# Personality Recognition on Social Media with Label Distribution Learning

Di Xue, Zheng Hong*, Shize Guo, Liang Gao, Lifa Wu,
Jinghua Zheng and Nan Zhao

***Abstract*—Personality is an important psychological construct accounting for individual differences in people. To reliably, validly, and efficiently recognize an individual's personality is a worthwhile goal, however, the traditional ways of personality assessment through self-report inventories or interviews conducted by psychologists are costly and less practical in social media domains, since they need the subjects to take active actions to cooperate. This paper proposes a method of Big Five personality recognition from microblog in Chinese language environments with a new machine learning paradigm named label distribution learning (LDL), which has never been previously reported to be used in personality recognition. One hundred and thirteen features are extracted from 994 active Sina Weibo users' profiles and micro-blogs. Eight LDL algorithms and nine non-trivial conventional machine learning algorithms are adopted to train the Big Five personality traits prediction models. Experimental results show that two of the proposed LDL approaches outperform the others in predictive ability, and the most predictive one also achieves relatively higher running efficiency among all the algorithms.**

***Index Terms*—Personality Recognition; Label Distribution Learning; Social Media Mining; Big Five Personality**

## I. INTRODUCTION

PERSONALITY is a psychological construct aimed at explaining the wide variety of human behaviors in terms of a few, stable and measurable individual characteristics [1]. It not only reflects an individual's consistent patterns of behavior, thought and interpersonal communication [2], but also influences important life aspects, including happiness, motivations, preferences, emotion, mental and physical health [3]. The study of personality is of central importance in psychology, and personality recognition [4] can also benefit many other applications, such as social network analysis [5], recommendation systems [6], deception detection [7], authorship attribution [8], sentiment analysis/opinion mining [9] and so on [10]. To reliably, validly, and efficiently recognize an individual's personality is a worthwhile goal, however, the traditional ways of personality assessment through self-report inventories or interviews conducted by psychologists are costly and less practical in social media domains [11], since they need the subjects to take active actions to cooperate.

Fortunately, with the development of various social media, enlisting modern computer science has real potential for advancing that endeavor [12]. The rich digital traces and self-disclosed personal information on social networking platforms render it possible to analyze the users' behaviors and infer their personality traits on the web [13], which have attracted much attention from different disciplines [14-16].

To date, various studies have been done to automatically recognize an individual's personality traits from publicly available information on social media (see section II for a review of these works). However, the achievements of the existing methodologies are not satisfactory, and there is much room for improvement. Specifically speaking, for one thing, the majority of the work used classification approaches, especially binary classification, to address the personality recognition problem, which simply split subjects into classes (e.g. above and below median with respect to a certain trait). This is not meaningful from a psychological point of view and is not useful for practical purposes. Because it can hardly provide convincing argument when emphasizing comparisons among individuals, which is exactly what people love to do. Therefore, representing the predicted traits in terms of continuous numerical scores would be more suitable and psychologically meaningful. For another, the bulk of the work has modeled each personality trait in isolation since the traits are supposed to be uncorrelated and independent for simplicity. In fact, psychological research has proved that there are correlations between the basic traits in Big Five model, and jointly modeling the traits might have better performance [1].

Given the above-mentioned limitations, we present a novel methodology of personality recognition based on a new machine learning paradigm named label distribution learning (LDL) [17], which assigns a label distribution rather than a single label or a relevant label set to each instance. LDL can deal with not only multiple labels of one instance, but also the different proportions that these labels account for in a full description of the instance. To the best of our knowledge, this kind of learning method has never been previously reported to be used in personality recognition.

Our study is carried out on Sina Weibo, the most popular microblogging website in mainland China [18]. Nine hundred and ninety-four active Sina Weibo users have been involved in our experiments, whose personality scores (i.e. the gold standard labels for objective evaluation) were obtained

Di Xue, Zheng Hong and Lifa Wu are with College of Command Information Systems, PLA University of Science and Technology, Nanjing 210002, China. e-mail: delia_xue@126.com; hz5215@163.com; wulifa@vip.163.com.

Shize Guo and Liang Gao are with the Institute of North Electronic Equipment, Beijing 100083, China. e-mail: nsfgsz@126.com; gaolleasy@163.com.

Jinghua Zheng is with Electronic Engineering Institute, Hefei 230037, China. e-mail: zhengjh1001@163.com.

Nan Zhao is with Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China. e-mail:zhaonan@psych.ac.cn.

* To whom correspondence should be addressed.

Manuscript received April 09, 2017.

through tests of the Big Five personality model [19]. Big Five personality model is the dominant paradigm in personality psychology as well as the widely accepted model in computing oriented personality research [1], which embraces five basic traits: openness to experience (O), conscientiousness (C), extraversion (E), agreeableness (A) and neuroticism (N) [20]. We transformed the Big Five personality recognition problem into a label distribution learning task, and adopted eight LDL algorithms to train the Big Five personality traits prediction models. Besides, nine non-trivial conventional algorithms were applied to train the baseline models based on Weka [1], a popular machine learning workbench. Experimental results show that the most predictive models were trained by two LDL algorithms named LD-SVR and SA-IIS. Besides, LD-SVR, which achieves the lowest prediction error, also occupies the preferable running efficiency when building prediction models.

The rest of the paper is organized as follows. Related work is discussed in Section II. In section III, we describe the proposed methodology of personality recognition in details, including description of the data set, feature extraction process and the process of utilizing LDL algorithms to automatically recognize an individual's Big Five personality traits. Experimental evaluation is presented in Section IV, followed by the conclusion and future work in section V.

## II. Related Work

Personality Recognition (PR) concerns the automatic inference of an individual's personality traits from various sources, which can be compared against gold standard labels obtained by means of personality tests [20].

The earliest efforts mainly focused on PR from written text. The first pioneering work was done by Argamon et al [21] about 10 years ago, which extracted word categories and relative frequency of function words from 2236 written essays of 1200 students as the input of Support Vector Machines (SVM) to discriminate between students at the opposite extremes of Extraversion and Neuroticism. The same data set and approach were used by Mairesse et al [22] to recognize individuals in the upper and lower half of the observed scores for all Big Five traits. They studied the effectiveness of different sets of textual features extracted from psychologically oriented text analysis tools (e.g. LIWC [2]) or psycholinguistic dictionary (e.g. MRC [3]), and found that the openness trait yields the best accuracy using SVM. The state of the art of PR from witten essays should be the approach proposed by Majumder et al [59], which adopted the Convolution Neural Network (CNN) to extract features and model document to solve the PR problem. The accuracy of this approach outperformed others for all big five personality traits.

Along with the popularization of online blogging platforms, attempts at PR tend to focus on blogs [23, 24], since blogs incline to concentrate on personal matters and experiences which may leave traces of the writer's personality [25]. In [23] and [24], Naive Bayes classifiers and SVM were applied to classify high and low scoring bloggers for Big Five traits by extracting the frequencies of N-grams (i.e. N-long word sequences) in the blogs as patterns, which is a bottom-up pattern extraction method other than a lexical approach adopted in [21, 22].

In more recent years, driven by the explosive increase of mobile and intelligent terminals and the penetration of social media, research on PR has evolved in two directions: PR via mobile and wearable devices and PR on social media. Scientists have carried out various PR studies from behavioral cues in face to face interactions, speeches, or videos collected by cameras, microphones [26-30], wearable sensors [31, 32], and mobile phones [33-35]. As for PR on social media, it is gaining increased research attention due to the rich self-disclosed personal information and emotional contents on social media, the proven high correlation between personality traits and users' digital traces of online activities [14, 36-42], as well as its potential in many computational applications [43].

