2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation

Personal Trait Analysis Using Word2vec Based on User-generated Text

Guanqun Sun*†, Ao Guo†, Jianhua Ma† and Jianguo Wei*
*School of Computer Software
Tianjin University, Tianjin, China 300350
{sunguanqun, jianguo}@tju.edu.cn
†Graduate School of Computer and Information Sciences
Hosei University, Tokyo, Japan 184-8584
{guanqun.sun.3g, guo.ao.33}@stu.hosei.ac.jp, jianhua@hosei.ac.jp

***Abstract*—Personal trait is to measure the habitual patterns of behavior, thought, and emotion. It differs over individuals and is comparatively stable over time, relatively consistent over situations. Personal trait is significant for it has a lot of applications, such as recommendation system, chatbot and human resource management. It is convenient to recognize personal trait through wearable devices, social media and so on. Traditionally, personal trait is measured in general categories such as Big Five, which contains five traits: extroversion, neuroticism, agreeableness, conscientiousness, and openness. However, it is too abstract to describe personal trait in five aspects. We need the personal trait measured in more specific aspects, such as trait of interest or affect. We can know a person better through the traits in specific aspects than in the traditional abstract ways. In this paper, we proposed a general method of measuring personal trait called Personal Trait Matrix including topic word extraction and the word representation by word2vec based on user-generated text. Then a case study is made with datasets called myPersonality. The diversity of affects and social interactions are measured. Next, the correlation between the trait and the personality of Big Five was analyzed and discussed. The results demonstrate that the proposed method can measure the personal trait in affect and social interactions.**

***Index Terms*—Word2vec, Personal Trait, User-generated Text, Big Five**

I. Introduction

The personal trait is related to the user profile, such as gender, age, etc. But individuals with the same user profile will also exhibit different traits. Unlike the state, the trait is comparatively stable over time. For example, a persons sentiment at a specific moment is a state, yet the positive emotional tendency the person showed is a trait. Studying personal trait has great importance. For example, the study on personal trait can help us improve the performance of the recommendation system by providing more information. We can not only rely on the item that the user has purchased, but also recommend the product according to the user's trait [1], [2]. A neural chatbot with personality is built by [3] which helps us better simulate a specific individual. A better understanding of personal trait can be used to detect and prevent psychological or neurological diseases [4]. In business human resource management, understanding the traits of individuals can make HR recruitment work more efficient [5], [6], [7].

The Big Five personality traits called the five-factor model (FFM) is an abstract taxonomy for personality traits [8]. Big Five contains five traits: extroversion, neuroticism, agreeableness, conscientiousness, and openness. The goals of the study to calculate personality are limited in Big Five personality. However, we should not justify the way we measure personal trait in general categories such as Big Five. Some more specific personal traits in specific aspects need to be measured [9]. For example, the diversity that a person exhibits in social interactions is a specific personal trait. The trait is reflected in whether an individual tends to participate extensively in different social interactions. If an individual used to be in contact with various family and friends, his social interactions could be relatively scattered. If an individual is only in connection with a small group of people, the result is the opposite.

There are a lot of researches on how to recognize personality. Some methods for estimating personality is introduced by [10], such as mobile and wearable devices, social media, questionnaires, etc. However, these studies focus on how to calculate personal traits without paying attention to its evolution over time. The personal trait is a relatively stable variable, but it is not static. For example, a youngster may have a wide range of interests. He is very interested in music, sports, etc. As time goes on, his distribution of interest may change a lot, and his hobby becomes concentrated on only one aspect. Therefore, the stability of a persons trait is also a problem needs to be investigated.

We propose Personal Trait Matrix to solve this problem. The method is based on the user-generated text and method of word2vec. Word2vec is a way to represent words in vectors. Word2vec created by [11] in 2013, has many advantages over the traditional method of word representation. An individuals trait can be reflected by its generated text. For example, a person is emotionally inclined to be positive. In his generated text, the probability of a positive word is greater than the probability of a negative word. On the one hand, word vectors as a multidimensional vector can make calculations more precise by performing operations on vectors. On the other hand, word vectors can be utilized as the input of machine learning algorithms in some research. We can also analyze

978-1-7281-4034-6/19/$31.00 ©2019 IEEE
DOI 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00213

personal traits based on the relationship between word vectors and the change of word vectors over time.

