法政大学学術機関リポジトリ
HOSEI UNIVERSITY REPOSITORY

Word2Vec-based Personal Trait Computing from User-generated Text

著者	Sun Guanqun
出版者	法政大学大学院情報科学研究科
journal or publication title	法政大学大学院紀要. 情報科学研究科編
volume	15
page range	1-6
year	2020-03-24
URL	http://doi.org/10.15002/00022718

Word2Vec-based Personal Trait Computing from User-generated Text

Guanqun Sun
Graduate School of Computer and Information Sciences
Hosei University
Tokyo 184-8584, Japan
guanqun.sun.3g@stu.hosei.ac.jp

***Abstract*—Personal trait is a habitual pattern for measuring behavior, thoughts, and emotions. It varies over individuals and is relatively stable in different situations over time. The personal trait is of great significance since it can be used in many applications, such as recommendation system, chatbot, and human resource management. Personal traits are easily recognized through wearable devices, social media, and the like. Most of the existing studies focus on user profile, behavior and personality. Specially, user profile and behavior are a person's manifestations that cannot accurately capture a person's internal characters. Personality is generally calculated by Big Five, which is obscure for non-psychologists. Generally, specific personal traits are especially critical in many aspects, such as disease detection, individual understanding, etc. Therefore, measuring more specific personal traits is essential. Given this, this paper proposes a word2vec-based general method for personal traits computing, which mainly includes topic word extraction, personal trait matrix generation, and personal trait computing. Furthermore, a case study is conducted to verify the effectiveness of the proposed method, and further analysis is provided to validate the methods.**

Keywords—Word2vec; Personal Trait; User-generated Text; Emotional Fluctuation; Sociable Tendency.

I. INTRODUCTION

Personal trait is considered a habitual pattern for measuring behavior, thoughts, and emotions [1]. Personal feature computing is used to calculate a series of personal characteristics, such as user profile, behavior and personality. User-generated text is any form of text that users post on an online platform such as social media. In the area of trait computing, a key issue is how to measure more specific personal traits (e.g., interest, emotional fluctuation) that are emerging and becoming more popular. For example, studies on personal traits can help improve the performance of the recommendation system by providing more information available. We can not only rely on the item purchased by the users but also recommend the products according to users' traits [2] [3]. In addition, a neural chatbot with personality [4] can help better simulate a specific individual, and a better understanding of personal traits can be used to detect and prevent psychological or neurological diseases [5], which also can be applied into human resource management for more efficient HR recruitment [6].

The common method of computing personal traits is to fit the labeled datasets by machine learning method [7]. The limitation of this method is that the specific traits cannot be computed without labeling. Word2vec, a distributed representation of words, can be used in personal traits computing from user-generated text. There are two advantages in this proposed computing method based on word2vec as follows. On the one hand, based on the word2vec, we can measure the distribution of multidimensional vectors in the vector space model to make the calculation more accurate. On the other hand, unlike physiological data and images, we can obtain texts for a long time to reflect the user's stable personal traits.

The existing research on user modeling mainly focuses on user profile, behavior and personality. Among them, personal traits are related to the user profile, such as gender and age, but individuals with the same user profile can also exhibit different traits. In addition, behavior is a person's external performance, which cannot be used to accurately infer a person's inner character. Thus, there are numbers of studies on how to recognize personality. Especially, some researchers have tried to predict the Big Five personality, which is an abstract taxonomy of personality. However, the Big Five is very abstract and incomprehensible to non-psychologists, so some more specific personal traits need to be measured.

The main contributions of this paper include the following three aspects. First, a word2vec-based personal trait computing method is proposed, which can be used to compute personal trait in various specific aspects. Second, we conducted a case study of computing emotional fluctuation and sociable tendency using the datasets of myPersonality, among them, emotional fluctuation refers to the tendency to experience emotions that change quickly [8], while sociable tendency is defined as the degree to which individuals tend to socialize in social groups such as family and friends. Third, we verified the validity of the experiment through manual evaluation.

In this paper, we review the relevant literature in section II. The section III details the method of word2vec-based personal trait computing. In section IV, a case study is conducted to verify the effectiveness of the proposed method. In the last section, the paper is summarized.