Taking advantages of corpora obtained from social media, such as Twitter, Facebook, Sina Weibo, a variety of work has been done to explore optimal feature space and test machine learning algorithms for PR [25, 44-48]. One of the earliest work was reported in 2011 by Golbeck et al [25], which exploited 77 features related to users' egocentric network, personal information, language usage, preferences and activities. They carried out regression analysis with M5' Rules and Gaussian Processes over 167 Facebook users assessed by the Big Five, and continued to apply a similar approach over 279 Twitter users in the same year [44]. Quercia et al [45] analyzed the relationship between personality and different types of Twitter users, and adopted M5 algorithm to recognize 335 users' Big Five traits simply based on three publicly available counts: follower, following and listed counts (i.e. the number of people that include the user in their reading list). They found that neuroticism is the hardest trait to predict. Alam et al [47] performed PR of Big Five traits by using *myPersonality* [4] corpus collected from Facebook by Celli et al [10]. They studied different classification methods, namely Sequential Minimal Optimization (SMO) for SVM, Bayesian Logistic Regression (BLR) and Multinomial Naive Bayes (MNB) sparse modeling, and found that MNB performed better than BLR and SMO for PR. Considering the difficulty in annotating the massive amounts of data generated in social media, Lima et al [60] proposed a semi-supervised learning approach and developed a unique system to recognize users' personlity traits based on groups of tweets, which promised the real possibility of working effectively with large dataset. The above-mentioned efforts were made on a single social networking site, however, Skowron et al [48] carried out PR experiments based on text, image, and users' meta data collected from two popular social networking sites, i.e., Twitter and Instagram, and reported that such joint analysis contributed to the decrease of the prediction error.

Besides the English corpora, PR has also been carried out in Chinese language environments [49-53], by gathering the

[1] http://www.cs.waikato.ac.nz/ml/weka/

[2] http://www.liwc.net

[3] http://www.psych.rl.ac.uk

[4] http://mypersonality.org

demographic information, usage statics and emotional states of 209 users on a Chinese social networking platform named RenRen, Bai et al applied C4.5 decision trees to recognize subjects in the low, middle or high of observed scores. Li et al [52] performed PR experiments over 547 Chinese active users of Sina Weibo. They extracted not only static features from users' profiles, privacy setting, self-expression and interpersonal behaviors but also dynamic features consisting of micro-blogging updates, @ mentions, use of apps and recordable browsing. Peng et al [53] performed experiments over 222 Chinese Facebook users to classify personality traits with SVM. They used a Chinese text segmentation tool named Jieba as the tokenizer, and found that the performance could be improved with the help of text segmentation and utilization of side information such as the number of friends.

Considering that almost all the scholars who worked in PR adopted different technology road maps, and standard dataset or common benchmarks are unavailable, Celli et al organized *Workshop on Computational Personality Recognition (Shared Task)* in 2013 [10] and 2014 [20] to define the state-of-the-art and provide corpora and tools for future standard evaluation of PR approaches. Besides, another shared task of personality recognition was organized under the umbrella of *Author Profiling task at PAN 2015* [54].

Overall, most of the past work in PR tended to propose classification approaches to address the personality recognition problem and modeled each personality trait in isolation. Although there were a few PR approaches testing the conventional multi-label machine learning algorithms, the reported differences between univariate and multivariate models were not significant [43]. In this paper, we introduce a new machine learning paradigm named Label Distribution Learning, which could deal with not only multiple labels of one instance, but also the different proportions that these labels account for in a full description of the instance. In addition, our approach can assign continuous scores to the subjects with respect to each personality trait of each user. Detailed methodology is presented in section III.

## III. Methodology

### A. Data Set

With over 297 million monthly active users and 132 million daily active users [55], Sina Weibo is one of the most popular microblogging website in mainland China [18]. Large quantities of users publicly report their statuses, share their experiences, and interact with others in daily life, and hundreds of millions of messages are posted each day on it [56]. It provides an ideal online platform for personality research and related applications, thus, our study is carried out on it.

To construct the dataset, we first recruited volunteers by sending invitation messages to active Sina Weibo users. About 1% of them responded to our invitation and participated in the questionnaire investigation through an online application named "XinLiDiTu", which was developed by Computational Cyber-Psychology Lab (CCPL)[5], Institute of Psychology, Chinese Academy of Sciences. The volenteers' profiles and status updates were crawled through Sina Weibo APIs, and their personality scores were tested by the popular Chinese-version Big Five Inventory (BFI), which contains 44 questions and was originally developed by Oliver John. Finally, we constructed a dataset involving 994 effectively labelled Sina Weibo users, whose total number of micro-blogs is greater than 532 and average count of micro-blogs updates per day is higher than 2.84.

The scale of BFI tested scores, mean value and standard deviation of the Sina Weibo users' personality scores are presented in Table I. The Pearson correlation coefficient among personality scores are presented in Table II, indicating that the Big Five traits are positively correlated with each other except the neuroticism dimension, which is negatively correlated with the other four personality traits. With a coefficient of -0.4817, the correlation relationship between neuroticism (N) and conscientiousness (C) is strongest in all, while the weakest relationship is between agreeableness (A) and extraversion (E). The distribution of personality scores are shown in Fig. 1, from which we can conculde that the personality scores of Sina Weibo users conform to normal distribution.

TABLE I
Statistics of the BFI tested Big Five personality scores

| Personality Trait | O | C | E | A | N |
|---|---|---|---|---|---|
| Mean | 35.63 | 28.32 | 23.88 | 32.86 | 25.12 |
| Stand Deviation | 6.18 | 5.81 | 5.48 | 5.05 | 5.43 |
| Scale | [10, 50] | [9, 45] | [8, 40] | [9, 45] | [8, 40] |

Note: O, C, E, A, N refer to the five dimensions of personality: openness to experience, conscientiousness, extraversion, agreeableness, neuroticism, respectively.

TABLE II
Pearson correlation coefficient among Big Five personality scores

| | O | C | E | A | N |
|---|---|---|---|---|---|
| O | 1 | 0.1911 | 0.2724 | 0.1322 | -0.1772 |
| C | | 1 | 0.2255 | 0.3197 | -0.4817 |
| E | | | 1 | O.1191 | -0.3811 |
| A | | | | 1 | -0.4626 |
| N | | | | | 1 |

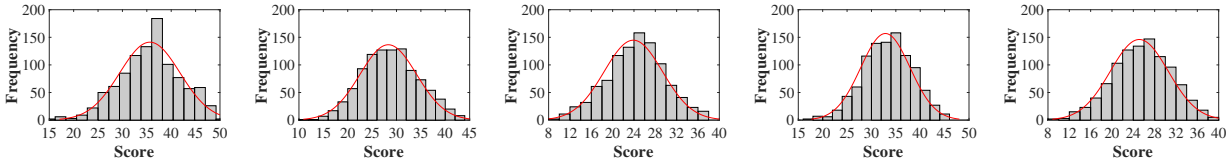
Note: O, C, E, A, N refer to the five dimensions of personality: openness to experience, conscientiousness, extraversion, agreeableness, neuroticism, respectively.

### B. Feature Extraction

It is noteworthy that this paper mainly focused on finding a better machine learning algorithm to solve the PR problem, rather than exploring the optimal feature set. Thus, the features should be the commonly accepted and relatively useful ones, while the feature extraction process we adopted tended to be the lightweight one.

