Predicting Personality On Social Media with Semi-supervised Learning

Dong Nie, Zengda Guan, Bibo Hao, Shuotian Bai, Tingshao Zhu*
{Institute of Psychology, University of Chinese Academy of Sciences}, Chinese Academy of Sciences
Beijing China
Email: tszhu@psych.ac.cn

***Abstract*—Personality research on social media is a hot topic recently due to the rapid development of social media as well as the central importance of personality study in psychology, but most studies are conducted on inadequate label samples. Our research aims to explore the usage of unlabeled samples to improve the prediction accuracy. By conducting n user study with 1792 users, we adopt local linear semi-supervised regression algorithm to predict the personality traits of Microblog users. Given a set of Microblog users' public information (e.g., number of followers) and a few labeled users, the task is to predict personality of other unlabeled users. The local linear semi-supervised regression algorithm has been employed to establish prediction model in this paper, and the experimental results demonstrate the usage of unlabeled data can improve the accuracy of prediction.**

***Keywords*—local linear kernel regression, unlabeled data, personality prediction**

I. Introduction

Personality can be defined as a set of characteristics which make a person unique, and the study of personality is of central importance in psychology. Among personality related researches, Big-Five theory is the mostly used one, which proposes five basic traits to form human personality: extraversion (E), agreeableness (A) , conscientiousness (C), neuroticism (N) and openness (O) [6]. Conventional personality research usually uses self-report inventory [17], which is less efficient. Though great efforts have been made by many scientists, personality prediction is still plagued by a number of problems, such as data collection and large-scale promotion. The rapid development of internet makes it possible to analyze behaviors and personality traits on the Web, which attracts much attention of scientists from different disciplines [8], [4], [16]. With the wide spread of Microblog nowadays, Sina Microblog (Weibo) becomes one of the most popular internet services in mainland China. More than 300 million has registered Weibo service [12], what's more, quite a large part of users spend much time on Weibo platform. As a consequence, Weibo plays a big role on people' normal life [5]. Thus, microblog provides an ideal online platform for personality research and relative application.

Much research has been done to identify the relationship between microblog usage and personality [7], [14], [8], but predicting personality for social media users is still on early stages. One of the biggest constraints is predictive accuracy as a result of limited training data. Specifically speaking, it is always a problem to acquire sufficient labeled data due to not only time-consuming and expensive, but also privacy concerns. On the contrary, it is much easier to obtain unlabeled data from social media platforms. Therefore, to exploit unlabeled data is a way to improve the predictive performance.

In recent years, there has been a substantial amount of work exploring how to incorporate unlabeled data into supervised learning, and several semi-supervised learning approaches have been proposed [3], [23], [20], [2], [22]. Successful applications have been made in many areas, such as computer vision [1], [21], and information retrieval [13]. Semi-supervised learning has also been used in the context of microblog classification [19]. In some cases, semi-supervised learning algorithms can outperform standard supervised algorithms, even the labeled data is inadequate. In our research, we obtain Microblog user data from Sina Microblog platform and propose a local linear semi-supervised model to incorporate information from unlabeled data. The traditional supervised learning merely focus on the labeled data, our method can exploit the cheaper unlabeled data to improve the accuracy. To solve the high dimension feature space problem, we employ an efficient feature selection strategy in this paper. The main contributions of this paper are as follows: 1)This paper manage to predict Sina Microblog users' personality traits through analyzing cyber behavioral characteristics. 2)The semi-supervised learning regression algorithm is employed to predict the personality traits. Due to the relative more unlabeled data involved in the model, our presented methods successfully improve the predictive accuracy.

The rest of this paper is organized as follows. In Section II, we introduce some related work. Section III will describe the dataset in detail. In Section IV, we present the detailed proposed methods. Experiment results will be presented in Section V, followed by discussion about the results VI. At last, we will make a conclusion in Section VII.

II. Related Work

Personality analysis based on social media has received considerable attention recently [14], [7], [11], [8]. They mainly collected internet data and corresponding labeled data, and then applied supervised learning approaches, such as, classification or regression, to build the model.

Qiu et al. [14] explored how personality was manifested and perceived in microblogs. 142 participants's personality traits were measured by Big Five Inventory, and their tweets were also collected correspondingly. They analyzed the collected data using statistical methods. Then they found that personality traits are associated with linguistic cues in microblogs and can be accurately judged by unknown others, for example, agreeableness and neuroticism can be reliably judged by unknown others on the basis of microblog content.