Therefore, in this paper, we review the related work first. Then we proposed a general method of measuring personal trait called personal feature matrix. The personal feature matrix is composed of three steps which are topic word generation, feature matrix generation, and trait analysis. First, we need to extract topic words from user-generated text, because it can reflect the user's personal trait. Next, we convert topic words into word vectors using word2vec. The word vectors construct the personal feature matrix. Furthermore, through the analysis of the personal feature matrix, the personal trait in a specific aspect is obtained. According to the proposed method above, we conducted a case study based on the social media data of text. We got the individuals diversity in the affect and social interaction. The correlation between the traits and the personality of Big Five were analyzed and discussed, for it can be a validation of our experiment.

II. RELATED WORK

A. Word2vec

Word2vec as a method of distributed representation is based on the distributional hypothesis which is proposed by [12]. The distributional hypothesis means to learn the meaning of a word by observing its context words. Word2vec, created in 2013 [11], is a method of using neural networks to generate word embeddings with a large corpus of text. There are two ways to generate a word vector. One is continuous bag-of-words (CBOW), by using neural networks it generates the vector of current word from the surrounding context words. The other one is called skip-gram. As contrary to CBOW, the skip-gram model can produce context words based on the current word.

As a representative of words, word2vec is widely used in the field of natural language processing. Word2vec can be used in document classification problem [13]. Handler used word2vec to study semantic similarity by computing the cosine distance in Wordnet [14]. Wrod2vec is also used to improve the performance of a classifier during Named Entity Recognition [15]. For sentiment analysis of text from different sources, word2vec has also been applied extensively [16], [17]. Also, word2vec as representations of words is applied to improve the consistency and coherence of Machine Translation [18].

The research above and this paper both regard word2vec as a state-of-the-art word embedding methods and use word2vec to map words with similar meaning to vectors with similar representation. But unlike these studies, we use word2vec to construct a vector space model (VSM) called Personal Trait Matrix to study the traits of an individual, rather than measuring the states of the individual at a specific moment.

B. Personal Trait

Personal trait is considered to measure the habitual patterns of behavior, thought, and emotion [19]. States do not last for a while but are expressed in a particular situation and at a special moment. However, the trait is different from the state. The trait can be regarded to be an aspect of personality. Personality is defined as the characteristic set of behaviors, cognitions, and emotional patterns that evolve from biological and environmental factors [20]. The trait differs over individuals and is comparatively stable over time, relatively consistent over situations.

The two famous approaches to study the trait are the three-factor model and Big Five personality traits. The three-factor model is also known as Eysenck Personality Questionnaire. It uses factor analysis to analyze the trait in terms of neuroticism, extraversion, and psychoticism [21]. The Big Five personality traits are abbreviated as the five-factor model (FFM) or the OCEAN model. The meaning of OCEAN can be described as openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. They are abstract descriptions of personality traits [8].

There are lots of papers that use different ways to calculate the trait. Some researchers use the mobile phone and wearable sensors to infer personal trait [22]. Wearable devices are used to collect behavioral evidence such as speaking activity, movement, proximity, face-to-face interactions, and position in the social network. The results of the experiment from 67 nurses show that these clues correlate with personality [23]. An experiment is done by collecting the data of the Electroencephalogram (EEG), Electrocardiogram (ECG), Galvanic Skin Response (GSR) and facial activity data. The data is labeled with Big Five personality scales and emotional self-ratings. Games are also used to study personal trait. The behaviors of 306 participants were analyzed in the video game [24]. It studied the relationship between flow state and character identification.

Lots of research has used data from social media. The three methods of Naive Bayes, k-nearest neighbor(KNN) and support-vector machine(SVM) were used for recognizing Big Five based on the datasets of myPersonality [25]. Convolutional Neural Network(CNN) model is selected by [26] to figure the personality of the Big Five. Based on the prole pictures on Facebook, the prediction of traits is implemented by a bag of visual words technique to extract features [27].

Compared with the research of [22], [23] and [24], we used text from social media instead of data from sensors or picture. The similarity between our research and [25], [26] is that the datasets of myPersonality is used. The difference is that [25] and [26] use machine learning to recognize Big Five, but we propose a general model that can measure personal trait in a specific aspect. The model called personal feature matrix can measure not only the traits of various aspects of an individual but also the changes of the personal trait over time.

