau \mid q_{\tau-c} : q_{\tau+d}) \tag{14}$$

The skip-gram model can predict the words in the context window of the target word. Figure 5 shows the structure of the skip-gram model. For a given readme sentence $w=\{q_1, q_2, \ldots, q_\varphi\}$, the objective function can be defined as:

$$OF = \frac{1}{\varphi}\sum_{\tau=1}^{\varphi}\sum_{-d\le j\le d, j\neq 0} logCP(q_{\tau+j} \mid q_\tau) \tag{15}$$

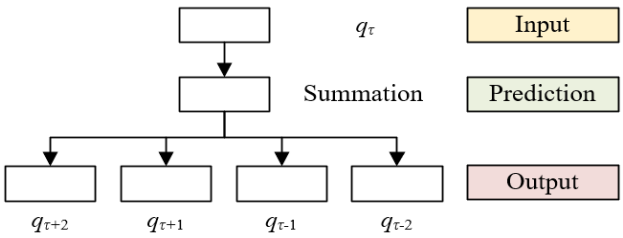

Figure 5. Structure of skip-gram model

According to the LIWC Dictionary (Linguistic Inquiry and Word Count), the differences among college sports majors in a certain type of words in social network readme texts can be quantified and used to prejudge and warn the presence of PC among these students.

The LIWC Dictionary contains the following types of words: personal pronouns like *You_d* and *I_d*; emotional features like *Anger_d*, *depression_d*, and *anxiety_d*; social words like *Military_d*, and *epidemic_d*; cognitive processes like *Money_d*, and *Love_d*. For the word sequence $w=\{q_1, q_2, \ldots, q_\varphi\}$ in a readme sentence of the length φ, the word frequency of word class d in LIWC Dictionary can be calculated by:

$$WFS(w,d) = \frac{1}{k}\sum_{i=1}^{k}\alpha(q_i, d) \tag{16}$$

$$\alpha(q_i,d)=\begin{cases}1 & if \quad q_i \in d\\ 0 & if \quad q_i \notin d\end{cases} \tag{17}$$

The statistical value varies greatly between word classes. The maximum statistical frequency $max_d(WFS(w, d))$ of each class of words in readme sentences can be normalized by:

$$WFS^*(w,d)=\frac{WFS(w,d)}{\max\limits_{d}(WFS^*(w,d))} \tag{18}$$

The empirical, physiological, and behavioral elements that affect the emotions of college sports majors all fluctuate over time. This paper tries to extract the emotional dynamics of the fluctuations, such as trajectories, modes, and laws, and defines the stability of emotional features of college sports majors as *STA*.

For a college sports major v publishing m^v posts, his/her readme sentences can be described as a set $W^v=\{w_1, w_2, \ldots, w_m\}$, and the publishing time of the sentences as $T^u=\{t_1, t_2, \ldots, t_m\}$. The statistic of word class d, i.e., the WFS feature of readme sentence i (w_i), can be described as $WFS(w_i, c)$. Then, the WFS feature stability $STA(v, d)$ based on the word class difference between two adjacent readme sentences can be expressed as:

$$STA(v,d)=\frac{\sum_{i=1}^{m^v-1}(WFS(w_{i+1},d)-WFS(w_i,d))^2}{m^v-1} \tag{19}$$

The emotional stability index $STA(v, d)$ abstracts the personal emotional change of a college sports major into the transformation between his/her adjacent readme sentences. The emotional change feature corresponding to word class d can be described by:

$$QOK(w_{i+1},d)=\frac{(WFS(w_{i+1},d)-WFS(w_i,d))^2}{t_{i+1}-t_i} \tag{20}$$

$$u_i^{QOK}=[QOK(w_i,d_1),\ldots,QOK(w_i,d_l)] \tag{21}$$

where, u_1^{QOK} is the zero vector for the first readme sentence $u_1^{QOK}$of college sports major v. All the readme sentences $W^v=\{w_1, w_2, \ldots, w_m\}$ can be mapped into a sequence of m^v eigenvectors $U^{QOK}=\{u^{QOK}_1, u^{QOK}_2, \ldots, u^{QOK}_m\}$. The emotional change feature characterizes the frequency stability of a class of emotional words used by a college sports major, and, to a certain extent, reflects the emotional fluctuations of him/her.