978-1-4799-4143-8/14 $31.00 © 2014 IEEE
DOI 10.1109/WI-IAT.2014.93

Kosinski et al. [11] researched on psychological variables on Facebook. Based on a dataset of over 58,000 volunteers who provided their Facebook Likes, detailed demographic profiles and the results of several psychometric tests, they used logistic/linear regression to predict individual psycho-demographic profiles from Likes. The model correctly discriminated between homosexual and heterosexual men. For the personality trait Openness, prediction accuracy was close to the test-retest accuracy of a standard personality test.

Gosling et al. [8] conducted experiments towards the manifestations of personality in Facebook usage. They built a simple mapping from SNS online behaviors to personality, also, they examined the personality with self-reported Facebook usage and observable profile information. They provided the correlation factor between personality and online behaviors. Although their research verified the correlation of online characteristics and personality labels, they did not further establish prediction model of personality, or give the quantitative indicators of personality and online behaviors.

Goldbeck et al. [7] presented a machine learning approach by which a user's personality can be accurately predicted through the publicly available information on their Facebook profile. As personality has been shown to be relevant to many types of interactions, they successfully utilized users'personality predicted from information on social media to improve the users' experiences with interfaces and with one another.

Various approaches have been employed to build manifestations between social media online data and personality traits, nevertheless, personality prediction needs further explored as usual. Therefore, we collected 1792 Microblog users' public information using Sina Microblog API, and gathered their corresponding personality traits. Features were extracted from the dataset, and feature selection methods were conducted to compact the feature space. A personality prediction model based on local liner semi-supervised regression model was established, and experiments were conducted to verify our proposed methods.

III. DATASET

The dataset consists of labeled Weibo users and huge unlabeled ones, and we acquire 1792 copies of Sina Microblog users' information together with personality scores as label.

The personality traits are measured in continuous value. Since only a few users are labeled, and much more users are un-labeled, it is very suitable to use semi-supervised learning techniques.

A. Data Collection

Using Sina Microblog API [1] , we first collected 999,9999 Microblog user IDs, then randomly chosen 100,000 user IDs, and crawled down their Microblog data. Using Weibo"@ function", we invited volunteers to complete big-five inventory online, and acquired 1792 copies of qualified questionnaires. Hence, we had 1792 copies of personality labeled data. The whole process took over two months, and volunteers had got reimbursement in return. The collected Microblog-user dataset (1792 copies of labeled Microblog data) is examined in our experiment.

[1] http://open.weibo.com

B. Feature Extraction

As our collected Microblog dataset is relatively simple (contain little behavior data), the preprocessing is straight forward. There are two types of features, summarized features from raw data directly and statistical features. We have totally extracted 47 features for each user from the Microblog data, and the complete feature list is presented in Table IV in Appendix A. The extracted features can be divided into several categories as bellow:

1) user's personal profile, such as nickname, address, gender, tags, birthday, personalized domain url, description and so on.
2) user's social circles, sucha sa friends, followers and mutual followers.
3) users' social activities, such as status, retweeted status, annotations, pictures and comments
4) users' social habit, for example, time to post status

For some features, we just process the original data directly, for example, the number of statuses. We ran a program to analyze users' descriptions, to identify each description as positive, neuter or negative. To calculate the time of creating statuses, we divide a whole day into 7 periods: 0:00-6:00, 6:00-8:00, 8:00-11:00, 11:00-13:00, 13:00-17:00, 17:00-20:00, and 20:00-24:00 , and they correspond to period0, period1, period2, period3, period4, period5 and period6 respectively. We then count the number of statuses that the user creates in each period.

After feature extraction, we normalize the data to make the data equally distributed as follows.

$$x = (x - MinValue)/(MaxValue - MinValue) \quad (1)$$

where x is the value of a dimension for a user, while $MinValue$ and $MaxValue$ respectively represent the maximum and the minimum value of this feature dimension for all users.

The original feature space consists of 47 dimensions. To improve the predictive performance and decrease the computational complexity, we adopt feature selection methods. After comparing several feature selection approaches via testing, we carry out stepwise regression [9] instead (most of feature selection methods are for classifiers which perform badly on regression tasks, while stepwise regression do a good job on regression problems [10]).

We used stepwise regression to the 47-dimension feature space for each five personality traits, and the features selected are listed in Table I (Note, features are represented by Feature ID, the meaning of Feature ID can refer to Table IV):

C. Personality-trait Assessment

As we received 1792 copies of effective big-five personality questionnaires, we calculated scores for each personality trait

from the 44 items in the inventory. In total, each subject is evaluated in five personality dimension, and the score of each dimension is scaled to interval $[0, 5]$.

