A peer-reviewed version of this preprint was published in PeerJ on 19 September 2016.

View the peer-reviewed version (peerj.com/articles/cs-81), which is the preferred citable publication unless you specifically need to cite this preprint.

Liu X, Zhu T. 2016. Deep learning for constructing microblog behavior representation to identify social media user's personality. PeerJ Computer Science 2:e81 https://doi.org/10.7717/peerj-cs.81

Deep learning for constructing microblog behavior representation to identify social media user's personality

Xiaoqian Liu, Tingshao Zhu

Due to the rapid development of information technology, Internet has become part of everyday life gradually. People would like to communicate with friends to share their opinions on social networks. The diverse social network behavior is an ideal users' personality traits reflection. Existing behavior analysis methods for personality prediction mostly extract behavior attributes with heuristic. Although they work fairly well, but it is hard to extend and maintain. In this paper, for personality prediction, we utilize deep learning algorithm to build feature learning model, which could unsupervised extract Linguistic Representation Feature Vector (LRFV) from text published on Sina Micro-blog actively. Compared with other feature extraction methods, LRFV, as an abstract representation of Micro-blog content, could describe use's semantic information more objectively and comprehensively. In the experiments, the personality prediction model is built using linear regression algorithm, and different attributes obtained through different feature extraction methods are taken as input of prediction model respectively. The results show that LRFV performs more excellently in micro-blog behavior description and improve the performance of personality prediction model.

Deep learning for constructing microblog behavior representation to identify social media user's personality

Xiaoqian Liu [*1], Tingshao Zhu[*1]

[1] Institute of Psychology, Chinese Academy of Sciences, Beijing, China

Corresponding Author:

Xiaoqian Liu [*1]

16 Lincui Road, Chaoyang District, Beijing, 100101, China

Email address: liuxiaoqian@psych.ac.cn

Tingshao Zhu[*1]

16 Lincui Road, Chaoyang District, Beijing, 100101, China

Email address: tszhu@psych.ac.cn

ABSTRACT

Due to the rapid development of information technology, Internet has become part of everyday life gradually. People would like to communicate with friends to share their opinions on social networks. The diverse social network behavior is an ideal users' personality traits reflection. Existing behavior analysis methods for personality prediction mostly extract behavior attributes with heuristic. Although they work fairly well, but it is hard to extend and maintain. In this paper, for personality prediction, we utilize deep learning algorithm to build feature learning model, which could unsupervised extract Linguistic Representation Feature Vector (LRFV) from text published on Sina Micro-blog actively. Compared with other feature extraction methods, LRFV, as an abstract representation of Micro-blog content, could describe use's semantic information more objectively and comprehensively. In the experiments, the personality prediction model is built using linear regression algorithm, and different attributes obtained through different feature extraction methods are taken as input of prediction model respectively. The results show that LRFV performs more excellently in micro-blog behavior description and improve the performance of personality prediction model.

Keywords: personality prediction, social media behavior, deep learning, feature learning

1 INTRODUCTION

Personality can be defined as a set of traits in behaviour, cognition and emotion which is distinctive among people [16]. In recent years, researchers have formed a consensus on personality structure, and proposed the Big Five factor model [18], which uses five broad domains or dimensions to describe human personality, including openness(O), conscientiousness(C), extraversion(E), agreeableness(A) and neuroticism(N) [6].

Traditionally, questionnaire has been widely used for personality assessment, especially the Big Five personality questionnaire. But the form of questionnaire may be inefficient on large population. Due to the rapid development of information technology, Internet becomes part of everyday life nowadays. People prefer expressing their thoughts and interacting with friends on social network platform. So researchers pay more and more attention to figuring out the correlation between users' behaviors on social network and their personality traits in order to realize automatical personality prediction by machine learning methods.

