Predicting Personality Traits of Chinese Users Based on Facebook Wall Posts

Kuei-Hsiang Peng, Li-Heng Liou, Cheng-Shang Chang, and Duan-Shin Lee
Institute of Communications Engineering
National Tsing Hua University
Hsinchu 300, Taiwan, R.O.C.
Email: 9761115@gmail.com; dacapo1142@gmail.com; cschang@ee.nthu.edu.tw; lds@cs.nthu.edu.tw

***Abstract*—Automatically recognizing personality based on historical action logs in online social networks is a promising method to infer a person's behaviors, and it has received a lot of attention lately as it might lead to the construction of a better personal recommendation system. However, very few previous works in the literature put their focus on predicting personality from *Chinese* texts. As Chinese texts are much more difficult to delimit than English texts, it poses more challenges in recognizing personality from Chinese texts.**

In this paper, we attempt to classify the personality traits from Chinese texts. We collected a dataset with posts and personality scores of 222 Facebook users who use Chinese as their main written language. Then, we used *Jieba*, a Chinese text segmentation tool, as the tokenizer for the task of text segmentation, and the Support Vector Machine (SVM) as the learning algorithm for personality classification. Our experimental results show that the performance in precision and recall can be significantly improved with the help of text segmentation. Moreover, exploiting side information, such as the number of friends, could improve the performance further. One interesting finding from our experiments is that extraverts seem to write more sentences and use more common words than introverts. This indicates that extraverts are more willing to share their mood and life with others than introverts.

I. Introduction

In recent years, research of users' behavior has gained a lot of attention as a way to predict a person's preferences. In the past, service providers (e.g., clerks of clothing stores) acquire the knowledge of customers' preferences by observing and interacting with customers and provide personalized shopping advices based on their experiences. However, this is not possible for online stores. Online stores can only recommend new items or hot-selling items to "all" customers. As such, customers (users) may need to view plenty of webpages to find out things they like. What if service providers know the preferences of their customers? A personal recommendation system with such knowledge can definitely be more effective in providing relevant information to its customers.

One of the widely used methods for modeling users' behavior in a personal recommender system is to trace shopping histories or browser histories of users. However, the information about a user collected from a clothing store may not be applicable in a music store. If a system can really "know" a user, then it can recommend items in different fields to that user! Personality is believed to be an important factor in determining individual variation in thoughts, emotions and behavior patterns. In addition to modeling users' behavior, we believe modeling users' personality can provide additional information and might lead to more commercial applications.

Online social networks and smartphones have entered people's daily life. People tend to spend more and more time on online social networks, posting texts, photos and even videos. Their actions on social media might reveal their personality, and it might be possible to automatically classify a person's personality trait by using their posts on online social networks. Researchers have done various attempts to recognize personality from articles on blogs or status updates on social media (see Section 2 for a review of these works). However, most research used *English* texts as their input sources. Very few of them put their focus on predicting personality from *Chinese* texts. Chinese is a quite different language from English and many analytical methods are not directly applicable to Chinese. This makes it more difficult to classify personality traits based on Chinese textual data.

In this paper, we attempt to classify the personality traits from Chinese texts. We collected a dataset with posts and personality scores of 222 Facebook users who use Chinese as their main written language. Then, we used *Jieba* [1], a Chinese text segmentation tool, as the tokenizer for the task of text segmentation, and the Support Vector Machine (SVM) as the learning algorithm for personality classification. Our experimental results show that the performance in precision and recall can be significantly improved with the help of text segmentation. Moreover, exploiting side information, such as the number of friends, could improve the performance further. One interesting finding from our experiments is that extraverts seem to write more sentences and use more common words than introverts. This indicates that extraverts are more willing to share their mood and life with others than introverts.

The rest of the paper is organized as follows. In Section 2, we provide the background information about personality recognizer, including the Big Five personality model and related works. In Section 3, we introduce the models used in our experiments. The data collection, experimental setup, and results are presented in Section 4. Finally, in Section 5 we conclude the paper by summarizing the findings of our work and describing possible directions for future work.

