IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 28, 2020, accepted May 14, 2020, date of publication May 20, 2020, date of current version June 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2995905

Classification of Proactive Personality: Text Mining Based on Weibo Text and Short-Answer Questions Text

PENG WANG[1], YUN YAN[1], YINGDONG SI[1], GANCHENG ZHU[2], XIANGPING ZHAN[1], JUN WANG[1], AND RUNSHENG PAN[1]

[1]School of Psychology, Shandong Normal University, Jinan 250300, China
[2]College of Computer Science and Technology, Jilin University, Changchun 130012, China

Corresponding author: Runsheng Pan (834019895@qq.com)

This work was supported by the 13th Five-year Plan of Education Sciences of Shandong Province under Grant BZZK201901.

ABSTRACT This study focused on the topic of predicting ''proactive personality''. With 901 participants selected by cluster sampling method, targeted short-answer questions text and participants' social media post text (Weibo) were obtained while participants' labels of proactive personality were evaluated by experts. In order to make classification, five machine learning algorithms included Support Vector Machine (SVM), XGBoost, K-Nearest-Neighbors (KNN), Naive Bayes (NB) and Logistic Regression (LR) were deployed. Seven different indicators, which include Accuracy (ACC), F1-score (F1), Sensitivity (SEN), Specificity (SPE), Positive Predictive Value (PPV), Negative Predictive Value (NPV) and Area under Curve (AUC), combined with hierarchical cross-validation were also used to make the comprehensive evaluation of models. With participants' Weibo text and short-answer questions text, we proposed a new approach to classify individuals' proactive personality based on text mining technology. The results showed that short-answer questions + Weibo text datasets had the best performance, followed by short-answer questions text datasets, while the outcome of Weibo text datasets were the worst. However, it is noteworthy that Weibo text has the highest average score on the SPE, which indicated that Weibo text played an important role in identifying individuals with low proactive personality. With Weibo text, SEN was also improved compared with only applying short-answer questions text. In addition, among all three datasets, the indicator SPE is always higher than SEN, indicating this text classification approach was more competent for identifying college students with low proactive personality. As for algorithms, Support Vector Machine and Logistic Regression showed steadier performance compared with other algorithms.

INDEX TERMS Machine learning, proactive personality, text mining.

I. INTRODUCTION

''How can we better achieve the success of our career?'' This question has raised extensive thinking and discussing, and the answer to this question is very tempting. In order to better answer this question, psychologists and management scientists have studied this topic from different perspectives. As a result, plenty of conclusions and suggestions were achieved. For instance, Judge and Bretz [1] believed it was necessary to analyze the relationship between individual success and personal intrinsic traits. Consistent with this suggestion, psychologists paid attention to the role of personality factors in individual success, especially how proactive personality affects individual success. Bateman and Crant [2] introduced the concept of proactive personality in the study of organizational behavior development, and analyzed the effect of proactive personality in the perspective of organizational development [3]. As described by its definition, proactive personality refers to individual who tends to position their roles and make efforts on one's own initiative in order to make changes to surrounding environment for better adaptation [4], [5]. Thus, individuals with proactive personality are more adaptable to changes of external environment. To be specific, they usually actively identify opportunities and take advantage of them, and then work hard towards goals for meaningful changes in their career [4], [6].

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

II. RELATED WORK

A. PROACTIVE PERSONALITY

In the research of proactive personality, Rita and Hans [7] reported significant positive correlation between proactive personality and self-reported employment behaviors among college students. Compared with individuals with low proactive personality, those who have high proactive personality usually have advantages in innovation capability. For instance, they are more willing to think, share and utilize their ability to improve their knowledge [8]. Besides that, proactive personality has positive effect on internal motivation [9], [10], which will urge individuals to actively make efforts to change their surrounding environment [11]. In this way, individuals can obtain better opportunities and outcomes in workplace, and more innovative behavior will be generated. Additionally, in the field of human resources, proactive personality has its predictive power in organizational behaviors. In Thompson's study [12], it had been proved that employees with high proactive personality had more self-examination in their daily work and can established good relationships with managers by actively changing surrounding environment, which improved their work performance [13]. What's more, employees with high proactive personality have their advantage in teamwork [14], that means they are more likely to get attention from their superiors, which will have important positive effect on their future career development [15].

B. TEXT MINING AND TEXT CLASSIFICATION

The technology of text Mining can discover, retrieve and extract information from a text corpus, which is usually too complicated for manual work [16]. To be more specific, text mining combines technologies such as natural language processing, artificial intelligence, information retrieval, and data mining to help understand complex written analytical processing systems [17], [18]. In the beginning, text mining was to provide government intelligence and security agencies to detect terrorist activities and other security threats. To improve its performance, text mining utilized text analysis components and technologies from external disciplines such as computer science [19], [20], management science, machine learning, and statistics [21]. These improvements are very important for carrying out practical research, and these techniques have since been widely used in related fields [22]. Nowadays, the accuracy of text mining and the ability to handle complex problems have been steadily improved.

Researchers in the field of psychology also deployed text mining as a tool for the purpose of analyzing psychological factors. In the field of emotional psychology, Neubaum *et al.* [23] confirmed the phenomenon of emotional contagion in online environment by analyzing the dynamic information of Facebook users. In the field of health psychology, Merchant *et al.* [24] used open word analysis technology to study language and personality, and achieved accurate prediction of mental health status for Internet users. In the field of personality psychology, Schwartz *et al.* [25] used a similar method to achieve accurate prediction of personal characteristics of network users; Kosinski *et al.* [26] used digital behavior records with dimension reduction and linear regression method to predict the user's sexual orientation; Chittaranjan *et al.* [27] studied the association between behavioral traits automatically extracted from smartphones usage and self-reported "big five" personality traits. In addition, some other studies have used big data text analysis to carry out the processing of data gathering [28], [29]. From these practical analysis results, it can be seen that text mining analysis improved the outcomes effectively. In short, Zhu *et al.* [30] summarized the basic research ideas of using big data for personality prediction. That was, analyzing the user's network behavior data, and then applying machine learning to build a personality feature prediction model. He believed that the psychological characteristics of individuals could be reflected by social network behaviors. Following this idea, Ren *et al.* [31] used Weibo texts to determine information about individual psychological characteristics, personality types, and social attitudes.

In addition, Text classification is an activity that labels natural language text with predefined categories [32]. It requires cross-disciplinary knowledge to build models and to improve accuracy of prediction [33]. Machine learning algorithms such as BP neural networks [34] and Bayesian theory [32], [35] were widely used in the process of text classification tasks. Text classification generally involves in text expression, selection of classifiers and evaluation of classification results. To be more specific, text expression can be divided into text preprocessing, statistics, feature extraction and other steps [16], [36]. Text preprocessing is an important step that can reduce the interference of noise and improve the accuracy of classification. And for the whole process, there are two most influential procedures: feature extraction and model training, which can directly affect the accuracy of classification [37].

C. MOTIVATION AND HYPOTHESIS

Firstly, the studies of individual personality in the field of psychology often analyze the personality traits of individuals from either measurement result or text content such as the diaries of individuals and the content of interviews. Although text content has its potential in the study of personality, data gathering and qualitative analysis of text content requires a lot of manpower. In order to improve the efficiency of this study, text mining analysis method was adopted to collect and process relevant data, which will effectively improve the efficiency and ensure the accuracy of the assessment.

In addition, in the measurement of proactive personality, most previous studies are performed with traditional psychological questionnaires [2], [6], [38], [39]. However, paper-and-pencil questionnaires have their shortcomings like high social desirability effects. Thus, this study will explore the possibility of predicting individuals' proactive personality with text from social media and targeted short-answer questions text. With participants' Weibo text and short-answer

questions text, we proposed a new approach to classify individuals' proactive personality based on text mining technology.

In conclusion, we believed that short-answer questions + Weibo text datasets will have better performance on predicting individuals' proactive personality.

III. METHOD

This study conformed with the code of ethics of the World Medical Association (Declaration of Helsinki) for experiments involving humans and was approved by the Ethics Committee of Shandong Normal University. Additionally, our research obtained written informed consent from the participants.