Nowadays, Internet is not just for communication, but also a platform for users to express their thoughts, ideas and feelings. Personality is expressed by users' behavior on the social network indirectly, which refers to a variety of operation on social network, such as comment, follow and like. In addition, text, punctuation and emoticon published by users can be regarded as one kind of social behavior. So, for automatic personality prediction, how to abstract these diverse and complex behaviors and acquire the digital representation of social network behaviors has become an critic problem. Existing behavior analysis methods are mostly based on some statistics rules, but artificial means have some disadvantages in objectivity and integrity. Generally, attributes are especially important for the performance of prediction model. A set of proper feature vectors could improve the effectiveness of prediction model to a certain extent. So, it is required that the attributes are not only the comprehensive and abstract description of individual's behavior characteristic, but also could reflect the diversity of different individuals' behaviors.

In this paper, we use deep learning algorithm to unsupervised extract LRFV actively from users' content published on Sina Micro-blog. Compared with other attributes obtained by artificially means, LRFV could represent users' linguistic behavior more objectively and comprehensively. There are two reasons of utilizing deep learning algorithm to investigate the correlation between users' linguistic behavior on social media and their personality traits. One is that deep learning algorithm could extract high-level abstract representation of data layer by layer by exploiting arithmetical operation and the architecture of model. It has been successfully applied in computer vision, object recognition and other research regions. Another is, the scale of social network data is huge and deep learning alg orithm can meet the computational demand of big data. Given all this, we do some preliminary study on constructing microblog behavior representation for personality prediction based on deep learning algorithm in this paper.

1.1 Related Work

At present, many researchers have paid attentions to the correlation between users' Internet behaviors and their personality traits. Qiu *et al.* [19] figured out the relationship between tweets delivered on Twitter and users' personality, and they found that some personality characteristics such as openness(O), extraversion(E) and agreeableness(A) are related to specific words used in tweets. Similarly, Vazire *et al.* [23] discovered that there is great relevance between users' specific Internet behaviors and their personality through studying users' behaviors on personal website. These conclusions can be explained as personality not only influences people's daily behaviors, but also plays an important role in users' Internet behaviors. With the rise of social media, more and more researchers begin to analyse uses' personality traits through social network data with the help of computer technology. Sibel *et al.* [21] predicted users' personality based on operational behaviors on Twitter utilizing linear regression model. Similarly, in [8], Jennifer *et al.* also used regression algorithm to build a personality prediction model, but they considered both of operational behaviors and linguistic behaviors. Ana *et al.* [14] used semi-supervised method to predict personality based on the attributes of linguistic behaviors extracted from tweets. Alvaro *et al.* [17] built users' personality prediction model according to their social interactions in Facebook by machine-learning methods, such as classification trees.

Although lots of researchers utilized machine learning methods to built personality prediction model and have gotten some achievements, but there are also some disadvantages need to be improved. First, in state of art methods, the behavior analysis method and behavior attributes extraction methods are mostly based on some experiential heuristic rules which are set artificially. The behavior attributes extracted manually by statistical methods may not be able to describe characteristics of behaviors comprehensively and objectively. Second, supervised and semi-supervised behavior feature extraction methods need a certain number of labeled data, but in the actual application, obtaining a large number of labeled data is difficult, time-consuming and high cost. So supervised and semi-supervised feature extraction methods are not suitable for a wide range of application.

1.2 Deep Learning

Deep learning is a set of algorithms in machine learning [1] [2], which owns a hierarchical structure in accordance with the biological characteristics of human brain. Deep learning algorithm is originated in artificial neural network, and it has been applied successfully in many artificial intelligence applications, such as face recognition [11], image classification [4], natural language processing [22] etc.. Recently, researchers are attempting to apply deep learning algorithm to other

research field. Lin *at el.* [12] [13] used Cross-media Auto-Encoder (CAE) to extract feature vector and identified users' psychological stress based on social network data. Due to the multi-layer structure and mathematics algorithm designed, deep learning algorithm can extract more abstract high-level representation from low-level feature through multiple non-linear transformations, and discover the distribution characteristics of data. In this paper, based on deep learning algorithm, we could train unsupervised linguistic behavior feature learning models for five dimensions of personality respectively. Through the feature learning models, the LRFV corresponding to each trait of personality can be learned actively from users' contents published on Sina Micro-blog. The LRFV could describe the users' linguistic behavior more objectively and comprehensively, and improve the accuracy of the personality prediction model.