III. PERSONAL TRAIT MATRIX GENERATION

In this section, Personal Trait Matrix will be introduced, and the method of building it will be described in detail. First, in general, the three steps of the generating Personal Trait Matrix are introduced. Then, the specific methods of extracting topic words and constructing matrix are described in detail separately.

A. Abstraction of Personal Trait Matrix Generation

As shown in Fig. 1, there are three steps to complete the generation of the matrix. The first step is to get user-generated text. The text is obtained from the users social media, conversations, or articles. It can reflect the user's personal information, for it is generated by the user. Therefore, we can use the text as a data source to infer personal trait. The second step is to extract topic words from text using lexicon, LDA (Latent Dirichlet Allocation) and so on. The third step is to use the word2vec method to represent each topic word in vector and put these vectors together to form a matrix. After analyzing this matrix, the personal trait can be obtained.

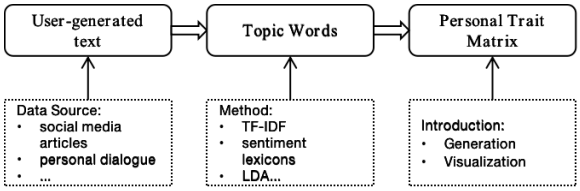

Fig. 1. Personal Trait Matrix generation.

We can get user-generated text from multiple sources, such as social media, article written by the user, dialogue from the instant messaging tool, and so on. For the text are subconsciously generated by the user, it can be used to analyze the personal traits of the user.

B. Topic Word Generation

Three methods for topic words extraction will be introduced, namely TF-IDF, lexicons and LDA.

a) TF-IDF: TF-IDF is short for term frequency-inverse document frequency. It is a mathematical statistic that is intended to reflect how important a word is to a document in a collection or corpus. IDF was able to give weight to the word frequency. TF-IDF is widely used in the field of information retrieval and text mining. The advantage of TF-IDF is that it is concise and can take into account the weight of words. Its disadvantage is that it extracts all kinds of words, not a specific aspect. TF-IDF cannot reflect the trait of a specific aspect of the user.

b) Lexicon: Lexicon, a catalogue of a languages words, is thought to include bound morphemes, which cannot stand alone as words [28]. The following are some commonly used lexicons, such as GI (the General Inquirer), LIWC (Linguistic Inquiry and Word Count), Bing Li U opinion lexicon, SentiWordNet, etc. In some studies, feature extraction was carried out by using sentiment lexicons for sentiment analysis, and the results of state of the art were obtained [29]. Unlike these studies, our study does not need to take advantage of the scale of emotional words, but simply extract words of a specific aspect. The advantage of lexicons is the variety that it can extract from different aspects according to different lexicons, especially for the extraction of words of affect. The disadvantage is that lexicons can only extract the word in the existence of it, and the words cannot be summed up abstractly.

c) Latent Dirichlet Allocation(LDA): Latent Dirichlet Allocation (LDA) is a generative statistical model. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a nite mixture over an underlying set of topics [30]. Research shows that LDA can extract topic words of interest from user-generated text. As an unsupervised learning method, LDA can extract topic words abstractly. Therefore, LDA is especially suitable for the extraction of abstract words such as words of interest.

Among these three methods of topic words extraction, TF-IDF cannot distinguish between different aspects of words, so it is not suitable for building trait matrix of specific aspects. The LDA method is suitable for extracting abstract words. The method of lexicons is more general, for different aspects of words can be extracted to reflect different aspects of personal features.

C. Personal Trait Matrix Generation

The Generation of Personal Trait Matrix has two steps. The first step is to convert all the topic words to vectors by word2vec. The second step is to combine these vectors into a matrix which reflects the personal traits of the user.

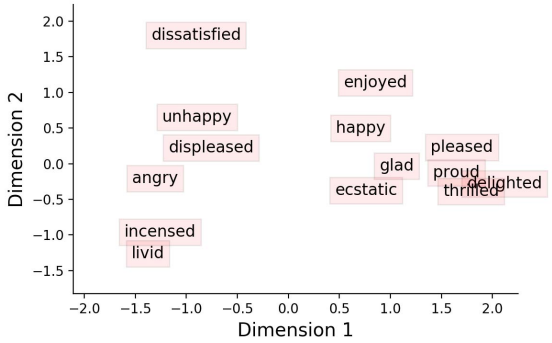

Fig. 2. Word2vec visualization.

