IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 22, 2019, accepted September 3, 2019, date of publication September 9, 2019,
date of current version September 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2940061

SMOTETomek-Based Resampling for Personality Recognition

ZHE WANG[1], CHUNHUA WU[1], KANGFENG ZHENG[1], XINXIN NIU[1], AND XIUJUAN WANG[2]
[1]School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]School of Computer Science, Beijing University of Technology, Beijing 100124, China

Corresponding author: Chunhua Wu (wuchunhua@bupt.edu.cn)

This work was supported in part by the National Key R&D Program of China under Grant 2017YFB0802703, in part by the National Natural Science Foundation of China under Grant 61602052, and in part by the BUPT Excellent Ph.D. Students Foundation under Grant CX2019231.

ABSTRACT The main challenge of user personality recognition is low accuracy resulting from small sample size and severe sample distribution imbalance. This paper analyzes the impact of imbalanced data distribution and positive and negative sample overlap on the machine learning classification model. The classification model is based on the data resampling technique, which can improve the classification accuracy. These problems can be solved once the data are effectively resampled. We present a personality prediction method based on particle swarm optimization (PSO) and synthetic minority oversampling technique+Tomek Link (SMOTETomek)resampling (PSO-SMOTETomek), which, apart from effective SMOTETomek resampling of data samples, is able to execute PSO feature optimization for each set of feature combinations. Validated by simulation, our analysis reveals that the PSO-SMOTETomek method is efficient under a small dataset, and the accuracy of personality recognition is improved by up to around 10%. The results are better than those of previous similar studies. The average accuracies of the plain text dataset and the non-plain text dataset are 75.34% and 78.78%, respectively. The average accuracies of the short text dataset and the long text dataset are 75.34% and 64.25%, respectively. From the experimental results, we found that short text has a better classification effect than long text. Plain text data can still have high personality discrimination accuracy, but there is no relevant external information. The proposed model is able to facilitate the design and implementation of a personality recognition system, and the model significantly outperforms existing state-of-the-art models.

INDEX TERMS Personality recognition, PSO-SMOTETomek, sample distribution imbalance.

I. INTRODUCTION

Social networks have become the most widely used communication and interaction tool between people in recent years. Personality is a combination of an individual's social behavior, emotion, motivation, and thought pattern characteristics, and makes a person unique. Our personality has a great impact on our lives, affecting our life choices, well-being, health, and preferences. Furthermore, personality is closely related to human behavior [1]. Personality enables the development of applications like personalized recommendations for products, music preference recognition, and prediction of teamwork performance [2], [3]. Thus, it is quite important to recognize a person's personality.

There are several personality models used in predicting personality, such as the Big Five Personality, Myers–Briggs Type Indicator (MBTI), Dominance Influence Steadiness Conscientiousness (DISC), and Strength Finder [4], [5]. MBTI is based on Carl Jung's typology theory and describes the four basic dimensions of human personality: extroversion/introversion, perception/intuition, thinking/feeling and judgment/perception [6]. Since MBTI is descriptive and analytical, it needs to be very good at analytical interpretation and psychoanalysis. DISC behavior evaluation is based on four main and critical behaviors: dominance, impact, stability and compatibility. DISC pays special attention to behavioral preferences. Compared to MBTI and DISC [7], the Strength Finder test has a greater limit on active strategies. It lacks an intuitive model for moving team members [7], [8]. Hence, after some considerations and reviewing the literature, the Big Five Personality is used in this study as it is the most popular

The associate editor coordinating the review of this manuscript and approving it for publication was Mario Luca Bernardi.

and precise in telling someone's personality traits. Moreover, personality is typic-ally formally described in terms of the Big Five Personality traits [9] with binary (yes/no) values:

- **Extroversion (EXT).** Is the person outgoing, talkative, and energetic versus reserved and solitary?
- **Neuroticism (NEU).** Is the person sensitive and nervous versus secure and confident?
- **Agreeableness (AGR).** Is the person trustworthy, straightforward, generous, and modest versus unreliable, complicated, meager, and boastful?
- **Conscientiousness (CON).** Is the person efficient and organized versus sloppy and careless?
- **Openness (OPN).** Is the person inventive and curious versus dogmatic and cautious?

There is an enormous interest in personality recognition. However, automatically classifying human personality traits through analyzing a user's text data is a challenging task considering the little existing research and low accuracy of current methods. In particular, text often reflects various aspects of a user's personality [10], but research on personality analysis of text data for social network users is limited and has low accuracy [1]–[9]. We analyzed the reasons for the low accuracy of user personality recognition. There are three main reasons: limited sample data, unbalanced data distribution, and serious positive and negative sample overlap.

For the three main problem mentioned above, we introduced a resampling technique named synthetic minority oversampling technique+Tomek Link (SMOTETomek), which combines undersampling and oversampling. SMOTETomek technique was applied using the library from imbalanced_learn, which includes the synthetic minority oversampling technique (SMOTE) function for oversampling as well as the Tomek Link function for undersampling. Due to the SMOTE is a classic oversampling technique that increases samples while Tomek Link is an undersampling technique for cleaning overlapping samples. Therefore, the SMOTETomek technique can solve the above problems very well. In the experimental part, we verified the effectiveness of the SMOTEToemk technique. In addition, our personality recognition model is built using some combinations of linguistic features with different feature processing approaches, as detailed in Section 3.4. We construct a particle swarm optimization (PSO) feature optimization algorithm to find the best combination of features. The PSO algorithm can select the optimal joint feature in the feature space of the particle swarm search [11], [12].

The contributions of this paper are summarized below.

In view of the above main problems, this paper proposes a user personality recognition method based on the SMOTETomek resampling technique. The goal of this research is to build an automatic personality recognition model based on the status of user-unbalanced text data. Hence, we propose a PSO-SMOTETomek method for personality recognition based on text so that the prediction model can be improved. The result shows that with the data balanced and features optimized, the model can classify personality traits with a relatively high accuracy.

The rest of the paper is organized as follows. Section 2 reviews prior studies, introducing relevant research methods and achievements. Section 3 describes the personality recognition model, which includes three modules. In section 4, we present the experiments and provide a discussion. Section 5 concludes the research and points out future work directions.

II. RELATED WORK

In earlier studies, self-reporting was the most widely used method of personality characterization, which meant that data were collected from users who filled out standardized questionnaires [13]. However, this method is subject to user subjectivity. Moreover, it has limitations in participant recruitment, feedback efficiency, and resource consumption [14].Therefore, the self-reporting method needs to be improved.

The correlation between users' social network activity and personality has been the focus of several studies in the last decade. Globeck intended to predict web users' personality traits through text features on Facebook and Twitter [15], [16]. Additional research in personality prediction based on Facebook statuses was done by using social network analysis (SNA) features, Linguistic Inquiry and Word Count (LIWC) features, and Structured Programming for Linguistic Cue Extraction (SPLICE) features [15]. The paper proposed a method of big five personality recognition (PR) from microblog in Chinese language environments with a new machine learning paradigm named label distribution learning (LDL) [17]. The goal of this paper is to investigate the predictability of the personality traits of Facebook users based on different feature and measures of the big five model [18]. Another study made a personality prediction system by using Twitter with LIWC and machine reading comprehension (MRC) features [19]. Most of the previous papers considered many external conditions [20], such as the network information and the time when the user published statuses; however, sometimes it was not known whether the external information of the user was accurate.

All of the aforementioned studies did personality prediction by using social media in English based on Big Five Personality models. Initially, the Big Five Personality were developed by several independent groups of researchers. Then, J.M. Digman made further advancements, and Lewis Goldberg later perfected it [21]. Recent research was conducted to make a personality prediction system using Twitter in Bahasa based on Big Five Personality models [22]. Other research on personality prediction was done using deep learning techniques to identify user personality based on Facebook social network statuses [23]. However, these studies were limited by small sample sizes and imbalanced data distributions.