A. PARTICIPANTS

Cluster sampling method was adopted to recruit college students as participants. As a result, 1671 students participated in a survey that contained 4 short-answer questions and Weibo ID inquires. Among them, 901 participants completed all 4 questions and provided valid Weibo ID. There are 100 males and 801 females in our sample, which include 226 juniors, 347 sophomores, and 328 freshmen. 307 of them are from one-child family. With the provided Weibo ID, all 901 participants' Weibo post was obtained with web crawler on the date of January 27, 2020. A total of 13,511 Weibo posts were collected. After removing non-original post (e.g. repost), 4,955 of Weibo posts were kept, with an average of 5.5 posts for each participate.

B. RESEARCH TOOLS

Four short-answer questions were set to reflect proactive personality. Since proactive personality is usually being discussed with occupational behavior, we designed short-answer questions for college students from the original definition of proactive personality and relevant expressions in different fields in order to make this research more meaningful for participants. The translated short-answer questions are as follows: (1) ''In your daily life, what would you do if your talents were constrained by the environment, and please explain your reason''; (2) ''In your life, which one do you prefer, accepting existed methods or seeking new approaches to solve problems, and please explain your reason''; (3) ''If leaving the pace of regular life and changing surrounding environment will increase the probability of making mistakes, will you still choose to make changes? Please explain your reason.'' (4) ''How do you think about your study and life in the next few years, and will it be connected with your future career?'' After reading these questions, participants were asked to answer all questions according to their actual thoughts and situations with about 60 words for each question.

C. PROCEDURE

1) TEXT PREPROCESSING

In text mining study, data preprocessing is necessary for better classification performance. In this study, we applied following preprocessing procedure: Firstly, for short-answer questions text, 12 psychology undergraduates and graduate students, who had been informed to transcribe text as its original condition including punctuation and no changes should be made during the transcription, completed the transcription works. For the Weibo text, we excluded non-original post includes advertising post and repost. Jieba package in Python [40] was utilized to perform words segment task, since all text was written in Chinese. Segmented words were compared with Harbin Institute of Technology stop word list [41] in order to remove pronouns, useless auxiliary words and punctuation. As a result, features in text will be more prominent.

2) TEXT FEATURE EXTRACTION

Term Frequency-Inverse Document Frequency (TF-IDF) is a widely-used technique for text feature extraction task in text mining practice. TF-IDF is a statistical method that evaluates the importance of a word to a document in a document set or a corpus. If the frequency of occurrence of TF in a document is high, and it rarely appears in other documents, it is considered to have good class discrimination ability. The formula of TF-IDF is as follows:

$$TF-IDF = TF \times IDF = tf(t,d) \times \log(N/nt + 0.01) \quad (1)$$

Among them, $tf_{(t,d)}$ is the word frequency of the feature term t in document d, N is the number of all training documents, and n_t is the number of documents where the feature term t appears in the training set. We used TF-IDF to perform weight analysis on the features to facilitate subsequent classifier training. Some studies may delete the weights of words with less than certain frequency, but we retained TF-IDF weights of all words. The reason is because the F-test was used for better feature selection later on. In the case of two categories of samples, even if low-frequency words appear, the features corresponding to the word will be filtered out due to the differences between features are not significant enough.

3) FEATURE SELECTION

In traditional statistics, F-test is widely used to test the difference between two or more levels of a variable. In this study, we applied F-test for feature selection since the principle of feature extraction is to distinguish a set of features that can better represent the meaning of text from another set that can't. To be more specific, labels of high and low proactive personality categories are corresponded to certain features that have a large difference, so in this way features can be distinguished with F-test. Also, p value is used as the threshold of extraction. The larger the differences between high score group and low score group are, the more accurate the classification task can be. In this study, we adopted 3 different levels of p value for comparison:

a) Retain all features with significance level less than or equal to 0.05, since p value less than or equal to 0.05 is a commonly used threshold for statistical inference in statistics. Generally, events with a probability of less than or equal to 5% are considered to be small-probability events. In this way, when the error rate is not higher than 5%, and the

difference between high score group and low score group is considered to be significant, it can be used for classification tasks.

b) Retain all features with significance level less than or equal to 0.1, since p value less than or equal to 0.1 is the threshold for statistical inference in some education measurement study [42]. When the error is not higher than 10%, and the difference between high score group and low score group is considered to be significant, it can be used for classification tasks.

c) The p value threshold was determined by exhaustive search. An exhaustive search for threshold p value from 0.05 to 0.1 was conducted, and the result of it was used for feature extraction standard.

4) MODEL TRAINING

In the process of model training, grid search with Cross-Validation (CV) was used to determine parameters. Grid search is a method that takes all possible values of parameters to construct a grid in a certain range in order to obtain the value with best performance [43].

After the values of parameters were obtained, we trained models and evaluated their performance with Stratified K-Fold cross-validation [44], [45]. In stratified K-Fold cross-validation, the data set is divided into K equal subsets, and the proportion of the sample category in each fold is the same as the proportion of the population category. In this study, based on previous experience and trial-and-error, K was determined to be 5. The whole process of this research can be illustrated as Figure 1, and described as follows:

a) Training set and validation set: Since five-fold cross-validation was applied, the original data was divided into five parts without repeated sampling. One was selected as the validation set each time, and the remaining text was used as the training set. Performance was calculated by the average performance of 5 trials.

b) Determination of "Classification Labels": with the work of collecting relevant literature on proactive personality done, we invited 16 experts in related fields to assist in this research. They were asked to score relative factors from 0 to 9. In this way, criterion for evaluating proactive personality was obtained. Then the weighted average score of each participant was calculated and sorted in descending order. The first 50% of participants were labeled as high proactive personality group(category), while the left participants were labeled as low proactive personality group.

c) Classifier selection: Based on previous experience, we selected 5 algorithms which were widely used in text mining study, including Support Vector Machine (SVM), XGBoost, K-Nearest-Neighbors (KNN), Naive Bayes (NB) and Logistic Regression (LR).

d) Model evaluation: 7 indicators include Accuracy (ACC), F1-score (F1), Sensitivity (SEN), Specificity (SPE), Positive Predictive Value (PPV), Negative Predictive Value (NPV) and Area under Curve (AUC) were calculated for comprehensive evaluation.

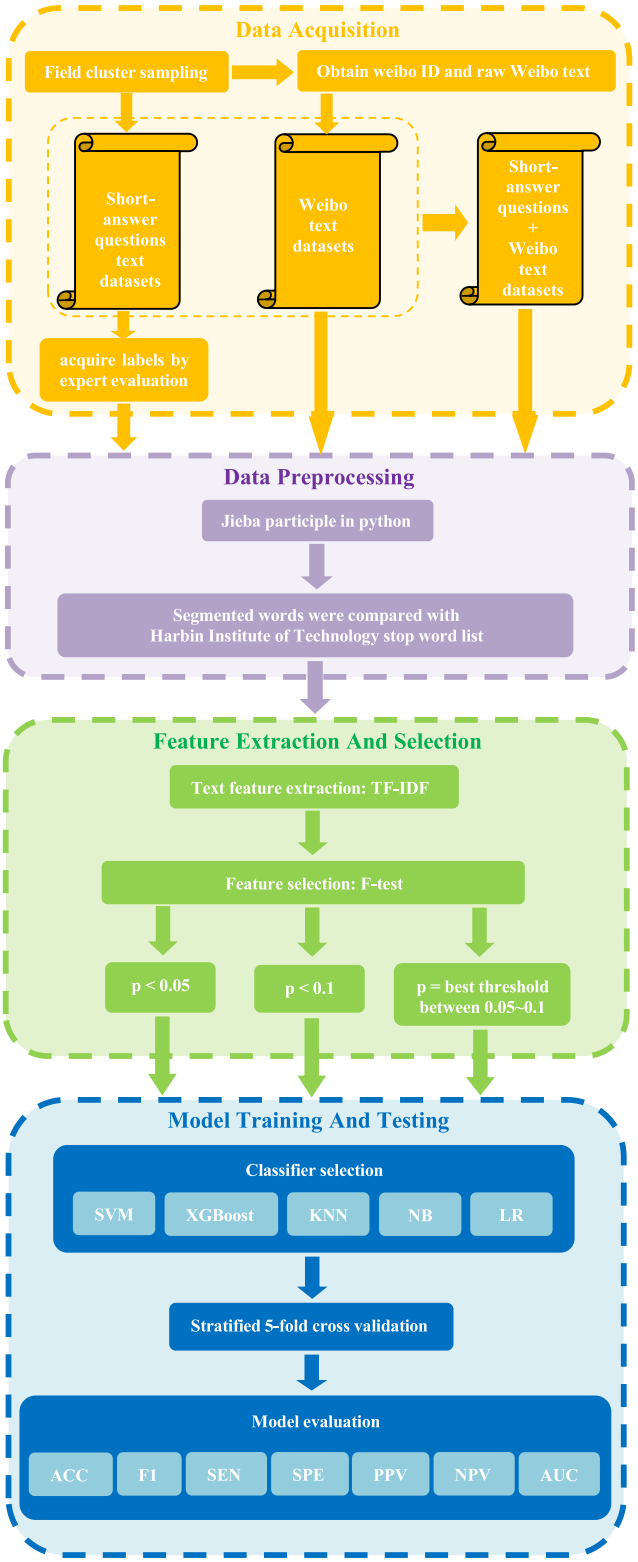

FIGURE 1. Research process.