II. Literature Review

Classification of personality traits has drawn much interest from psychologists and even researchers in other fields. For several decades of research, psychologists have reached a consensus on the *Big Five* model of personality, i.e., extraversion, neuroticism, agreeableness, conscientiousness and openness (see [2] for a historical review of the development of the model and [3], [4], [5] for more detailed explanations of this model). This five-factor structure of personality has been found on a wide range of people from different cultural backgrounds and age ranges. As such, we adopt such a model as our personality model. In particular, we will focus on *extraversion* that measures the tendency towards obtaining gratification from what is outside the self [6]. Extraverts are usually active in interacting with others. They are passionate, talkative, energized and enjoying in social activities. Introverts tend to be quiet and keep a low profile. They seldom attend social activities. Although they are willing to interact with close friends, they are more interested in solitary activities such as painting, reading, and writing.

Text mining has been the focus of sentiment analysis and opinion mining in the literature. In particular, Otterbacher [7] inferred the gender of movie reviewers by using logistic regression to explore writing style and content. Fu et. al [8] identified reasons why users like or dislike an app from the weights of words assigned by the linear regression model. Oberlander and Nowson [9] classified author's personality from weblog texts by using the n-grams as the features and the Naïve Bayes algorithm as the classification algorithm. They performed experiments on the authors with the highest and lowest scores and reported how to automatically select features that yield the best performance.

Recently, there is a workshop on Computational Personality Recognition [10] that released two datasets, annotated with gold standard personality labels, for participants of the workshop to evaluate features and learning techniques. The works in the workshop are of particular interest to us. In particular, Markovikj et al. [11] predicted personality by exploiting various features extracted from the Facebook data. The features included Facebook profile data (e.g., age and gender), statistical data for user's activities (e.g., number of likes), linguistic features (e.g., word count), Part Of Speech tags, word emotional values (AFINN) and word intensity scale (H4Lvd). They achieved good performance by using a ranking algorithm for feature selection. Their work shows the importance of feature selection in personality recognition. Also, Tomlinson et al. [12] predicted conscientiousness by exploiting the nuances on the usages of the verbs, in other words, measuring the specificity and objectivity of the verbs taken from WordNet and Senti-WordNet. They reported an accuracy of 68% when only classifying the outliers (i.e. users with scores at least one standard deviation from the median). Alam et al. [13] used unigrams as features and predicted personality by different classification methods such as Support Vector Machine, Bayesian Logistic Regression and Multinomial Naïve Bayes. They reported the best accuracy of 61.79% with Multinomial Naïve Bayes.

All the works mentioned above, however, were conducted with English texts. Most of the methods for extracting features cannot be directly applied for Chinese textual data since delimitation of Chinese sentences is quite different from that of English. Moreover, many features in the literature were extracted by using dictionaries which are only available in English. Motivated by all these, in this paper we conduct experiments to see whether personality can be predicted by using Chinese texts. For this, we first collect personality data from 222 Facebook users in Taiwan and then make an attempt to classify their personality by using their Facebook posts. We use a Chinese segmentation algorithm as the tokenizer for feature extraction and the Support Vector Machine as our learning algorithm. These methods will be described in details in the next section.

III. Methods

In this section we describe the methods used for processing and classifying the Chinese textual data. Unlike image data, which can be directly represented by numeric values without losing too much information, textual data could lose a lot of information (e.g., the order of words in a sentence) when represented by numeric values. Therefore, the method used to process the textual data plays an important role in textual data analysis.

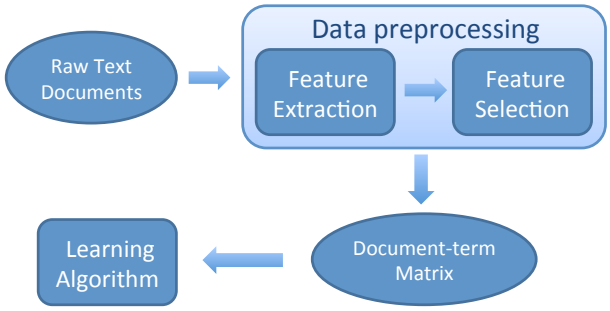

Fig. 1. **The generic procedure for text classification.**

As in most research of text analysis (see e.g., [14]), we process Chinese texts as follows (see Figure 1): We first read in the raw text documents. Then we use feature extraction and selection methods to construct the vector representation. Finally, we apply an appropriate classification algorithm to the document vectors. The details of each step will be introduced later in this section.