The purpose of this study is to improve the personality recognition accuracy by analyzing the relationship between

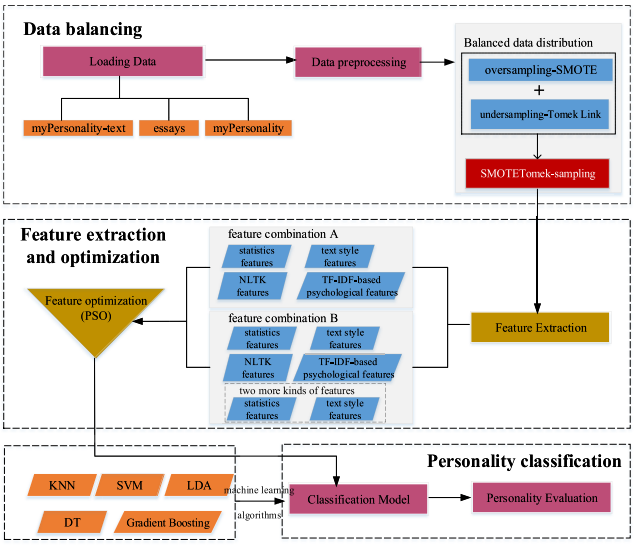

FIGURE 1. Workflow of the experimental pipeline.

personality characteristics and user text data distributions using three public datasets in which the user has an external information dataset that can be used for comparative experiments. Unlike in previous studies, we are not only concerned with the type of user data, but also with the issue of data distribution. Hence, this paper develops a personality recognition model of text based on the PSO-SMOTETomek method for social network users. This is described in detail in section 3.

III. METHODOLOGY

In this section, we introduce the features selection and optimization, and the data balancing; then, we discuss the implementation of our system in detail. We also present the fundamental knowledge of the PSO and resampling technique, aiming at clearly explaining our proposed PSO feature optimization model and SMOTETomek technique.

A. PERSONALITY RECOGNITION MODEL

Our method includes three modules: data balancing, feature extraction optimization and personality classification. Fig.1 illustrates this in detail. Due to the unbalanced distribution of the data samples, our first module is the data balancing module; it includes data loading and data preconditioning. To balance the data distribution, we adopted the SMOTETomek algorithm, which combines the oversampling method SMOTE and undersampling method Tomek Link. The second module is feature extraction and optimization. In this paper, two sets of features (combination A and combination B) are extracted, and PSO optimization is performed on each. In the last module, in order to better verify the method proposed in this paper, we used five traditional machine learning algorithms and 10-fold validation evaluation for user personality recognition.

TABLE 1. Distributions of MyPersonality and MyPersonality_text datasets.

Value	OPN	CON	EXT	AGR	NEU
Yes	146	107	82	113	77
No	59	98	123	92	128

TABLE 2. Distributions of essays dataset.

Value	OPN	CON	EXT	AGR	NEU
Yes	1271	1253	1276	1310	1233
No	1196	1214	1191	1157	1234

B. BALANCING THE DATA

1) DATASET

The experimental data used in this study are divided into three different datasets: myPersonality [24], myPersonality_text, and essays [25]. The first dataset obtained from myPersonality consists of data from 250 Facebook users with approximately 10,000 statuses, with the given personality label based on the Big Five Personality traits. It is a complete dataset of social network users, including users' text information and external information (such as the time of posting, and network size). The second dataset, myPersonality_text, is the plain text data of myPersonality, with the users' external information removed.

Since some statistical features are based on state updates, users with fewer state updates are unable to provide enough information for our model. Therefore, only users with at least five status updates are selected. After data selection, the datasets myPersonality and myPersonality_text consist of 9009 status updates written by 205 Facebook users. The personality type distributions of the myPersonality and myPersonality_text datasets are presented in Table 1.

The final dataset, essays, contains 2,468 anonymous essays tagged with the authors' personality traits: EXT, NEU, AGR, CON, and OPN. We removed from the dataset one essay that contained only the text "Err:508", and we experimented with the remaining 2,467 essays. essays is a plain text library longer than myPersonality_text. The personality type distributions of the essays dataset is presented in Table 2.

2) PREPROCESSING

All of the data in English went through the preprocessing stage before it could be processed. Preprocessing steps consist of removing URLs, symbols, names, spaces, lowering case, stemming, and removing stop words. Data in Bahasa went through additional preprocessing process: slang words or non-standard words were manually replaced, and then the text was translated into English. For the essays dataset, preprocessing steps also included sentence splitting, data cleaning, and unification.

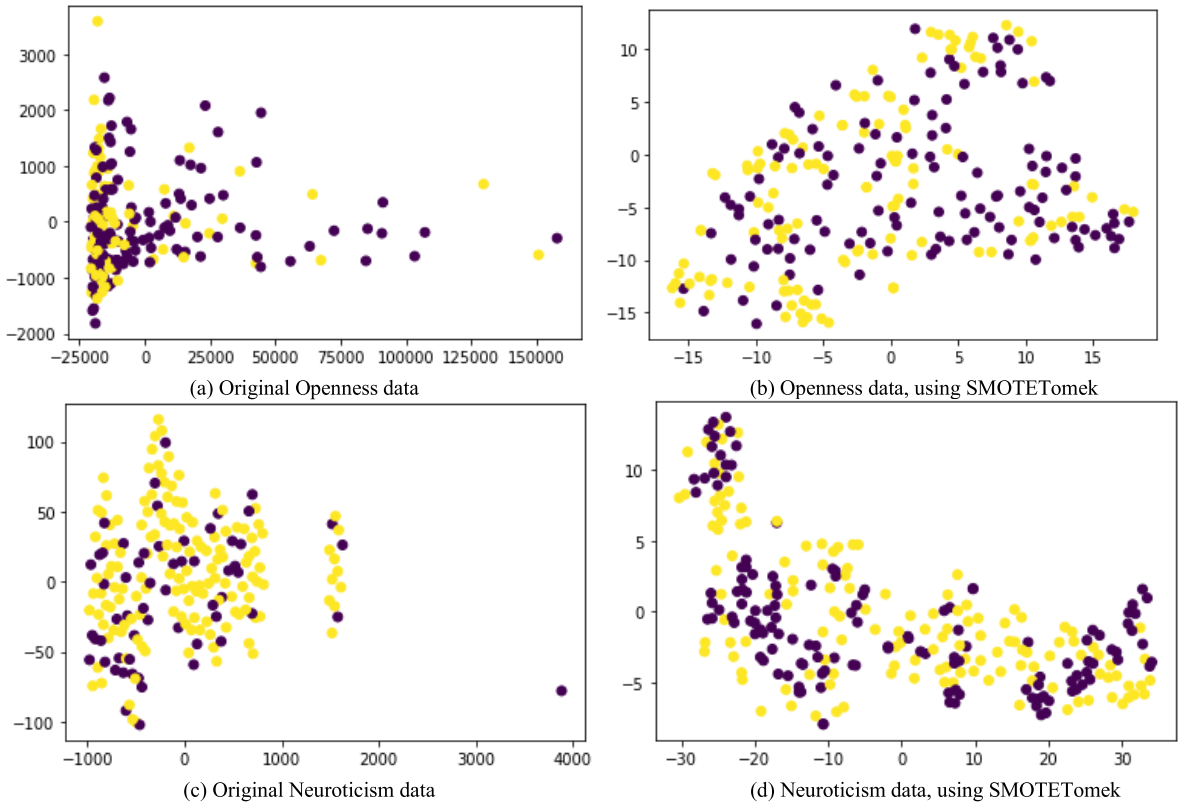

FIGURE 2. T-SNE visualization of the Openness data and Neuroticism data distribution.