2 DATASET

In this paper, we utilize deep learning algorithm to construct unsupervised feature learning model which can extract Linguistic Representation Feature Vector (LRFV) from users' contents published on Sina Micro-blog actively and objectively. Next, five personality prediction models corresponding to five personality traits are built using linear regression algorithm based on LRFV. We conduct preliminary experiments on relatively small data as pre-study of exploring the feasibility of using deep learning algorithm to investigate the correlation between user's social network behaviors and his personality.

2.1 Data collection

Nowadays, users prefer to expressing their attitudes and feelings through social network. Therefore, the linguistic information on social network is more significant for analysing users' personality characteristics. In this paper, we pay more attention to the correlation between users' linguistic behaviors on Sina Micro-blog and their personalities. According to the latest statistics, by the end of Dec. 2014, the total number of registered users of Sina Micro-blog has exceeded 500 million. On the 2015 spring festival's eve, the number of daily active users is more than 1 billion firstly. It can be said that Sina Micro-blog is one of the most popular social network platforms in China currently. Similar to Facebook and Twitter, Sina Micro-blog users can post blogs to share what they saw and heard. Through Sina Micro-blog, people express their inner thoughts and ideas, follow friends or someone they want to pay attention to, and comment or repost blogs they interested in or agreed with.

For data collection, we firstly released the experiment recruitment information on Sina Micro-blog. In totally, 2385 volunteers were recruited to participate in our experiments. They have to accomplished the Big Five questionnaire [24] online and authorized us to obtain the public personal information and all blogs. Collecting volunteers' IDs of Sina Micro-blog, we crawled their micro-blog data through Sina Micro-blog API. The micro-blog data collected consists the users' all blogs and their basic status information, such as age, gender, province, personal description and so on. The whole process of subjects recruitment and data collection lasted nearly two months. Through the preliminary screening, we obtained 1552 valid samples finally. When filtering invalid and noisy data, we designed some heuristic rules as follows:

- If the total number of one's micro-blogs is more than 500, this volunteer is a valid sample. This rule can ensure that the volunteer is an active user.

- In order to ensure the authenticity of the results of questionnaire, we set several polygraph questions in the questionnaire. The samples with unqualified questionnaires were removed.

- When the volunteers filled out the questionnaire online, the time they costed on each question were recorded. If the answering time was too short, the corresponding volunteer was considered as an invalid sample. In our experiments, we set the the answering time should longer than 2 seconds.

2.2 Data for linguistic behavior feature learning

Through iteration and calculation layer by layer, deep learning algorithm can mine the internal connection and intrinsical characteristics of linguistic information on social network platform. Assuming the text in micro-blogs could reflect users' personality characteristics, for each trait of personality, we build a linguistic behavior feature learning model based on deep learning algorithm to extract the corresponding LRFV from users' expressions in Sina Micro-blog.

Linguistic Inquiry and Word Count (LIWC) is a kind of language statistical analysis software, which has been widely used by many researches to extract attributes of English contents from Twitter and Facebook [8] [9]. In order to meet the demands of simple Chinese semantic analysis, we developed a simplified Chinese psychological linguistic analysis dictionary for Sina Micro-blog (SCLIWC) [7]. This dictionary was built based on LIWC 2007 [20] and the traditional Chinese version of LIWC (CLIWC) [10]. Besides referring to the original LIWC, we added five thousand words which are most frequently used in Sina Micro-blog into this dictionary. The words in dictionary are classified into 88 categories according to emotion and meaning, such as positive word, negative word, family, money, punctuation etc. Through analysis and observation, we found that in some dimensions of personality, users of different scores show great differences in the number of using words belonging to positive emotion, negative emotion and some other categories in the dictionary.