A. Text feature extraction algorithms

Feature extraction is the main procedure to extract meaningful words as the features to represent each document based on the *bag-of-words model*. It contains two main steps: tokenizing and counting. In the tokenizing step, we use a Chinese text segmentation tool instead of the simple tokenizer in the scikit-learn toolkit [15]. In the counting step, we use term frequency (TF) and term frequency-inverse document frequency (TF-IDF) as our two weighted schemes.

1) Bag-of-words model: The bag-of-words model is a commonly used document representation method in the field of information retrieval and natural language processing. In this model, each text document is viewed as a bag of words, discarding the word order and grammar. A corpus, or a collection of documents, therefore, is viewed as a bag of N unique words. As a result, a text document can be simply represented by an N-dimension vector with the occurrence of a word as the value of each entry. Such a matrix is called a *document-term matrix*. This concept is also the same as the vector space model proposed by Salton, Wong, and Yang [16].

2) Chinese text segmentation: Unlike words in English texts or other western language texts that can be divided by spaces, words in Chinese texts cannot be easily delimited by computers. Therefore, we need to apply an alternative tokenization method for Chinese textual data. Since we implemented the personality recognizer in Python, the open source "Jieba Chinese Text Segmentation" algorithm is the most suitable choice for our experiments [1].

3) Weighted schemes: In the term frequency (TF) weighted scheme (see e.g., the book [17]), the value of each entry in the document-term matrix represents the term frequency of each feature in a document. The term frequency here can be simply the occurrence count. If a feature also occurs in many documents, that feature might be treated as a noise. As such, in the *term frequency-inverse document frequency* (TF-IDF) weighted scheme, the weight of a feature that occurs in many documents is further reduced. As a result, TF-IDF is expected to eliminate noises or unimportant information in the data and thus can enlarge the representative features for each document.

B. Feature selection algorithms

A large number of extracted features will result in a high-dimensional document vector and thus increases the difficulty to classify the document. As such, it is important to select an appropriate set of features. Here we use two feature selection algorithms: the chi-squared test and the recursive feature elimination algorithm [18].

1) Chi-squared test: The chi-squared test (see e.g., the book [19]), also referred to the χ^2-test, is a statistical test in which the sampling distribution is a chi-squared distribution when the null hypothesis is true. There are several kinds of chi-squared tests used for different purposes. In most scientific articles, it is usually referred to Pearson's chi-squared test if the test is mentioned without any additional explanation/specific statement. Pearson's chi-squared test is known as two types of purposes: tests of goodness of fit and tests of independence. In this paper, we use it as the tests of independence (between a feature and a label).

2) Recursive Feature Elimination: Recursive Feature Elimination (RFE) that utilizes the Support Vector Machine methods was first proposed by Guyon et. al and originally applied to cancer classification [18]. Since SVM-RFE yields good performance of feature selection in other research fields, it has become a popular approach nowadays. The concept of RFE is to recursively construct a model and remove features with low weights until the desired number of selected features is reached. This algorithm has been implemented in the scikit-learn toolkit [15].

C. Classification algorithms

Support vector machine (SVM) is a supervised learning algorithm and has been widely utilized in classification problems. Given a set of training instances labeled with one of two classes, the goal of a SVM model is to find a hyperplane that can maximize the margin between two classes, that is, to separate two classes as far as possible. In this paper, we used the open source LIBSVM [20].

IV. Experiments and results

In this section, we provide the details for our data collection, the basic statistical information of the dataset, and the results from various experiments.