Steps such as removing names, stop words, and stemming used the NLTK library. A total of 152 stop words were removed in the experiments. Further data processing was done manually by using written regex and codes.

3) BALANCED DATA DISTRIBUTION

As shown in Table 1, the distribution of positive and negative sample data of the Openness trait is 2.4 (yes): 1 (not). The distribution of positive and negative sample data of the Neuroticism trait is 1 (yes): 1.6 (no). The purpose of resampling is to balance the data. The traditional resampling techniques are undersampling and oversampling. However, the disadvantage of undersampling is that it will lose the information of most samples and cannot make full use of the existing information. Meanwhile, the disadvantage of the oversampling technique is that there are too many repeated samples, which may lead to overfitting of the classifier. Previous works have paid little attention to the problem of imbalanced datasets in myPersonality and my Personality_text, where the number of positive and negative samples is very uneven; these imbalanced datasets can cause classifier decision boundaries to be biased toward the majority class.

For imbalanced datasets, our paper introduced a resampling technique named SMOTETomek, which combines undersampling and oversampling. SMOTETomek is a good way to avoid the disadvantages of the SMOTE and the Tomek Link technique. The SMOTETomek technique is applied using the library from imbalanced_learn, and included an SMOTE function for oversampling as well as a Tomek Link function for undersampling. The algorithm flow of the SMOTETomek method is to combine SMOTE and Tomek Link to form a pipeline. The standard flow is as follows:

Step 1: For a dataset D with an unbalanced data distribution, it uses the SMOTE method to obtain an extended dataset D' by generating many new minority samples.

Step2: Tomek Link pairs in dataset D' are removed using the Tomek Link method.

In this paper, the personality traits of Openness and Neuroticism are taken as examples. In order to see the change in data distribution more intuitively, we used the high-dimensional data dimensionality reduction algorithm t-distributed stochastic neighbor embedding (t-SNE) to map the data distribution in two-dimensional space. The data distribution before and after resampling is shown in Fig. 2. As can be seen from Fig.2 (a) and (b), the overlap of Openness data is a serious problem. After using SMOTETomek technique, Openness data distribution is balanced and overlapping data is effectively reduced. Fig.2 (c) and (d) show that the distribution of positive and negative samples of Neuroticism data is unbalanced, SMOTETomek technique can balance the neu data distribution well.

C. FEATURE EXTRACTION AND OPTIMIZATION

1) FEATURE EXTRACTION

This study uses six scenarios with different features to compare the results and capabilities of the proposed method.

The main reason for this is to investigate the suitability and performance of the various features for personality modeling. The different feature combinations are used for different types of datasets in this paper. For plain text datasets, such as myPersonality_text and essays, we used four kinds of numeric features (feature combination A). The four kinds of features are described as follows:

- **NLTK features:** Additional language analysis is performed using the Brown corpus contained in the NLTK toolkit [26], resulting in an expanded set of linguistic features that included the number of words and the average number of words in a particular grammatical category.
- **Text style features:** In [27], style information is successfully used to identify different people's writing styles. Since human behavior is a reflection of personality, it is feasible to use text style features to identify people with different personality traits. Text style features include exclamation tokens, punctuation tokens, feature tokens, and uppercase and lowercase word tokens. The feature values are displayed along with the frequency at which these tokens appear.
- **Psychological features:** LIWC only considers the importance of words in the same category. Since the text length of each user is inconsistent, simply calculating the frequency of a word will be inaccurate for personality recognition. In order to eliminate the influence of text length inconsistency, we increase the frequency of vocabulary, and calculate the TF-IDF value of psychological vocabulary in the LIWC dictionary [28].
- **Statistics features:** There are six features not included in the above categories: the total number of statuses per user, number of capitalized words, number of capital letters, number of words that are used more than once, number of URLs, and number of occurrences of string PROPNAME-a string used in the data to replace proper names of persons for anonymization.

In addition to the linguistic features described above, for the myPersonality dataset, we also utilized time features and social network features because of the non-plain text information distribution. Hence, we used six kinds of numeric features (feature combination B). These two additional features are described below:

- **Time features:** There are six features related to the time of the status update: (1) frequency of status updates per day, (2) number of statuses posted between 00:00-6:00 am, (3) number of statuses posted between 6:00-11:00 am, (4) number of statuses posted between 11:00-16:00, (5) number of statuses posted between 16:00-21:00, and (6) number of statuses posted between 21:00-00:00. In this study, we assume that all the times are based on one time zone.
- **Social network features:** There are seven features related to the social network of the user: (1) NETWORKSIZE, (2) BETWEENNESS, (3) NBETWEENNESS, (4) DENSITY, (5) BROKERAGE, (6) NBROKERAGE, and (7) TRANSITIVITY. For more information about these measures, see [29].

2) FEATURE OPTIMIZATION

Feature optimization based on PSO improves algorithm performance. Particle swarm optimization is an optimization method based on swarm intelligence. The particle swarm algorithm simulates a bird in a flock by designing a massless particle. The particle has only two properties: velocity and position, where velocity represents the velocity of movement, and position represents the direction of movement.

Individual optimal solution (P_{best}) is the historically optimal positional information found for each particle. The global optimal solution (g_{best}) is the optimal position information found from all individual optimal solutions. Each particle searches for the optimal solution separately in the search space and records it as the current individual extreme value (P_{best}). Each particle shares the individual extreme value with other particles in the whole particle swarm to find the optimal individual extreme value as the current global optimal solution (g_{best}) of the whole particle swarm. All particles in the particle swarm adjust their velocity and position according to the current individual extremum found by themselves and the current global optimal solution shared by the whole particle swarm [30]. The improved position is used to guide the future movement of the colony [31]. The particle swarm is initialized as a group of random particles (random solutions), and the search process is repeated until the optimal location of the swarm is finally found.

Suppose that in a D-dimensional target search space, there are N particles forming a community, where the position of the *i-th* particle is represented by a D-dimensional vector as $X_i = (x_{i1}, x_{i2},\ldots x_{iD})$, $i = 1,2,\ldots N$, and the velocity of the *i-th* particle uses a D-dimensional vector, recorded as $V_i = (v_{i1}, v_{i2},\ldots v_{iD})$, $i = 1,2,\ldots, N$. In each iteration, the particle updates itself by tracking two extremes (P_{best}, g_{best}); the first is the optimal solution found by the particle itself, which is called the extremum of the individual. The optimal position that the *i-th* particle has searched so far is called the P_{best}, defined as $P_{best} = (p_{i1}, p_{i2},\ldots p_{iD})$, $i = 1,2,\ldots N$. The other extreme is the optimal solution currently found for the entire population. This extreme is the global optimal position, and is defined as $g_{best} = (g_1, g_2,\ldots g_D)$.

After finding the two optimal values P_{best} and g_{best}, at iteration k, the particles update their speed and position according to the following equations:

$$v_{iD}^{k} = wv_{iD}^{k-1} + c_1 r_1 \left(pbest_{iD} - x_{iD}^{k-1}\right) + c_2 r_2 \left(gbest_D - x_{iD}^{k-1}\right) \quad (1)$$

$$x_{iD}^{k} = x_{iD}^{k-1} + v_{iD}^{k} \quad (2)$$

TABLE 3. Experimental scenarios for traditional machine learning methods.