IV. METHODS

In our research, the personality traits are all measured by real-value scores between 0 and 5, thus the problem is the prediction of continuous value, in other words a model fitting process. Accordingly, we adopt local linear kernel regression as the basic approach. To take advantage of huge unlabeled data, we mainly build models based on a local linear semi-supervised regression approach [15], we also improve the algorithm to make it adjust to our personality prediction tasks.

The labeled dataset is denoted as $\{(X_1, Y_1), (X_2, Y_2), ... (X_n, Y_n)\}$, where X_i is a d-dimension feature vector, and Y_i is the true value for this point. Correspondingly, $\{X_{n+1}, ...X_m\}$ is the unlabeled dataset, and $m \gg n$.

A. Local Linear Kernel Regression Model

Local linear kernel regression is a broadly used nonparametric regression approach. The idea is to fit locally a straight line (or a hyperplane for higher dimensions), and not the constant (horizontal line). The local linear regression can be defined as a weighted least square problem

$$\min \sum_i \left(Y_i - m - (X_i - x)^T \beta\right)^2 K\left(\frac{X_i - x}{h}\right) \tag{2}$$

with respect to m and β, where $K(x)$ is the kernel function and h is bandwidth. And the notation T in the top right corner means transpose. Let $\widehat{m} = \widehat{m}(x)$ and $\beta = \widehat{\beta}(x)$ be solutions to the above minimization problem. Let $M(x) = \left(m(x), \beta(x)^T\right)^T$. Let

$$W_x = diag\left(K\left((X_1 - x)/h\right), ..., K\left((X_n - x)/h\right)\right)$$

and

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, X_x = \begin{pmatrix} 1 & X_1 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{pmatrix}$$

Then $\widehat{M}(x) = \left(\widehat{m}(x), \widehat{\beta}(x)^T\right)^T$ solves the following minimization problem:

$$\min (Y - X_x M)^T W_x (Y - X_x M) \tag{3}$$

with respect to M. That is

$$\widehat{M}(x) = \left(X_x^T W_x X_x\right)^{-1} X_x^T W_x Y \tag{4}$$

TABLE I. SUBSET FEATURES SELECTED BY STEPWISE REGRESSION MODEL FOR EACH PERSONALITY TRAIT

Personality	**Features Selected**
E	**V9, V11, V22, V23, V30**
A	**V10, V13, V16, V22, V30**
C	**V9, V10, V11, V13, V30**
N	**V1, V10, V13, V15, V16, V30, V34, V41**
O	**V2, V11, V14, V23, V38**

B. Local Linear Semi-supervised Regression Model

For most semi-supervised classification algorithms, the cluster assumption has been mentioned quite often: points on the same structure (typically referred to as a cluster) or nearby points are likely to have the same label. This assumption of clustering can not be completely transferred to regression due to continuous value, but a somewhat similar "smoothness" assumption can be reached: the value of the regression function is expected not to "jump" or change suddenly [15]. Specific to our task, we expect users who are similar to each other to have close personality values.

A very intuitive way to instantiate this assumption in semi-supervised regression is by finding estimates $\widehat{m}(x)$ that minimize the following objective function (subject to the constraint that $\widehat{\mathrm{m}}(X_i) = Y_i$ for the labeled data):

$$\sum_{i,j} w_{ij} \left(\widehat{m}(X_i) - \widehat{m}(X_j)\right)^2 \tag{5}$$

where $\widehat{m}(X_i)$ is the estimated value of the function at example X_i, w_{ij} is a measure of the similarity between examples X_i and X_j, Y_i is the value of the function at X_i (only defined on the labeled examples).

This is exactly the objective function minimized by the Gaussian Fields algorithm [23]. However, the Equation 5 is locally constant, which becomes a drawback when it comes to regression. Thus, we have to modify the above regularization term to locally linear one

$$\sum_{i,j} w_{ij} \left(\widehat{m}(X_i) - X_{ji}^T M(X_j)\right)^2 \tag{6}$$

We add the regularization term in Equation 6 to the supervised local linear kernel regression in Equation 2, and then the problem becomes

$$\begin{aligned} &\min \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{n} \left(Y_i - X_{ji}^T M(X_j)\right)^2 K\left(\frac{X_i - X_j}{h}\right) \\ &+ \frac{\mu}{2} \sum_{i,j} w_{ij} \left(\widehat{m}(X_i) - X_{ji}^T M(X_j)\right)^2 \end{aligned} \tag{7}$$