II. RELATED WORK

Distributed representation of word which is proposed by [9] is a method of distributed representation based on the distributional hypothesis. The distributional hypothesis means to learn the meaning of a word by observing its context words. Word2vec, created in 2013 [10], is a method of using neural networks to generate word embeddings with a large corpus of text.

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

---

Supervisor: Prof. Jianhua Ma

moment. Personal trait is considered to measure the habitual patterns of behavior, thought, and emotion [1]. 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.

Many personal characteristics can be measured. Many researchers have calculated the above types of personal information: profile, behavior, personality. Many studies have predicted user profiles. Based on word clusters and embeddings, occupational class was predicted for a public user profile [11]. Age and gender were predicted as social variables from data collected through an online game [12]. In addition to predicting gender and age, the degree of religiosity and IT background status were also predicted, in which different knowledge sources are used [13]. Demographic attributes such as age and gender were studied based on his or her full-text items based on word2vec word embeddings [14]. Unlike user profile, some researchers have studied user behavior through social media. A total of 53,226 Facebook user profiles were analyzed to study the relations between personality types and user preferences in multiple entertainment domains [15]. Personality is defined as the characteristic set of behaviors, cognitions, and emotional patterns that evolve from biological and environmental factors [16]. 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 [17].

The trait is reflected in whether an individual tends to participate extensively in different sociable tendency. For example, the diversity that a person exhibits in sociable tendency is a specific personality. If an individual used to be in contact with various family and friends, his sociable tendency could be relatively scattered. If an individual is only in connection with a small group of people, the result is the opposite. Compared with profile, behavior, and personality, the personal trait has a unique advantage to let us know a person better. The personal trait is a more in-depth description of a person than the profile. The personal trait is a higher level than behavior. The personal trait can reflect a more stable character of a person.

III. Word2vec-based Personal Trait Computing

In this section, we introduce the computing process of personal trait here. First, we introduce the process of personal trait computing. Second, we describe the specific techniques of extracting topic words and the generation of the personal trait matrix in detail separately. Finally, we show the method of personal trait calculation.

A. Process of Word2Vec-based Personal Trait Computing

There are several types of word representation. Among them, one-hot encoding and distributed representation are often used. In distributed representation, word2vec is the typical one. Suppose we need to represent the following three words: “happy”, “funny”, “sad”. In Equation (1), these words are represented by the one-hot encoding, in which every word was represented as a vector with all 0s one 1. However, one-hot encoding has the following limitations. First, since there are 13 million tokens for the English language. We need at least 13 million dimensions vectors to represent all tokens. Second, some unknown tokens cannot be represented in advance. Third, the relationship between words cannot be expressed. For example, “Happy” and “funny” are related in meaning. In one-hot encoding, we can’t see their relationship.

$$w^{happy} = \begin{bmatrix}1\\0\\0\end{bmatrix}, w^{funny} = \begin{bmatrix}0\\1\\0\end{bmatrix}, w^{sad} = \begin{bmatrix}0\\0\\1\end{bmatrix} \quad (1)$$

Fig. 1 was a visualization of some example words about sentiment using method of SVD (Singular Value Decomposition). The word vectors were obtained from Googles pretrained word2vec model. In the model, every vector had 300 dimensions. In Fig. 1 we could see that the words of positive emotions gather on the right side, and the words of negative emotions gather on the left.

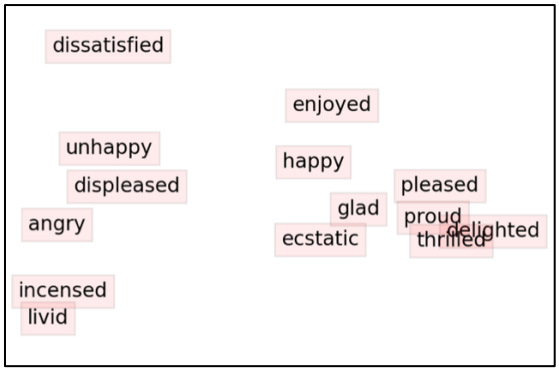

Fig. 1. Example of word2vec distribution.

Personal Trait Matrix, constructed by word2vector, had the following benefits in describing the personal trait. On the one hand, word2vec was a more accurate way to describe the personal trait. In some traditional method of measuring personal trait, the result was one-dimensional. Like the words about positive sentiment, happy and satisfied could be measured with a scale to represent the extent of the positive sentiment. But in the method of word2vec, each word was represented by a multidimensional vector which can be more accurate. A more accurate and solid result could be obtained. On the other hand, 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.