After obtaining topic words, the words will be transformed into a vector through word2vec. Fig. 2 is a visualization of some words about sentiment using method of SVD. The word vectors are obtained from Googles pre-trained word2vec model. In the model, every vector have 300 dimensions. In Fig. 2 we can see that the words of positive emotions gather on the right side, and the words of negative emotions gather on the left. It shows that each word vector not only represents a kind of sentiment but also reflects the relationship between topic words. If the word vectors are close to each other in vector space, the meaning of the words will be relatively similar.

Personal Trait Matrix, constructed by word2vector, has the following benefits in describing personal traits.

Word2vec is a more accurate way to describe personal trait. In some traditional method of measuring personal trait, the result is one-dimensional. Like the words about positive sentiment, happy and satisfied can be measured with a scale to represent the extent of the positive sentiment. But in

the method of word2vec, each word is represented by a multidimensional vector which can be more accurate. A more accurate and solid result can be obtained.

The word2vec provide a new way to explore personal trait. For example, through the relationship of vectors in multidimensional space, whether the vectors are accumulated or not can be obtained by computing the distance of the vectors. Thus, the personal trait can be inferred from the vectors. Also, the evolution of Personal Trait Matrix over time can be observed form the change of the vectors.

IV. PERSONAL TRAIT ANALYSIS

In this section, the analysis of Personal Trait Matrix will be discussed. First, we describe what Personal Trait Matrix is. Then, the features of Personal Trait Matrix and personal trait calculation are introduced.

A. Personal Trait Matrix of Different Individuals

An example is given to show why Personal Trait Matrix can describe personal traits. we suppose there are two individuals A and B. According to the method described in the previous chapter, individual As topic words of interest are obtained as football, worldcup, basketball, sport, cake. Individual A is a person who is interested in sports. And individual Bs topic words of interest are obtained as jazz, running, impressionism, film, cake, bread. They are mostly related to music. Individual Bs interests are extensive. The topic words are converted into word vectors using word2vec. Then, these word vectors are put together to form a matrix, which is the Personal Trait Matrix of Interest. The Personal Trait Matrix of Interest of individual A and individual B is visualized by SVD.

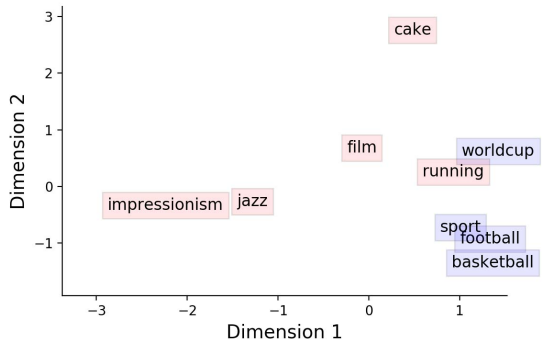

Fig. 3. Comparison of individual trait description from two different individuals.

As is shown in Fig. 3, the topic words on a red background belong to individual A, and the topic words on a blue background belong to individual B. The word vectors of A, such as worldcup, sport, football and basketball are gathered in the lower right corner of the image. However, the distribution of the word vector of B is very scattered. The distribution reflects the different characteristics of individual A and individual B in interest. The interest of A is focused on the topic of sport. Compared to individual A, individual B has a more extensive interest. In this example, we can learn that the Personal Trait Matrix can represent personal characteristics.

B. Features of Personal Trait Matrix

The Personal Trait Matrix has three features, namely multimodal, incremental and dynamic.

a) Multimodal: Using data from multimodal sources provide the possibility of understanding natural phenomena deeply. Multimodal data can be used during the construction of the Personal Trait Matrix. For audio data, it can be converted to text. For image data, machine learning can be used to obtain topic words from it.

b) Incremental: Personal Trait Matrix is incremental due to the increasing data. As more and more user data are acquired, more topic words are obtained. Therefore, more word vectors are added to the Personal Trait Matrix. As the Personal Trait Matrix grows, it can measure the individual more accurately.

c) Dynamic: As data is periodically updated, the Personal Trait Matrix changes asynchronously over time. As the matrix changes, the personal trait may change. For example, an individual may show a wide range of interest for a certain period of time and show a not extensive interest in the next period. Such dynamic changes can be inferred from the Personal Trait Matrix.