with respect to M, where M has the same meaning with Equation 3, $X_{ji}^T = \begin{pmatrix} 1 & X_i - X_j \end{pmatrix}$ and μ is a regularization constant. And the solution to Equation 7 is

$$\widehat{M}(x) = \left(X_x^T W_x X_x + \mu\Delta\right)^{-1} X_x^T W_x Y \tag{8}$$

where $\Delta = \mathrm{D} - \mathrm{W}$, D is a diag matrix with

$$\mathrm{D_i} = \frac{1}{2} \sum_j w_{ij} \left(e_1 e_1^T + X_{ij} X_{ij}^T\right) \tag{9}$$

and $W = [w_{ij}]$.

However, the solution given by Equation 8 is not applicable for all data. When the data is sparse enough, there can be no result. To improve the generalization of the solution, we add a

Tiknohov regularizer [18] term to Equation 7, and Equation 10 will help solve the problem.

$$\begin{aligned}\min \frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(Y_i - X_{ji}^T M\left(X_j\right)\right)^2 K\left(\frac{X_i - X_j}{h}\right) \\ +\frac{\mu}{2}\sum_{i,j} w_{ij}\left(\widehat{m}\left(X_i\right) - X_{ji}^T M\left(X_j\right)\right)^2 + \frac{\lambda}{2}\|M\|^2\end{aligned} \tag{10}$$

We take the derivative of the loss function in Equation 10 with respect to M, and then set the derivative to zero. The final solution can be represented by Equation 11

$$\widehat{M}\left(x\right) = \left(X_x^T W_x X_x + \mu\Delta + \lambda I\right)^{-1} X_x^T W_x Y \tag{11}$$

where I is an identity matrix.

V. Results

Our experiments aim to identify users' personality traits via their public information on Microblog platforms, two regression models stated in Section IV are adopted to solve the problems. There are totally 1792 users in our dataset, and each user is expressed by a feature vector as well as personality trait score vector.

Due to the local linear semi-supervised regression model, we have to test our model on different sizes of training set. Therefore, we randomly sample from the dataset according to the training set size, the original dataset is divided into two disjointed training and test sets at every training set size, the training set is regarded as the labeled data and the test sets is seen as unlabeled data. Mean Absolute Error (MAE) is adopted as evaluation criterion:

$$MAE = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left|\widehat{m}\left(X_i\right) - Y_i\right| \tag{12}$$

A. Regression Accuracy

We implement the local linear kernel regression algorithm and local linear semi-supervised regression algorithm in matlab respectively. We have also employed two baseline methods which are broadly used regression methods in the field of psychology. The baseline algorithms as follows:

1) MVR: Mean value regression is to predict the value of test set to be mean value of training set:

$$Y = \frac{1}{n}\sum_{j=1}^{n} Y_j \tag{13}$$

2) LR: Linear regression is a broadly used parametric model, it calculates model parameter β via the training set, and then the prediction value can be computed by the trained model:

$$Y_i = \beta_1 X_{i1} + ... + \beta_p X_{id} + \varepsilon_i = X_i^T\beta + \varepsilon_i, i = 1, ...n \tag{14}$$

As some regression approaches may predict the personality trait scores beyond $[0, 5]$, we can bound the values in the interval $[0, 5]$ using the following trick:

$$\widehat{Y}_i = \begin{cases} \overline{Y} + S & if\widehat{\mathrm{Y}}_i > 5 \\ \widehat{Y}_i & if 0 \leq \widehat{\mathrm{Y}}_i \leq 5 \\ \overline{Y} - S & if\widehat{\mathrm{Y}}_i < 0 \end{cases} \tag{15}$$

where $\overline{Y}$ is sample average and S^2 is sample variation.

Every testing is repeated 5 times. As there are five dimension of personality traits, we take the third dimension (C) for an example here (in fact, the other dimensions can also be chosen to present here), and the results for C are listed in Table II, the results for other personality traits are followed in Appendix A.

TABLE II. MAE ACHIEVED BY EACH ALGORITHM

Training Set Size	MVR	LR	LLKR	LLSSR
100	0.6124	0.8064	0.5097	**0.4950**
300	0.6123	0.7871	0.4870	**0.4686**
600	0.6130	0.7815	0.4879	**0.4592**
900	0.6118	0.7793	0.4807	**0.4478**
1200	0.6129	0.7757	0.4671	**0.4430**
1500	0.6118	0.7584	0.4752	**0.4292**
1700	0.6125	0.7668	0.4452	**0.4258**

"LLKR" refers to local linear kernel regression method, and "LLSSR" refers to local linear semi-supervised regression approach. The bandwidth selection for LLKR and LLSSR will be detailed later, and the regularization constant μ and λ for LLSSR is set by 0.1.