According to SCLIWC [7], the users' usage degree of words in blogs could be computed in 88 categories. In order to obtain the usage characteristics of social media text in the temporal domain, we divide the time by week firstly. For the i^{th} word category of SCLIWC, the usage frequency within the j^{th} week f_j^i (i=1,2,...,88) is calculated by Equation 1, in which, i denotes the serial number of category, and j denotes the serial number of week. We collect all the text published in Sina Micro-blog during recent three years (Jun.2012~Jun.2015), and there are 156 weeks in total. So, corresponding to each category of SCLIWC, the vector $f^i = \{f_1^i, f_2^i, \ldots, f_{156}^i\}$ is the digital representation of the i^{th} category in temporal domain.

$$f_j^i = \frac{The\ number\ of\ words\ belongs\ to\ the\ i^{th}\ category\ of\ SCLIWC\ in\ j^{th}\ week}{The\ total\ number\ of\ words\ in\ blogs\ in\ j^{th}\ week} \tag{1}$$

Then, we utilize Fast Fourier Transform(FFT) [15] to obtain the varying characteristics of social media text usage in temporal space. Fourier Transform is a special integral transform, which could convert the original temporal signal into frequency domain signal which is easily analyzed. FFT is the fast algorithm of Discrete Fourier Transform (DFT), defined by

$$X(k) = DFT[x(n)] = \sum_{n=0}^{N-1} x(n) W_N^{kn},\ k = 0, 1, \ldots, N-1 \tag{2}$$

$$W_N = e^{-j\frac{2\pi}{N}} \tag{3}$$

In order to extract the temporal information from massive high-dimensional digital vectors, Fourier time-series analysis is considered. Concretely, we conduct FFT for each vector. Through FFT, the amplitudes calculated include frequency information, and former 8 maximum amplitudes are selected to constitute a vector as the representation of each word category. Finally, linking the vectors of each category in series, we can obtained a linguistic vector of 704 length corresponding to each user ID.

In our experiment, we use 1552 users' blogs published in 3 years as data for preliminary study. Each user's linguistic behavior is represented as vector form through FFT based on SCLIWC.

2.3 Data for personality prediction

In order to verify the deep learning algorithm is an effective method for extracting the representation of user's Sina Micro-blog linguistic behaviors, we build personality prediction model based on linguistic behavior feature vectors. The personality prediction model is constructed by linear regression algorithm. For each volunteer, five linguistic behavior feature vectors corresponding to five traits of personality are obtained by feature learning models respectively. The training process of personality prediction model is supervised, so users' five scores of five personality traits in the Big Five questionnaire are taken as their labels of the corresponding linguistic behavior feature vectors.

3 METHODS

3.1 Unsupervised feature learning based on Stacked Autoencoders

Feature learning can be seen as a process of dimensionality reduction. In order to improve the computational efficiency, for all traits of personality, we utilize the relatively simpler form of artificial neural network, autoencoder [1]. Fig 1 shows the basic structure of an autoencoder. Basiclly, for an autoencoder, the input and output own the same dimensions, both of them can be taken as x, but through mathematical transformation, the input and output may be not completely equal. In Fig 1, x denotes input and x' denotes output. The hidden layer is encoded through x, and can be decoded to form x'. When training an autoencoder, the input vector x will be mapped into a different representation y by Equation 4, in which y is the variable in hidden layer in Fig 1.

$$y = f_\theta(x) = s(Wx+b) \tag{4}$$

In Equation 4, $\{W,b\}$ are parameters which can be obtained through training. In addition, a reconstructed vector x' in input vector space could be obtained by mapping the result of hidden layer y back through a mapping function,

$$x' = g_{\theta'}(x) = s'(W'y+b') \tag{5}$$

If we want the mapping result y is another representation of input x, it is assumed that the input x and the reconstructed x' are the same. According to this assumption, the training process of an autoencoder could be conducted and the parameters of autoencoder are adjusted according to

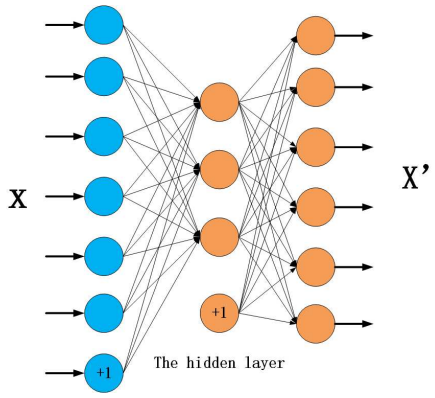


Figure 1. The basic structure of an autoencoder.

minimize the error value between x and x', as shown in Fig 2. Due to the error is directly computed based on the comparison between the original input and the reconstruction obtained, so the whole training process is unsupervised.