C. Personal Trait Calculation

Personality Trait of Diversity is the condition of having or being composed of differing elements in a specific aspect of an individual. As is shown in Fig. 3, the word vectors of individual A are gathered in the lower right corner of the image. It means that the interest of individual A is not extensive. However, the distribution of the word vectors of individual B shows that the interest of individual B is very extensive. The personal trait of interest diversity is used to measure such difference.

Also, affective diversity can be measured in a similar way. If ones emotions remain in a calm state of neither excitement nor depression. In this case, the affective diversity scale is lower. In the opposite case, the scale of affective diversity is higher.

Among methods of measuring the similarity between two words in the vector space model, to compute the cosine similarity between the word vectors is the most common way [31]. Cosine function is widely used to measure features in vector space model [32].

Inner product of the vectors is used to measure the cosine of the angle between them. Let $\mathbf{a} \cdot \mathbf{b} \in \mathbb{R}^{\mathbb{D}}$ be word vectors in a D-dimensional vector space. $\mathbf{a} \cdot \mathbf{b}$ is the dot product of vector $\mathbf{a}$ and vector $\mathbf{b}$. $\|a\|$ and $\|b\|$ is the ℓ_2-norm of them. The cosine formula is shown in (1).

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|a\| \, \|b\|} \tag{1}$$

Let the Personal Trait Matrix consist of n vectors. Let V_i and V_j be the vectors in the matrix. $d_{i,j}$ represents the cosine

distance of V_i and V_j. $d_{i,j}$ equals to one minus $\cos(V_i, V_j)$. $\sum_{j=1}^{n}(d_{ij} - d_{ii})$ represents all the cosine distance of V_i and other vectors in the matrix. Equation (2) represents the diversity of the vectors in the matrix.

$$Div = \sum_{i=1}^{n} \frac{\sum_{j=1}^{n}(d_{ij} - d_{ij})}{n} \quad (2)$$

In addition to measuring diversity, other kinds of traits can also be obtained similarly. For example, we can infer the stability of affect by observing the change of affect through the changes of word vectors over time.

V. CASE STUDIES ON PERSONAL TRAIT ANALYSIS

In this section, a case study will be implemented and analyzed according to the Personal Trait Matrix presented in the previous section including datasets, methodology, and experimental results.

A. Datasets

The myPersonality, issued by David Stillwell in 2007, was a popular Facebook application. Users can do psychometric tests in myPersonality and get the results immediately. At the same time, if the user allows his results of psychological test and other information on Facebook to be obtained. That is how the information from Facebook and psychometric test results together formed the labeled datasets. A subset dataset of the myPersonality is used in this paper. It contains the text data of personal status and the score of Big Five tests including 250 users. The test of Big Five reflects the individuals Big Five personality traits. The descriptive statistics of Big Five is described in Table I. The mean value, standard deviation, maximum, and minimum can be observed from it.

TABLE I
DESCRIPTIVE STATISTICS OF DATASETS

Descriptive Statistic	Big Five				
	OPN	***CON***	***EXT***	***AGR***	***NEU***
Mean	4.079	3.523	3.292	3.600	2.627
Standard Deviation	0.575	0.740	0.861	0.671	0.077
Minimun	2.250	1.450	1.330	1.650	1.250
Maximum	5.000	5.000	5.000	5.000	4.750

As we stated before, a lot of research studied on how to recognize Big Five personality from social media [25], [26]. Unlike these studies, what we focus on is not how to calculate personality but the personal trait of a specific aspect. Besides, we need to analyze the relationship between personal trait and the Big Five personality as a validation to the experiment. That is why the datasets of text from social media labeled with Big Five is selected.

Another important thing about the datasets is the preprocessing of the data. Here, the data preprocessing is divided into two steps. The first step is the processing of the text of status. In the datasets, one user corresponds to many Facebook statuses. We combine a lot of Facebook statuses of a user into a piece of text. The combination helps to the subsequent steps of the topic generation. The second step in preprocessing the data is to preprocess the words. Normalization of the text is needed to make it be unified into lowercase. Also, we need to perform stemming and lemmatization on the words. Let the words shift to the original form. In this paper, we use the NLTK tool to complete the processing of words. NLTK, the abbreviation of Natural Language Toolkit, is a natural language processing (NLP) toolbox for English.

B. Methodology

As explained in the preceding section, there are several ways to extract topic words as feature words, such as TF-IDF, lexicons or LDA. In this case study, we use a lexicon called LIWC2015 to extract topic words.