As shown in Table II, the LR method works worst over all the training sets, it indicates personality prediction cannot simply use a linear fitting model. MVR is the simplest model with smallest computational complexity, while it provides a better results than LR. In reality, personality follows a approximate normal distribution in the crowd, thus, the results for MVR is not so bad. LLKR and LLSSR both outperform the two baseline methods for a lot, and LLSSR achieves a nearly 2.5% to 7% reduction of MAE significantly.

In the field of psychology, there is another important assessment criterion to measure the model called Relative Absolute Error (RAE), RAE is defined by comparing the model's MAE with the MAE of mean value model:

$$RAE = \frac{MAE_{yours}}{MAE_{mvr}} \tag{16}$$

Here we depict the RAE for three methods over all the training sets in Table III and Figure 1

TABLE III. RAE% ACHIEVED BY EACH ALGORITHM

Training Set Size	LR	LLKR	LLSSR
100	131.68	83.23	**80.83**
300	126.91	79.54	**77.05**
600	127.49	79.59	**74.91**
900	127.38	78.57	**73.19**
1200	128.19	76.21	**72.28**
1500	123.96	77.67	**70.15**
1700	125.19	72.69	**69.52**

From Figure 1, it can be intuitive to see that LLKR and LLSSR both put up a good performance, and LLSSR outperforms LLKR a little. The data in Table III shows that RAE for

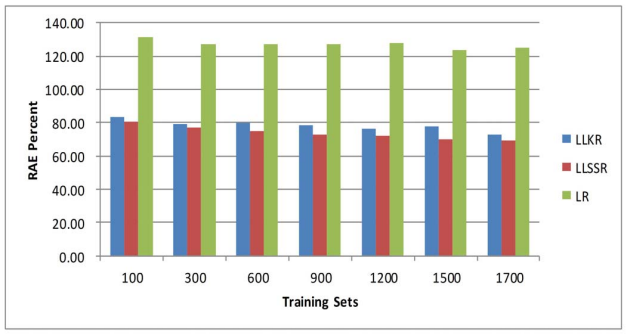

Fig. 1. RAE percent for the three methods (baseline is 100%)

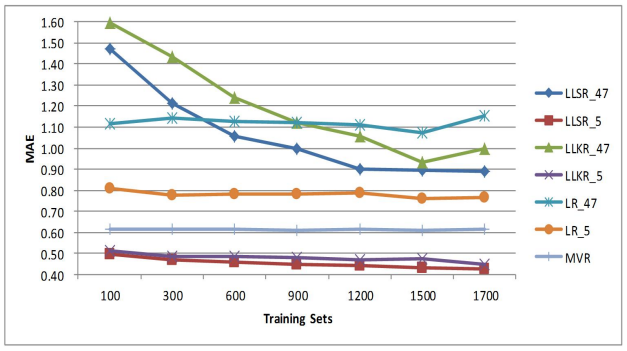

Fig. 2. MAE comparison of models trained with original feature space (47) and selected feature subset space (5)

LLKR are always higher than 75% until the training set size increase to 1700, while RAE for LLSSR decreases to 74.91% when the training set size is 600 (Correspondingly, test set size is 1192), and the lowest RAE for LLSSR reaches 69.52%. It is a great achievement in psychology when the RAE is less than 75%, thus, it is worthwhile to make use of unlabeled data with extra computation complexities.

Experimental results on other four personality traits also indicate that LLKR and LLSSR performs fairly well on personality prediction.

B. Parameter Selection

For local linear kernel regression (LLKR) approach, we employ 2 order Guassian kernel function in this paper, and select bandwidths using least square cross validation (LSCV) method:

$$CV_{-m}(h) = \sum_{i=1}^{n} \left(Y_i - \widehat{m}_{-i}(X_i)\right)^2 \tag{17}$$

Minimize the above equation and combine some experience, the approximate solution for the optimized bandwidth is

$$h_o = 1.06\widehat{\sigma}(X)\, n^{\frac{-1}{4+d}} \tag{18}$$

where $\widehat{\sigma}(X)$ is standard deviation for training set. n is the training set size and d is the dimension of feature space.