D. MACHINE LEARNING CLASSIFICATION ALGORITHM

1) SUPPORT VECTOR MACHINE (SVM)

SVM was firstly proposed by Vapnik [46], which aimed to find the hyperplane with the largest spacing as a proportional

classification boundary. It is based on structural risk minimization (SRM) principle in the statistical learning theory, and it has outstanding generalization performance [47]–[49]. Recently, kernel functions such as linear kernel, polynomial kernel, radial basis kernel (RBF), fourier kernel, and spline kernelis are introduced into SVM to solve the inner product operation in high dimensional space, so as to deal with the nonlinear classification task well.

Non-stationary kernel like polynomial kernel is well suited for problems where all the training data is normalized. But due to the consideration of time cost, we didn't choose non-stationary kernel. In this study, linear kernel, sigmoid kernel and radial basis kernel (RBF) of SVM were used. The Linear kernel is the simplest kernel function. It is given by the inner product <x, y> plus an optional constant c. Kernel algorithms using a linear kernel are often equivalent to their non-kernel counterparts. The sigmoid kernel comes from the neural networks field, where the bipolar sigmoid function is often used as an activation function for artificial neurons. It is interesting to note that an SVM model using a sigmoid kernel function is equivalent to a two-layer, perceptron neural network. This kernel was quite popular for support vector machines due to its origin from neural network theory. Also, despite being only conditionally positive definite, it has been found to perform well in practice. In addition, the radial basis function is a kind of scalar function symmetrical along the radial direction, which is usually defined as a monotonic function of the Euclidean distance between any point x in space and a certain center x_c. It is similar to the gaussian distribution, so it is also called the gaussian kernel function, which can map the original features to infinite dimensions.

2) XGBOOST

The scalable end-to-end tree boosting system called XGBoost, which is characterized by fast computation and good performance, has been widely used for data scientists to achieve state-of-the-art results in many machine learning tournaments [50], [51]. The most important factor behind the success of XGBoost is its scalability in all scenarios. The system runs more than ten times faster than existing popular solutions on a single machine and scales to billions of examples in distributed or memory-limited settings.

The idea of the algorithm is to continuously add trees and continuously perform feature splitting to grow a tree. Each time as a tree is added, the model is actually learning a new function to fit the residuals of the last prediction. When k trees are obtained after training, the score of a sample is predicted. In fact, according to the characteristics of this sample, a corresponding leaf node will fall in each tree, and each leaf node corresponds to a score. Scores corresponding to each tree add up to the predicted value of the sample. As a non-parametric model for supervised learning, the selection of XGboost parameters depends on the training data used in the model [52]. The objective function is

$$L(\phi) = \sum_i l(\hat{y}_i - y_i) + \sum_k \Omega(f_k) \tag{2}$$

where i represents the i-th sample and, $l(\hat{y}_i - y_i)$ represents the prediction error of the i-th sample. $\sum_k \Omega(f_k)$ represents the function of the complexity of the tree. The smaller the complexity, the lower the generalization ability. The expression is

$$\Omega(ft) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3}$$

3) K NEAREST NEIGHBORS (KNN)

The classifier KNN was proposed by Cover and Hart in 1968, its performance has been proved to be excellent with large sample size [53]. The error rate of KNN can reach Bayes optimization in very mild conditions [54]. KNN is an instance-based algorithm which means there were no such 'model' trained, classification is made based on comparison between instances in training data and cases. That makes KNN become very sensitive to the number of features, irrelative features can influence the accuracy of prediction greatly [55]. Thus, feature extraction process is very import for KNN algorithm. The process of KNN algorithm can be described as:

a) Calculate distance (etc. European distance, Manhattan distance, cosine Angle distance) between known cases.

b) Sort cases by distance in increasing order.

c) Select k points that have smallest distance from the current case.

d) Determine the frequency of k points in categories.

e) The most frequent categories among k points will be the output.

4) NAIVE BAYES

Naive Bayes is a simple but well-used classifier based on statistics. In text mining, decisions are made based on the presence or absence of certain features [56]. That means probability of being to a certain class was assigned to each feature based on training data. After all probability being calculated, decision can be made based on the presence of features in testing set. The term ''naive'' means all features will be treated independently. In another word, the frequency of features in testing set will not be taken into account, and it assumes all features will present independently. Giving training set as $D = \{d1, d2, \cdots\cdots, dn\}$ with corresponding category as $X = \{x1, x2, \cdots\cdots, xd\}$, variable parameters as $Y = \{y1, y2, \cdots\cdots, ym\}$, the prior probabilities of Y will be $P_{prior} = P(Y)$ and posterior probability of Y will be $P_{post} = P(Y|X)$. In this way, $P_{priort} = P(Y|X)$ can be calculated as

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)} \tag{4}$$

Since features are independent from each other, we can have

$$P(X|Y = y) = \prod_{i=1}^{d} P(X_i|Y = y) \tag{5}$$

Thus, posterior probability of Y can be calculated as

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{d} P(X_i|Y)}{P(X)} \tag{6}$$

5) LOGISTIC REGRESSION

Logistic regression is a generalized linear regression, which can be used to achieve classification or prediction by constructing a regression function [57]. Logistic regression model is a classifier that focuses on the binary classification problem [58], and it can also handle multi-classification problems. Logistic regression maps any input value to the [0, 1] and gets a predicted value in linear regression. Then, map this value to the Sigmoid function, and use the predicted value as the x-axis variable and the y-axis as a probability.

The logistic regression function is

$$g(z) = \frac{1}{1+e^{-z}} \tag{7}$$

and the prediction function is

$$\mathrm{h}\theta(x) = g(\theta^{\mathrm{T}}\mathrm{x}) = \frac{1}{1+e^{-\theta^{\mathrm{T}}\mathrm{x}}} \tag{8}$$

The value of θ has a special meaning, it presents the probability that $\mathrm{h}\theta(x)$ is 1. Therefore, the probability that the classification result of the input x is category 1 and category 0 is

$$P(y=1|x;\theta) = h\theta(x) \tag{9}$$

$$P(y=0|x;\theta) = 1 - h\theta(x) \tag{10}$$

which can be written together as

$$P(y|x;\theta) = (\mathrm{h}\theta(x))^{y}(1-\mathrm{h}\theta(x))^{1-y} \tag{11}$$

The likelihood function is

$$L(\theta)\prod_{i=1}^{m} P(y^{(i)}\left|x^{(i)};\theta\right.) = \prod_{i=1}^{m}(h\theta(x^{(i)}))^{y^{(i)}}(1-h\theta(x^{(i)}))^{1-y^{(i)}} \tag{12}$$

and the log-likelihood function is

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{m}(y^{(i)}\log h\theta(x^{(i)}))\log(1-h\theta(x^{(i)})) \tag{13}$$

At this time, gradient rise is used to find the θ when $l(\theta)$ is maximized.

$$J(\theta) = -\frac{1}{m}l(\theta) \tag{14}$$

E. INDICATORS OF CLASSIFICATION

To evaluate the performance of prediction, we deployed several indicators related to confusion matrix (Table 1) plus area under curve (AUC), in order to evaluate our models in a comprehensive way. Among these indicators, accuracy (ACC) (formula 15) indicates the percentage of individuals that had been classified correctly. Meanwhile sensitivity (SEN) (formula 16) suggests the percentage of high level proactive personality individuals that had been classified correctly; specificity (SPE) (formula 17) shows the percentage of low level proactive personality individuals that had been classified correctly; these two indicators are meaningful under many situations because they are not influenced by unbalanced distribution between high level individuals and low level individuals. Positive Predictive Value (PPV) (formula 18) reveals the percentage of predicted high level individuals that had been classified as high correctly; Negative Predictive Value (NPV) (formula 19) indicates the percentage of predicted low level individuals that had been classified as low correctly. AUC represents the area under receiver operating characteristic curve. The value of AUC ranges from 0.5 to 1. F1-score (F1) (formula 20) represents the weighted harmonic mean of PPV and SEN. For all indicators above, the higher the results are, the more reliable the prediction can be. All these indicators were calculated under 3 different level of datasets, which were short-answer questions text datasets, Weibo text datasets and short-answer questions + Weibo text datasets.