We extract three categories of features from users' profiles and their posted micro-blogs, namely profile-based static features, profile-based dynamic features and content-based micro-blogs features. The profile-based static features specifically

[5]http://http://ccpl.psych.ac.cn




(a) Openness to experience (b) Conscientiousness (c) Extraversion (d) Agreeableness (e) Neuroticism

Fig. 1: Distribution of personality scores on Big Five traits on Sina Weibo dataset

refer to those that may experience little changes over time, such as gender, address, nickname, while the profile-based dynamic features are those that may suffer obvious changes over time, such as the number of followers or followings. These profile-based features are directly obtained from users' public profiles. The content-based micro-blogs features refer to those extracted from the raw data of users' posted micro-blogs, including linguistic features, psychological features, and so on. For each user, we combined all the micro-blogs he/she has posted as a single document, and extracted the content-based micro-blogs features by utilizing *TextMind* [57], a Chinese language psychological analysis system developed by CCPL. In total, 113 dimensions of features are obtained, please see Table A1 in APPENDIX for details.

After the feature extraction, we normalize the feature space by min-max normalization as Eq. (1) to rescale the range of all variables so that all of them belong to the same range,

$$f^* = \frac{f - Min}{Max - Min} \tag{1}$$

where $f^*$ and $f$ are respectively the normalized and the original value of a certain dimension of feature for a user, $Min$ and $Max$ are respectively the minimum and maximum value of the corresponding dimension of feature for all users.

### C. Personality Recognition

To address the problem of PR, we adopt a new machine learning paradigm named label distribution learning (LDL) [17], which was recently proposed and has never been previously reported to be used in personality recognition. Each instance in LDL is associated with a *label distribution* (i.e. a real-valued vector) that covers a certain number of labels. The real-valued elements in the label distribution are defined as the *description degrees* of the labels, representing the degree to which each label describes the instance. Formally, let $\boldsymbol{x}_i$ denotes the $i$-th instance, $y_j$ denotes the $j$-th label, then the description degree of the label $y_j$ to the instance $\boldsymbol{x}_i$ can be represented by $d_{\boldsymbol{x}_i}^{y_j}$, and the label distribution of $\boldsymbol{x}_i$ can be denoted by $\vec{D}_i = \left[d_{\boldsymbol{x}_i}^{y_1}, d_{\boldsymbol{x}_i}^{y_2}, ...d_{\boldsymbol{x}_i}^{y_j}, ..., d_{\boldsymbol{x}_i}^{y_c}\right]$, where $c$ represents the total number of possible labels [17]. The above analysis implies that LDL can deal with not only multiple labels of one instance but also the different proportions that these labels account for in a fully description of the instance, while in multi-label learning (MLL), each label just equally describes the instance. Thus, LDL can annotate the instance in a more natural way and is a more general learning framework, which includes MLL as a special case [17].

In our research, we consider that individuals' personality could be regarded as the mixtures of basic personality traits and allows different intensities in each trait, which implies that LDL may be helpful to the PR problem. However, it is noteworthy that the description degree in LDL needs to satisfy two constraints, that is $d_{\boldsymbol{x}_i}^{y_j} \in [0, 1]$ and $\sum_j d_{\boldsymbol{x}_i}^{y_j} = 1$, which means that using the labels involved in the label distribution could always fully describe the instance [58]. Given the fact that the existing personality model (e.g. Big Five Model) has already been widely recognized as powerful enough in describing one's personality, we do have reason to believe that using the basic personality traits could fully describe an individual's personality, or at least certain aspects of one's personality. This exactly means that the description degrees of the basic personality traits could satisfy the constraint that$\sum_y d_{\boldsymbol{x}}^{y} = 1$, provided a normalization of personality scores is carried out. Therefore, it is possible for us to transform the PR problem into a label distribution learning task, so as to achieve the goal of predicting the personality traits as continuous scores and improve the performance of PR. The process of utilizing LDL algorithms to automatically recognize an individual's personality traits can be divided into three steps: (1) Generating label distribution, (2) Constructing label distribution prediction model, (3) Calculating personality scores.

*1) Generating Label Distribution:* Let $X = \mathbb{R}^m$represents the user space, $Y = \{y_1, y_2, ..., y_j, ...y_c\}$ represents the set of personality traits. For an individual $\boldsymbol{x}_i$ and a certain basic trait $y_j$, the description degree $d_{\boldsymbol{x}_i}^{y_j}$ that describes the relative intensity of trait $y_j$ to instance $\boldsymbol{x}_i$ could be obtained from Eq. (2),

$$d_{\boldsymbol{x}_i}^{y_j} = \frac{s_{\boldsymbol{x}_i}^{y_j}}{\sum_j s_{\boldsymbol{x}_i}^{y_j}} \tag{2}$$

where $s_{\boldsymbol{x}_i}^{y_j}$ is the original personality score of $\boldsymbol{x}_i$ in trait $y_j$ tested by Big Five Inventory. The vector containing the obtained description degrees of the basic traits, i.e. $\vec{D}_i = \left[d_{\boldsymbol{x}_i}^{y_1}, d_{\boldsymbol{x}_i}^{y_2}, ...d_{\boldsymbol{x}_i}^{y_j}, ..., d_{\boldsymbol{x}_i}^{y_c}\right]$, would exactly be the label distribution associated with the instance $\boldsymbol{x}_i$. Then, the original data set could be transformed into a new format, which could be denoted by $T = \{(\boldsymbol{x}_1, \vec{D}_1), (\boldsymbol{x}_2, \vec{D}_2), ..., (\boldsymbol{x}_n, \vec{D}_n)\}$.

*2) Constructing Label Distribution Prediction Model:* Since the PR problem has been transformed into a LDL task, we further construct the prediction model of the description degrees corresponding to Big Five personality traits with LDL algorithms. Specifically, given the processed training set $T$, we adopt all eight existing LDL algorithms to learn a mapping from $\boldsymbol{x} \in \mathbb{R}^m$ to $\vec{D} = \mathbb{R}^c$. The adopted algorithms

include PT-Bayes, PT-SVM, AA-$k$NN, AA-BP, LD-SVRSA-IIS, SA-BFGS and CPNN. A brief introduction of the adopted algorithms is as follows, and details can be found in [17, 58, 61].

**PT-Bayes** and **PT-SVM** are respectively developed from the existing popular Bayes classifier and SVM by transforming an LDL problem into an SLL (Single Label Learning) problem. Specifically, each training example $\left(\boldsymbol{x_i}, \vec{D}_i\right)$ is transformed into $c$ single-label examples $(\boldsymbol{x_i}, y_j)$ with the weight $d_{\boldsymbol{x}_i}^{y_j}$, where $i = 1, ..., n$ and $j = 1, ..., c$. The training set is then resampled to a standard single-label training set including $c \times n$ examples, and the SLL algorithms (i.e. Bayes classifier and SVM) can be applied to the training set. As for the output, the confidence/probability calculated by the learner for each label $y_j$ is regarded as the description degree of the corresponding label, i.e., $d_{\boldsymbol{x}}^{y_j} = P\left(y_j|\boldsymbol{x}\right)$ [17].

**AA-$k$NN**, **AA-BP** and **LD-SVR** are obtained by extending the existing popular learning algorithms $k$-NN, Back Propagation (BP) neural network and standard SVR, respectively. For AA-$k$NN, given a new instance $\boldsymbol{x}$, the $k$ nearest neighbors of $\boldsymbol{x}$ are firstly found in the training set. Then, the mean of the label distributions of all the $k$ nearest neighbors is calculated as the label distribution of $\boldsymbol{x}$. As to AA-BP, it is a neural network that has $m$ (the dimensionality of $\boldsymbol{x}$) input units which receive $\boldsymbol{x}$, and $c$ (the number of different labels) output units each of which outputs the description degree of a label $y_j$. The target of the algorithm is to minimize the sum-squared error of the output of the neural network compared with the real label distributions. Besides, the softmax activation function is used in each output unit to make sure the output of the neural network $\boldsymbol{z} = [z_1, z_2, ..., z_c]$ satisfies that $z_j \in [0, 1]$ for $j = 1, 2, ..., c$ and $\sum_j z_j = 1$ [17]. With regard to LD-SVR, its basic idea is to fit a sigmoid function to each component of the label distribution simultaneously by a multi-output support vector machine, so that it could solve the multivariate output problem and probability output problem simultaneously, rather than one by one [58]. Please refer to [58] for detailed analysis and derivation.