We introduce the abstraction process of personal trait computing in Fig. 2. The first step is to extract topic word from the user-generated text. We could get the user-generated text from multiple sources, such as social media, Life Diary, SMS log, and so on. Since the text is generated unconsciously by users, from which we can get more objective results. We think TF-IDF, Lexicon, and LDA can be used to extract topic words. These methods could help us to extract specific topic words. The second step was to generate the personal trait matrix. In this step, we proposed the definition of personal trait matrix and analyzed its features. We also explained why it could embody personal traits. Based on the topic words from the first step, we converted topic words into vectors and formed a matrix. The matrix reflect the personal traits of some aspect. The third step is to calculate the personal trait. We use the word2vec method to represent each topic word in vector and put these vectors together to form a matrix. By calculating the

distribution of vectors in the matrix, we could infer the personal trait.

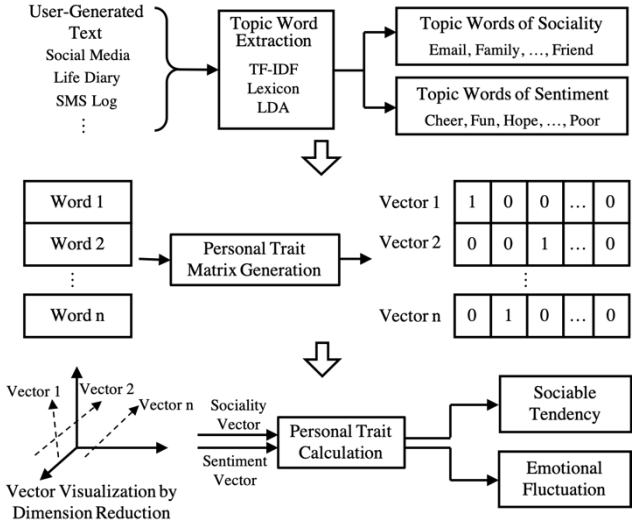

Fig. 2. Process of word2vec-based personal trait computing.

B. Topic Word Extraction

We use three methods for topic words extraction, namely TF-IDF, lexicons, and LDA. 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. Lexicon, a catalog of a language words, is thought to include bound morphemes, which cannot stand alone as words [18]. The lexicons, such as GI (the General Inquirer), LIWC (Linguistic Inquiry and Word Count), Bing Li U opinion lexicon, SentiWordNet, etc are commonly used. 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 mixture over an underlying set of topics [19].

We choose lexicon as the method of topic word extraction. For 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.

TABLE I. 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
Perceptual Process	acid, appear, appeared, audible, beautiful, beauty, bitter, bitterly, black, blonde, blue, bright

LIWC 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 [20]. 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 I.

C. Personal Trait Matrix Generation

Personal trait matrix is an important method for calculating personal characteristics. In this part, we introduce the generation of personal trait matrix from process, advantage, and principle. Personal trait matrix is a matrix composed of user's topic word vectors in specific aspects. Personal trait matrix generation refers to the process of transforming topic words into word vectors and forming matrices. The generation of personal trait matrix had two steps. The first step was to convert all the topic words to vectors by word2vec. The second step was to combine these vectors into a matrix which reflects the personal traits of the user.

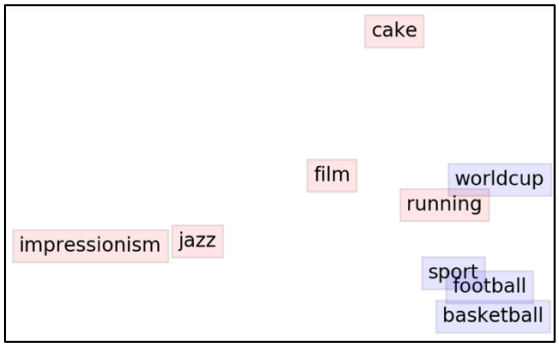

Fig. 3. Comparison of topic words from two individuals.

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 A's topic words of interest are obtained as football, worldcup, basketball, sport, and cake. Individual A is a person who is interested in sports. And individual B's topic words of interest are obtained as jazz, running, impressionism, film, cake, and 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. As is shown in Fig. 3, the topic words on a red background belong to individual B, and the topic words on a blue background belong to individual A. 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.