TABLE 1. Confusion matrix.

	Actual Category	
	C1	C2
Predicted Category C_1	*a*	*b*
Predicted Category C_2	*c*	*d*

(***a*** represents high level proactive personality individuals that has been predicted as high correctly; ***b*** represents low level individuals that has been predicted as high wrong; ***c*** represents high level individuals that has been predicted as low wrong; ***d*** represents low level individuals that has been predicted as low correctly)

$$ACC = \frac{\mathrm{a}+d}{a+b+c+d} \tag{15}$$

$$SEN = \frac{\mathrm{a}}{a+c} \tag{16}$$

$$SPE = \frac{d}{b+d} \tag{17}$$

$$PPV = \frac{\mathrm{a}}{a+b} \tag{18}$$

$$NPV = \frac{d}{c+d} \tag{19}$$

$$F1 = 2\times\frac{PPV\times SEN}{PPV+SEN} \tag{20}$$

IV. RESULTS

A. RELATIONSHIP BETWEEN NUMBER OF ACCUMULATED FEATURES AND P VALUE

Figure 2 to 4 illustrates the tendency between number of accumulated features and p value under 3 different text datasets. As is reflected by the figures, the tendencies in these charts are similar, p value dropped while the number of extracted features decreased, especially when it comes to the range

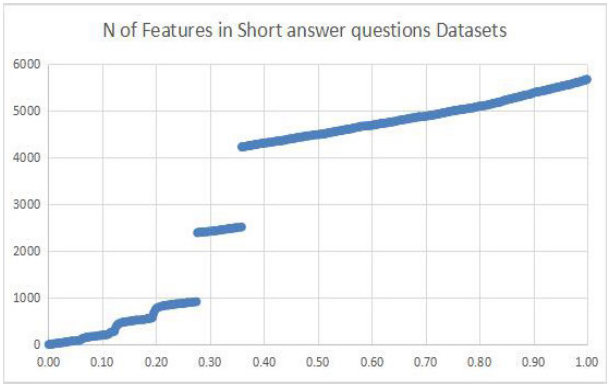

FIGURE 2. Relationship between number of accumulated features and *p* value under short-answer questions text (This chart illustrates the relationship between number of accumulated features and *p* value under short-answer questions text. X-axis represents *p* value while y-axis represents number of accumulated features).

FIGURE 3. Relationship between number of accumulated features and *p* value under Weibo text (This chart illustrates the relationship between number of accumulated features and *p* value under Weibo text. X-axis represents *p* value while y-axis represents number of accumulated features).

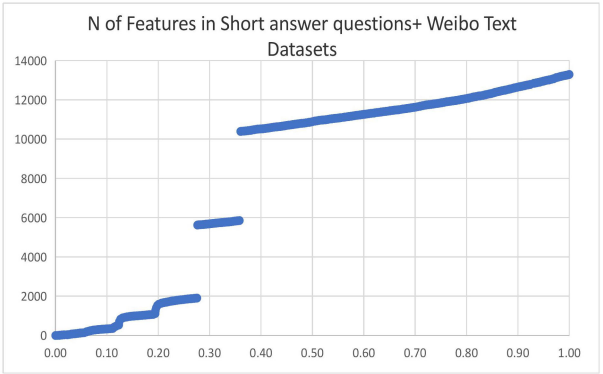

FIGURE 4. Relationship between number of accumulated features and *p* value under short-answer questions + Weibo text (This chart illustrates the relationship between number of accumulated features and *p* value under short-answer questions + Weibo text. X-axis represents *p* value while y-axis represents number of accumulated features).

from 0.27 to 0.35. It showed that feature extraction process can improve the accuracy of prediction. Among 3 text datasets, the short-answer questions text had least number of features, since the information of answers was limited in the frame of questions; following by Weibo text, which has more abundant information compared with short-answer questions text; and it is not surprising that short-answer questions + Weibo text datasets had most features. After we performed exhaustive search in the range between 0.05 to 0.1, p value of 0.0762 performed best for short-answer questions text, while 0.0866 is the best one for Weibo text and 0.0868 for short-answer questions + Weibo text.

TABLE 2. Results of short-answer questions text datasets at 0.05 feature extraction significance level.

	ACC	F1	AUC	SEN	SPE	PPV	NPV
SVM	0.723	0.619	0.795	0.493	0.916	0.832	0.682
XGBoost	0.663	0.601	0.712	0.556	0.753	0.655	0.668
KNN	0.656	0.543	0.681	0.449	0.830	0.702	0.642
NaiveBayes	0.696	0.537	**0.815**	0.386	**0.957**	**0.884**	0.649
Logistic Regression	**0.737**	**0.682**	0.804	**0.619**	0.836	0.762	**0.723**

TABLE 3. Results of Weibo text datasets at 0.05 feature extraction significance level.

	ACC	F1	AUC	SEN	SPE	PPV	NPV
SVM	0.644	0.423	0.693	0.286	0.945	0.815	0.611
XGBoost	0.597	0.328	0.603	0.216	0.918	0.695	0.582
KNN	0.624	0.539	0.661	0.553	0.683	0.688	0.689
NaiveBayes	0.564	**0.676**	0.606	**0.993**	0.202	0.512	**0.969**
Logistic Regression	**0.653**	0.431	**0.711**	0.289	**0.959**	**0.855**	0.616

TABLE 4. Results of short-answer questions + Weibo text datasets at 0.05 feature extraction significance level.

	ACC	F1	AUC	SEN	SPE	PPV	NPV
SVM	**0.768**	0.704	**0.855**	0.604	0.906	0.845	0.731
XGBoost	0.669	0.614	0.730	0.575	0.748	0.662	0.676
KNN	0.728	0.701	0.774	**0.694**	0.757	0.711	0.745
NaiveBayes	0.751	0.671	0.852	0.556	**0.916**	**0.850**	0.710
Logistic Regression	0.766	**0.726**	0.845	0.680	0.838	0.781	**0.757**

B. RESULT OF CLASSIFICATION

1) RESULT OF CLASSIFICATION UNDER P VALUE = 0.05

As illustrated by Table 2 to 4, when features were extracted under the standard of p value = 0.05, short-answer questions + Weibo text had best performance on ACC, AUC and F1, which are 0.768 and 0.726 with SVM algorithm and 0.855 with LR algorithm; Weibo text had best performance on SEN, NPV and SPE, which are 0.993, 0.959 with NB algorithm and 0.969 with LR algorithm; short-answer questions text had best performance on PPV with NB algorithm, which is 0.884. From the perspective of algorithms, Logistic Regression and Naïve Bayes outperformed other algorithms.

2) RESULT OF CLASSIFICATION UNDER P VALUE = 0.1

As illustrated by Table 5 to 7, when features were extracted under the standard of p value = 0.1, short-answer questions + Weibo text had best performance on ACC, F1, AUC, NPV,

TABLE 5. Results of short-answer questions text datasets at 0.1 feature extraction significance level.

	ACC	F1	AUC	SEN	SPE	PPV	NPV
SVM	**0.775**	0.707	0.850	0.595	0.926	0.875	0.731
XGBoost	0.667	0.619	0.709	0.592	0.730	0.649	0.680
KNN	0.687	0.611	0.740	0.541	0.810	0.708	0.678
NaiveBayes	0.729	0.612	**0.876**	0.469	**0.949**	**0.886**	0.680
Logistic Regression	0.768	**0.729**	0.849	**0.682**	0.841	0.784	**0.758**

TABLE 6. Results of Weibo text datasets at 0.1 feature extraction significance level.