Dataset Type	Scenario	Machine Learning							
		Features		Feature Selection(PSO)		Resampling			
		Combination A	Combination B	Yes	No	Without Resampling	SMOTE	Tomek	SMOTETomek
plain text dataset	1	√			√	√			
	2	√		√		√			
	3	√		√					√
non-plain text dataset	4		√		√	√			
	5		√	√		√			
	6		√	√					√
	7		√	√			√		
	8		√	√				√	

SMOTETomek: synthetic minority oversampling technique +Tomek Link.

where w is the inertia weight, and c_1 and c_2 are learning factors, also known as acceleration constants. c_1 and c_2 are usually set to a random value within the range 0 to 4, usually, $c_1 = c_2 = 2$. r_1 and r_2 are uniform random numbers in the range 0 to 1. The standard particle swarm algorithm flow is as follows:

Step1: Initializing particle swarm, including particle population size, position x_i and velocity v_i of each particle.

Step2: The F1-socre metric of machine learning methods using the selected features in the training set is used to measure the results of the particles as $Fit[i]$.

Step3: For each particle, compare its fitness value $Fit[i]$ with the individual optimal solution P_{best}. If $Fit[i] > P_{best}$, replace P_{best} with $Fit[i]$.

Step4: For each particle, compare its fitness value $Fit[i]$ with the global optimal solution g_{best}. If $Fit[i] > g_{best}$, replace g_{best} with $Fit[i]$.

Step5: Update the velocity and position of the particles according to (1) and (2).

Step6: Exit the loop if the error is good enough or the maximum number of cycles is reached, otherwise, return to Step2.

D. PERSONALITY CLASSIFICATION

This study uses some machine learning algorithms that have been widely used in previous studies: the decision tree (DT), support vector machine (SVM), k-nearest neighbor (KNN), gradient boosting, and linear discriminant analysis (LDA). In this experiment, we treat all of the extracted features as original features and compare them with the best features obtained by the PSO method. Finally, original features are compared with the features selected by the PSO-SMOTETomek method. For model validation, we use a 10-fold cross validation technique using Python libraries. The 10-fold cross validation selected 10% of the dataset for testing data and 90% of the dataset for training data. The accuracy of the final experimental result is the max of 10 times 10-fold cross validation.

We conduct a series of tests under various scenarios to understand the accuracy of each method and each algorithm in predicting personality type. Experiments are done by adding a data processing method to compare accuracy of personality recognition. The first method is resampling, where we balance the data distribution by using the SMOTETomek method. The second method is feature selection, which attempts to filter or delete features that are considered to have low correlation with personality traits by using the PSO method. We do some experiments to compare the methods presented in this paper. Table 3 is a breakdown of experimental scenarios to be performed on traditional machine learning.

IV. EXPERIMENT AND DISCUSSION

A. PERFORMANCE FOR 3 DATASETS IN OUR PAPER

All of the classification results obtained by using traditional machine learning can be seen in Table 4, 5, 6, and 7. In Table 7, we report the highest accuracy and F1-score of each trait for each dataset with each method. The calculation formulas of F1-score and accuracy are as follows:

$$\boldsymbol{accuracy} = (\boldsymbol{a} + \boldsymbol{b})/(\boldsymbol{a} + \boldsymbol{b} + \boldsymbol{c} + \boldsymbol{d}) \quad (3)$$

$$\boldsymbol{precision} = \boldsymbol{a}/(\boldsymbol{a} + \boldsymbol{c}) \quad (4)$$

$$\boldsymbol{recall} = \boldsymbol{a}/(\boldsymbol{a} + \boldsymbol{c}) \quad (5)$$

$$\boldsymbol{F1\_score} = 2 * \boldsymbol{precision} * \boldsymbol{recall}/(\boldsymbol{precision} + \boldsymbol{recall}) \quad (6)$$

where a indicates the number of the positive samples which are correctly predicted. b is the number of negative samples

TABLE 4. Traditional machine learning classification results of accuracy and F1-score on the MyPersonality dataset.

Features	Algorithm	OPN	CON	EXT	AGR	NEU	Accuracy Average
All features (Scenarios 4)	SVM	**71.50%**/0.84	58.50%/0.55	64.50%/0.27	59.50%/0.60	61.00%/0.33	63.00%
	DT	65.00%/0.67	58.50%/**0.60**	65.00%/**0.64**	**62.50%**/0.59	64.00%/0.28	63.00%
	KNN	71.00%/0.79	**59.50%**/0.57	63.50%/0.62	60.50%/**0.70**	61.50%/0.60	63.20%
	LDA	65.00%/0.75	59.00%/0.50	**68.00%**/0.63	54.50%/0.67	61.50%/**0.62**	61.60%
	Gradient Boosting	70.00%/**0.90**	58.00%/0.47	64.00%/0.40	58.50%/0.57	**67.50%**/0.51	**63.60%**
	Accuracy Average	**68.50%**	58.70%	65.00%	59.10%	63.10%	—
PSO (Scenarios 5)	SVM	**76.00%**/0.73	62.00%/0.61	70.00%/0.57	64.00%/0.57	67.00%/0.44	67.80%
	DT	72.50%/0.75	60.50%/0.63	**72.50%**/0.60	64.00%/**0.75**	70.00%/0.31	66.75%
	KNN	78.50%/**0.82**	**65.00%/0.63**	73.50%/**0.73**	**74.50%**/0.64	68.00%/0.22	**71.90%**
	LDA	67.00%/0.74	61.00%/0.63	64.50%/0.66	58.00%/0.50	66.00%/0.30	63.30%
	Gradient Boosting	73.50%/0.83	64.00%/0.50	70.00%/0.47	60.50%/0.32	**73.50%/0.73**	68.30%
	Accuracy Average	**73.50%**	62.50%	69.50%	64.20%	68.90%	—
PSO-SMOTETomek (Scenarios 6)	SVM	**89.57%/0.90**	**68.00%/0.66**	82.00%/0.71	73.89%/0.74	80.42%/0.81	**78.78%**
	DT	78.70%/0.65	65.00%/0.22	75.50%/0.76	66.67%0.70	73.75%/0.76	71.92%
	KNN	79.57%/0.77	64.67%/0.44	73.00%/0.47	**77.23%/0.74**	73.75%/0.71	73.64%
	LDA	68.26%/0.43	62.00%/0.65	71.50%/0.55	64.44%/0.66	71.25%/0.69	67.49%
	Gradient Boosting	82.61%/0.80	66.00%/0.57	**82.50%/0.84**	68.33%/0.53	**82.50%/0.81**	76.39%
	Accuracy Average	**79.74%**	65.13%	76.90%	70.11%	76.33%	—
PSO-SMOTE (Scenarios 7)	SVM	**83.50%**/0.76	**62.80%/0.64**	71.20%/0.61	62.72%/0.57	74.40%/0.69	**71.52%**
	DT	72.41%/**0.83**	58.09%/0.31	70.00%/0.63	59.09%/**0.69**	66.70%/0.67	65.26%
	KNN	80.34%/0.71	60.00%/0.62	70.00%/0.78	**72.45%**/0.47	69.60%/**0.80**	70.48%
	LDA	77.58%/0.61	61.43%/0.62	67.91%/**0.83**	56.36%/0.52	70.00%/0.78	66.66%
	Gradient Boosting	77.51%/0.66	56.19%/0.25	**72.50%**/0.69	62.72%/0.57	**73.60%**/0.67	68.50%
	Accuracy Average	**78.87%**	60.00%	70.00%	62.67%	70.86%	—
PSO-Tomek (Scenarios 8)	SVM	**70.00%/0.86**	**63.52%**/0.40	68.89%/0.61	56.11%/0.55	64.40%/0.40	**64.58%**
	DT	62.70%/0.69	64.70%/**0.63**	66.70%/**0.77**	57.78%/0.61	61.11%/0.61	62.60%
	KNN	66.11%/0.63	59.40%/0.50	62.22%/0.50	**72.77%/0.67**	61.67%/**0.82**	64.43%
	LDA	62.77%/0.78	61.76%/0.56	64.44%/0.53	57.22%/0.50	63.89%/0.31	62.02%
	Gradient Boosting	66.80%/0.61	58.82%/0.50	**69.44%**/0.57	58.89%/0.42	**64.50%**/0.40	63.69%
	Accuracy Average	**65.68%**	62.00%	66.00%	60.55%	63.11%	—

SMOTETomek: synthetic minority oversampling technique +Tomek Link.
The bold values significance that the values are bigger when the methods are compare.

which are correctly predicted. c represents the number of negative samples which are wrongly predicted. d is the number of positive samples which are wrongly predicted.