For local linear semi-supervised regression (LLSSR) method, we also use 2 order Guassain kernel function to measure the similarity. The selection of h refers to h_o computed by LLKR, combining regularization constant μ, we conduct grid search over $h \in \{1.001, 1.01, 1.1, 2\}\, h_o$, $\mu \in \{0.001, 0.01, 0.1, 1\}$ and $\lambda \in \{0.001, 0.01, 0.1, 1\}$.

C. Impact of Feature Selection

To test the performance of feature selection, we conduct experiments with the original 47-dimension feature sets and selected feature subsets. We take C dimension. After conducting stepwise regression model to select features, the new subset contains 5 features. All the four approaches are tested on both original feature space and selected subset space. The analysis results are shown in Figure 2.

From Figure 2, all methods run on the new selected subset feature space outperform on the original feature space. If there is no feature selection, MVR performs the best. After selecting feature, LLSSR and LLKR produce better predictive results than MVR. As to LR model, feature selection also improve the performance though it is still worse than MVR model. Feature selection is found to have a greater impact on LLKR and LLSSR, and it decreases more MAE on the two methods than on LR. The phenomenon may arise from the use of kernel function in LLKR and LLSSR models. The experiments are also conducted on the other personality traits, and the impact of feature selection are similar with C dimension.

VI. Discussion

A. Rationality about Cluster Assumption in This Paper

To use information from unlabeled data, we propose local linear semi-supervised regression approach to build the predictive models. This approach holds under the assumption that nearby points are likely to have the same value. Specifically, it presumes that people with similar Microblog behavior are likely to have nearby personality trait scores. This assumption surely holds over social behavior and personality, it is the basic assumption for personality research in psychology. Though behavior on social media is not the main part of social behavior, much researches have already shown the correlation between social media behavior and personality traits, in other words, similar social media behavior reflects approximate personality, and people with similar personality are likely to perform approximately. Under this assumption, [8] successfully manifest personality using online social networks behaviors and public facebook profile. [7] also achieve good performance to predict personality from social media. Thus, it is rational to make such a cluster assumption in our paper.

From Section V, we know that the Linear local semi-supervised regression model reduces Mean Absolute Error (MAE) by 2.5% to 7% on the basis of local linear kernel regression. It proves the correctness of our assumption in some degree.

B. Personality and Microblog Features

The Relative Absolute Error (RAE) achieved by local linear semi-supervised regression model approximates 75% when the training-test rate is greater than $1/2$. This result indicates the success of our methods. Nevertheless, we don't even know how the specific features can affect personality prediction. With the selected features in Table I, we will analyze it as follows.

At first, we present the characteristics for the five personality traits.

1) Extroversion: outgoing, amicable, assertive. Extraversion tends to be manifested in outgoing, talkative, energetic behavior.
2) Conscientiousness: thorough, careful, persevering. Conscientious individuals are generally hard working and reliable.
3) Agreeableness: kind, sympathetic, cooperative, warm and considerate. People who score high in agreeableness are peace-keepers who are generally optimistic and trusting of others.
4) Neuroticism: anxiety, moodiness, worry, envy and jealousy. Individuals who score high on neuroticism are more likely than the average to experience such feelings as anxiety, anger, envy, guilt, and depressed mood.
5) Openness to Experience: active imagination, aesthetic sensitivity, attentiveness to inner feelings, preference for variety and intellectual curiosity. People high in openness tend to be artistic and sophisticated in taste and appreciate more liberal political views, ideas, and experiences.

As listed above, extraversion tends to be manifested in outgoing, talkative, energetic behavior in daily life. Their performance on social media also supports the assertion. Compared with introverts, the extraverts incline to make their Microblog profile as complete as possible (V9), take a longer self-description (V11) and present more tags (V22) for themselves. The more micro-medal (V23) they obtained, the more active they are on the Microblog platforms. Thus, all these features are beneficial for distinguishing the extraversion extent, and that is the reason that they are chosen.

Agreeableness are kind and cooperative in social life. Correspondingly, individuals high in agreeableness also perform warm and considerate. They have relatively more mutual followers (V13), and the sentiment of their self-description are more positive.

Conscientiousness can be expressed as thorough, careful and persevering. Conscientious individuals are much different from immoral individuals with the following behaviors: the complete degree of personal profile (V9), number of mutual followers (V13). We also find the conscientious people have shorter screen name (V10).

Neuroticism is the stability of emotion. People high in neuroticism are more likely to suffer mental health. In our research, there are totally 8 features related to this dimension, three are related to status publishing period (V30, V34, V41). It seems that people with high degree neuroticism prefer to post statuses in early morning (0:00-6:00), and they are also found to publish more statuses (V1).