PC risk rating

Based on support vector machine (SVM), this paper rates the PC risk of each social network readme sentence of college sports majors. Through SVM classification, each readme sentence was assigned a label of PC risk level. Figure 6 shows the principle of SVM.

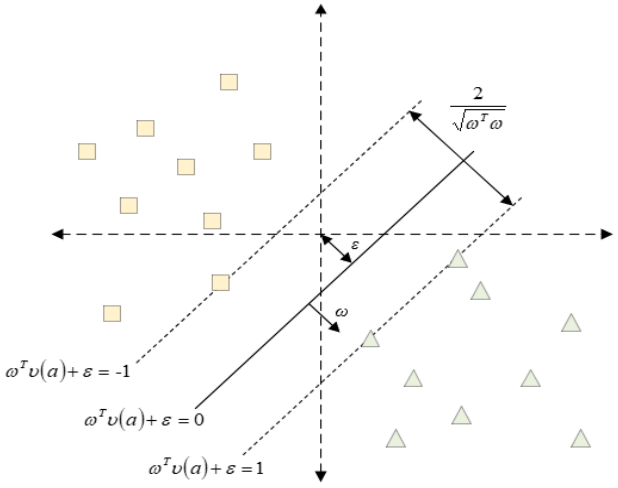

Figure 6. Principle of SVM

The training vectors of social network readme sentences were defined as $a_i \in R^h$, i =1, 2…, m. For the binary classification problem (presence/absence of PC), the predicted class was assumed to be $b_i=\in \{0, 1\}^N$, and the transformed feature space be $\upsilon(a)$. Then, the hyperplane linear model can be defined as:

$$b=\omega^T\upsilon(a)+\varepsilon \tag{22}$$

After the transformation of feature space, the classification problem is equivalent to solving the following formula:

$$\begin{cases}\min\limits_{\omega,\varepsilon,\gamma}\frac{1}{2}\omega^T\omega+D_U\sum_{i=1}^{m}\gamma_i\\ b^T(\omega^T\upsilon(a_i)+\varepsilon)\geq 1-\gamma_i\\ \gamma_i\geq 0, i=1,\ldots,m\end{cases} \tag{23}$$

Let g be all vectors; D_U be the upper bound; W be an m-the order semidefinite matrix. Then, we have:

$$\begin{cases}\min\limits_{\beta}\frac{1}{2}\beta^T\Psi\beta-g^T\beta\\ b^T\beta=0\\ 0\leq\beta_i\leq D_U, i=1,\ldots,m\end{cases} \tag{24}$$

Let $L(a_i,a_j)=\upsilon(a_i)^T\upsilon(a_j)$ be the kernel of the SVM model. W can be defined by:

$$W_{ij}=b_ib_jL(a_i,a_j) \tag{25}$$

Longitudinal feature analysis

This paper mainly aims to summarize the readme texts of each college sports major, which appear when the student is willing to seek help, based on the existing data on such texts in social network, and predict whether he/she is

willing to seek PC intervention.

Different from the longitudinal analysis of traditional psychology, this paper redivides all readme texts based on the time window, and extracts and analyzes the longitudinal features. The purpose is to identify the periods where college sports majors seek help for PC intervention. If help-seeking words appear in the readme sentences of college sports major v during the time window of the length Φ, then the readme sentence sequence of him/her in the previous Φ days will be treated as a positive sample W^+. Meanwhile, the readme sentence sequence of a college sports major that does not send any help-seeking text in that period will be treated as a negative sample W^-. Then, a sample $W=<w_1, w_2, \ldots, w_r>$ containing r readme sentences is either a positive sample W^+ or a negative sample W^-. Firstly, the u_i^{FV}, $FV \in \{\psi_{TF\text{-}IDF}, WFS, QOK\}$ of each readme sentence needs to be computed. Then, the mean of all readme sentences in W should be solved in three characteristic dimensions: $\psi_{TF\text{-}IDF}$, WFS, and $QOK3$:

$$u^{FV} = \frac{1}{r}\sum_{i=1}^{r} u_i^{FV} \tag{26}$$

In W, the number of readme sentences with the crisis level DJ is denoted as $PN^*(DJ)$; the mean and variance of this eigenvalues on W^+ and W^- are denoted as λ and θ, respectively; the PC risk of readme sentence w is denoted as DJ_w, $DJ_w \in \{I, II, III, IV, V\}$. After feature accumulation and Z-score normalization of the readme sentences in the time window, the number w of readme sentences with the crisis level DJ in W can be valued on DJ as:

$$PN(DJ) = \sum_{w \in W} \alpha(DJ_w = DJ) \tag{27}$$

$$PN^*(k) = \frac{1}{\theta}\left|PN(DJ) - \lambda\right| \tag{28}$$

Eventually, the PC risk features of W can be expressed as:

$$u^{PCR} = \left[PN(I), PN(II), PN(III), PN(IV), PN(V)\right] \tag{29}$$

Considering the total posts and interval between adjacent posts of college sports majors, this paper proposes an adaptive optimization scheme for the time window. Suppose the readme sentence set $W^v = \{w_1, w_2, w_3, \ldots, w_{mv}\}$ of a college sports major contains m^v readme sentences published with an interval of $\Delta\tau$. The adaptive time window φ_{ATP} of the student equals the median of $\Delta\tau$:

$$\varphi_{ATP} = ME\left(\Delta\tau * ln\, m^v\right) \tag{30}$$

The adaptive time window helps to divide the data of every college sports major more in line with his/her personal attributes. A short time window is suitable for the students with a high posting frequency, while a long-time window is suitable for the students with a low positing frequency. Hence, the adaptive time window ensures the accuracy of the PC prediction for college sports majors.

The time windows must contain sufficient information, while meeting timeliness requirement. The quality of time windows can be measured by timeliness-information ratio (TIR).

For the post set $W^v = \{w_1, w_2, w_3, \ldots, w_{mv}\}$ of a college sports major, W^v can be divided into $W^v = \{W_1, W_2, W_3, \ldots, W_m\}$, following the proposed division scheme. The help-seeking sentences in the set can be described by $W_m = \{w_i, w_{i+1}, w_{i+2}, w^{HELP}, \ldots, w_j\}$. The response time for these sentences is the time difference Φ_{RES} between the first sentence W_1 in the time window and W^{HELP}. Besides, the total number of posts m_{TW} was defined as the information volume in the time window. Then, TIR can be computed by:

$$TIR = \frac{\Phi_{RES} * \Phi}{(m_{TW})^2} \tag{31}$$

Linguistic feature can characterize the wording habits of each college sports major and help to extract the features of the period when he/she is willing to seek PC intervention. Let $WFS(W, d)$ be the eigenvalue of word class d in W; $v(d)$ be the mean of d for $WFS(W, d)$ in all sentence samples; $A^+ = \{W_1^+, W_2^+, \ldots, W_{m+}^+\}$ be the set of m^+ positive samples; $A^- = \{W_1^-, W_2^-, \ldots, W_{m-}^-\}$ be the set of m^- negative samples. Then, the density of word class feature d in positive and negative samples can be respectively calculated by:

$$\sigma^+(d) = \frac{1}{m^+}\sum_{i=1}^{m^+}\left(WFS\left(W_i^+, d\right)\right) \tag{32}$$

$$\sigma^-(d) = \frac{1}{m^-}\sum_{i=1}^{m^-}\left(WFS\left(W_i^-, d\right)\right) \tag{33}$$

Formulas (32) and (33) show that $\sigma^+(d)$ and $\sigma^-(d)$ are the mean of $WFS(W, d)$ on W^+ and W^-, respectively. The standard deviation ξ of word class d on W^+ and W^- can be respectively calculated by:

$$\xi^2(d) = \frac{\left(\sigma^-(d) - \lambda(d)\right)^2 + \left(\sigma^+(d) - \lambda(d)\right)^2}{2} \tag{34}$$

$$\lambda(d) = \frac{\sum_{i=1}^{m^+}\left(WFS\left(W_i^+, d\right)\right) + \sum_{i=1}^{m^-}\left(WFS\left(W_i^-, d\right)\right)}{m^+ + m^-} \tag{35}$$

The smaller the ξ, the smaller the difference of the word class on different sentence samples. Then, word class d can be sorted by the standard deviation.