For the three datasets in this paper, essays dataset is no problem of unbalanced data and small amount of data. my Personality_text is a plain text dataset from my Personality. Hence, in order to verify the effectiveness of the PSO-SMOTETomek method, take the myPersonality dataset as an example, we added oversampling and under-sampling experiments, in which oversampling adopted the classical SMOTE technique and under-sampling adopted Tomek Link technique. Table 4 shows the result obtained on the my Personality dataset. The highest accuracy and F1-score are in scenario number 6. The highest accuracy and F1-score are 89.57% and 0.90, respectively, obtained from SVM algorithm. The highest average accuracy is 78.78%, obtained by the SVM algorithm. The highest average accuracy for all of the traits is 79.74%, obtained from Openness (OPN). Compared with the methods of PSO-SMOTE and PSO-Tomek, the experimental results show that the PSO-SMOTETomek method achieves the highest accuracy and F1-score.

Table 5 shows the result obtained by using the my Personality_text dataset. The accuracy and F1-score are highest in scenario numbers 3. The highest accuracy and F1-score are 88.15% and 0.82, respectively, obtained by the SVM algorithm. The highest average accuracy is 75.34%, obtained by the SVM algorithm. The highest average accuracy among all traits is 77.48% and is obtained for Openness (OPN). Table 6 shows the result obtained on the essays dataset. The accuracy and F1-score are highest in scenarios number 2. The highest accuracy is 67.58%,

TABLE 5. Traditional machine learning classification results of accuracy and F1-score on the MyPersonality_text dataset.

Features	Algorithm	OPN	CON	EXT	AGR	NEU	Accuracy Average
All features (Scenarios 1)	SVM	67.00%/0.70	**60.00%/0.64**	60.50%/**0.60**	59.00%/0.60	62.50%/**0.57**	61.80%
	DT	67.00%/0.69	59.00%/0.63	61.00%/0.48	**62.00%**/0.56	63.50%/0.38	**62.50%**
	KNN	**69.00%/0.80**	57.50%/0.63	**60.50%**/0.48	60.00%/0.56	63.00%/0.17	62.00%
	LDA	64.00%/0.76	58.50%/0.62	57.00%/0.60	56.50%/**0.66**	65.00%/0.40	60.20%
	Gradient Boosting	65.50%/0.67	55.00%/0.59	58.50%/0.40	57.00%/0.65	**66.50%**/0.30	60.50%
	Accuracy Average	**66.50%**	58.00%	59.50%	58.90%	64.00%	—
PSO (Scenarios 2)	SVM	73.50%/**0.78**	**64.50%**/0.27	64.00%/0.60	62.50%/0.55	68.50%/**0.62**	66.6%
	DT	72.50%/0.67	61.00%/0.48	**67.00%/0.67**	**64.50%**/0.46	68.00%/0.50	66.6%
	KNN	**74.50%**/0.69	62.50%/0.60	65.00%/0.30	63.50%/0.67	68.50%/0.31	**66.8%**
	LDA	69.00%/0.67	61.00%/0.53	63.00%/0.31	59.50%/0.35	**70.50%**/0.36	64.6%
	Gradient Boosting	71.00%/0.68	60.50%/**0.65**	61.50%/0.36	61.50%/**0.69**	70.00%/0.29	64.9%
	Accuracy Average	**72.0%**	61.9%	64.0%	62.3%	69.1%	—
PSO-SMOTETomek (Scenarios 3)	SVM	**88.15%/0.82**	63.13%/0.59	70.00%0.80	**74.12%/0.76**	**81.30%0.75**	**75.34%**
	DT	78.15%/0.79	60.63%/0.55	68.42%/0.62	65.29%/0.63	75.22%/0.60	69.34%
	KNN	77.78%/0.73	63.75%0.60	**70.53%0.78**	67.65%/0.66	76.96%/0.72	71.33%
	LDA	61.48%/0.73	56.88%/0.60	63.68%/0.78	62.36%/0.38	76.52%/0.57	64.18%
	Gradient Boosting	81.85%/0.76	**65.00%/0.67**	68.95%/0.64	64.72%/0.40	76.09%/0.62	71.32%
	Accuracy Average	**77.48%**	61.88%	68.00%	66.83%	77.22%	—

SMOTETomek: synthetic minority oversampling technique +Tomek Link.
The bold values significance that the values are bigger when the methods are compare.

TABLE 6. Traditional machine learning classification results of accuracy and F1-score on the essays dataset.

Features	Algorithm	OPN	CON	EXT	AGR	NEU	Accuracy Average
All features (Scenarios 1)	SVM	**58.38%/0.64**	55.94%/0.59	55.20%/0.59	54.95%/0.53	**56.88%**/0.56	**56.27%**
	DT	57.49%/0.56	**56.83%**/0.55	55.57%/0.54	53.25%/**0.61**	57.31%/0.55	56.09%
	KNN	53.86%/0.58	53.79%/**0.62**	54.26%/0.57	55.00%/0.60	55.00%/**0.57**	54.38%
	LDA	55.00%/0.56	55.16%/0.59	55.50%/**0.62**	**56.11%**/0.55	56.60%/0.55	56.00%
	Gradient Boosting	56.68%/0.54	55.66%/0.53	**56.19%**/0.57	54.26%/0.54	56.60%/0.53	55.88%
	Accuracy Average	56.28%	55.48%	55.34%	54.71%	**56.48%**	—
PSO (Scenarios 2)	SVM	66.94%/0.62	63.93%/0.64	61.88%/0.56	61.47%/0.60	64.93%/0.58	63.83%
	DT	65.11%/0.61	**65.50%/0.65**	63.11%/0.59	62.65%/0.55	62.29%/0.60	63.74%
	KNN	65.58%0.59	64.03%/0.63	62.70%/0.58	63.11%/0.59	64.34%/0.52	63.95%
	LDA	66.53%/0.60	63.94%/0.56	62.30%/0.60	**64.75%/0.62**	64.35%/0.62	**64.37%**
	Gradient Boosting	**67.58%/0.70**	64.94%/0.63	**64.34%/0.62**	64.30%/0.57	**65.53%/0.63**	65.34%
	Accuracy Average	**66.35%**	64.47%	62.87%	63.26%	64.29%	—

The bold values significance that the values are bigger when the methods are compare.

obtained by the Gradient Boosting algorithm. The highest average accuracy is 65.34%, obtained by the Gradient Boosting algorithm. The highest F1-score is 0.70, obtained from the Gradient Boosting algorithm. The highest average accuracy for all of the traits is 66.35% obtained from Openness (OPN).