Openness shows the richness of active imagination, attentiveness to inner feelings and curiosity about new things. People with high score in openness are often like to know what happens to others, the performance of more friends (they follow more other people, V11) on Microblog supports this point. They are also found to be active on Microblog (V2, V14).

VII. Conclusion

In this paper, we have investigated personality traits identification based on Microblog users' public information, and local linear kernel approach and local linear semi-supervised method are employed. Due to high dimensional feature space, stepwise regression method is used for feature selection. Experiment was conducted, and the results demonstrate our models perform well.

In future, we continue to collect users data on Sina Microblog, and invite more participants to acquire more labeled data to improve the accuracy of our models. Furthermore, we intend to take Microblog content (for example, nlp related feature) into account, and explore better feature selection methods. We also plan to conduct further research on the behavior patterns of other psychological attributes.

VIII. Acknowledgments

The authors gratefully acknowledges the generous support from National High-tech R&D Program of China (2013AA01A606), NSFC (61070115), Key Research Program of CAS (KJZD-EW-L04), Strategic Priority Research Program (XDA06030800) and 100-Talent Project (Y2CX093006) from Chinese Academy of Sciences.

References

[1] Balcan, M.-F, Blum.A, Choi, J. P. P.and Lafferty, Pantano.B, Rwebangira, M. R., and X. Zhu. Person identification in webcam images: An application of semi-supervised learning. In *ICML 2005 Workshop on Learning with Partially Classified Training Data*, 2005.

[2] M. Belkin and V. Niyogi, P.and Sindhwani. On manifold regularization. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2005.

[3] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proc. 18th International Conf. on Machine Learning*, 2001.

[4] T. Buchanan and J. L. Smith. Using the internet for psychological research: Personality testing on the world wide web. *British Journal of Psychology*, 90(1):125–144, 1999.

[5] B. Cao. Sina's weibo outlook buoys internet stock gains: China overnight. Technical report, Bloomberg, 2012.

[6] D. Funder. Personality. *Annu. Rev. Psychol.*, 52:197–221, 2001.

[7] J. Golbeck, C. Robles, and K. Turner. Prediciting personality with social media. In *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems*, pages 253–262. ACM, 2011.

[8] S. Gosling, A. Augustine, S. Vazire, N. Holtzman, and S. Gaddis. Manifestations of personality in online social networks: Self-reported facebook related behaviors and observable prole information. *Cyberpsychology behavior and social networking*, 14:483–488, 2011.

[9] R. R. Hocking. The analysis and selection of variables in linear regression. *Biometrics*, 32, 1976.

[10] M. Karagiannopoulos, D. Anyfantis, S. Kotsiantis, and P. Pintelas. Feature selection for regression problems. *Proceedings of HERCMA07*, 2007.

[11] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

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

[13] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Learning to classify text from labeled and unlabeled documents. In *AAAI-98, 15th Conference of the American Association for Artificial Intelligence*, pages 792–799, 1998.

[14] L. Qiu, H. Lin, J. Ramsay, and F. Yang. You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality*, 46:710–718, December 2012.

[15] M. R. Rwebangira and J. Lafferty. Local linear semi-supervised regression. School of Computer Science Carnegie Mellon University, Pittsburgh, PA, 2009, 2009.

[16] C. Sumner, M.S., and A. Byers. Determining personality traits and privacy concerns from facebook activity. *Black Hat Briefings*, 11, 2011.

[17] E. Thompson. Development and validation of an international english big-five mini-markers. *Personality and Individual Differences*, 45(6):542–548, 2008.

[18] A. N. Tikhonov. Regularization of incorrectly posed problems. In *Soviet Math. Dokl*, volume 4, pages 1624–1627, 1963.

[19] H. Zheng, N. Kaji, N. Yoshinaga, and M. Toyoda. A study on microblog classification based on information publicness. In *DEIM Forum*, 2012.

[20] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schlkopf. Learning with local and global consistency. In *Advances in Neural Information Processing System*, 2004.

[21] Z.-H. Zhou, K.-J. Chen, and H.-B. Dai. Enhancing relevance feedback in image retrieval using unlabeled data. *ACM Transactions on Information Systems*, 24:219–244, 2006.

[22] Z.-H. Zhou and M. Li. Semi-supervised regression with co-training. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2005.

[23] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *The 20th International Conference on Machine Learning (ICML)*, 2003.

APPENDIX

The extracted features are depicted in Table IV.