A. Data collection

To evaluate the accuracy of the Chinese personality recognizer, we need to collect a dataset with both Chinese texts and personality scores of users. For this, we contacted Facebook friends of ours. Since most of them are not sufficiently motivated to complete a general personality test, we designed a short online questionnaire based on the Big Five Mini-Makers (Saucier, 1994) [21] and TIPI (Ten Item Personality Inventory) [3]. In the questionnaire, we gave a description of each personality dimension and let the respondents to evaluate themselves. The personality scores are measured by a 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree). Facebook identifier of each user was also obtained in order to collect the posts of each user.

Originally, we collected personality scores and Facebook profile data with a total number of 264 users. Open-source Facebook graph API was utilized to access the posts created by those users. We collected all the posts of these users up to May 2014. However, some of them had posts less than 10 and these were discarded. As a result, the dataset with 222 users was used in our experiments.

B. Statistical characteristics of the dataset

In Table I, we present the overall average scores of personality in the Big Five model. As shown in the table, the average scores of agreeableness and openness to experience are high. This might be due to sampling bias as most respondents are students or graduates from our university and they are within the same age range.

Personality	Average Score
Extraversion	3.34
Neuroticism	2.68
Agreeableness	4.06
Conscientious	3.19
Openness to Experience	3.85

TABLE I
Average score of each personality

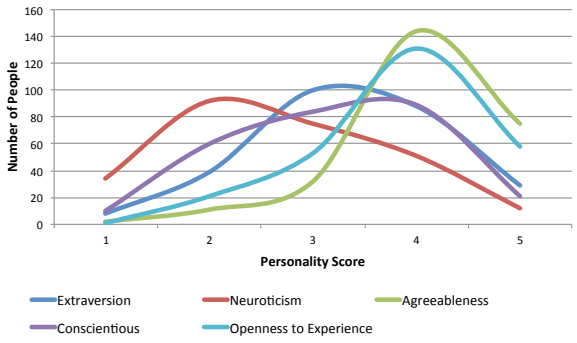

Fig. 2. **The distribution of personality.**

The score distribution of personality is depicted in Figure 2. The distribution of extraversion and conscientious are close to the normal distribution, and that of neuroticism is slightly towards left, whereas over half of the people rated on 4 in the dimensions of agreeableness and openness to experience.

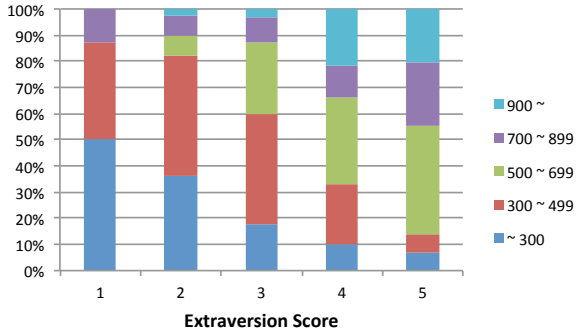

Fig. 3. **The extraversion score vs. the number of friends.**

From the Big Five model we can infer that extraverts are likely to have more friends than introverts. This is also supported by the statistical results presented in Figure 3. Users with less than 500 friends were likely to score low in extraversion, while almost all the users with more than 900 friends scored high in extraversion.

C. Evaluation Metrics

In classification problems, we can predict the labels of instances in the testing set by using the trained model and then compare the predicted label with the actual label to evaluate the performance of the model. Let tp (resp. tn) be the number of positive (resp. negative) instances that are *correctly* predicted as positive (resp. negative). Also, let fp (resp. fn) be the number of negative (resp. positive) instances that are *incorrectly* predicted as positive (resp. negative). Suppose there are a total number of T instances. Then *accuracy* (Acc) is defined as $(tp + tn)/T$, *precision* (Pre) is defined as $tp/(tp+fp)$, *recall* (Rec) is defined as $tp/(tp+fn)$, *negative predictive value* (NPV) is defined as $tn/(tn + fn)$, and *true negative rate* (TNR) is defined as $tn/(fp + tn)$.