	ACC	F1	AUC	SEN	SPE	PPV	NPV
SVM	0.713	0.609	0.790	0.493	0.898	0.801	0.678
XGBoost	0.636	0.470	0.665	0.356	0.871	0.693	0.617
KNN	0.679	0.578	0.742	0.500	0.830	0.750	0.671
NaiveBayes	0.736	0.626	**0.847**	0.485	**0.947**	**0.887**	0.686
Logistic Regression	**0.737**	**0.641**	0.818	**0.517**	0.922	0.850	**0.695**

TABLE 7. Results of short-answer questions + Weibo text datasets at 0.1 feature extraction significance level.

	ACC	F1	AUC	SEN	SPE	PPV	NPV
SVM	**0.842**	**0.811**	**0.919**	0.740	0.928	0.897	**0.810**
XGBoost	0.658	0.620	0.698	0.609	0.699	0.632	0.680
KNN	0.743	0.702	0.736	0.660	0.812	0.753	0.739
NaiveBayes	0.796	0.736	0.915	0.624	**0.941**	**0.899**	0.748
Logistic Regression	0.805	0.778	0.885	**0.750**	0.851	0.809	0.802

TABLE 8. Results of short-answer questions text datasets at 0.0762 feature extraction significance level.

	ACC	F1	AUC	SEN	SPE	PPV	NPV
SVM	0.776	0.708	0.851	0.595	0.928	0.875	0.731
XGBoost	0.683	0.620	0.736	0.565	0.781	0.687	0.681
KNN	0.709	0.648	0.765	0.588	0.812	0.729	0.701
NaiveBayes	0.733	0.606	**0.883**	0.452	**0.969**	**0.925**	0.677
Logistic Regression	**0.777**	**0.737**	0.855	**0.682**	0.857	0.802	**0.762**

SEN and PPV, which are 0.842, 0.811, 0.919, 0.750 with SVM algorithm plus 0.899 and 0.810 with LR algorithm; short-answer questions text had best performance on SPE with NB algorithm, which is 0.949. Weibo text did not show significant advantages compared with other two datasets. From the perspective of algorithms, SVM, Naïve Bayes and Logistic Regression had better performance.

3) RESULT OF CLASSIFICATION UNDER *P* VALUE = BEST THRESHOLD

As illustrated by Table 8 to 10, after exhaustive search for best threshold, short-answer questions + Weibo text had best performance on ACC, F1, NPV, AUC and SEN, which are 0.840, 0.808, 0.922 with SVM algorithm, 0.745 with NB algorithm, plus 0.807 with LR algorithm; short-answer questions text had best performance on SPE and PPV with NB algorithm, which are 0.969 and 0.925. Weibo text did not show significant advantage compared with other two datasets.

TABLE 9. Results of Weibo text datasets at 0.0866 feature extraction significance level.

	ACC	F1	AUC	SEN	SPE	PPV	NPV
SVM	0.708	0.598	0.788	0.480	0.900	0.798	0.674
XGBoost	0.633	0.465	0.664	0.354	0.867	0.684	0.616
KNN	0.679	0.576	0.742	0.498	0.832	0.753	0.671
NaiveBayes	**0.739**	**0.634**	**0.844**	0.497	**0.943**	**0.882**	**0.691**
Logistic Regression	0.730	0.626	0.814	**0.500**	0.924	0.847	0.688

TABLE 10. Results of short-answer questions + Weibo text datasets at 0.0868 feature extraction significance level.

	ACC	F1	AUC	SEN	SPE	PPV	NPV
SVM	**0.840**	**0.808**	0.917	0.736	0.928	0.897	**0.807**
XGBoost	0.662	0.612	0.716	0.585	0.726	0.644	0.675
KNN	0.766	0.722	0.758	0.665	0.851	0.793	0.751
NaiveBayes	0.794	0.731	**0.922**	0.614	**0.945**	**0.904**	0.744
Logistic Regression	0.807	0.779	0.881	**0.745**	0.859	0.818	0.801

From the perspective of algorithms, SVM, Naïve Bayes and Logistic Regression had better performance.

When we compared 3 datasets vertically, the average ACC of them on 5 algorithms are 0.736, 0.698 and 0.774 respectively, which indicated that all 3 datasets had acceptable result of classification. Besides that, short-answer questions + Weibo text datasets showed clear advantage compared with other two datasets on classification, which indicated the combination of them can help to make more accurate prediction. When we compared the result between short-answer questions text datasets and Weibo text datasets, short-answer questions text datasets were superior to Weibo text datasets on all indicators except SEN and NPV under 0.05 p value, which means short-answer questions text had better predictive effect. For PPV and NPV with short-answer questions + Weibo text datasets, they ranged from 0.850 to 0.904 and 0.757 to 0.810 respectively, which means both of them reflected high confidence of making correct classification.

At the same time, features with high weight were extracted as Table 11, these words were crucial in classifying proactive personality individuals.

4) COMPARISON BETWEEN CLASSIFIERS UNDER BEST THRESHOLD *P* VALUE

Figure 5 illustrates the average result of 3 datasets after extracting features under best threshold p value; figure 6 illustrates the average performance of 5 classifiers after extracting features under best threshold p value. After calculating the average performance, we can find that short-answer questions + Weibo text datasets had best performance among 3 datasets, following by short-answer questions text datasets; Weibo text datasets had worst predictive accuracy. However, it is worth mentioning that Weibo datasets have best performance on SPE indicator. From the perspective of indicators, all three datasets had steady performance on the indicators of

TABLE 11. High weigh words under 3 datasets (English version).

Short-answer questions text	Weibo text	Short-answer questions + Weibo text
Go to school, keep improving, not resigned to, professional courses, boring personality, habit, internet, space, preference, living, well-prepared, company, consider both sides, introversion, pursue, ordinary, alone, pressure, shining, development, only want to, qualified, team, geography, review, more, university teacher, brain, relieved, seek for opportunities, try my best, fantasy, constrained by, setback, accept, teacher occupation, freshness, boring, not the right time yet, common fame is seldom to blame, learning from, make mistakes, graduate student, staffing, network, PhD exam, cost, career planning, certificate, job-hopping, easily, achieve, have a perfect mastery of.	All one's life, everything, get off, rainy, not eat, embarrassed, enjoy, cost, anyone, confidence, admire, fulfilling, justice, winter, share, first snow, cheer up, driver, group photo, desk mate, hear of, quarrel, people without dream, preference, light of summer, aunt, college student, failure, strive, good-looking, sis, alone, comfort, childhood, try, as much as one likes, work, senior fellow apprentice, happy, mindset, interest, mature, torment, photograph, educate, new semester, lemon, subside, disappear, sober, life-loving, love, knowledge, dead tried, commemorate, cate, freedom, Anglo-Saxon, strawberry, performance, star chaser, youth, bread, Korean, restaurant, mark.	a pool of stagnant water, keep improving, not resigned to, characteristic, secondary school teacher, internet, well-prepared, justices, winter, first snow, ordinary, driver, surrounding environment, light of summer, good-looking, alone, smooth and steady, loneliness, dust, try my best, happy, smile, express, talent, suffer, relax, freshness, common fame is seldom to blame, learning from, loving, watch TV series, graduate student, destiny, qualified, jogging, youth, let nature take its course, color, proud, efficient

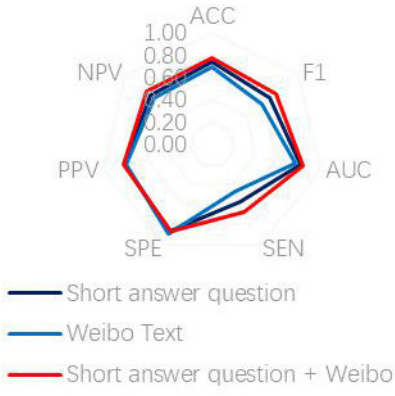

FIGURE 5. Average performance between different text datasets This figure illustrates the average result of 3 datasets after extracting features under best threshold *p* value).

PPV and SPE, while the combination of datasets improved SEN a lot.