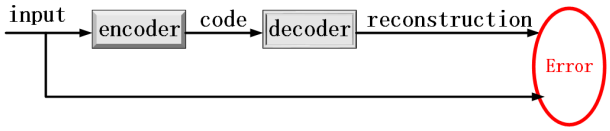


Figure 2. The training principle diagram of an autoencoder.

Several autoencoders are stacked to initialize the deep architectures layer by layer as Fig 3. Let the hidden layer of k^{th} layer be used as the input of $(k+1)^{th}$ layer. Based on the layer-wise algorithm [3], $(k+1)^{th}$ layer will be trained after the completion of training k^{th} layer. The number of layer would be decided according to the optimal value of many experiments. Then, we take the output of the last layer as the abstract representation of the original linguistic behavior information. In our experiments, 1552 users' content information of Sina Micro-blog are used as training dataset, and the unsupervised feature learning models corresponding different personality traits are trained respectively. That is, we could obtain five feature learning models in total. For each trait, there will be corresponding linguistic behavior feature vectors extracted from social network behavior data actively.

Finally, based on the Big Five questionnaire, for each user, we could obtained five scores (S_A, S_C, S_E, S_N, S_O) corresponding to "A", "C", "E", "N", "O" five dimensions respectively. These scores are used to label corresponding linguistic behavior feature vectors for personality prediction models.

3.2 Personality prediction model based on linear regression

Personality prediction is a supervised process. The linguistic behavior feature vectors are labeled by the corresponding scores of the Big Five questionnaire. For five traits of personality, we utilized the linear regression algorithm to build five personality prediction models in totally.

Take one trait of personality as an example, the linguistic behavior feature vectors are represented by

$$X = \{X_i \mid X_i = (x_{i1}, x_{i2}, \ldots, x_{im})\}_{i=1}^{n}, \tag{6}$$

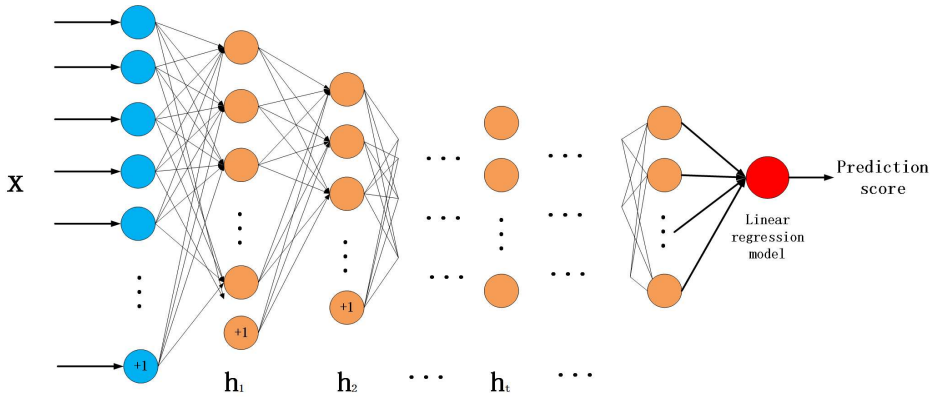


Figure 3. The deep architecture of Stacked Autoencoders

in which, n is the number of samples, $n = 1552$, and m denotes the number of dimensions of the input vector. The scores of the Big Five questionnaire are taken as the labels,

$$Y = \{y_i\}_{i=1}^{n} \tag{7}$$

The general form of linear regression is

$$y_i = \omega_1 x_{i1} + \omega_2 x_{i2} + \ldots + \omega_m x_{im} + \varepsilon_i, (i = 1, 2, \ldots, n) \tag{8}$$

We trained five personality prediction models based on linear regression algorithm using corresponding linguistic behavior feature vectors and labels.