D. Experimental setup

The procedure of the experiment is illustrated in Figure 1 in Section 3. We implemented our personality recognizer in Python using the scikit-learn toolkit [15]. Since Facebook is a social media providing a platform for people to interact with their friends, we focused on the classification of *extraversion* which is most related to social activities in all the five dimensions of the Big Five model. The original scores of extraversion are discrete values from 1 to 5. In order for binary classification, we label users with scores from 4 to 5 as *extraverts* and users with scores from 1 to 3 as *introverts*. As a result, there are 94 extraverts and 128 introverts. For the text part, we put all the posts created by the same user into a text, and regarded it as a document. By doing so, we constructed the dataset in which each instance consists of a Chinese text document and a corresponding binary label for extraversion.

Next, we randomly split the original dataset into two groups: a training set and a testing set. A training set consists of 199 instances (90%) and a testing set consists of 23 instances (10%). For the training set, we extracted features by utilizing the Jieba Chinese text segmentation as the tokenizer. The resulting text segments were dropped if they occur only once in the entire corpus. Then, a feature selection algorithm was applied to filter out the features uncorrelated to the label classes. The remaining ones were regarded as the features representing each document and thus we constructed a document-term matrix for the training set. For the testing set, we used the same features as the training set to construct a document-term matrix.

Finally, we used the open-source LIBSVM [20] with RBF kernel to classify extraversion traits. To evaluate the performance, we used the training set to come up with a binary classifier for extraversion and then made predictions with that classifier for the testing set. For all experiments, we repeated the whole procedure (with random splitting of the dataset) ten times to obtain the average performance for each performance metric.

E. The effect of tokenizers

As mentioned before, in English texts words are delimited by spaces. If we use the simple tokenizer in the scikit-learn, then sentences are delimited in Chinese texts. As a result, the simple tokenizer in the scikit-learn can only extract sentences as tokens, which cannot represent a Chinese document well. An alternative is to use the Jieba Chinese segmentation tool as the tokenizer. Using the same feature selection and learning algorithms, our experimental results show that using Jieba is significantly better than using the simple tokenizer in the scikit-learn. In particular, the precision can be improved by 60% when using Jieba. This shows the importance of choosing a Chinese text segmentation tool.

F. The effect of the document-term matrix representation

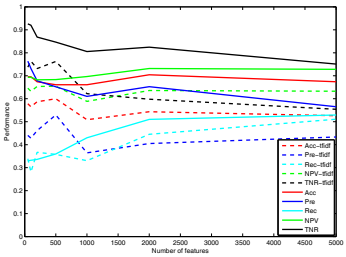

Fig. 4. **Performance comparison of the TF scheme (solid line) and the TF-IDF weighed scheme (dotted line).**

Figure 4 presents the performances of classifying extraversion by using the TF scheme and the TF-IDF scheme for the document-term matrix representation. It shows that the TF scheme outperforms the TF-IDF scheme by a wide margin and the accuracy of the TF-IDF weighted scheme is nearly a random guess when the number of features is above 1000. The results are surprising at first glance since the TF-IDF weighted scheme empirically can eliminate noises or unimportant information in the data and thus should achieve better performances than the basic counting scheme. However, this is not the case here as commonly used words (that are often regarded as "noises") are important features for our classification problem (see Section IV-I for further explanations).

G. The effect of feature selection

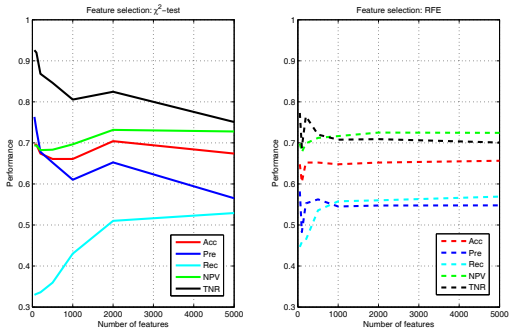

Fig. 5. **Performance comparison of the χ^2-test (left) and the RFE scheme (right).**

In this experiment, we employed two feature selection methods. One is the χ^2-test and the other is the Recursive Feature Elimination (RFE). We selected K best features according to the calculated χ^2 scores (for χ^2-test) or feature weights (for RFE). Then we changed K from 100 to 5000 to observe the variation of prediction performances.