TABLE II
CATEGORIES AND WORDS OF LIWC

Category	Examples
Function Words	a, about, above, absolutely, across, actually, after, again, against, ahead, almost, along, already
Affect Words	cheer, cheerful, cheers, coldly, comforting, concerned, confidence, cool, crazy, created
Social Words	admit, admits, admitted, admitting, adult, advice, ally, army, ask, awkward, awkwardness, babies
Cognitive Process	absolute, accept, accepted, acknowledge, adjust, admit, admitted, affect, affecting, against, allow
Cognitive Process	acid, appear, appeared, audible, beautiful, beauty, bitter, bitterly, black, blonde, blue, bright

LIWC, an abbreviation for Linguistic Inquiry and Word Count, is developed to provide an efficient and effective method for studying the various emotional, cognitive, and structural components present in individuals verbal and written speech samples [33]. As a vocabulary classification dictionary, LIWC2015 is the third version, and it contains approximately 6,400 words. Each word is assigned to one or more categories. As an example, some Categories and words belonging to it are shown in Table II.

Firstly, the generation of topic words is done by LIWC lexicons. We first get the vocabulary database of LIWC2015 and restore the vocabulary to the original form. The vocabulary related to affect and social interaction are taken as lexicons. By comparing the words in datasets with the words in lexicons, we extract the corresponding words to form a list of personal topic words. For example, if a user's status is I admit that I am cheerful. The word admit will be extracted as word of social interaction. And the word cheerful will be extracted as a word of affect. For the statistical significance of the data, users with less than ten topic words are treated as outliers. These outliers are not included in subsequent calculations.

Secondly, we implement the generation of the word vector to form the Personal Trait Matrix. As mentioned in the previous chapter, we convert topic words into word vectors through word2vec. Googles pre-trained word2vec with 300 dimensions is used here. This tool which is developed in 2013 gives a useful implementation of the continuous bag-of-words and skip-gram architectures to computing vector

representations of words. Because we have normalized the words before, most of the words can find the corresponding word vector. Word vectors of an individual are put together to form the Personal Trait Matrix.

Thirdly, we analyze the Personal Trait Matrix. Since we have extracted the words of two specific aspects, which are affect and social interaction in LIWC. Depending on 1 and 2 in the previous chapter, we perform the calculation of the diversity for each matrix. Also, we calculate the Pearson Correlation of diversity and the scale of Big Five personality to explore the relationship between the personal traits and personality.

C. Experimental Results

We implement the experiment based on the dataseta called myPersonality with about 250 users. LIWC2015 is used as the lexicon to generate topic words. Googles pre-trained word2vec with 300 dimensions is adopted to represent topic words in vectors. And based on the distance of the vectors, we can compute the diversity of the vectors as the traits of affect and social interaction.

Fig. 4 is a bar chart about the diversity of affection. The horizontal axis represents different individuals, and the height of the vertical axis reflects the level of the diversity scale. Among them, the maximum of the diversity is 254.62, the minimum is 6.07, and the average amount is 39.13. If some individuals have a high scale of diversity of affect, this indicates that they show a richer emotion. For example, they may feel happy sometimes and sometimes feel sad.

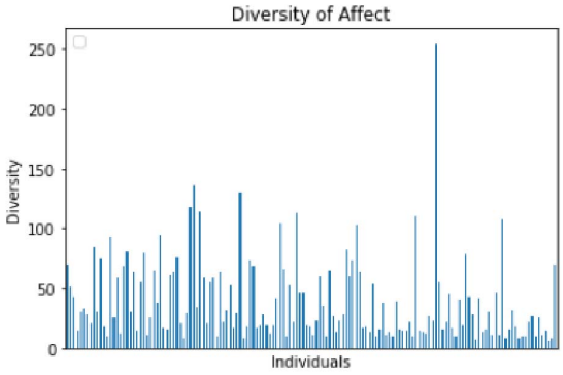

Fig. 4. Diversity of social affect.

Fig. 5 is a bar chart about the diversity of social interaction. Also, the horizontal axis represents different individuals, and the height of the vertical axis reflects the scale of the diversity. Among them, the maximum of the diversity is 185.96, the minimum is 6.04. It has an average of 40.05 and a standard deviation of 34.88. If some individuals have high scales of diversity of social interaction, this means they will participate in a broader range of social interactions. For example, some individuals may often be in contact with different family and friends, while others may be the opposite.