4 RESULTS

In Experiments, we collect 1552 users' Sina Micro-blog data in total. Users' linguistic behaviors are quantified based on SCLIWC, and the temporal characteristics are calculated through FFT. Then, we utilize deep learning algorithm to construct feature learning models, which could extract objective and comprehensive representation of linguistic behaviors from the temporal sequence. Finally, personality prediction model is trained by linear regression algorithm based on linguistic behavior feature vectors.

4.1 Evaluation measures

In this paper, we conducted preliminary study about constructing Micro-blog behavior representation for predicting social media user's personality. The five dimensions of personality are all tested. We use Pearson correlation coefficient (r) and Root Mean Square Error ($RMSE$) to measure the quality of different behavior feature representation methods. The computational formulas of two measurements are shown in Equation 9 and 10 respectively. In Equation 9, $Cov(Y, Y^{'})$ denotes the covariance of Y and $Y^{'}$, and $Var(Y)$ and $Var(Y^{'})$ represents the variances of the real score Y and prediction score $Y^{'}$ respectively. when $r > 0$, it means the results of questionnaire and prediction model are positive correlation. On the contrary, $r < 0$ means negative correlation. The absolute value is greater, the higher is the degree of correlation. In psychology research, $r \in [0.2, 0.4]$ presents there are weak correlation between two variables and $r \in [0.4, 0.6]$ indicates two variables are moderate

correlative. In Equation 10, i is the sequence number of sample and n is the total number of samples, $n = 1552$. In the Big Five questionnaire used in our experiments, there are 44 questions in all. The score ranges of "A", "C", "E", "N", "O" are $[9,45]$, $[8,40]$, $[9,45]$, $[8,40]$, $[10,50]$ respectively. The value of *RMSE* shows the average difference between our prediction results and the scores of questionnaire. The smaller is the value of *RMSE*, the better is the performance of prediction model.

$$r = Cor(Y,Y^{'}) = \frac{Cov(Y,Y^{'})}{\sqrt{Var(Y)Var(Y^{'})}} \tag{9}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i^{'})^2}{n}} \tag{10}$$

4.2 Experiment results

In comparison experiments, we utilized three different kinds of attributes to train and build the personality prediction model respectively. The three kinds of attributes including the attributes selected by artificial statistical method without feature selection (denoted by Attribute 1), the attributes selected from Attribute 1 by Principal Component Analysis (PCA) [5] (denoted by Attribute 2) and linguistic behavior feature vector obtained based on Stacked Antoencoders (SAE) (denoted by Attribute SAE). For different kinds of attributes, the personality prediction models are all built by linear regression algorithm. In order to obtain the stable model and prevent occurrence of overfitting, for each dimension of personality, we use 10-fold cross validation and run over 10 randomized experiments. Finally, the mean of 10 randomized experiments' results is recorded as the final prediction result. The comparison of prediction results of five personality dimensions using three kinds of attributes are shown in Tables 1 and 2. The letters in subscript "a", "c", "e", "n", "o" indicate different personality dimensions respectively.

Table 1. The comparison of prediction results in Pearson correlation coefficient

Attributes	r_a	r_c	r_e	r_n	r_o
Attributes 1	0.1012	0.1849	0.1044	0.0832	0.181
Attributes 2	0.102	0.2166	0.1049	0.1235	0.1871
Attributes SAE	**0.2583**	**0.4001**	**0.3503**	**0.3245**	**0.4238**

Table 2. The Comparison of prediction results in RMSE

Attributes	$RMSE_a$	$RMSE_c$	$RMSE_e$	$RMSE_n$	$RMSE_o$
Attributes 1	5.6538	6.1335	4.9197	6.5591	7.0195
Attributes 2	5.1628	5.6181	5.6781	5.9426	6.4579
Attributes SAE	**4.7753**	**5.339**	**4.8043**	**5.6188**	**5.1587**

5 DISCUSSION

This study explore the relevance between users' personality and their social network behaviors. The feature learning models are built to unsupervised extract the representations of social network linguistic behaviors. Compared with the attributes obtained by supervised behavior analysis methods, the linguistic behavior feature extracted by unsupervised feature learning method is more objective, efficient, comprehensive and universal. In addition, based on the linguistic behavior feature vectors, the accuracy of the personality prediction model could be improved.