The results are shown in Figure 5. For the χ^2-test case (on the left), the curve for accuracy has two peaks: one is when the number of selected feature is around 50-100 and the other is around 2000. The best average prediction accuracy is 70.36%. As the number of selected feature grows, the average precision decreases while the average recall increases. The best average precision is 76.33% with 50 selected features. We also calculated the negative predictive value (NPV) and true negative rate (TNR) to evaluate the performance for predicting users with low extraversion scores. When the number of selected features is between 50 and 100, we can successfully classify 92% of users who scored low in extraversion. For the RFE case on the right side of Figure 5, the value of each performance metric seems to be stable when the number of features is above 500.

H. The effect of using the number of friends as an additional feature

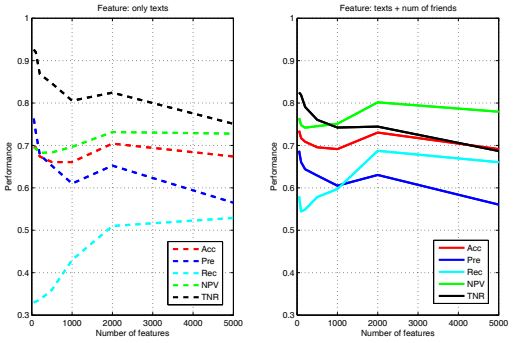

Fig. 6. **Performance comparison of using the number of friends as an additional feature: using texts only (left) and using both texts and the number of friends (right).**

Since the number of friends appears to be correlated with extraversion (as shown in Figure 3), we also added this feature into our consideration. As shown in Figure 6, the overall performances can be improved by considering both the text and the number of friends. In particular, recall is significantly improved and the accuracy can achieve 73.5%.

I. The selected features of these experiments

χ^2-test	Simple Tokenizer	TF-IDF	RFE
\n	成就達成(reach achievement)	\n	一起(together)
!	卡馬救星(Karma saver)	大老二(Big Two)	原來(originally)
我(I)	學習經營之道(learn to manager)	名([first] place)	哭(cry)
都(all)	我的餐廳(my restaurant)	手牌(Poker hands)	實在(truly)
有(have)	哈哈哈(hahaha)	!	快來(come quickly)
好(good)	QQ(crying face)	人渣(asshole)	打卡(check into place)
我們(we)	贏了(win)	戴牙套(wear braces)	超可愛的(very cute)
真的(really)	開始玩開心廚房啦(start to play..)	包機(charter flight)	我想(I think)
QQ(crying face)	混亂(chaotic)	卑賤(menial)	笑(smile)
今天(today)	T_T(crying face)	成就(achievement)	當然(sure)
大家(everybody)	正在(is [doing sth.])	稀有(rare)	棒(great)
說 (say)	感謝(thanks)	演過(acted)	總是(always)

Fig. 7. **The selected features.**

in Figure 7, we list a few selected Chinese text features (with English translation on the right) of each experiment. In the first column, we show the features selected by using the χ^2-test scheme with Jieba as the tokenizer and the TF scheme

as the representation of the document-term matrix. Such a scheme has the best performance in all our experiments. In the second column, we replace the tokenizer by the simple tokenizer in the scikit-learn. In the third column, we replace the TF scheme by the TF-IDF scheme. As we can see from the table, the selected features (words) of the χ^2-test scheme occur more frequently and are shorter than those of the other schemes. The selected features from using the simple tokenizer has the longest word length among all the schemes. This may reduce the correlation between documents and thus makes classification more difficult. The selected features of the TF-IDF scheme are more unfrequent words comparing to those of the others. This may be the reason why it has the worst performance (accuracy only 50-60%). The features of the RFE scheme are roughly common words with middle word length. Its accuracy is about 65%. In summary, extraverts tend to use commonly used words in their posts. This may be the reason why the χ^2-test scheme achieve the best performance among all schemes.

In addition, from the table, the feature "\n" appears on the top of two columns. As "\n" appears when a sentence is completed, it indicates that a document contains many sentences if the document has a large occurrence count of the feature "\n." We also found that many people scoring high in extraversion had a large value of this feature. It seems that extraverts tend to create a post with more sentences or create posts more frequently than introverts do.