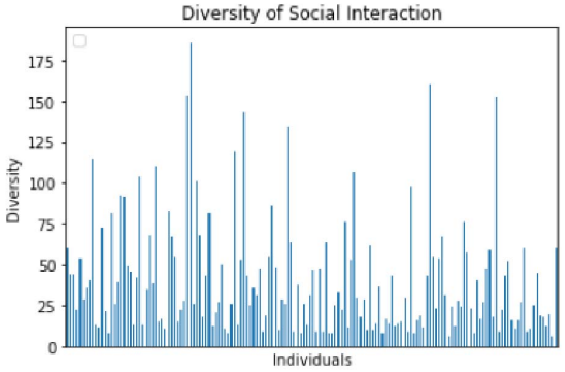

Fig. 5. Diversity of social interaction.

Our research goal is to calculate diversity of affect and social interaction from the text. Compared with Bif Five, these traits are more specific projections of an Individual. These specific traits must have some relevance with the abstract feature Big Five. Therefore We calculate the correlation between Diversity of Affect and Social Interaction and Big Five.

We can see the relationship between the personal traits and personality from TABLE III. The diversity of affect and social interaction are regarded as personal traits. The personality is represented by the scale of Big Five. The diversity of affect and social interaction both have a strong correlation with Extraversion. The result can be explained by common sense that an extroverted person is often emotionally rich and prefers to have more social interactions.

TABLE III
RELATIONSHIEP BETWEEN PERSONAL TRAIT AND BIG FIVE

Scale of Diversity	**Big Five**				
	OPN	***CON***	***EXT***	***AGR***	***NEU***
Affect	0.0046	0.0911	0.1580	0.0339	0.0356
Social Interaction	0.0578	0.0335	0.1442	0.0276	0.0305

From the above analysis, we can see that diversity of affect and social interaction are the specific traits of individuals. There are certain correlations between them and Big Five, which can also be explained. Therefore, these results of the case studies suggest that the Personal Trait Matrix is a feasible way to study personal traits. On the one hand, diversity is a trait that reflects the individuals stability, not a state at some point. Each individual exhibits a different diversity means that each show different personal traits. On the other hand, we find that there is a specific correlation between personal traits and personality. The relationship can be seen as the validation of the case study, which indicates that the Personal Trait Matrix can reflect personal traits.

Also, there are two reasons for the lack of a strong correlation. One is because Big Five personality is a very abstract description of personal traits. The diversity we calculate is a subdivision in the personal trait. Having a very strong

correlation between the two is difficult. Another reason is a lack of large volume of the corpus.

VI. CONCLUSION AND FUTURE WORK

In this paper, we first describe what is personal trait, and then proposed a model called Personal Trait Matrix to measure the personal traits in specific aspects through user-generated text. First, topic words should be extracted by means such as LDA or lexicons. Then, we need to use word2vec to generate the word vectors and to form the trait matrix. We implemented a case study to calculate the diversity of different individuals in both the affect and social interaction. We found that there are significant differences in the diversity of different individuals. We also found that diversity correlates with Big Five personality. The experiment results show that diversity is a kind of personal traits and our proposed model of Personal Trait Matrix can effectively measure personal trait.

In the future study, improvement can be made in three aspects. First, from the perspective of datasets, we can label the data through some questionnaires and other information; the results of the experiment can be evaluated. We can use multi-modal data for fusion, such as data from pictures, audio, etc. Second, more methods can be used to extract topic words. Third, the case study of more personal traits can be measured such stability of interest over time.

ACKNOWLEDGMENT

This work is partially supported by the Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research (No.18K11408).

REFERENCES

[1] B. Ferwerda and M. Schedl, "Personality-based user modeling for music recommender systems," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2016, pp. 254–257.

[2] R. P. Karumur, T. T. Nguyen, and J. A. Konstan, "Personality, user preferences and behavior in recommender systems," *Information Systems Frontiers*, vol. 20, no. 6, pp. 1241–1265, 2018.

[3] H. Nguyen, D. Morales, and T. Chin, "A neural chatbot with personality," 2017.

[4] A. R. Sutin, A. B. Zonderman, L. Ferrucci, and A. Terracciano, "Personality traits and chronic disease: Implications for adult personality development," *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 68, no. 6, pp. 912–920, 2013.