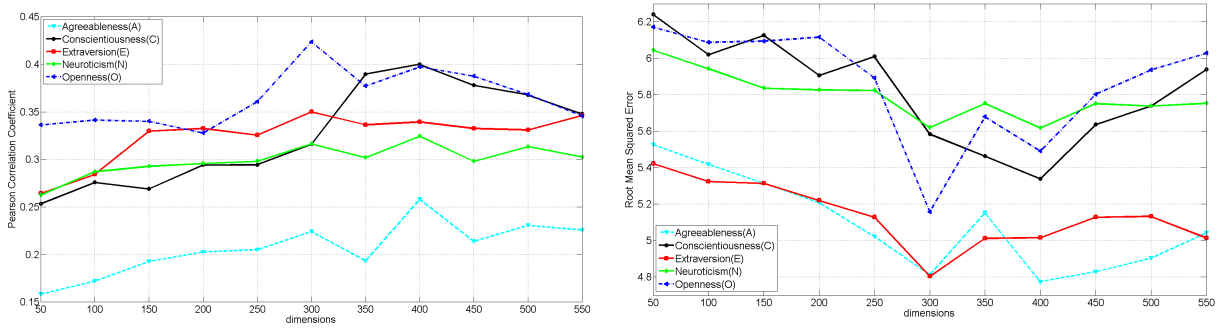
5.1 The performance of personality prediction model

The results in Tables 1 and 2 show that the linguistic behavior feature vectors learned through Stacked Autoencoders perform better than other attributes in both Pearson correlation coefficient and RMSE. When using Attribute SAE, the Pearson correlation coefficients of "A", "E", "N", "O" are all more than 0.2, which mean that there are weak correlation between the results of personality prediction models and questionnaire scores. For "C" and "O", $r_c = 0.4001$ and $r_o = 0.4238$, which means that the prediction results of "C" and "O" correlate with the results of questionnaire moderately. It is concluded that personality prediction based on the linguistic behavior in social network is feasible. Besides, the traits of conscientiousness and openness could be reflected in the network linguistic behavior more obviously.

Compared with other feature extraction methods, our proposed method performs better. When using Attributes 1, the prediction results r are all less than 0.2. When using Attributes 2, except for "C", others are also less than 0.2. Similarly, considering *RMSE* of every personality traits, the prediction model also obtain better results based on the linguistic behavior feature vectors.

5.2 Parameters selection

In each Stacked Autoencoders model, the sigmoid function is used as activation function of hidden layers. For each personality trait, the dimensionality of linguistic behavior feature vector is set according to the optimal result of prediction model obtained from repeated experiments. For each personality trait, the comparison of r and *RMSE* when using linguistic behavior feature vectors with different dimensionality are presented in Figs 4(a) and 4(b) respectively. For "A", "C" and "N", prediction models perform better when the dimensionality of feature vector is 400. For "E" and "O", we could obtain the better results when the dimensionality of feature vector is 300.



(a) The comparison of r (b) The comparison of *RMSE*

Figure 4. The comparison of prediction results using linguistic feature vectors with different dimensionality. (a)The comparison of r. (b)The comparison of *RMSE*.

5.3 Differences in modeling performance across personality traits

Through analysing the results of experiments, we summarize that Agreeableness correlate with users' social network linguistic behaviors relative weakly than the other personality traits. The correlation between openness and users' social network linguistic behaviors is highest of all. We could identify whether the users own higher scores in openness or not through their blogs published in social network platform. Probably because the person with high scores in openness usually prefer expressing their thoughts and feelings publicly. Similarly, conscientiousness is moderately correlate with social network linguistic behaviors. And for conscientiousness, there are significant differences of using the words belonging to the categories of family, positive emotion and so on.