**SA-IIS**, **SA-BFGS** and **CPNN** are specially designed for the LDL problem. Since label distribution shares the same constraints with probability distribution (i.e. $d_{\boldsymbol{x}}^{y} \in [0, 1]$ and $\sum_y d_{\boldsymbol{x}}^{y} = 1$), many statistical theories and methods can be applied to label distribution. Specifically, $d_{\boldsymbol{x}}^{y}$ can be represented by the form of conditional probability, i.e. $d_{\boldsymbol{x}}^{y} = P(y|\boldsymbol{x})$. Suppose $P(y|\boldsymbol{x})$ is a parametric model $P(y|\boldsymbol{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the parameter vector. Given the training set $T$, the goal of LDL is to find the $\boldsymbol{\theta}$ that can generate a distribution similar to $\vec{D}_i$ given the instance $\boldsymbol{x}_i$. In order to directly solve this optimization problem, SA-IIS and SA-BFGS both assume the parametric model $P(y|\boldsymbol{x}; \boldsymbol{\theta})$ to be the maximum entropy model, and SA-IIS uses a strategy similar to Improved Iterative Scaling (IIS) to maximize the likelihood of the maximum entropy model, while SA-BFGS takes a quasi-Newton method BFGS as its optimization algorithm[17]. As to CPNN (conditional probability neural network), it models $P(y|\boldsymbol{x})$ by a three layer neural network, whose input includes both $\boldsymbol{x}$ and $y$, and output is a single value that is expected to be $P(y|\boldsymbol{x})$ [61].

*3) Calculating Personality Scores:* Utilizing the model trained by LDL algorithms, we can predict the label distribution of the unseen users' personality traits. Then the target personality scores of a certain instance $\boldsymbol{x}_i$ could be obtained through Eq. (3) by fetching the processed description degrees from the predicted label distribution, where $\left(s_{\boldsymbol{x}_i}^{y_j}\right)^*$ is the ultimate predicted scores of $\boldsymbol{x}_i$ in trait $y_j$, $\left(d_{\boldsymbol{x}_i}^{y_j}\right)^*$ is the predicted description degree of trait $y_j$ to instance $\boldsymbol{x}_i$.

$$\left(s_{\boldsymbol{x}_i}^{y_j}\right)^* = \left(d_{\boldsymbol{x}_i}^{y_j}\right)^* \cdot AproxSum \tag{3}$$

*AproxSum* is the approximate sum of all the traits involved in LDL algorithm for each instance $\boldsymbol{x}_i$, which could be computed as the median, mode or mean value of each training instances' sum of personality scores. Eq. (4) indicates the computing method of *AproxSum*, where $f$ refers to the median, mode or mean value function, $\sum_j s_{\boldsymbol{x}_k}^{y_j}$ denoted the sum of instance $\boldsymbol{x}_k$'s observed personality scores, and $t$ indicates the total number of instances in the training set.

$$AproxSum = f\left(\sum_j s_{\boldsymbol{x}_1}^{y_j}, ..., \sum_j s_{\boldsymbol{x}_k}^{y_j}, ..., \sum_j s_{\boldsymbol{x}_t}^{y_j}\right) \tag{4}$$

## IV. EXPERIMENTS

We conducted experimental evaluations in both predictive ability and running efficiency of the PR approaches based on Sina Weibo dataset. By using the whole 113-dimension feature space, all eight LDL algorithms were employed to build personality recognition models with the aid of a toolkit named *LDLPackage_v1.2*[6]. Note that, to identify the impact of parameter *AproxSum*, each LDL algorithm was applied three times with varied *AproxSum*. Further considering that the personality scores of users follow the normal distribution, thus, we only test the arithmetic mean, median and mode value of each training instances' sum of personality scores.

Besides, nine non-trivial conventional algorithms, which have been widely adopted in previous PR research, were applied in our evaluation experiments as baselines based on Weka, a popular machine-learning workbench. Specifically, they are M'5 Rules, ZeroR, Random Forest, Random Tree, Gaussian Processes, Linear Regression, Simple Linear Regression, Support Vector Regression (SVR), and Multi-layer Perceptron (MLP) neural network. All the experiments were carried out with 10-fold cross validation, and each time a single fold was used for testing while the other 9 folds were used for training.

### A. Evaluation of Predictive Ability

In our experiments, the predictive ability of the personality recognition approaches was evaluated by MAE (Mean Absolute Error), a frequently used measure of differences between the predicted score and the observed score tested by Big Five Inventory in PR research. It can be calculated using Eq. (5), where $n$ denotes the number of unseen instances, $\left(s_{\boldsymbol{x}_i}^{y_j}\right)^*$ the

[6]http://ldl.herokuapp.com

predicted value for trait $y_j$, and $\left(s_{\boldsymbol{x}_i}^{y_j}\right)$ the observed one. Since MAE is a measure of error, thus, the lower, the better. We carried out the model training and testing experiments and calculated the MAE of each model. The average results over a 10-fold cross validation achieved by different algorithms are presented in Table III.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\left(s_{\boldsymbol{x}_i}^{y_j}\right)^* - \left(s_{\boldsymbol{x}_i}^{y_j}\right)\right| \tag{5}$$

As we can see from Table III, the LDL algorithm named LD-SVR performs the best among all the algorithms, since all three PR approaches that adopted LD-SVR algorithm achieve lower prediction errors than the others, including the baselines and other LDL approaches. Among the top three PR approaches, the LD-SVR approach which adopted the median value of each training instances' sum of personality scores as *AproxSum* (i.e., the LD-SVR.median approach indicated in Table III) achieves the lowest prediction error in all the Big Five traits except extraversion (E), and takes the first place with an average MAE of 4.26. The other LD-SVR approach which adopted mean value as *AproxSum* (i.e., LD-SVR.mean) comes off second best with a minor disadvantage in average MAE compared with LDL-SVR.median. The standard Support Vector Regression algorithm performs much worse than LD-SVR, which proves the advantages of the LD-SVR algorithm specially designed for the label distribution learning task.

Another LDL algorithm named SA-IIS ranks second only to LD-SVR, and outperforms the best baseline in all dimensions of personality traits except the openness to experience (O) one. All three SA-IIS approaches achieve relatively good prediction accuracy, whose average MAEs are no larger than 4.388, and the best performance is achieved by the one whose *AproxSum* was set to the median.