TABLE 7. Comparison of accuracy results on 3 datasets, using PSO and PSO-SMOTETomek.

features	dataset	OPN	CON	EXT	AGR	NEU	Accuracy Average
All features	myPersonality_text	69.00%/0.80	60.00%/0.64	60.50%/0.60	62.00%/0.66	66.50%/0.57	63.60%
	essays	58.38%/0.64	56.83%/0.62	56.19%/0.57	56.11%/0.55	56.88%/0.57	56.88%
	myPersonality	71.50%/0.90	59.50%/0.60	68.00%/0.64	62.50%/0.70	67.50%/0.62	65.80%
PSO	myPersonality_text	74.50%/0.78	64.50%/0.65	67.00%/0.67	64.50%/0.69	70.50%/0.62	68.20%
	essays	**66.94%/0.70**	**65.50%/0.65**	**64.34%/0.62**	**64.75%/0.62**	**65.53%/0.63**	**65.41%**
	myPersonality	76.00%/0.82	65.00%/0.63	72.50%/0.73	73.50%/0.75	71.90%/0.73	71.78%
PSO-SMOTETomek	myPersonality_text	**88.15%/0.82**	**65.00%/0.67**	**70.53%/0.78**	**74.12%/0.76**	**81.30%/0.75**	**75.82%**
	essays	—	—	—	—	—	—
	myPersonality	**89.57%/0.90**	**68.00%/0.66**	**82.50%/0.84**	**77.23%/0.74**	**82.50%/0.81**	**79.96%**

SMOTETomek: synthetic minority oversampling technique +Tomek Link.
The bold values significance that the values are bigger when the methods are compared.

As can be seen from the experiment results, first, the PSO-SMOTETomek method is the best method among all the methods. Second, for the myPersonality and my Personality_text datasets, SVM model is the algorithm with the best performance among the five algorithms. However, for the essays dataset, Gradient Boosting algorithm is the model with the best performance among the five algorithms. Finally, in the three datasets of this paper, the highest average accuracy for all of the traits obtained from Openness (OPN).

Table 7 shows that personality traits are well predicted using feature optimization methods and the PSO-SMOTETomek method. Both PSO-SMOTETomek and PSO contribute to the classification results, and PSO-SMOTETomek performs significantly better than PSO alone. For the PSO method, the best average accuracy on the three datasets (myPersonality_text, essays, my Personality) are 68.20%, 65.41%, and 71.78%, respectively. For the PSO-SMOTETomek method, the best results are 75.82%, 65.41%, and 79.96%, the best F1-score are 0.82, 0.70, and 0.90, respectively. This is shown in Fig.3.

At the end of the experiment, we calculated the time cost of the experiment. We calculated the training time and test time sum of the five personality programs executed by each model on 3 datasets. From Table 8, for the 3 datasets, first, the PSO method takes the shortest time, because the PSO method has the effect of selecting the optimal feature, which reduces the dimension of the feature. Second, for the datasets myPersonality_text and myPersonality, the maximum time consumed by the PSO-SMOTETomek method over the PSO method are 0.0069s/0.0022s and 0.0330s/0.0008s, respectively. Hence, the PSO-SMOTETomek method is not much more time-consuming than the PSO method, besides the PSO-SMOTETomek method solves the problem of data imbalance by adding data.

By observing all average accuracies and F1-score from experiments on traditional machine learning, in the three

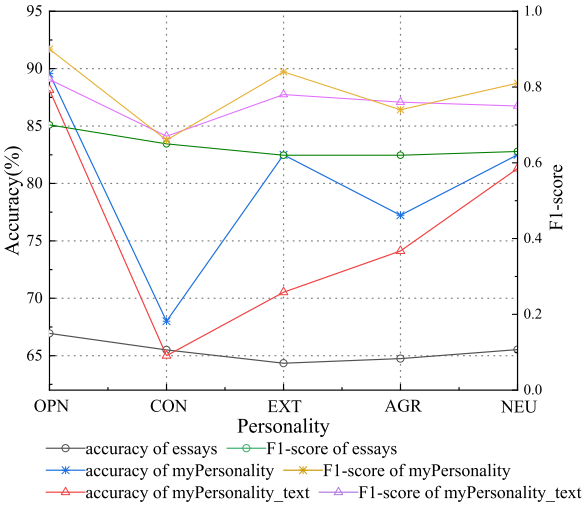

FIGURE 3. Best accuracy and F1-score results on essays, MyPersonality, MyPersonality_text.

different methods, all of the average accuracies and F1-score for each dataset are quite different. From the average accuracy results based on traits, we can see that the PSO-SMOTETomek method has the highest average accuracies and F1-score on the three datasets while the PSO method had the second highest average accuracy and F1-score. The result may be different in other research, as it heavily depends on the dataset and the classification model and its features used for the prediction.

B. COMPARISON WITH SAME TYPE OF DATASETS

The performances of three different methods using five basic machine learning algorithms on the three datasets are presented in Table 4-7. The accuracies on the three datasets are presented in Fig.4.

From Fig.4 (a), for the two plain text datasets, essays and my Personality_text, first, when comparing the classifier

TABLE 8. Comparison of time results on 5 algorithms, using PSO and PSO-SMOTETomek (unit:s).

features	dataset	SVM	DT	KNN	LDA	Gradient Boosting
All features	myPersonality_text	0.0244/0.0035	0.0600/0.0012	0.0067/0.0075	0.0243/0.0015	0.2850/0.0035
	essays	2.9700/0.2890	1.0400/0.0021	0.0774/0.2600	0.1170/0.0022	2.0827/0.0031
	myPersonality	0.0327/0.0034	0.0409/0.0012	0.0046/0.0021	0.0344/0.0021	0.3920/0.0031
PSO	myPersonality_text	**0.0158/0.0024**	**0.0258/0.0020**	**0.0065/0.0038**	**0.0178/0.0019**	**0.2245/0.0022**
	essays	**2.1974/0.1855**	**0.3972/0.0027**	**0.0372/0.1942**	**0.0676/0.0017**	**0.9337/0.0048**
	myPersonality	**0.0208/0.0034**	**0.0308/0.0016**	**0.0055/0.0072**	**0.0209/0.0012**	**0.3170/0.0020**
PSO-SMOTETomek	myPersonality_text	0.0195/0.0026	0.0327/0.0018	0.0072/0.0060	0.0194/0.0015	0.2258/0.0031
	essays	—	—	—	—	—
	myPersonality	0.0225/0.0025	0.0324/0.0019	0.0056/0.0067	0.0221/0.0016	0.3500/0.0028

SMOTETomek: synthetic minority oversampling technique +Tomek Link.

The bold values significance that the values are smaller when the methods are compared.