6 CONCLUSIONS

In this paper, we utilized deep learning algorithm to investigate the correlations between users' personality traits and their social network linguistic behaviors. Firstly, the linguistic behavior feature vectors are unsupervised extracted using Stacked Autoencoders models actively. Then, the personality prediction models are built based on the linguistic behavior feature vectors by linear regression algorithm. Our comparison experiments are conducted on three different kinds of attributes, and the results show that the linguistic behavior feature vectors could represent users' social network linguistic behavior objectively, comprehensively and universally and improve the performance of personality prediction models.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the generous support from Young Talent Research Fund (Y4CX103005) from Institute of Psychology Chinese Academy of Sciences, NSFC (61070115), Strategic Priority Research Program (XDA06030800) and 100-Talent Project (Y2CX093006) from Chinese Academy of Sciences.

REFERENCES

[1] Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127.

[2] Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1789–1828.

[3] Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007). Greedy layerwise training of deep networks. *Advances in Neural Information Processing Systems*, pages 153–160.

[4] Ciresan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3642–3649.

[5] Dunteman, G. H. (1989). *Principal components analysis*. Number 69. Sage.

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

[7] Gao, R., Hao, B., Li, H., Gao, Y., and Zhu, T. (2013). Developing simplified chinese psychological linguistic analysis dictionary for microblog. In *International Conference on Brain Health Informatics*.

[8] Golbeck, J., Robles, C., Edmondson, M., and Turner, K. (2011a). Predicting personality from

twitter. In *IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and IEEE Third Inernational Conference on Social Computing*, pages 149–156.

[9] Golbeck, J., Robles, C., and Turner, K. (2011b). Predicting personality with social media. *Extended Abstracts on Human Factors in Computing Systems*, pages 253–262.

[10] Huang, C. L., Chung, C. K., Hui, N., Lin, Y. C., Yi-Tai, S., Lam, B. C. P., Chen, W. C., Bond, M. H., and Pennebaker, J. W. (2012a). The development of the chinese linguistic inquiry and word count dictionary. *Chinese Journal of Psychology*, 55:185–201.

[11] Huang, G. B., Lee, H., and Learned-miller, E. (2012b). Learning hierarchical representations for face verification with convolutional deep belief networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2518–2525.

[12] Huijie, L., Jia, J., Quan, G., Yuanyuan, X., Jie, H., Lianhong, C., and Ling, F. (2014a). Psychological stress detection from cross-media microblog data using deep sparse neural network. In *Proceedings of IEEE International Conference on Multimedia Expo*.

[13] Huijie, L., Jia, J., Quan, G., Yuanyuan, X., Qi, L., Jie, H., Lianhong, C., and Ling, F. (2014b). User-level psychological stress detection from social media using deep neural network. In *Proceedings of the ACM International Conference on Multimedia*, pages 507–516.

[14] Lima., A. C. E. S. and de Castro, L. N. (2013). Multi-label semi-supervised classification applied to personality prediction in tweets. In *The 11th Brazilian Congress on Computational Intelligence*, pages 195–203.

[15] Loan, C. V. (1992). Computational frameworks for the fast fourier transform. *SIAM*, 10.

[16] Mischel, W., Shoda, Y., and Ayduk, O. (2007). *Introduction to personality: Toward an integration.* 8th ed. Wiley Press.

[17] Ortigosa, A., Carro, R. M., and Quiroga, J. I. (2013). Predicting user personality by mining social interactions in facebook. *Journal of Computer and System Sciences*, 80(1):57–71.

[18] P.T.Costa and R.R.McCrae (1992). Revised neo personality inventory and neo five-factor inventory (neo-ffi) manual. *Odessa, FL: Psychological Assessment Resources*.

[19] Qiu, L., H.Lin, J.Ramsay, and F.Yang (2012). You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality*, 46(6):710–718.

[20] R, T. Y. and W, P. J. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

[21] Sibel, A. and Golbeck, J. (2014). Predicting personality with social behavior: a comparative study. *Social Network Analysis and Mining*, 4:159.

[22] Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing*.

[23] Vazire, S. and S.D.Gosling (2004). e-Perceptions: personality impressions based on personal websites. *Journal of personality and social psychology*, 87(1):123.

[24] Vittorio, C. G., Claudio, B., Laura, B., and Marco, P. (1993). The "big five questionnaire": A new questionnaire to assess the five factor model. *Personality and individual differences*, 15(3):281–288.