From the perspective of classifiers, SVM, LR and NB algorithms had their advantage respectively. Among them, SVM and LR showed steady performance while NB had best performance on AUC, SPE and PPV. Yet, the performance of NB on other indicators was not so desirable. The performance

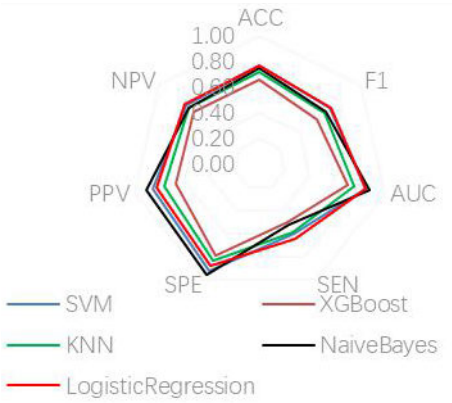

FIGURE 6. Average performance between classifiers (This figure illustrates the average performance of 5 classifiers after extracting features under best threshold *p* value).

of KNN and XGboost algorithms was unsatisfactory, which means they might not be suitable for text classification in this study. From the perspective of indicators, SPE was the highest among all indicators, following by AUC; while SEN was the lowest.

V. DISCUSSION

In the field of psychology, previous in-depth studies of personality often analyzed individual personality traits from individuals' freely expressing diaries and interview content [59], [60]. However, as the sample size increases, qualitative analysis of all subjects requires a lot of manpower and material resources. In addition, in the measurement of proactive personality, most studies chose to use traditional scales [2], [6], [38], [39]. As a result, shortcomings of self-reported scales like social desirability effect are inevitable. The evaluation process of this study added a new approach to the field of personality measurement.

As we mentioned, individual's proactive behavior will have a very significant effect on both oneself and organizations [6], [8], [9], [10], [13], [14], [61]. As Frese [62] mentioned in his research, there are two ways to measure individual's proactive personality, which included behavioral interview and self-reported approach. Abstractly, they represented objective approach and subjective approach personality evaluation. Francesca [63] once mentioned in educational research that the concept of variable analysis could be used as a reference for demographic variables, which had a very important role in improving the accuracy of data. He [37] mentioned that the self-report text and the PTSD symptom scale can be combined to analyze the initiative of the personnel, and when the Bayesian theory was integrated into it, a good evaluation measurement result was obtained. This study analyzed subjective material in an objective way, and this kind of combination shows great potential due to its accurate prediction and convenient approach. In addition, with applying text from social media, we greatly improve the ecological validity of proactive personality measurement.

In the process of data mining, there were three major innovation parts in our study. Firstly, this study analyzed 3 datasets, and most previous research chose only social media text. Secondly, features were extracted based on significance level. Thirdly, 5 classifiers by 5-fold cross validation were deployed and compared with 7 comprehensive indicators. As was mentioned by He [37], an accuracy of 0.700 was reached when using NB to identify PTSD individuals. The reason why accuracy among 3 datasets in this study is higher than He's study is because of following possible reasons. (1) In the processing of collecting text data, the number of words was controlled within 150, while the sample size was reasonable according to Liliya *et al.* [64]. (2) Classifiers like SVM are more compatible with medium sample size in text classification task [49]. (3) Besides that, our datasets contained 4 short-answer questions text and average of 5.5 Weibo post from each participant, which are more abundant compared with previous research.

In the comparison between 3 datasets, the combined datasets showed the best performance. This result is consistent with our expect, the increasing of information among quantity and types can help in classifying proactive individuals. The reason why short-answer questions text was superior to Weibo text is possibly because short-answer questions were set with provision for measuring proactive personality, whereas Weibo text is open-minded. Therefore, the noise in Weibo text can disturb classifications [65]. However, it is worth noting that Weibo text can enhance the predictive effect of short-answer questions text, which means the daily life status reflected by Weibo can help in predicting personality. Given that Weibo text has its universality, we suggest similar future text mining research could consider using targeted short-answer questions text plus social media text to improve the accuracy of classification. Besides that, the indicator SPE were higher than SEN in all 3 datasets, which means text mining is more competent in identifying individuals with low proactive personality.

For the comparison between algorithms, as we mentioned above, SVM and LR showed good performance on all indicators while NB algorithm had advantages on AUC, SPE and PPV indicators. Consistent with previous studies, this result proved linear algorithms like SVM and LR to be suitable for binary problems in text classification domain. For example, in the study of Liu *et al.* [66] which used Weibo text as predictor, SVM had the best performance in identifying individuals with suicide risk; in Sun's [67] comparative study, standard SVM algorithm often learned the best decision surface in most test case. They also suggested that threshold value for SVM algorithm's performance is crucial. Considering the result of this study to be acceptable, using p value as threshold should be a good approach in deciding the threshold. In contrast, non-linear classifiers like KNN and XGBoost didn't show good performance in this study, suggesting that overfitting might be one of the causes. The unstable performance of NB might be due to the characteristic of the algorithm itself, which assumes that attributes should be independent [68]. This conditional independence assumption makes NBC to be very sensitive to form of input. Thus, we would like to suggest future studies to consider linear classifiers when it comes to binary personality classification problems, especially when the sample size is limited.

TABLE 12. High weigh words under 3 datasets (Chinese version).

Short-answer questions text	Weibo text	Short-answer questions + Weibo text
上学, 不断完善, 不甘, 专业课, 个性乏味, 习惯, 互联网, 余地, 偏好, 做人, 充分准备, 公司, 兼顾, 内向, 努力争取, 千篇一律, 单独, 压力, 发光, 发展, 只想, 合格, 团队, 地理, 复习, 多一些, 大学老师, 大脑, 安心, 寻找机会, 尽人事, 幻想, 拘泥于, 挫折, 接受, 教师职业, 新鲜感, 无聊, 时机未到, 枪打出头鸟, 汲取, 犯错误, 研究生, 编制, 网络, 考博, 耗费, 职业规划, 证书, 跳槽, 轻易, 达成, 过硬	一辈子, 万事, 下车, 下雨, 不吃, 不好意思, 享受, 代价, 任何人, 信心, 倾心, 充实, 公平, 冬天, 分享, 初雪, 加油, 司机, 合照, 同桌, 听说, 吵架, 咸鱼, 喜好, 夏之光, 大妈, 大学生, 失败, 奋斗, 好看, 姐妹, 孤独, 安慰, 小时候, 尝试, 尽情, 工作, 师哥, 开心, 心态, 感兴趣, 成熟, 折磨, 拍照, 教养, 新学期, 柠檬, 沉淀, 消失, 清醒, 热爱生活, 爱着, 知识, 累死, 纪念, 美食, 自由, 英美, 草莓, 表演, 追星, 青春, 面包, 韩国, 餐厅, 马克	一潭死水, 不断完善, 不甘, 个性, 中学老师, 互联网, 充分准备, 公平, 冬天, 初雪, 千篇一律, 司机, 周围环境, 夏之光, 好看, 孤独, 安安稳稳, 寂寞, 尘土, 尽人事, 开心, 微笑, 快递, 才华, 承受, 放松, 新鲜感, 枪打出头鸟, 汲取, 爱着, 看剧, 研究生, 缘分, 胜任, 跑步, 青春, 顺其自然, 颜色, 骄傲, 高效

VI. CONCLUSION

In conclusion, text mining technology showed great value in predicting individuals' proactive personality, especially for identifying individuals with low proactive personality, and this can be very valuable for career education practice in high school and college. The best accuracy and specificity reached 0.842 and 0.969 respectively. The form of data used in this study is innovative, few previous studies combine short-answer questions text and social media text together for text mining purpose. The supplementary effect of social media text in predicting individuals' personality is noteworthy, which is reasonable to assume that this kind of

effect not only works in predicting proactive personality, but also can be valuable for predicting other traits. Additionally, features extraction based on p value had been proved to be an effective approach in data preprocessing when dealing with text material. Last but not least, support vector machine and logistic regression showed steady performance in this text mining study, we would like to recommend researchers to give priority to these two algorithms in similar study.

APPENDIX

The high weight words under the three datasets in Chinese version are as shown in Table 12.

ACKNOWLEDGMENT

(Peng Wang, Yingdong Si, Gancheng Zhu, Xiangping Zhan, and Jun Wang are co-first authors.)