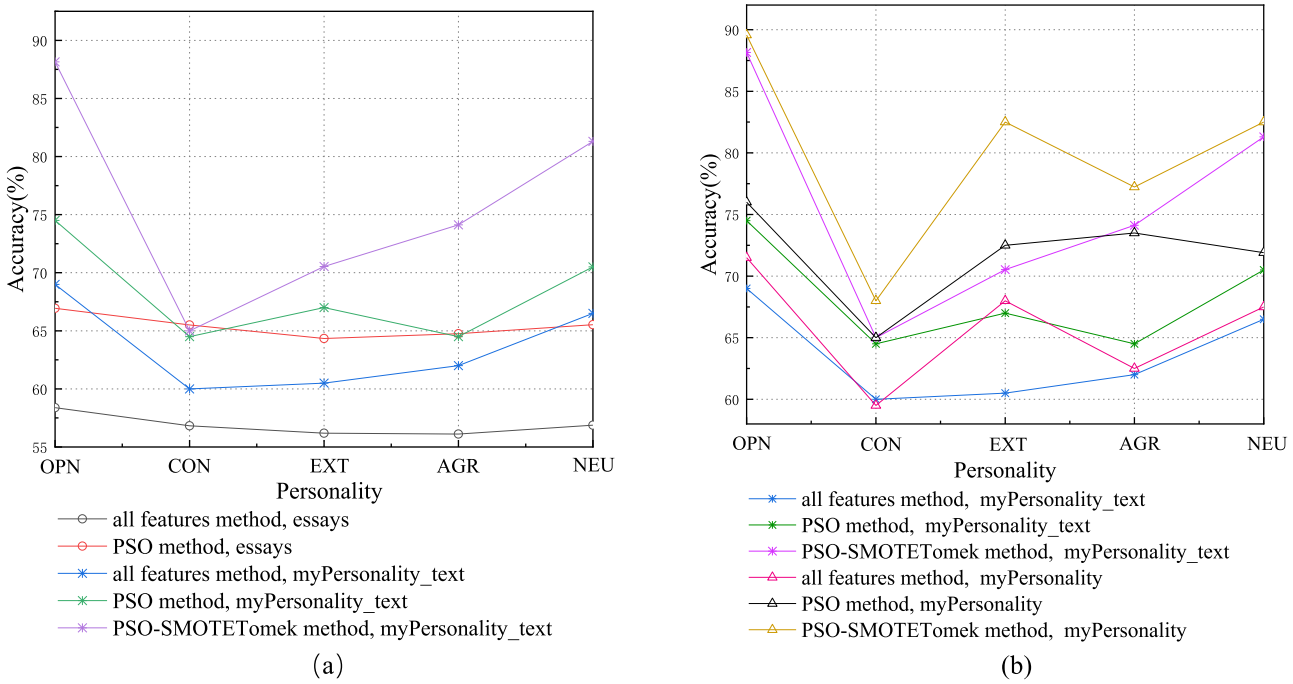

FIGURE 4. (a) Comparision of the best results of essays and MyPersonality_text, (b) comparision of the best results on MyPersonality and MyPersonality_text.

performances with and without PSO, we can conclude that these five machine learning methods achieve better accuracy when using PSO. Second, the myPersonality_text dataset provides significantly better accuracy than the essays dataset. In other words, the short text dataset (myPersonality_text) is more accurate in this experiment. Finally, the accuracy on my Personality_text is highest when the PSO-SMOTETomek method was used.

In the same way, from Fig.4 (b), we can see the following. First, when comparing the classifier performances with and without users' external information, the accuracy on my Personality is better than that on my Personality_text, which means that the users' external information has a certain role in personality recognition. However, from the experimental data point of view, it is not as important as we think because the accuracy on the two datasets is not very different. Second, PSO and resampling can also significantly improve the accuracy. Finally, the accuracies on myPersonality_text and myPersonality are highest when the PSO-SMOTETomek method was used.

C. COMPARISON WITH EXISTING PERSONALITY PREDICTION METHODS

In our paper, considering that Navonil Majumder *et al.*'s study [9] on personality recognition is most closely related to our work, we compare our results on the essays dataset

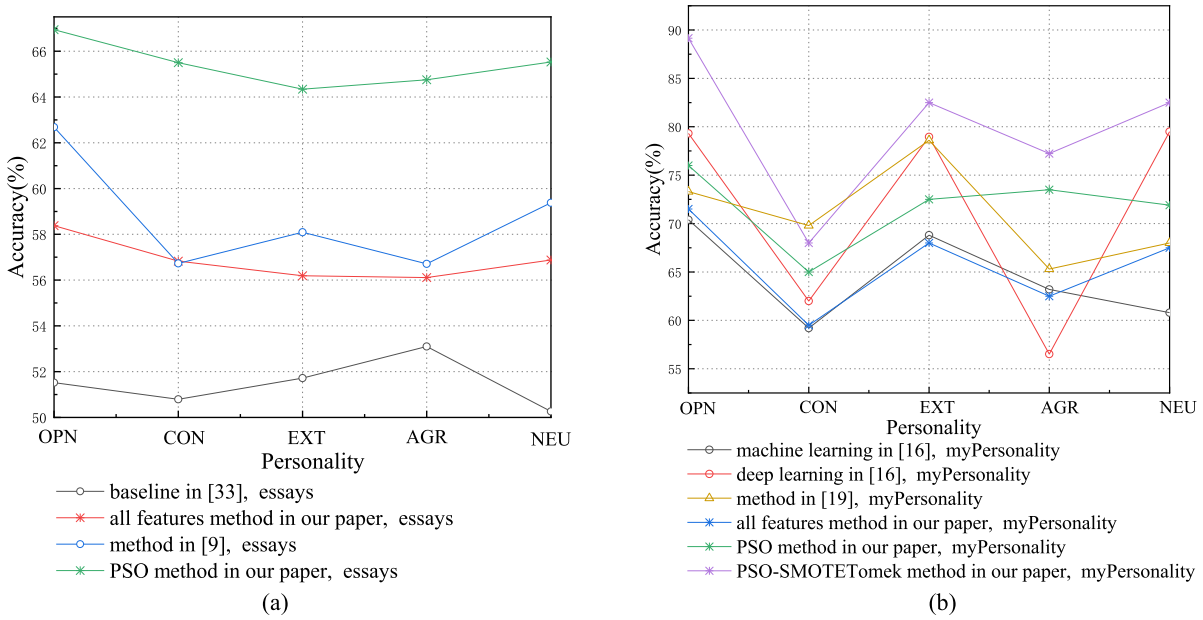

FIGURE 5. (a) Comparison with the best results of Navonil Majumder's work; (b) comparison with the best results of Tommy Tandera's work.

with the best results of their work. From Fig.5 (a), the results show that the PSO feature optimization algorithm can extract a better combination of features, which contributes to higher performance with our model; moreover, and the accuracy of personality recognition on essays is improved by around 4%–8%.

For the myPersonality dataset, we compared the algorithms proposed by [16], [18] with different approaches to machine learning. In this paper, we introduced a resampling technique to try to balance the number of minority and majority classes in the dataset. After performing resampling of the original dataset using the PSO-SMOTETomek method, we applied SVM, DT, KNN, gradient boosting, and LDA as classifiers. The comparison of results is illustrated in Fig.5(b). The accuracy of personality recognition on myPersonality is improved by around 3%-10% compared to methods in [16], [18].

The greatest improvement of accuracy occurred when using PSO and PSO-SMOTETomek. In summary, PSO-SMOTETomek can improve the accuracy of classifiers on the minority class and the overall accuracy. Comparing the machine learning methods, on the whole, SVM had the best performance. Without the PSO method, DT outperforms SVM on accuracy; meanwhile, when using the resampling method, KNN outperforms SVM. The experimental results also show that the proposed method can greatly improve the accuracy compared with other methods [9], [16], [18], [32].

V. CONCLUSION

Considering the evaluation results in Table 4-7 and Fig. 3-5, for the three datasets, the methods of PSO and PSO-SMOTETomek outperform several existing methods because it improves accuracy. These results may have occurred because PSO can optimize features and SMOTETomek pays particular attention to the positive and negative sample overlap region-those most likely to be misclassified.

The average accuracies of the my Personality, my Personality_text and essays datasets are 78.78%, 75.34%, and 65.34%, respectively. For the datasets of my Personality and my Personality_text, the highest accuracies obtained using the SVM algorithm are 88.15% and 89.57%, respectively. For the essays dataset, the highest accuracies obtained using the Gradient Boosting algorithm is 67.58%. From the experimental results, we can see that PSO-SMOTETomek method is efficient under a small dataset. More importantly, we found that compared with long text dataset, short text dataset has better classification effect. In the absence of user external information, plain text data can still have high personality recognition accuracy.

In a future study, we plan to use a data enhancement algorithm, deep learning architectures, and other processes to improve the proposed prediction method.

ACKNOWLEDGMENT

The authors would like to thank LetPub(www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

REFERENCES

[1] L. Li, A. Li, B. Hao, Z. Guan, and T. Zhu, "Predicting active users' personality based on micro-blogging behaviors," *PLoS ONE*, vol. 9, no. 5, Jan. 2014, Art. no. e98489.
[2] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski, "Workshop on computational personality recognition," in *Proc. 7th Int. AAAI Conf. Weblogs Social Media*, Boston, MA, USA, Jun. 2013, pp. 2–5.
[3] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M.-F. Moens, and M. De Cock, "Computational personality recognition in social media," *User Model. User-Adapted Interact.*, vol. 26, nos. 2–3, pp. 109–142, 2016.
[4] N. Ahmad and J. Siddique, "Personality assessment using Twitter tweets," *Procedia Comput. Sci.*, vol. 112, pp. 1964–1973, Jan. 2017.