D. Personal Trait Calculation

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

fluctuation if one's emotions remain in a calm state of neither excitement nor depression. In this case, the level of emotional fluctuation is lower. In the opposite case, the degree of emotional fluctuation 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 [21]. The cosine function is widely used to measure features in vector space model [22].

Inner product of the vectors is used to measure the cosine of the angle between them. Let a, b $\in \mathbb{R}^D$ be word vectors in a D-dimensional vector space. a · b is the dot product of vector a and vector b . $\|a\|$ and $\|b\|$ is the ℓ_2-norm of them. The cosine formula is shown in (2).

$$\cos(a, b) = \frac{a \cdot b}{\|a\|\|b\|} \quad (2)$$

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 $\cos(V_i, V_j)$. Cosine distance is expressed as $1 - \cos(V_i, V_j)$. $\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}$ represents sum of the cosine distance of V_i and other vectors in the matrix. n^2 represents the number of vectors that have been computed. (3) represents the average distance of the vectors in the matrix.

$$Dis = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}}{n(n-1)} \quad (3)$$

Next, we give an example to illustrate the usage of (1) and (2). We assume that there are two individuals, A and B. Using the method introduced before, we extract topic words of A as "cheerful", "funny" and "happy". Similarly, the topic words of B we get are "cheerful" and "sad". Here, we assume that the word vector corresponding to "cheerful" obtained by word2vec is $V_1 = [1,1,0]$. Similarly, we assume that "funny", "happy" and "sad" correspond to word vectors $V_2 = [1,2,0], V_3 = [1,3,0]$ and $V_4 = [0,0,1]$ respectively.

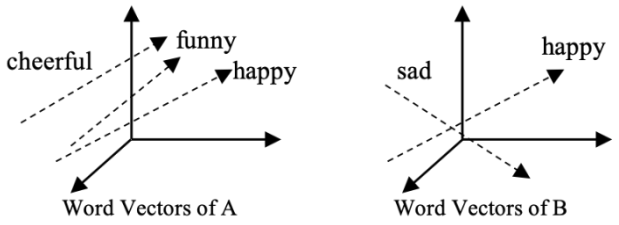

Fig. 4. Sentiment vectors of individuals A and B.

We assume that personal trait matrix of A and B is visualized by dimension reduction as shown in the Fig. 4. In Fig. 4, we can see that the average distance of the word vectors of A is less than that of the word vectors of B. In addition to measuring emotional fluctuation, we can infer other kinds of traits in the same way. For example, we can obtain a sociable tendency by observing the distribution of sociable words.

IV. Case Studies on Personal Trait Computing

In this section, a case study is implemented and analyzed according to the method presented in the previous section, including emotional fluctuation and sociable tendency, datasets, experimental process and evaluation method, computed results, and evaluation.

A. Emotional Fluctuation and Sociable Tendency

In this experiment, our objectives are to compute emotional fluctuation and sociable tendency.

Emotional fluctuation refers to the tendency to experience emotions that change quickly [8]. Here, we define Emotional fluctuation as the degree to which an individual's emotions vary across time. If an individual is not particularly happy or sad for some time and remains calm, the degree of emotion fluctuation is relatively low. In contrast, during this period, if a person experiences many emotional states such as cheerful, funny, etc., the degree of emotional fluctuation is relatively high.

Sociable tendency is defined as the degree to which individuals tend to socialize in social groups such as family and friends. If an individual has extensive social contact with family and friends over a period of time, the degree of sociable tendency is higher for him. On the contrary, during this period, if an individual only contacts a few friends, the degree of sociable tendency is relatively low.

B. Datasets

The myPersonality, issued by David Stillwell in 2007, was a popular Facebook application. We selected a subset dataset of the myPersonality in this thesis. It contained the text data of personal status and the score of Big Five tests, including 250 users.

An important thing about the datasets was the preprocessing of the data. As a result, we divided the data preprocessing into two steps. The first step was the processing of the text of status. In the datasets, one user corresponded to many Facebook statuses. We combined a lot of Facebook statuses of a user into a piece of text. The combination helped to the subsequent steps of the topic generation. The second step in preprocessing the data was to preprocess the words. Normalization of the text was to unify it into lowercase. Also, we performed stemming and lemmatization on the word, which made the words shift to the original form. In this paper, we used 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.