Experiments and Results Analysis

The existing studies have completed the design of PC prevention system based on data mining, and provided the testing method, process, and result. Using the graphical user interface (GUI) tool of MATLAB 2014a, this paper designs and synthetizes three different kernel programs for

psychological data mining and embeds the synthetized program into the psychological management system, aiming to improve the effectiveness of data mining in PC prevention.

To clearly understand the activity of college sports majors in social network, this paper firstly counts the posting intervals in their social network behavior dataset. As shown in Figure 7, most college sports majors posted texts with an interval of 2-4d. Overall, their posting intervals are roughly normally distributed.

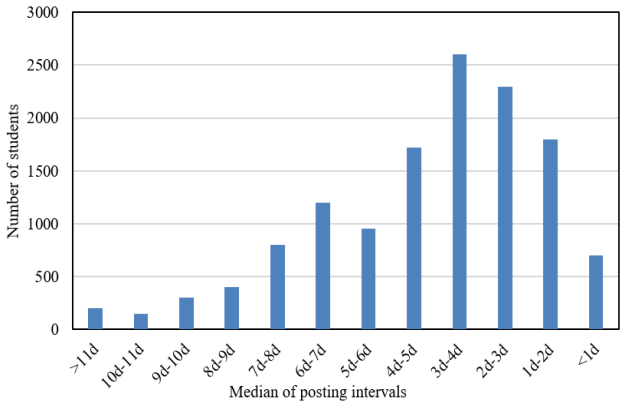

Figure 7. Distribution of median of posting intervals

All readme text data were divided by four fixed time windows: 1d, 2d, 4d, and 8d. The time windows for college sports majors with a long positing interval mostly contain only one readme text each. To grasp the trend of the attributes of each student in the time window, this paper carries out longitudinal feature analysis. If the time window is too small, it would be difficult to identify the emotional change or fluctuation in readme texts, which dampens the effect of longitudinal analysis. If the time window is too large, the readme texts will be weakly correlated, and could not be allocated to the same class for analysis.

Table 1

TIR values of different fixed time window schemes

Time window	1d	2d	4d	8d	φATP
ΦRES	1.027	1.439	2.516	4.218	2.757
mTW	1.065	1.872	4.518	5.376	4.253
TIR	0.854	1.246	0.769	1.453	0.575

The readme texts of every college sports major were divided by different fixed time window schemes, and the mean TIR in each time window was computed as the predicted score of PC intervention for the student. The higher the score, the less ideal the time window scheme for PC prediction. Table 1 lists the TIR values of different fixed time window schemes. The response times Φ_{RES} of the schemes were basically around 2d, belonging to the reasonable observational cycle for psychological diagnosis and treatment. With the extension of the time window, the mean window length was on the rise. In general, the adaptive time window φ_{ATP} scheme achieved the best effect. This scheme can configure time windows in line with the posting habits of college sports majors.

Table 2

Distribution of some word classes

Word class d	$\sigma+(d)$	$\sigma-(d)$	Standard deviation
Personal pronouns	0.653	0.271	0.235
Faith class	0.846	0.523	0.195
Appellation class	0.758	0.486	0.187
Anxiety class	0.671	0.459	0.125
Anger class	0.842	0.571	0.124
Association class	0.475	0.653	0.172
Depression class	0.718	0.476	0.115
Money class	0.736	0.579	0.085
Emotion class	0.833	0.658	0.076
Paranoid class	0.651	0.465	0.082

The word classes d was sorted by the ξ value computed by formula (34). Some of the results are given in Table 2. The word classes in the table are highly discriminatory in the collected positive and negative samples, and valuable for judging whether the object suffers from PC. When a college sports major is willing to seek PC intervention, he/she used anger class and association class words more frequently and used the more intimate appellation class and faith class words less frequently. In this way, the emotional tendency of the student could be fully demonstrated for that period. Moreover, the readme texts of the student mentioned emotion class, money class, anxiety class, and paranoid class words with a high frequency. In addition, physical health and interpersonal relationship are the main inducers of college sports majors seek for PC intervention.