Ranking after the LD-SVR and SA-IIS approaches with an average MAE of 4.438, Random Forest becomes the most accurate prediction model among all the baselines. However, the LDL approaches named CPNN, including CPNN.mean, CPNN.median and CPNN.mode, still outperform the Random Forest approach in the personality dimension of conscientiousness (C) with average MAEs no larger than 4.606, while the LDL approach named SA-BFGS.median achieves lower prediction errors than Random Forest in the extraversion (E) dimension. Even so, the performances of CPNN and SA-BFGS seem to be not that satisfactory compared with LD-SVR. The reason might be two-fold. For one thing, while CPNN and

TABLE III
Average mean absolute error (MAE) results for Big Five personality traits prediction using various LDL and baseline approaches on Sina Weibo dataset

| Approach Category | Approach Name | O | C | E | A | N | Average |
|---|---|---|---|---|---|---|---|
| LDL | AA-BP.mean | 5.394 | 4.949 | 4.807 | 4.505 | 4.741 | 4.879 |
| | AA-BP.median | 5.187 | 5.007 | 4.733 | 4.599 | 4.698 | 4.845 |
| | AA-BP.mode | 5.111 | 5.024 | 4.705 | 4.739 | 4.698 | 4.855 |
| | AA-$k$NN.mean | 4.991 | 4.768 | 4.467 | 4.178 | 4.700 | 4.621 |
| | AA-$k$NN.median | 5.015 | 4.781 | 4.457 | 4.213 | 4.620 | 4.617 |
| | AA-$k$NN.mode | 5.045 | 4.775 | 4.526 | 4.221 | 4.676 | 4.649 |
| | SA-BFGS.mean | 4.930 | 4.840 | 4.337 | 4.218 | 4.611 | 4.587 |
| | SA-BFGS.median | 4.868 | 4.833 | 4.285 * | 4.252 | 4.727 | 4.593 |
| | SA-BFGS.mode | 5.027 | 4.868 | 4.349 | 4.256 | 4.756 | 4.651 |
| | CPNN.mean | 7.057 | 4.542 * | 5.176 | 4.208 | 4.934 | 5.184 |
| | CPNN.median | 7.218 | 4.576 * | 5.253 | 4.173 | 5.046 | 5.253 |
| | CPNN.mode | 6.797 | 4.606 * | 5.159 | 4.104 | 4.968 | 5.127 |
| | SA-IIS.mean | 4.850 | 4.529 * | 4.226 * | 3.997 * | 4.247 * | 4.370 * |
| | SA-IIS.median | 4.856 | 4.521 * | 4.224 * | 3.985 * | 4.247 * | 4.367 * |
| | SA-IIS.mode | 4.851 | 4.576 * | 4.230 * | 4.011 * | 4.273 * | 4.388 * |
| | LD-SVR.mean | 4.691 * | 4.482 * | **4.086** * | 3.935 * | 4.132 * | 4.265 * |
| | LD-SVR.median | **4.690** * | **4.480** * | 4.092 * | **3.930** * | **4.116** * | **4.262** * |
| | LD-SVR.mode | 4.714 * | 4.512 * | 4.094 * | 3.946 * | 4.198 * | 4.293 * |
| | PT-Bayes.mean | 44.639 | 35.437 | 44.418 | 41.993 | 39.344 | 41.166 |
| | PT-Bayes.median | 44.818 | 37.495 | 40.609 | 42.605 | 34.818 | 40.069 |
| | PT-Bayes.mode | 41.151 | 33.602 | 47.241 | 41.211 | 42.219 | 41.085 |
| | PT-SVM.mean | 7.322 | 6.013 | 7.452 | 6.728 | 5.752 | 6.653 |
| | PT-SVM.median | 8.576 | 6.193 | 6.429 | 5.295 | 5.528 | 6.404 |
| | PT-SVM.mode | 7.616 | 5.989 | 6.773 | 5.720 | 5.031 | 6.226 |
| Baseline | Gaussian Processes | 4.887 | 4.684 | 4.308 | 4.103 | 4.323 | 4.461 |
| | Linear Regression | 5.181 | 4.999 | 4.536 | 4.368 | 4.617 | 4.740 |
| | M'5 Rules | 5.256 | 5.030 | 4.614 | 4.659 | 4.749 | 4.861 |
| | Random Forest | 4.891 | 4.608 # | 4.293 | 4.059 # | 4.342 | 4.438 # |
| | Rand Tree | 6.700 | 6.591 | 6.136 | 5.377 | 6.024 | 6.166 |
| | Simple Linear Regression | 4.936 | 4.627 | 4.293 # | 4.124 | 4.320 | 4.460 |
| | ZeroR | 4.955 | 4.689 | 4.362 | 4.088 | 4.367 | 4.492 |
| | Support Vector Regression | 4.840 # | 4.654 | 4.334 | 4.102 | 4.318 # | 4.450 |
| | Multi-layer Perceptron | 5.537 | 5.061 | 4.735 | 4.552 | 4.727 | 4.922 |

Note: O, C, E, A, N refer to the five dimensions of personality: openness to experience, conscientiousness, extraversion, agreeableness, neuroticism, respectively. In each column, the lowest MAEs among all the approaches, including LDLs and the baselines, are typeset in bold, the lowest MAEs among the baselines are denoted by a # sign, and all the MAEs of LDL approaches which outperform the baselines are denoted by a * sign.

SA-BFGS seek to directly minimize the distance between the predicted and ground truth distributions, LD-SVR takes advantage of the large margin regression by a support vector machine. For another, application of the kernel trick renders it possible for LD-SVR to solve the problem in a higher-dimensional thus more discriminative feature space without loss of computational feasibility.

The predictive performances of AA-BP and AA-$k$NN approaches are fair, while the PT-SVM and PT-Bayes approaches unfortunately underperform the baselines with larger MAEs. In fact, these four LDL algorithms' underperformance could be explained by their straightforward design strategies. Specifically speaking, PT-SVM and PT-Bayes are exactly the SVM and Bayes algorithms which are applied to solve the LDL problem by simply transforming it into a single label learning problem (i.e., change the training examples into weighted single-label examples), while the AA-$k$NN and AA-BP are designed by simply extending existing $k$NN and backpropagation (BP) neural network without too many essential improvements. In other words, these four LDL algorithms, by nature, take little correlations between the labels of the training instance. From this aspect, their poor performance is understandable.

As for the impact of parameter *AproxSum*, if we disregard the performance differences among different LDL approaches, and only focus on the MAEs of the approaches with same LDL algorithm but varied *AproxSum*s, we may find that the approaches that adopted median value as *AproxSum* seems to be a little more effective when calculating the ultimate personality scores from the predicted label distribution, according to the results presented in Table III. However, given that the users' personality scores follow the normal distributions as we mentioned above, the differences among the arithmetic mean, mode and median values of each training instance's sum of personality scores are not significant, resulting in similar MAEs among the approaches which adopted same LDL algorithm but varied *AproxSum*. Therefore, we may conclude that, what value (i.e., mean, median or mode) to be set as the *AproxSum*, in fact, has little effect on the prediction accuracy of LDL approaches.

### B. Evaluation of Efficiency

Besides effectiveness, the running efficiency of PR approach is also significant since high-efficiency ones could be more adaptive for various application scenarios. To evaluate the efficiency of the proposed approach, we record the time taken to build personality recognition models by different LDL algorithms as well as the baselines. Runtimes are averaged over a 10-fold cross validation, and details are presented in Table IV .

It is noteworthy that LDL algorithms are able to predict all the Big Five traits of a given user at once, while the baselines need to build five independent prediction models, one for each trait. Therefore, the average runtime of the baselines presented in Table IV is the sum of the average time taken to build all the five prediction models.

As we can see from Table IV, the approaches which adopt the same LDL algorithm but different *AproxSum* achieve similar running efficiency. This is easily accountable since their computation complexity of building the label distribution models are the same, and the running time of computing the mean, median or mode value actually won't make a big difference. The LD-SVR algorithm, which outperforms all the other algorithms in predictive ability, achieves relatively higher running efficiency. It is much more efficient than the SA-IIS algorithm, which ranks second in predictive accuracy. ZeroR, a baseline algorithm that performs not so good in MAE, achieves the shortest overall runtime, i.e. 0.02 seconds, while the most predictive baseline algorithm, Random Forest, needs 5 seconds to build five prediction models for all the five personality traits.