C. Experimental Process and Evaluation Method

In this case study, we used a lexicon called LIWC2015 to extract topic words.

Firstly, the generation of topic words was done by LIWC lexicons. We first got the vocabulary database of LIWC2015 and restored the vocabulary to the original form. The vocabulary related to affect and social interaction were taken as lexicons. By comparing the words in datasets with the words in lexicons, we extracted the corresponding words to form a list of personal topic words. Secondly, we implemented the generation of the word vector to form the Personal Trait Matrix. As mentioned in the previous chapter, we converted topic words into word vectors through word2vec. We used Googles pre-trained word2vec with 300 dimensions. Word vectors of an individual were put together to form the Personal Trait Matrix. Thirdly, we analyzed the Personal Trait Matrix. We extracted the words of emotional fluctuation and sociable tendency. Depending on Equation (2) and (3) in the previous chapter, we performed the calculation of the distribution for each matrix. Also, we calculated the Pearson Correlation of distribution and the scale of the Big Five personality to

explore the relationship between personal traits and personality.

Since our calculation process is unsupervised, the computing results were not labeled in the dataset. Under such circumstances, it becomes very important to verify the experimental results. Only by validating the experimental results can we demonstrate the validity of the proposed method. In supervised learning, computational objectives are labeled in datasets. However, there are some problems in this way of labeling. On the one hand, tagging data through questionnaires is a control experiment. In this state, it becomes unnatural for the experimenter to answer questions. It increases the randomness of the experimental results. Secondly, since experimenters want to protect their privacy, and they may intentionally fail to answer questions correctly.

In this research, we will choose the way of manual evaluation by others. Such an evaluation method has the following advantages. First, the objective evaluation of others is often more accurate than the subjective evaluation of oneself. Second, unlike filling out questionnaires, allowing others to evaluate can control the environment and eliminate experimental interference. Third, for the calculation results of unsupervised learning, manual evaluation can be used.

D. Computed Results

Based on the distance of the vectors, we can compute the diversity of the vectors as the traits of sentiment consistency and sociable tendency.

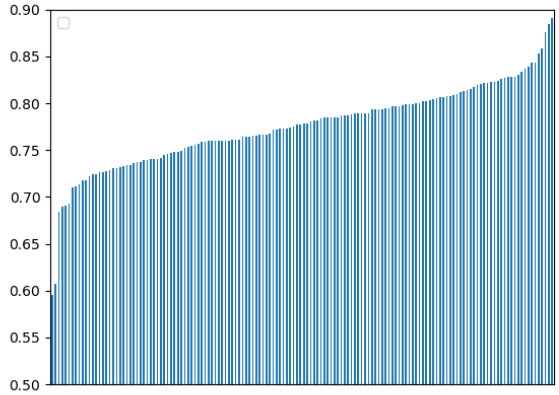

Fig. 5. Emotional fluctuation of different individuals.

Fig. 5 is a bar chart about emotional fluctuation. The horizontal axis represents different individuals of 148, and the height of the vertical axis reflect the level of the emotional fluctuation scale. We normalize the level of emotional fluctuation to 0-1. We regard the maximum of emotional fluctuation as 1 and the minimum as 0. The average amount is 0.775 and the standard deviation is 0.044.

Fig. 6 is a bar chart about the diversity of sociable tendency. Also, the horizontal axis represents different individuals of 156, and the height of the vertical axis reflects the sociable tendency. We normalize the level of sociable tendency to 0-1. We regard the maximum of sociable tendency as 1 and the minimum as 0. The average amount is 0.727 and the standard deviation is 0.055.

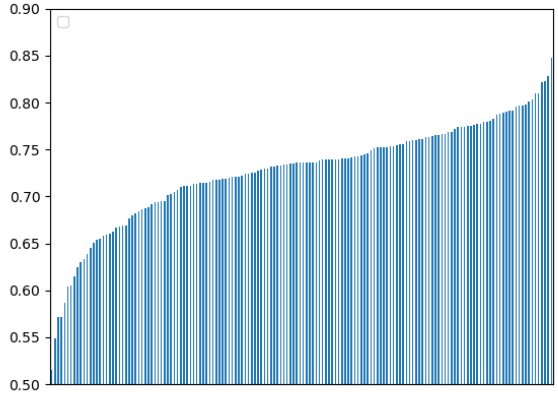

Fig. 6. Sociable tendency of different individuals.