[5] I.-S. Oh, S. Kim, and C. H. Van Iddekinge, "Taking it to another level: Do personality-based human capital resources matter to firm performance?" *Journal of Applied Psychology*, vol. 100, no. 3, p. 935, 2015.

[6] X. Wang, L. T. Yang, L. Kuang, X. Liu, Q. Zhang, and M. J. Deen, "A tensor-based big-data-driven routing recommendation approach for heterogeneous networks," *IEEE Network*, vol. 33, no. 1, pp. 64–69, 2018.

[7] X. Wang, L. T. Yang, X. Chen, M. J. Deen, and J. Jin, "Improved multi-order distributed hosvd with its incremental computing for smart city services," *IEEE Transactions on Sustainable Computing*, 2018.

[8] S. Rothmann and E. P. Coetzer, "The big five personality dimensions and job performance," *SA Journal of Industrial Psychology*, vol. 29, no. 1, pp. 68–74, 2003.

[9] X. Wang, L. T. Yang, H. Liu, and M. J. Deen, "A big data-as-a-service framework: State-of-the-art and perspectives," *IEEE Transactions on Big Data*, vol. 4, no. 3, pp. 325–340, 2017.

[10] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[12] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[13] A. Ahmad and M. R. Amin, "Bengali word embeddings and it's application in solving document classification problem," in *2016 19th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2016, pp. 425–430.

[14] A. Handler, "An empirical study of semantic similarity in wordnet and word2vec," 2014.

[15] S. K. Sienčnik, "Adapting word2vec to named entity recognition," in *Proceedings of the 20th nordic conference of computational linguistics, nodalida 2015, may 11-13, 2015, vilnius, lithuania*, no. 109. Linköping University Electronic Press, 2015, pp. 239–243.

[16] C. Dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 69–78.

[17] A. Severyn and A. Moschitti, "Twitter sentiment analysis with deep convolutional neural networks," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015, pp. 959–962.

[18] E. M. Garcia, J. Tiedemann, C. España-Bonet, and L. Màrquez, "Word's vector representations meet machine translation," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2014, pp. 132–134.

[19] S. M. Kassin, *Essentials of psychology*. Prentice Hall, 2003.

[20] P. J. Corr and G. Matthews, *The Cambridge handbook of personality psychology*. Cambridge University Press Cambridge, UK:, 2009.

[21] H. J. Eysenck and S. B. G. Eysenck, *Manual of the Eysenck Personality Questionnaire (junior and adult)*. Hodder and Stoughton, 1975.

[22] D. O. Olguın, P. A. Gloor, and A. S. Pentland, "Capturing individual and group behavior with wearable sensors," in *Proceedings of the 2009 aaai spring symposium on human behavior modeling, SSS*, vol. 9, 2009.

[23] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "Ascertain: Emotion and personality recognition using commercial sensors," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, 2018.

[24] A. R. B. Soutter and M. Hitchens, "The relationship between character identification and flow state within video games," *Computers in Human Behavior*, vol. 55, pp. 1030–1038, 2016.

[25] B. Y. Pratama and R. Sarno, "Personality classification based on twitter text using naive bayes, knn and svm," in *2015 International Conference on Data and Software Engineering (ICoDSE)*. IEEE, 2015, pp. 170–174.

[26] M. Yuan, Y. Chen, Y. Huang, and W. Lu, "Behavioral and metabolic phenotype indicate personality in zebrafish (danio rerio)," *Frontiers in physiology*, vol. 9, p. 653, 2018.

[27] F. Celli, E. Bruni, and B. Lepri, "Automatic personality and interaction style recognition from facebook profile pictures," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 1101–1104.

[28] D. Sandra and M. Taft, *Morphological structure, lexical representation and lexical access*. Taylor & Francis, 1994.

[29] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[30] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[31] M. Faruqui, Y. Tsvetkov, P. Rastogi, and C. Dyer, "Problems with evaluation of word embeddings using word similarity tasks," *arXiv preprint arXiv:1605.02276*, 2016.

[32] G. Sidorov, A. Gelbukh, H. Gómez-Adorno, and D. Pinto, "Soft similarity and soft cosine measure: Similarity of features in vector space model," *Computación y Sistemas*, vol. 18, no. 3, pp. 491–504, 2014.

[33] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," Tech. Rep., 2015.