TABLE IV
Average time (measured in seconds) taken to build personality recognition models by LDL and baseline approaches on Sina Weibo dataset

| Approach Category | Approach Name | Runtimes |
|---|---|---|
| LDL | AA-BP.mean | 23.49 |
| | AA-BP.median | 27.03 |
| | AA-BP.mode | 24.11 |
| | AA-$k$NN.mean | 0.16 |
| | AA-$k$NN.median | 0.20 |
| | AA-$k$NN.mode | 0.16 |
| | SA-BFGS.mean | 28.42 |
| | SA-BFGS.median | 23.51 |
| | SA-BFGS.mode | 28.37 |
| | CPNN.mean | 13.15 |
| | CPNN.median | 15.22 |
| | CPNN.mode | 13.66 |
| | SA-IIS.mean | 17.69 * |
| | SA-IIS.median | 16.46 * |
| | SA-IIS.mode | 17.30* |
| | LD-SVR.mean | 0.38 * |
| | LD-SVR.median | 0.28 * |
| | LD-SVR.mode | 0.31* |
| | PT-Bayes.mean | 0.05 |
| | PT-Bayes.median | 0.06 |
| | PT-Bayes.mode | 0.05 |
| | PT-SVM.mean | 49.36 |
| | PT-SVM.median | 50.39 |
| | PT-SVM.mode | 51.40 |
| Baseline | Gaussian Processes | 9.08 |
| | Linear Regression | 1.62 |
| | M'5 Rules | 5.90 |
| | Random Forest | 5.00 # |
| | Random Tree | 0.14 |
| | SimpleLinearRegression | 0.08 |
| | ZeroR | **0.02** |
| | Support Vector Regression | 0.33 |
| | Multi-layer Perceptron | 21.37 |

Note: The runtimes of LDL approaches which outperform the baselines in average MAE are denoted by a * sign. The runtimes of the most predictive baselines are denoted by a # sign. The shortest runtime is highlighted in bold.

## V. Conclusions and Future Work

Personality recognition on social media is an emerging research field that consists of the automatic inference of users' personality traits from publicly available information on online social platform. In this paper, we introduce a new machine learning paradigm named label distribution learning into this field, and propose a method of Big Five personality recognition from microblog in Chinese language environments with label

distribution learning. One hundred and thirteen content features are extracted from 994 active Sina Weibo users' profiles, and eight LDL algorithms are adopted to train the Big Five personality prediction models. Experimental results show that the LDL algorithms, especially the ones named LD-SVR and SA-IIS, perform better than the traditional regression methods and the most predictive LD-SVR also achieves higher running efficiency. In the future, we plan to explore optimal feature space by applying deep learning techniques on larger data sets so as to further improve the prediction accuracy of the LDL approaches in personality recognition.

## APPENDIX

In the appendix, all the features adopted by our PR research are listed. Features No.1 to No.6 are those profile-based static features, and No.7 to No.11 are profile-based dynamic ones. The rests are the content-based micro-blogs features extracted by Textmind.

TABLE A1 Adopted Features

| Category | Feature Description |
|---|---|
| Profile-based | 1. Length of Nickname |
| | 2. Gender |
| | 3. Province |
| | 4. City |
| | 5. Language: traditional/simplified Chinese |
| | 6. Length of SelfDescription |
| | 7. NUM Friends |
| | 8. NUM Followers |
| | 9. NUM Followings |
| | 10. NUM Statuses |
| | 11. NUM Favourites |
| Content-based | 12. NUM Function Wd |
| | 13. NUM Pronoun |
| | 14. NUM PersonalPronoun |
| | 15. NUM NonPersonalPronoun |
| | 16. NUM "I" |
| | 17. NUM "We" |
| | 18. NUM "You" |
| | 19. NUM "She" or "He" |
| | 20. NUM "They" |
| | 21. NUM Verb |
| | 22. NUM Article |
| | 23. NUM Specific Article |
| | 24. NUM Adverb |
| | 25. NUM Preposition |
| | 26. NUM Conjunction |
| | 27. NUM Interjunction |
| | 28. NUM Auxilary |
| | 29. NUM Numeral |
| | 30. NUM QuantityUnit |
| | 31. NUM Digit |
| | 32. NUM Post Position |
| | 33. NUM 2ndPersPlural |

TABLE A1 (continued)

| Category | Feature Description |
|---|---|
| Content-based | 34. NUM Sentimental Wd |
| | 35. NUM ProgrammeRelated Wd |
| | 36. NUM Multifunction Wd |
| | 37. NUM PastTense |
| | 38. NUM PresentTense |
| | 39. NUM FutureTense |
| | 40. NUM TenseMark |
| | 41. NUM PastTenseMark |
| | 42. NUM PresentMark |
| | 43. NUM FutureTenseMark |
| | 44. NUM Negative Wd |
| | 45. NUM Swear |
| | 46. NUM Social Wd |
| | 47. NUM FamilyRelated Wd |
| | 48. NUM FriendRelated Wd |
| | 49. NUM PositiveSentimental Wd |
| | 50. NUM NegativeSentimental Wd |
| | 51. NUM HumansRelated Wd |
| | 52. NUM InhibitionRelated Wd |
| | 53. NUM AnxietyRelated Wd |
| | 54. NUM AngerRelated Wd |
| | 55. NUM BodyRelated Wd |
| | 56. NUM InsightRelated Wd |
| | 57. NUM CausalityRelated Wd |
| | 58. NUM DiscrepanceRelated Wd |
| | 59. NUM CognitiveProcess Wd |
| | 60. NUM Certain Wd |
| | 61. NUM Sad Wd |
| | 62. NUM Inclusive Wd |
| | 63. NUM Exclusive Wd |
| | 64. NUM PerceptionRelated Wd |
| | 65. NUM Visual Wd |
| | 66. NUM Auditory Wd |
| | 67. NUM Sensory Wd |
| | 68. NUM PhysiologyRelated Wd |
| | 69. NUM HealthyRelated Wd |
| | 70. NUM Tentative Wd |
| | 71. NUM SexRelated Wd |
| | 72. NUM IngestionRelated Wd |
| | 73. NUM Rlative Wd |
| | 74. NUM MotionRelated Wd |
| | 75. NUM SpaceRelated Wd |
| | 76. NUM TimeRelated Wd |
| | 77. NUM WorkRelated Wd |
| | 78. NUM HomeRelated Wd |
| | 79. NUM LeisureRelated Wd |
| | 80. NUM Achievement Wd |
| | 81. NUM MoneyRelated Wd |
| | 82. NUM ReligionRelated Wd |
| | 83. NUM DeathRelated Wd |
| | 84. NUM Assent Wd |
| | 85. NUM Pause |

TABLE A1 (continued)

| Category | Feature Description |
|---|---|
| Content-based | 86. NUM Psychology Wd |
| | 87. NUM Filler |
| | 88. NUM Loving Wd |
| | 89. NUM PastRelated Wd |
| | 90. NUM NowRelated Wd |
| | 91. NUM FutureRelated Wd |
| | 92. NUM Period |
| | 93. NUM Comma |
| | 94. NUM Colon |
| | 95. NUM Semicolon |
| | 96. NUM QuestionMark |
| | 97. NUM ExclamationMark |
| | 98. NUM Dash |
| | 99. NUM Quote |
| | 100. NUM Apostrophe |
| | 101. NUM Parenthesis |
| | 102. NUM Other Punctuations |
| | 103. Total NUM Wd |
| | 104. NUM Words Per Sentence |
| | 105. RATE Dictionary Coverage |
| | 106. RATE Numeral |
| | 107. RATE Six-Character Wd |
| | 108. RATE Four-Character Wd |
| | 109. RATE Latin Wd |
| | 110. NUM @ Mention |
| | 111. NUM Emotion |
| | 112. NUM HashTag |
| | 113. NUM URL |

Note: "NUM", "RATE and "Wd" are abbreviations for "Num of", "Rate of" and "Word", respectively.