REFERENCES

[1] T. A. Judge and R. D. Bretz, "Political influence behavior and career success," *J. Manage.*, vol. 20, no. 1, pp. 43–65, Apr. 1994.
[2] T. S. Bateman and J. M. Crant, "The proactive component of organizational behavior," *J. Organ. Behav.*, vol. 14, pp. 103–118, May 1993.
[3] J. Xie and M. Yan, "Active coping or avoidance? The effect of proactive personality on the relationship between workplace ostracism and organizational citizenship behavior," *Acta Psychologica Sinica*, vol. 48, no. 10, p. 1314, 2016.
[4] J. M. Crant, "Proactive behavior in organizations," *J. Manage.*, vol. 26, no. 3, pp. 435–462, Jun. 2000.
[5] C.-H. Wu, H. Deng, and Y. Li, "Enhancing a sense of competence at work by engaging in proactive behavior: The role of proactive personality," *J. Happiness Stud.*, vol. 19, no. 3, pp. 801–816, Mar. 2018.
[6] Y. Zhang and F. Yang, "Proactive personality: Mechanisms and future directions," *Adv. Psychol. Sci.*, vol. 25, no. 9, p. 1544, 2017.
[7] C. Rita and D. W. Hans, "Determinants of graduates' preparatory job search behaviour: A competitive test of proactive personality and expectancy-value theory," *Psychologica Belgica*, vol. 42, no. 4, pp. 251–266, 2002.
[8] Z. G. Zhang, C. P. Yu, and Y. J. Li, "The Relationship among Proactive Personality, Knowledge Sharing and Employee' s Innovation Behavior," *Manage. Rev.*, vol. 28, no. 4, pp. 124–134, 2016.
[9] I. Setti, P. Dordoni, B. Piccoli, M. Bellotto, and P. Argentero, "Proactive personality and training motivation among older workers," *Eur. J. Training Develop.*, vol. 39, no. 8, pp. 681–699, Sep. 2015.
[10] W.-H. Chang, K.-C. Chen, Y.-K. Yang, P.-S. Chen, R.-B. Lu, T.-L. Yeh, C. S.-M. Wang, and I.-H. Lee, "Association between auditory p300, psychopathology, and memory function in drug-naïve schizophrenia," *Kaohsiung J. Med. Sci.*, vol. 30, no. 3, pp. 133–138, Mar. 2014.
[11] C.-Y. Chiu, B. P. Owens, and P. E. Tesluk, "Initiating and utilizing shared leadership in teams: The role of leader humility, team proactive personality, and team performance capability.," *J. Appl. Psychol.*, vol. 101, no. 12, pp. 1705–1720, Dec. 2016.
[12] J. A. Thompson, "Proactive personality and job performance: A social capital Perspective.," *J. Appl. Psychol.*, vol. 90, no. 5, pp. 1011–1017, 2005.
[13] P. Q. Viet and T. A. Tuan, "The impact of proactive personality on job performance through job crafting: The case of vietcombank in ho chi minh city," *Bus. Econ. Res.*, vol. 8, no. 3, p. 149, Aug. 2018.
[14] S. Sun and H. I. van Emmerik, "Are proactive personalities always beneficial political skill as a moderator," *J. Appl. Psychol.*, vol. 100, no. 3, pp. 966–975, 2015.
[15] U. K. Bind and S. K. Parker, "Proactive work behavior: Forward-thinking and change-oriented action in organizations," *APA Handbook Ind. Organizational Psychol.*, vol. 2, pp. 567–598, Oct. 2011.
[16] E. E. Chen and S. P. Wojcik, "Supplemental material for a practical guide to big data research in psychology," *Psychol. Methods*, vol. 21, no. 4, pp. 458–474, 2016.
[17] S. M. Al-Daihani and A. Abrahams, "A text mining analysis of academic Libraries' tweets," *J. Academic Librarianship*, vol. 42, no. 2, pp. 135–143, Mar. 2016.
[18] A. Rzhetsky, M. Seringhaus, and M. B. Gerstein, "Getting started in text mining: Part two," *PLoS Comput. Biol.*, vol. 5, no. 7, pp. 7–9, 2009.
[19] A. Joorabchi, M. English, and A. E. Mahdi, "Text mining stackoverflow: An insight into challenges and subject-related difficulties faced by computer science learners," *J. Enterprise Inf. Manage.*, vol. 29, no. 2, pp. 255–275, Mar. 2016.
[20] M. N. Kassim, M. A. Maarof, A. Zainal, and A. A. Wahab, "Enhanced rules application order to stem affixation, reduplication and compounding words in malay texts," in *Pacific Rim Knowledge Acquisition Workshop*. Cham, Switzerland: Springer, 2016.
[21] G. G. Miner, *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. London, U.K.: Academic, 2012.
[22] J. Silge and D. Robinson, "Tidytext: Text mining and analysis using tidy data principles in r," *J. Open Source Softw.*, vol. 1, no. 3, p. 37, Jul. 2016.
[23] G. Neubaum, L. Rösner, A. M. Rosenthal-von der Pätten, and N. C. Krämer, "Psychosocial functions of social media usage in a disaster situation: A multi-methodological approach," *Comput. Hum. Behav.*, vol. 34, pp. 28–38, May 2014.
[24] R. M. Merchant, "Evaluating the predictability of medical conditions from social media posts," *PLoS ONE*, vol. 14, no. 6, pp. 1–12, 2019.
[25] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PLoS ONE*, vol. 8, no. 9, Sep. 2013, Art. no. e73791.
[26] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 15, pp. 5802–5805, Apr. 2013.
[27] G. Chittaranjan, J. Blom, and D. Gatica-Perez, "Mining large-scale smartphone data for personality studies," *Pers. Ubiquitous Comput.*, vol. 17, no. 3, pp. 433–450, Mar. 2013.
[28] A. Li, D. Jiao, and T. Zhu, "Detecting depression stigma on social media: A linguistic analysis," *J. Affect. Disorders*, vol. 232, pp. 358–362, May 2018.
[29] Y. Yuan, B. Li, D. Jiao, and T. Zhu, "The personality analysis of characters in vernacular novels by SC-LIWC," *Lect. Notes Comput. Sci.*, vol. 10745, pp. 400–409, May 2018.
[30] T. S. Zhu, J. Y. Wang, N. Zhao, and X. Q. Liu, "Reform on psychological research in big data age," *J. Xinjiang Normal Univ.*, vol. 36, no. 4, pp. 100–107, 2015.
[31] X. P. Ren, X. R. Ma, Y. Zhou, and T. S. Zhu, "The difference between the number of followers / followers on Weibo," in *Proc. 20th Nat. Conf. Psychol.-Psychol. Nat. Mental Health*, 2017.
[32] A. Rahman and U. Qamar, "A Bayesian classifiers based combination model for automatic text classification," in *Proc. 7th IEEE Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Aug. 2016, pp. 63–67.
[33] X. F. Fang, "Application of association rule data mining technology based on linear linked list in digital library," *China New Commun.*, vol. 16, pp. 15–20, Oct. 2017.
[34] J. Chen, S. Yan, and K.-C. Wong, "Verbal aggression detection on Twitter comments: Convolutional neural network for short-text sentiment analysis," *Neural Comput. Appl.*, vol. 2018, pp. 1–10, Mar. 2018.
[35] H. Gao, X. Zeng, and C. Yao, "Application of improved distributed naive Bayesian algorithms in text classification," *J. Supercomput.*, vol. 75, no. 9, pp. 5831–5847, Sep. 2019.
[36] S. Albitar, S. Fournier, and B. Espinasse, "An effective TF/IDF-based text-to-text semantic similarity measure for text classification," *Lect. Notes Artif. Intell. Lect. Notes*, vol. 8786, pp. 105–114, Jun. 2014.
[37] Q. He, "Text mining and IRT for psychiatric and psychological assessment," Ph.D. dissertation, Dept. Behavioural, Manage. Social Sci., Univ. Twente, Enschede, The Netherlands, 2013.
[38] K. J. Qu, R. H. Ju, and Q. Q. Zhang, "The relationships among proactive personality,career decision-making self-efficacy and career exploration in college students," *Psychol. Develop. Edu.*, vol. 31, no. 4, pp. 445–450, 2015.
[39] J. Plomp, M. Tims, J. Akkermans, S. N. Khapova, P. G. W. Jansen, and A. B. Bakker, "Career competencies and job crafting: How proactive employees influence their well-being," *Career Develop. Int.*, vol. 21, no. 6, pp. 587–602, Oct. 2016.
[40] *Jieba*. Accessed: Feb. 20, 2020. [Online]. Available: https://github.com/fxsjy/jieba