Mean Absolute Error (MAE) for the other four personality traits are presented as follows: Table V for E, Table VI for A, Table VII for N and Table VIII for O.

TABLE V. MAE ACHIEVED BY EACH ALGORITHM FOR PERSONALITY TRAIT E

Training Set Size	MVR	LR	LLKR	LLSSR
100	0.6664	0.7780	0.5655	**0.5399**
300	0.6659	0.7829	0.5392	**0.5140**
600	0.6663	0.7756	0.5393	**0.4997**
900	0.6661	0.7538	0.5312	**0.4916**
1200	0.6664	0.7661	0.5179	**0.4843**
1500	0.6671	0.7526	0.5054	**0.4705**
1700	0.6685	0.8365	0.4996	**0.4682**

TABLE VI. MAE ACHIEVED BY EACH ALGORITHM FOR PERSONALITY TRAIT A

Training Set Size	MVR	LR	LLKR	LLSSR
100	0.5398	0.8084	0.4502	**0.4363**
300	0.5367	0.7954	0.4224	**0.4149**
600	0.5401	0.7733	0.4194	**0.4039**
900	0.5380	0.7683	0.4208	**0.3938**
1200	0.5383	0.7528	0.4210	**0.3896**
1500	0.5392	0.7489	0.3977	**0.3774**
1700	0.5385	0.7192	0.4195	**0.3833**

TABLE VII. MAE ACHIEVED BY EACH ALGORITHM FOR PERSONALITY TRAIT N

Training Set Size	MVR	LR	LLKR	LLSSR
100	0.7085	0.8856	0.5989	**0.5730**
300	0.7084	0.8659	0.5774	**0.5466**
600	0.7091	0.8651	0.5628	**0.5313**
900	0.7090	0.8862	0.5516	**0.5180**
1200	0.7085	0.8861	0.5478	**0.5080**
1500	0.7085	0.8667	0.5368	**0.4960**
1700	0.7090	0.8232	**0.4652**	0.5044

TABLE VIII. MAE ACHIEVED BY EACH ALGORITHM FOR PERSONALITY TRAIT O

Training Set Size	MVR	LR	LLKR	LLSSR
100	0.6224	0.8353	0.5044	**0.5021**
300	0.6226	0.8396	0.4938	**0.4811**
600	0.6229	0.8294	0.4858	**0.4667**
900	0.6219	0.8402	0.4779	**0.4597**
1200	0.6228	0.8338	0.4699	**0.4508**
1500	0.6229	0.8235	0.4549	**0.4374**
1700	0.6205	0.9278	0.4460	**0.4398**

TABLE IV. EXTRACTED FEATURES

Feature ID	Feature Name	Description
V1	original_status_count	The number of user's original statuses
V2	status_count	The total number of user's statuses (including retweeted statuses)
V3	picture_count	The total number of user's pictures (pictures in all statuses)
V4	repost_status_count	The total number of user's reposted statuses
V5	comment_count	The total number of user's comments
V6	annotation_count	The total number of user's annotations
V7	original_status_rate	The rate of original statuses
V8	comment_average	The average number comments for each status
V9	profile_degree	The complete degree of personal profile
V10	screen_name_length	The length of screen name
V11	description_length	The length of user's description
V12	followers_count	The number of user's followers
V13	bi_followers_count	The number of user's mutual followers
V14	friends_count	The number of user's friends
V15	fav_status_count	The number of user's favourite statuses
V16	description_evaluation	The sentiment evaluation of user's description
V17	original_pic_count	The total number of user's original pictures
V18	original_pic_rate	The rate of user's original pictures
V19	original_pic_average	The average number original pictures for each status
V20	annotation_average	The average number of user's annotations for each status
V21	domain_url	Whether the user have a personalized domain address
V22	tags_count	The total number of tags
V23	medal_count	The total number of micro-medal on the user's account
V24	microblog_level	The current level of user's microblog account
V25	first_status_period	The period user most likely to give first status per day
V26	last_status_period	The period user most likely to give last status per day
V27	fav_status_period	The period user most likely to give most statuses per day
V28	first_p0	The days user created first status between 0:00 and 6:00
V29	first_p1	The days user created first status between 6:00 and 8:00
...	...	...
V33	first_p6	The days user created first status between 20:00 and 24:00
V34	last_p0	The days user created last status between 0:00 and 6:00
...	...	...
V40	last_p6	The days user created last status between 20:00 and 24:00
V41	fav_p0	The days user created most statuses between 0:00 and 6:00
...	...	...
V47	fav_p6	The days user created most statuses between 20:00 and 24:00