## ACKNOWLEDGMENT

This work was funded by the scientific research funds of PLA under Grant AWS13J003. We would like to thank the Computational Cyber-Psychology Lab (CCPL), Institute of Psychology, Chinese Academy of Sciences for generously supporting our research.

## REFERENCES

[1] A. Vinciarelli and G. Mohammadi, A survey of personality computing, *IEEE Transactions on Affective Computing*, vol. 5, pp. 273-291, Jul-Sep 2014.

[2] D. C. Funder, Personality, *Annual Review of Psychology*, vol. 52, pp. 197-221, 2001.

[3] G. W. Allport, *Personality: A psychological interpretation*, New York: Henry Holt, 1937.

[4] L. Zhang, X. L. Huang, T. L. Liu, A. Li, Z. X. Chen, and T. S. Zhu, Using linguistic features to estimate suicide probability of Chinese microblog users, in *Human Centered Computing*, ed: Springer, 2014, pp. 549-559.

[5] F. Celli and L. Rossi, The role of emotional stability in Twitter conversations, in *Proceedings of the Workshop on Semantic Analysis in Social Media 2012*, Avignon, France, 2012, pp. 10-17.

[6] A. Roshchina, J. Cardiff, and P. Rosso, A comparative evaluation of personality estimation algorithms for the twin recommender system, in *Proceedings of the 3th International Workshop on Search and Mining User-Generated Contents*, Glasgow, UK, 2011, pp. 11-17.

[7] F. Enos, S. Benus, R. L. Cautin, M. Graciarena, J. Hirschberg, and E. Shriberg, Personality factors in human deception detection: comparing human to machine performance, in *Proceedings of the INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing*, Pittsburgh, Pennsylvania, 2006.

[8] K. Luyckx and W. Daelemans, Personae: a corpus for author and personality prediction from text, in *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco, 2008, pp. 2981-2987.

[9] J. Golbeck and D. Hansen, Computing political preference among Twitter followers, *Social Networks*, vol. 36, pp. 1105-1108, 2011.

[10] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski, Workshop on computational personality recognition: shared task, in *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, Boston, Massachusetts, USA, 2013.

[11] D. Nie, Z. D. Guan, B. B. Hao, S. T. Bai, and T. S. Zhu, Predicting personality on social media with semi-supervised learning, in *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies -Volume 02*, Warsaw, Poland, 2014, pp. 158-165.

[12] A. G. Wright, Current directions in personality science and the potential for advances through computing, *IEEE Transactions on Affective Computing*, vol. 5, pp. 292-296, 2014.

[13] L. Rainie and B. Wellman, *Networked: the new social operating system*, MIT Press, 2014.

[14] S. D. Gosling, A. A. Augustine, S. Vazire, N. Holtzman, and S. Gaddis, Manifestations of personality in online social networks: self-reported FacebookRelated behaviors and observable profile information, *Cyberpsychology Behavior and Social Networking*, vol. 14, pp. 483-488, 2011.

[15] T. Buchanan and J. L. Smith, Using the Internet for psychological research: Personality testing on the World Wide Web, *British Journal of Psychology*, vol. 90, pp. 125144, 1999.

[16] C. Sumner, A. Byers, and M. Shearing, Determining personality traits & privacy concerns from Facebook activity, *Black Hat Briefings*, vol. 11, pp. 197-221, 2011.

[17] X. Geng, Label distribution learning, *IEEE Transactions on Knowledge and Data Engineering*, 2016.

[18] S. Millward, China's forgotten 3rd Twitter clone hits 260 million users, Technical report, *techinasia.com*, 2012 .

[19] P. T. Costa and R. R. McCrae, The revised NEO personality inventory (NEO-PI-R), *The SAGE handbook of personality theory and assessment*, vol. 2, pp. 179-198, 2008.

[20] F. Celli, B. Lepri, J. I. Biel, D. Gatica-Perez, G. Riccardi, and F. Pianesi, The workshop on computational personality recognition 2014, in *Proceedings of the 2014 ACM Conference on Multimedia*, Orlando, Florida, USA, 2014, pp. 1245-1246.

[21] A. Shlomo, K. Moshe, S. Dhawle, and J. W. Pennebaker, Lexical predictors of personality type, in *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*, St. Louis, Missouri, USA, 2005.

[22] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, Using linguistic cues for the automatic recognition of personality in conversation and text, *Journal of Artificial Intelligence Research*, vol. 30, pp. 457-500, 2010.

[23] J. Oberlander and S. Nowson, Whose thumb is it anyway? Classifying author personality from weblog text, in *Proceedings of the COLING/ACL on Main conference poster sessions*, Sydney, Australia, 2006, pp. 627-634.

[24] S. Nowson and J. Oberlander, Identifying more bloggers: towards large scale personality classification of personal weblogs, in *Proceedings of the 2007 International Conference on Weblogs and Social Media*, Boulder, Colorado, USA, 2007.

[25] J. Golbeck, C. Robles, and K. Turner, Predicting personality with social media, in *Proceedings of the CHI'11 Extended Abstracts on Human Factors in Computing Systems*, Vancouver, BC, Canada, 2011, pp. 253-262.

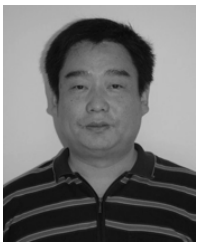
[26] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, Multimodal recognition of personality traits in social interactions, in *Proceedings of the 10th International Conference on Multimodal Interfaces*, Chania, Greece, 2008, pp. 53-60.

[27] L. M. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe, Please, tell me about yourself: automatic personality assessment using short self-presentations, in *Proceedings of the 13th international conference on multimodal interfaces*, Alicante, Spain, 2011, pp. 255-262.