Table 3

Distribution of QOK eigenvalues

Word class c	Mean QOK on $W+$	Mean QOK on $W-$	Standard QOK deviation
Association class	0.431	0.075	0.223
Anger class	0.528	0.175	0.162
Emotion class	0.436	0.279	0.121
Faith class	0.045	0.268	0.118
Money class	0.437	0.375	0.115
Anxiety class	0.568	0.352	0.113
Paranoid class	0.293	0.127	0.086
Depression class	0.015	0.123	0.052
Physical health class	0.126	0.034	0.047
Interpersonal relationship class	0.0974	0.015	0.036

Compared with Table 2, Table 3 contains two new word classes, namely, physical health and interpersonal relationship, in terms of the distribution of QOK eigenvalues. There was no significant difference between negative and positive readme texts in the two tables, in the classes like personal pronouns and appellation. Hence, association class and appellation class words appear frequently when a college sports major is willing to seek PC intervention. Only in such a period, could the student talk about the causes of negative emotions, and name the associated figures. The emotion class words fluctuated more violently on negative samples than on positive samples, suggesting that a college sports major often complain or talk out to others when he/she is not willing to seek PC intervention. From the angle of linguistics and statistics, it is necessary to issue a PC warning if a college sports major published the above classes of words or the relevant words in his/her readme sentences in social network.

Table 4

Prediction results of PC intervention with different features

Time window	Metric	$\psi_{TF\text{-}IDF}$	WFS	QOK	PCR
1d	Precision	0.556	0.572	0.527	0.482
	Recall	0.637	0.518	0.482	0.453
	Composite index	0.592	0.542	0.476	0.472
2d	Precision	0.618	0.668	0.561	0.486
	Recall	0.525	0.523	0.572	0.572
	Composite index	0.634	0.674	0.583	0.516
4d	Precision	0.687	0.652	0.654	0.534
	Recall	0.542	0.541	0.618	0.525
	Composite index	0.686	0.618	0.627	0.533
8d	Precision	0.672	0.625	0.634	0.572
	Recall	0.547	0.607	0.528	0.581
	Composite index	0.628	0.561	0.572	0.563

To verify the influence of different features ($\psi_{TF\text{-}IDF}$, *WFS, QOK, QOK*) with different time windows on the prediction results of PC intervention, the readme texts were divided by 1d, 2d, 4d, and 8d, respectively, and the prediction results with different features were compared through experiments. Table 4 contrasts the classification effects of the four features under the four fixed time windows. It can be learned that the evaluation effect of $\psi_{TF\text{-}IDF}$ was better than that of WFS. The different classification effects between the two features arise from the fact that short samples contain fewer information and the corpus for training is not complete. The QOK feature, which characterizes the emotional change of college sports majors, had a poor evaluation effect in relatively short time windows. But the performance improved with the length of the fixed time window. PC risk feature PCR changed similarly.

Table 5

Prediction results with adaptive time window and a fixed time window

Time window	Metric	$\psi_{TF\text{-}IDF}$	WFS	QOK	PCR
8d	Precision	0.647	0.682	0.652	0.672
0.6	Recall	0.572	0.608	0.572	0.654
	Composite index	0.626	0.649	0.618	0.612
Adaptive time window	Precision	0.681	0.642	0.624	0.539
	Recall	0.645	0.647	0.611	0.527
	Composite index	0.672	0.641	0.605	0.563

According to the social network posts, the adaptive time window was determined for each college sports major. On this basis, the eigenvalues of his/her posts were processed, and classified. Table 5 compares the prediction results with the adaptive time window and a fixed time window. Compared with the fixed time window, the adaptive time window achieved relatively good classification results based on most features ($\psi_{TF\text{-}IDF}$, *WFS, QOK, PCR*). Considering the different posting habits of college sports majors, the adaptive time window can better balance the text contents, properly lengthen or shorten the statistical windows, and ensure every time window contain enough posts, despite that some college sports majors are not so active. The adaptive time window performed better on emotional fluctuation features like *QOK* and *PCR* than on $\psi_{TF\text{-}IDF}$ and *WFS*.