[41] Q. Guan, S. Deng, and H. Wang, "Chinese stop words for text clustering: A Comparative Study," *Data Anal. Know1edge Discovery*, vol. 3, pp. 76–84, Apr. 2017.
[42] X. P. Deng, "Relationship between teachers' behavior, parent-child interaction and children's creativity in preschool," Ph.D. dissertation, School Psychol., Dongbei Normal Univ., Dongbei, China, 2013.
[43] C. Liu, S. Q. Yin, M. Zhang, Y. Zeng, and J. Y. Liu, "An improved grid search algorithm for parameters optimization on SVM," *Appl. Mech. Mater.*, vols. 644–650, pp. 2216–2219, Sep. 2014.
[44] A. Comelli, A. Stefano, V. Benfante, and G. Russo, "Normal and abnormal tissue classification in positron emission tomography oncological studies," *Pattern Recognit. Image Anal.*, vol. 28, no. 1, pp. 106–113, Jan. 2018.
[45] S. Armand, E. Watelain, E. Roux, M. Mercier, and F.-X. Lepoutre, "Linking clinical measurements and kinematic gait patterns of toe-walking using fuzzy decision trees," *Gait Posture*, vol. 25, no. 3, pp. 475–484, Mar. 2007.
[46] V. Cherkassky, "The nature of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 8, no. 6, p. 1564, Nov. 1997.
[47] V. Vapnik, *Statistical Learning Theory*, vol. 2. New York, NY, USA: Wiley, 1998.
[48] Z. Yin and J. Hou, "Recent advances on SVM based fault diagnosis and process monitoring in complicated industrial processes," *Neurocomputing*, vol. 174, pp. 643–650, Jan. 2016.
[49] J. X. Han and H. C. He, "SVMDT classifier and its application research in text classification," *Appl. Res. Comput.*, vol. 21, no. 1, pp. 23–24, 2004.
[50] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
[51] R. Zhang, B. Li, and B. Jiao, "Application of XGboost Algorithm in Bearing Fault Diagnosis," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 490, no. 7, pp. 1–5, 2019.
[52] W. Z. Wang, "Electricity consumption prediction using XGBoost based on discrete wavelet transform," in *Proc. 2nd Int. Conf. Artif. Intell. Eng. Appl. (AIEA)*, 2017, pp. 1–14.
[53] P. E. Hart, "The condensed nearest neighbor rule," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 3, pp. 515–516, May 1968.
[54] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, "Learning k for kNN Classification," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, pp. 1–9, 2017.
[55] J. Chen, T. Xiao, J. Sheng, and A. Teredesai, "Gender prediction on a real life blog data set using LSI and KNN," in *Proc. IEEE 7th Annu. Comput. Commun. Work. Conf. (CCWC)*, Jan. 2017, pp. 1–6.
[56] V. Bijalwan, V. Kumar, P. Kumari, and J. Pascual, "KNN based machine learning approach for text and document mining," *Int. J. Database Theory Appl.*, vol. 7, no. 1, pp. 61–70, Feb. 2014.
[57] L. Yu, L. Wang, Y. Shao, L. Guo, and B. Cui, "GLM+: An efficient system for generalized linear models," in *Proc. IEEE Int. Conf. Big Data Smart Comput. BigComp*, May 2018, pp. 293–300.
[58] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Introduction to the Logistic Regression Model*. Hoboken, NJ, USA: Wiley, 2000.
[59] E. Preti, R. Di Pierro, G. Costantini, I. M. A. Benzi, C. De Panfilis, and F. Madeddu, "Using the structured interview of personality organization for DSM–5 level of personality functioning rating performed by inexperienced raters," *J. Personality Assessment*, vol. 100, no. 6, pp. 621–629, Nov. 2018.
[60] A. Eunice, G. Jonathan, R. Chris, T. Gemma, and W. Alison, "The influence of personality disorder on outcome in adolescent self-harm," *Brit. J. Psychiatry J. Mental Sci.*, vol. 207, no. 4, p. 313, 2015.
[61] J. A. Jorgenson, "Tools of the trade: Creativity, innovation, influence, and advocacy," *Amer. J. Health-Syst. Pharmacy*, vol. 75, no. 11, pp. 785–794, Jun. 2018.
[62] M. Frese, D. Fay, T. Hilburger, K. Leng, and A. Tag, "The concept of personal initiative: Operationalization, reliability and validity in two german samples," *J. Occupational Organizational Psychol.*, vol. 70, no. 2, pp. 139–161, Jun. 1997.
[63] M. Francesca, *Chemical Evolution of Irregular Galaxies*. Berlin, Germany: Springer, 2012.
[64] L. Demidova, E. Nikulchev, and Y. Sokolova, "The SVM classifier based on the modified particle swarm optimization," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 2, pp. 16–24, 2016.
[65] S. C. Ding, Y. Wang, and X. Li, "SVM-based Chinese Microblog Sentiment Analysis," *Inf. Document. Services*, vol. 21, no. 1, pp. 23–24, 2016.
[66] X. Liu, X. Liu, J. Sun, N. X. Yu, B. Sun, Q. Li, and T. Zhu, "Proactive suicide prevention online (PSPO): Machine identification and crisis management for chinese social media users with suicidal thoughts and behaviors," *J. Med. Internet Res.*, vol. 21, no. 5, May 2019, Art. no. e11705.
[67] A. Sun, E.-P. Lim, and Y. Liu, "On strategies for imbalanced text classification using SVM: A comparative study," *Decis. Support Syst.*, vol. 48, no. 1, pp. 191–201, Dec. 2009.
[68] S. Eyheramendy, D. Lewis, and D. Madigan, "On the naive Bayes model for text categorization," in *Proc. 9th Int. Workshop Artif. Intell. Statist.*, Mar. 2003.

PENG WANG received the Ph.D. degree in psychological statistics and measurement with Jiangxi Normal University. He is currently a Professor with the School of Psychology, Shandong Normal University. He is also the Director of the Shandong Career Planning and Guiding Committee (SCPGC) and the Standing Director of the New Basic Career Education Research Center, Shandong Normal University. He is also working on integrating career fields with big data. He has published various articles and chapters on these subjects. He has been involved in data mining research with Tsinghua University, as an Advanced Visiting Scholar. His research focuses on the big data psychology, modern measurement theory, and career planning. He was awarded as the Young Talents of Dongyue Scholars at Shandong Normal University, from 2018 to 2023.

YUN YAN is currently pursuing the bachelor's degree with the School of Psychology, Shandong Normal University. She is also working with Prof. Peng Wang's Group on big data research. Her main focuses of research is big data psychology. She believes big data can not only provide more diversified and heterogeneous samples for psychology research, but also free researchers from the limitations of time and space, and avoid social expectation effects as much as possible. Also, it avoids the complex and unrelated interference that the research subject is subjected to during the test.

YINGDONG SI is currently pursuing the master's degree with Shandong Normal University. His tutor is Prof. Peng Wang, who mainly follows Prof. Wang's career and network research. Since enrolling in 2017, he has been studying statistical methods and is skilled in using structural equation models to solve psychological problems. In addition, he has gained in cyberbullying and Internet addiction. He has participated in publishing articles on Internet addiction in important foreign journals.

GANCHENG ZHU received the bachelor's degree from Shandong Normal University, China, in 2019. He is currently pursuing the master's degree with Jilin University. His favorites were both a psychometric theory named Item Response Theory (IRT) and big data in applied psychology. He have already published his first article about IRT in CSSCI and chapter on big data in psychology under the guidance of Prof. Wang when he was a Senior, the book Looking at the world with Big Data: Middle School and Big Data Culture (Wang, 2018). He is also quite dedicated to Bioinformatics and Computational Analysis.

XIANGPING ZHAN received the bachelor's degree from Southwest University, China, in 2018. She is currently pursuing the master's degree with Shandong Normal University. Her research interests include big data psychology and text mining. She is interested in regarding big data method as a research tool, such as Weibo and WeChat in China, to analyze people's online psychology and behavior through social media. She followed her tutor Prof. Peng Wang to carry on big data research.

JUN WANG received the bachelor's degree from Qingdao University, China, in 2016. He is currently pursuing the master's degree with Shandong Normal University. His research focuses is psychology of big data. His research direction is big data psychology, which mainly uses the method of big data to collect, process and analyze data, so as to increase the understanding of social phenomena and public psychology. He followed his tutor Prof. Peng Wang to participate in a number of national and provincial projects.

RUNSHENG PAN is currently pursuing the master's degree in applied psychology with Shandong Normal University, under the guidance of Prof. Peng Wang. With passion towards psychology, his current research interests are in pathological internet use and adaptation of college students.