[5] W. Bleidorn and C. J. Hopwood, "Using machine learning to advance personality assessment and theory," *Personality Social Psychol. Rev.*, vol. 23, no. 2, pp. 190–203, 2019.
[6] M.-R. Oh and S.-B. Won, "A study on the relational analysis between GEOPIA and MBTI preference index," *J. Korea Academia-Ind. Cooperation Soc.*, vol. 19, no. 7, pp. 325–336, 2018.
[7] S. J. Hunt, "Communication and change: A qualitative study of law enforcement team DISC personality traits," Grand Canyon Univ., Phoenix, AZ, USA, 2018.
[8] P. D. Harms, "Gallup strengths finder," in *Encyclopedia of Personality and Individual Differences*. 2017, pp. 1–3.
[9] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intell. Syst.*, vol. 32, no. 2, pp. 74–79, Mar./Apr. 2017.
[10] M. Fallahnezhad, M. Vali, and M. Khalili, "Automatic personality recognition from reading text speech," in *Proc. Iranian Conf. Elect. Eng. (ICEE)*, May 2017, pp. 18–23.
[11] K. Chen, F. Zhou, Y. Wang, and L. Yin, "An ameliorated particle swarm optimizer for solving numerical optimization problems," *Appl. Soft Comput.*, vol. 73, pp. 482–496, Dec. 2018.
[12] Q. Wu, Z. Ma, J. Fan, G. Xu, and Y. Shen, "A feature selection method based on hybrid improved binary quantum particle swarm optimization," *IEEE Access*, vol. 7, pp. 80588–80601, 2019.
[13] L. R. Goldberg, "The structure of phenotypic personality traits," *Amer. Psycholog.*, vol. 48, no. 1, pp. 26–34, 1993.
[14] M. C. Y. Shun, M. C. Yan, S. Zhiqi, and A. Bo, "Learning personality modeling for regulating learning feedback," in *Proc. IEEE 15th Int. Conf. Adv. Learn. Technol.*, Jul. 2015, pp. 355–357.
[15] J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media," in *Proc. Extended Abstr. Hum. Factors Comput. Syst.*, May 2011, pp. 253–262.
[16] H. T. Tandera, D. Suhartono, R. Wongso, and Y. L. Prasetio, "Personality prediction system from Facebook users," *Procedia Comput. Sci.*, vol. 116, pp. 604–611, Jan. 2017.
[17] D. Xue, Z. Hong, S. Guo, L. Gao, L. Wu, J. Zheng, and N. Zhao, "Personality recognition on social media with label distribution learning," *IEEE Access*, vol. 5, pp. 13478–13488, 2017.
[18] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Personality predictions based on user behavior on the Facebook social media platform," *IEEE Access*, vol. 6, pp. 61959–61969, 2018.
[19] V. Ong, A. D. S. Rahmanto, Williem, D. Suhartono, A. E. Nugroho, E. W. Andangsari, and M. N. Suprayogi, "Personality prediction based on Twitter information in Bahasa Indonesia," in *Proc. Federated Conf. Comput. Sci. Inf. Syst. (FedCSIS)*, 2017, pp. 367–372.
[20] M. Vaidhya, B. Shrestha, B. Sainju, K. Khaniya, and A. Shakya, "Personality traits analysis from Facebook data," in *Proc. 21st Int. Comput. Sci. Eng. Conf. (ICSEC)*, Nov. 2017, pp. 1–5.
[21] T. A. Judge, C. A. Higgins, C. J. Thoresen, and M. R. Barrick, "The big five personality traits, general mental ability, and career success across the life span," *Personnel Psychol.*, vol. 52, no. 3, pp. 621–652, 1999.
[22] J. M. Digman, "Personality structure: Emergence of the five-factor model," *Annu. Rev. Psychol.*, vol. 41, no. 1, pp. 417–440, 1990.
[23] N. Febrianto, I. Prasetia, and A. Wijaya, "Pembuatan sistem prediksi kepribadian 'the big five traits' dari media sosial Twitter," Univ. Bina Nusantara, Jakarta, Indonesia, Tech. Rep., 2015.
[24] M. Kosinski, S. C. Matz, S. D. Gosling, V. Popov, and D. Stillwell, "Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines," *Amer. Psychologist*, vol. 70, no. 6, p. 543, 2015.
[25] E. P. Tighe, J. C. Ureta, B. A. L. Pollo, C. K. Cheng, and R. de Dios Bulos, "Personality trait classification of essays with the application of feature reduction," in *Proc. SAAIP IJCAI*, 2016, pp. 22–28.
[26] D. Markovikj, S. Gievska, M. Kosinski, and D. J. Stillwell, "Mining Facebook data for predictive personality modeling," in *Proc. 7th Int. AAAI Conf. Weblogs Social Media*, Jun. 2013, pp. 23–26.
[27] W. Wang, J. Liu, S. Yang, and Z. Guo, "Typography with decor: Intelligent text style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5889–5897.
[28] Y. Mao, D. Zhang, C. Wu, K. Zheng, and X. Wang, "Feature analysis and optimisation for computational personality recognition," in *Proc. IEEE 4th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2018, pp. 2410–2414.
[29] G. Farnadi, S. Zoghbi, M.-F. Moens, and M. De Cock, "Recognising personality traits using Facebook status updates," in *Proc. 7th Int. AAAI Conf. Weblogs Social Media*, Jun. 2013, pp. 14–18.
[30] M. Amoozegar and B. Minaei-Bidgoli, "Optimizing multi-objective PSO based feature selection method using a feature elitism mechanism," *Expert Syst. Appl.*, vol. 113, pp. 499–514, Dec. 2018.
[31] A. A. Nagar, F. Han, Q.-H. Ling, and S. Mehta, "An improved hybrid method combining gravitational search algorithm with dynamic multi swarm particle swarm optimization," *IEEE Access*, vol. 7, pp. 50388–50399, 2019.
[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

ZHE WANG received the master's degree from Henan Polytechnic University, Henan, China, in 2014. She is currently pursuing the Ph.D. degree with the School of Cyberspace Security, Beijing University of Posts and Telecommunications. Her main research interests include network security optimization and defense combined with the disciplines of computer science and social engineering.

CHUNHUA WU received the Ph.D. degree from the Beijing University of Posts and Telecommunications, in 2008, where she is currently a Teacher with the Information Security Center, School of Cyberspace Security. Her research interests include network and information security, applied more than seven important projects.

KANGFENG ZHENG received the Ph.D. degree from the Beijing University of Posts and Telecommunications, in 2006, where he is currently a Professor with the Information Security Center, School of Cyberspace Security. He has published over 50 technical articles in international conferences and journals. His research interests include networking and system security, network information processing, and network coding. He has presided over a number of national scientific research projects and received a number of national and provincial awards.

XINXIN NIU received the Ph.D. degree from The Chinese University of Hong Kong. She is currently a Professor and the Director of the Information Security Center, School of Cyberspace Security, Beijing University of Posts and Telecommunications. Her research interests include information security, information hiding, digital watermarking, digital content, and its security. She has presided over a number of national scientific research projects and received a number of national and provincial awards.

XIUJUAN WANG received the Ph.D. degree in information and signal processing from the Beijing University of Posts and Telecommunications, in July 2006. She is currently an Instructor Lecturer with the Faculty of Information Technology, Beijing University of Technology. Her research interests include information and signal processing, network security, and network coding.