[28] G. Mohammadi and A. Vinciarelli, Automatic personality perception: prediction of trait attribution based on prosodic features, *IEEE Transactions on Affective Computing*, vol. 3, pp. 273-284, 2012.
[29] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe, Connecting meeting behavior with extraversiona systematic study, *IEEE Transactions on Affective Computing*, vol. 3, pp. 443-455, 2012.
[30] O. Aran and D. Gatica-Perez, Cross-domain personality prediction: from video blogs to small group meetings, in *Proceedings of the 15th ACM International Conference on Multimodal Interaction*, Sydney, Australia, 2013, pp. 127-130.
[31] D. O. Olgun, P. A. Gloor, and A. S. Pentland, Capturing individual and group behavior with wearable sensors, in *Proceedings of the AAAI Spring Symposium on Human Behavior Modeling*, Stanford, California, USA, 2009.
[32] K. Kalimeri, B. Lepri, and F. Pianesi, Going beyond traits: multimodal classification of personality states in the wild, in *Proceedings of the 15th ACM on International conference on multimodal interaction*, Sydney, Australia, 2013, pp. 27-34.
[33] J. Staiano, B. Lepri, N. Aharony, F. Pianesi, N. Sebe, and A. Pentland, Friends don't lie: inferring personality traits from social network structure, in *Proceedings of the 14th International Conference on Ubiquitous Computing*, UbiComp 2012, Pittsburgh, PA, USA, 2012, pp. 321-330.
[34] G. Chittaranjan, J. Blom, and D. Gatica-Perez, Mining large-scale smartphone data for personality studies, *Personal and Ubiquitous Computing*, vol. 17, pp. 433-450, 2013.
[35] Y. A. de Montjoye, J. Quoidbach, F. Robic, and A. S. Pentland, Predicting personality using novel mobile phone-based metrics, in *Proceedings of the 6th international conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, Washington, DC,USA, 2013, pp. 48-55.
[36] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, et al., Personality, gender, and age in the language of social media: the open-vocabulary approach, *PloS ONE*, vol. 8, p. e73791, 2013.
[37] M. Polonsky, Online social networks and insights into marketing communications, *Journal of Internet Commerce*, vol. 6, pp. 55-72, 2007.
[38] P. A. Rosen and D. H. Kluemper, The Impact of the Big Five personality traits on the acceptance of social networking website, presented at the 14th Americas Conference on Information Systems,, Toronto, Ontario, Canada, 2008.
[39] J. Schrammel, C. Ffel, and M. Tscheligi, Personality traits, usage patterns and information disclosure in online communities, in *Proceedings of the 23rd Annual Conference on Human Computer Interaction*, HCI 2009, Cambridge, UK, 2009, pp. 169-174.
[40] M. Selfhout, W. Burk, S. Branje, J. Denissen, M. V. Aken, and W. Meeus, Emerging late adolescent friendship networks and Big Five personality traits: a social network approach, *Journal of Personality*, vol. 78, pp. 509-38, 2010.
[41] A. Li, Z. Yan, and T. S. Zhu, Self-report versus web-log: which one is better to predict personality of website users?, *International Journal of Cyber Behavior, Psychology and Learning*, vol. 3, pp. 44-54, 2013.
[42] M. Kosinski, D. Stillwell, and T. Graepel, Private traits and attributes are predictable from digital records of human behavior, in *Proceedings of the National Academy of Sciences*, vol. 110, pp. 5802-5805, 2013.
[43] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, et al., Computational personality recognition in social media, *User Modeling and User-Adapted Interaction*, vol. 26, pp. 109-142, Jun 2016.
[44] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, Predicting personality from Twitter, in *Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, Boston, Massachusetts, USA, 2011, pp. 149-156.
[45] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, Our Twitter profiles, our selves: predicting personality with twitter, in *Proceedings of the 2011 IEEE International Conference on Privacy, Security, Risk and Trust and 2011 IEEE International Conference on Social Computing*, Boston, MA, USA, 2011, pp. 180-185.
[46] G. Farnadi, S. Zoghbi, M. F. Moens, and M. De Cock, Recognising personality traits using Facebook status updates, in *Proceedings of the Workshop on Computational Personality Recognition at the 7th International AAAI Conference on Weblogs and Social Media*, Boston, Massachusetts USA, 2013, pp. 14-18.
[47] F. Alam, E. A. Stepanov, and G. Riccardi, Personality traits recognition on social network-Facebook, in *Proceedings of the 2013 International Conference on Weblogs and Social Media*, Cambridge, MA, USA, 2013, pp. 6-9.
[48] M. Skowron, B. Ferwerda, M. Tkali, and M. Schedl, Fusing social media cues: personality prediction from Twitter and Instagram, in *Proceedings of the 25th International Conference Companion on World Wide Web*, Montreal, Canada, 2016, pp. 107-108.
[49] S. T. Bai, T. S. Zhu, and L. Cheng, Big-five personality prediction based on user behaviors at social network sites, *Computer Science*, vol. 8, pp. e2682-e2682, 2012.
[50] S. T. Bai, B. B. Hao, A. Li, S. Yuan, R. Gao, and T. S. Zhu, Predicting big five personality traits of Microblog users, *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 501-508, 2013.
[51] D. Nie, L. Li, and T. S. Zhu, Conscientiousness measurement from Weibo's public information, in *IAPR International Workshop on Partially Supervised Learning*, Nanjing, China, 2013, pp. 58-67.
[52] L. Li, A. Li, B. B. Hao, Z. D. Guan, T. S. Zhu, and C. Liu, Predicting active users' personality based on micro-blogging behaviors, *PLoS ONE*, vol. 9, p. e84997, 2014.
[53] K. H. Peng, L. H. Liou, C. S. Chang, and D. S. Lee, Predicting personality traits of Chinese users based on Facebook wall posts, in *Proceedings of the 24th Wireless and Optical Communication Conference*, Taipei, Taiwan, 2015, pp. 9-14.
[54] F. Rangel, P. Rosso, M. Potthast, B. Stein, and W. Daelemans, Overview of the 3rd author profiling task at PAN 2015, in *Proceedings of the Conference and Labs of the Evaluation forum*, Toulouse, France, 2015.
[55] Sina Tech. (20 December, 2016), Sina Weibo released financial reports of the third quarter of 2016, Available: http://tech.sina.com.cn/i/2016-11-22/doc-ifxxwrwh4878257.shtml
[56] B. Cao. Sina's Weibo outlook buoys Internet stock gains: China overnight, Technical report, *Bloomberg*, 2012.
[57] R. Gao, B. Hao, H. Li, Y. Gao, and T. Zhu, Developing simplified Chinese psychological linguistic analysis dictionary for microblog, in *Proceedings of the International Conference on Brain and Health Informatics*, Maebashi, Japan, 2013, pp. 359-368.
[58] X. Geng and P. Hou, Pre-release prediction of crowd opinion on movies by label distribution learning, in *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, 2015, pp. 3511-3517.
[59] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, Deep Learning-Based Document Modeling for Personality Detection from Text, in *IEEE Intelligent Systems*, vol. 32, pp. 74-79, 2017.
[60] A. C. E. Lima and L. N. De Castro, A multi-label, semi-supervised classification approach applied to personality prediction in social media, in *Neural Networks*, vol. 58, pp. 122-130, 2014.
[61] X. Geng, C. Yin, Z H. Zhou, Facial age estimation by learning from label distributions, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 2401-2412, 2013.

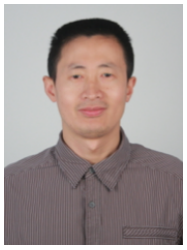
**Di Xue** received her B.E. and M.S. degrees from PLA University of Science and Technology in 2012 and 2015, respectively. Now she is a Ph.D. student of this university. Her research fields concern information security, social network analysis and data mining.

**Zheng Hong** received his Ph.D. from PLA University of Science and Technology in 2007. Now he is an associate professor in the university. His research fields concern network security and protocol reverse engineering.



**ShiZe Guo** received the Ph.D. degree from Harbin Institute of Technology in 1989 and the M.S., B.S. degrees from Ordnance Engineering College, China, in 1991 and 1988, respectively. He is currently a researcher in the Institute of North Electronic Equipment and a professor in the Beijing University of Post and Telecommunications. His main research interests include information technology and information security.

**Liang Gao** received the B.S., M.S. and Ph.D. degrees from National University of Defense Technology, in 2004, 2007 and 2012, respectively. He is currently an engineer in the Institute of North Electronic Equipment. His main research interests include network science, machine learning, and data mining.



**Lifa Wu** received his Ph.D. from Nanjing University in 1998. He is currently a professor in PLA University of Science and Technology. His research fields concern network security and information security.

**Jinghua Zheng** received her M.S. degrees from Electronic Engineering Institute, in 2005. Now she is studying for the Ph.D. degree in this university. Her main research interests include machine learning and data mining.

**Nan Zhao** received the Sc.D. degree from Institute of Psychology, Chinese Academy of Sciences in 2014, and B.S. degree from East China Normal University in 2009. Now he is an assistant professor in Institute of Psychology, Chinese Academy of Sciences. His main research interests include Computational cyber psychology, human-computer interaction, driving behavior, machine learning and data mining.