Conclusions

This paper proposes and validates a PC intervention and prediction method based on big data analysis. Firstly, a massive amount of PC evaluation data was collected from a psychological census database and subjected to data mining; a PC evaluation model was established based on the DT algorithm. Next, a PC intervention prediction method was proposed based on social network readme texts. Drawing on the big data about the students' social network behaviors, the features were extracted from readme texts, the PC risk was rated, and the longitudinal features were analyzed. Through experiments, the TIRs of different fixed time window schemes were compared, and the prediction result of a fixed time window scheme was contrasted with that of the proposed adaptive time window. The comparison shows our adaptive time window boasts the best effects. Further, QOK eigenvalue distribution was derived, as well as the prediction results on PC intervention with different features. The results were analyzed to provide a reference for the design and implementation of PC intervention measures for college sports majors.

In addition, this paper introduces natural language processing to psychological health problem and predicts the proper time for PC intervention based on the readme texts of college students. This strategy facilitates the communication between administrators with college students and promotes the treatment of psychological diseases.

References

Cai, D. (2017). The mental health education system of college Students based on the computer platform. *AGRO FOOD INDUSTRY HI-TECH*, *28*(1), 735-739.

Chang, Q., & Xinyi, L. (2017). Psychological crisis of college students in a computer network environment. *AGRO FOOD INDUSTRY HI-TECH*, *28*(1), 1963-1966.

Ding, J., & Yang, L. (2021). College students' mental health evaluation method based on WLLE and MLP. 2021 2nd International Conference on Artificial Intelligence and Information Systems,

Fang, L. (2021). The Prevention and Analysis of College Students' Psychological Crisis Based on Data Mining Technology. In M. Atiquzzaman, N. Yen, & Z. Xu, *Big Data Analytics for Cyber-Physical System in Smart City* Singapore.

Jia, X. (2020). Psychological Characteristics of Special Groups of College Students Based on Artificial Intelligence and Big Data Technology. International conference on Big Data Analytics for Cyber-Physical-Systems,

Jiang, D., & Wang, B. (2017). Study on the intervention model of psychological health education of college students based on stress coping. *Boletin Tecnico/Technical Bulletin*, *55*(18), 70-78.

Jiang, P. (2018). *Research on College Students' Psychological Crisis Intervention in the Context of Big Data* Proceedings of the 2018 International Conference on Big Data and Education, Honolulu, HI, USA.

Liu, J., Shi, G., Zhou, J., & Yao, Q. (2021). Prediction of College Students' Psychological Crisis Based on Data Mining. *Mobile Information Systems*, *2021*, 1-7. https://doi.org/https://doi.org/10.1155/2021/9979770

Luo, Y., Dai, H., Shen, Q., Li, J., Sun, Y., & Zheng, P. (2016). Comparative Analysis of the Psychological Health Education Condition of College Students. International Conference on Man-Machine-Environment System Engineering,

McKinley, C. J., & Ruppel, E. K. (2014). Exploring how perceived threat and self-efficacy contribute to college students' use and perceptions of online mental health resources. *Computers in Human Behavior*, *34*, 101-109. https://doi.org/https://doi.org/10.1016/j.chb.2014.01.038

Rongfeng, H., & Chaomin, G. (2021). Development of psychological health education modes for college students based on information technology. 2021 2nd International Conference on Education, Knowledge and Information Management (ICEKIM),

Tian, Q., Wang, R., Li, S., Wang, W., Wu, O., Li, F., & Jiao, P. (2021). College Students' Psychological Health Analysis Based on Multitask Gaussian Graphical Models. *Complexity*, *2021*, 1-17. https://doi.org/https://doi.org/10.1155/2021/5710459

Wang, M., Song, Y., & Qin, H. (2017). Experimental study on the influence of exercise intervention on the physical and mental health of contemporary college students. *Boletin Tecnico/Technical Bulletin*, *55*(1), 456-461.

Xiong, W. (2020). Identification and Early Warning of College Students' Psychological Crisis Based on Big Data. In *Data Processing Techniques and Applications for Cyber-Physical Systems (DPTA 2019)* (pp. 23-28). Springer. https://doi.org/https://doi.org/10.1007/978-981-15-1468-5_4

Yue, W., & Fangli, L. (2018). Study on the correlation of mental health, resilience and stress events of college students. Proceedings of the 6th International Conference on Information and Education Technology,

Zhang, Y. (2020). Research on College Students' Psychological Crisis Early Warning Method Based on Fuzzy Clustering Algorithm. 2020 International Conference on Computers, Information Processing and Advanced Education (CIPAE),