V. Conclusion

Recognizing personality is not only an interesting subject but also a challenging task. Based on our experiments, we summarize our findings as follows:

(i) A good Chinese text segmentation tool plays an important role for processing Chinese documents.

(ii) The TF scheme based on the occurrence count seems to be a basic but reliable scheme for the representation of a document-term matrix.

(iii) Feature selection can lead to significant performance improvements. Using a small proportion of all features can still yield a good result or even better than using a large proportion of the features.

(iv) A large number of friends and a large number of sentences in the posts are positively correlated to extraversion.

(v) Extraverts seem to write more sentences than introverts do. Moreover, extraverts tend to use commonly used words in their posts. This indicates that extraverts are more willing to share their mood and life with others than introverts.

Our best accuracy of classifying extraversion is 73.5%. This is done by using the Jieba Chinese text segmentation as the tokenizer, the TF scheme for the representation of the document-term matrix, the χ^2-test as the feature selection algorithm, and the SVM as the learning algorithm. One possible way to improve the performance is to extract features from a text using a wider variety of methods such as looking up a dictionary of emotions. Another extension is to incorporate our classifier for extraversion into a personal recommendation system. It would be of interest to see whether the performance of the recommender can be improved. In the paper, we only conducted experiments for *extraversion*. It is not clear whether the other four factors can be accurately classified by using the Facebook posts.

References

[1] "Jieba chinese text segmentation." https://github.com/fxsjy/jieba, accessed July 21, 2014.

[2] L. R. Goldberg, "The structure of phenotypic personality traits." *American psychologist*, vol. 48, no. 1, p. 26, 1993.

[3] S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr, "A very brief measure of the big-five personality domains," *Journal of Research in personality*, vol. 37, no. 6, pp. 504–528, 2003.

[4] R. Hogan, J. A. Johnson, and S. R. Briggs, *Handbook of personality psychology*. Elsevier, 1997.

[5] P. T. Costa and R. R. McCrae, "Normal personality assessment in clinical practice: The neo personality inventory." *Psychological assessment*, vol. 4, no. 1, p. 5, 1992.

[6] "Extroversion." http://www.merriam-webster.com/dictionary/extroversion, accessed July 21, 2014.

[7] J. Otterbacher, "Inferring gender of movie reviewers: exploiting writing style, content and metadata," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 369–378.

[8] B. Fu, J. Lin, L. Li, C. Faloutsos, J. Hong, and N. Sadeh, "Why people hate your app: Making sense of user feedback in a mobile app store," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1276–1284.

[9] 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*. Association for Computational Linguistics, 2006, pp. 627–634.

[10] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski, "Workshop on computational personality recognition: Shared task," in *Proc of Workshop on Computational Personality Recognition, AAAI Press, Melon Park, CA*, 2013.

[11] D. Markovikj, S. Gievska, M. Kosinski, and D. Stillwell, "Mining facebook data for predictive personality modeling," in *Proceedings of the 7th international AAAI conference on Weblogs and Social Media (ICWSM 2013), Boston, MA, USA*, 2013.

[12] M. T. Tomlinson, D. Hinote, and D. B. Bracewell, "Predicting conscientiousness through semantic analysis of facebook posts," *Proc of Workshop on Computational Personality Recognition, AAAI Press, Melon Park, CA*, 2013.

[13] F. Alam, E. A. Stepanov, and G. Riccardi, "Personality traits recognition on social network-facebook," in *Proc of Workshop on Computational Personality Recognition, AAAI Press, Melon Park, CA*, 2013, pp. 6–9.

[14] B. Baharudin, L. H. Lee, and K. Khan, "A review of machine learning algorithms for text-documents classification," *Journal of advances in information technology*, vol. 1, no. 1, pp. 4–20, 2010.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[17] A. Rajaraman and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press, 2011.

[18] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[19] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons,, 1999.

[20] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[21] G. Saucier, "Mini-markers: A brief version of goldberg's unipolar big-five markers," *Journal of personality assessment*, vol. 63, no. 3, pp. 506–516, 1994.