To explore the relationship between specific features like Emotional Fluctuation and abstract features like Big Five, we calculated the correlation between emotional fluctuation and sociable tendency and Big Five. We can see the relationship between the personal traits and personality from TABLE II. The emotional fluctuation and sociable tendency are regarded as personal traits. The personality is represented by the scale of Big Five. The emotional fluctuation and sociable tendency both have a strong correlation with Neuroticism. The result can be explained by common sense that a sensitive person prefers to have emotional fluctuations.

TABLE II. CORRELATION BETWEEN PERSONAL TRAIT AND BIG FIVE

Correlation	**Big Five**				
	OPN	***CON***	***EXT***	***AGR***	***NEU***
Emotional Fluctuation	0.114	-0.078	0.012	-0.137	0.111
Sociable Tendency	-0.184	-0.026	-0.040	0.037	0.105

There are two reasons for the lack of a strong correlation. One is because the Big Five personality is a very abstract description of personal traits. The diversity we calculate is a subdivision in the personal traits. Having a robust correlation between the two is difficult. The other reason is the lack of accurate labeling data for it is labeled by psychological questionnaires.

E. Evaluation

In this part, we evaluated the experimental results manually. The purpose of this verification is to test the proposed method of word2vec-based personal trait computing.

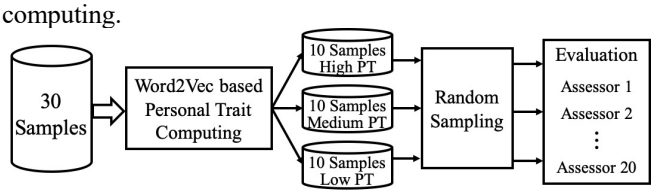

Fig. 7. Vectors of individual A and B.

As shown in Fig. 7, ten samples were extracted from the three parts of the calculation results: high, middle and low. We obtained three groups of samples, 10 in each group, representing the high, middle and low levels of our calculation results. We allocated 10 samples for each group to 20 evaluators, so that each sample was evaluated twice on average. If the evaluation results of a sample are consistent with our calculation results, the evaluation is correct. We have

made statistics on the evaluation results, and the results are shown in TABLE III.

TABLE III. RESULT OF ACCURACY RATE

Accuracy Rate	Emotional Fluctuation	Sociable Tendency
High Level	86.6%	76.7%
Medium Level	80.0%	83.3%
Low Level	83.3%	63.3%
Average	83.4%	74.4%

The number of people for each level is 20.

We can draw the following conclusions. On the one hand, as specific personal traits, the emotional fluctuation and sociable tendency calculated by us can be identified and understood by people. Since the personal traits we computed can be reflected in daily life examples. On the other hand, the results of the manual evaluation show that the method of calculating traits by constructing personal trait matrix is effective.

V. CONCLUSION AND FUTURE WORK

In this paper, we first describe the difference between profile, behavior, personality, and personal trait, and then proposed a word2vec-based personal trait computing method using 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 emotional fluctuation and sociable tendency. We found that there are significant differences in different individuals. We also found that emotional fluctuation and sociable tendency correlates with Big Five personality. Manual verification proves that the proposed method is effective. The experiment results show that the method of word2vec-based personal trait computing can effectively calculate 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 conducted to compute the stability of interest over time.

REFERENCES

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

[2] 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.

[3] 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.

[4] H. Nguyen, D. Morales, and T. Chin, “A neural chatbot with personality,” 2017.

[5] 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.

[6] 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.

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

[8] Kukk and K. Akkermann, “Fluctuations in negative emotions predict binge eating both in women and men: An experience sampling study,” Eating Disorders, vol. 25, no. 1, pp. 65–79, 2017.

[9] Z. S. Harris, “Distributional structure,” Word, vol. 10, no. 2-3, pp. 146162, 1954.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” ICLR Workshop arXiv:1301.3781, 2013.

[11] D. Preot¸iuc-Pietro, V. Lampos, and N. Aletras, “An analysis of the user occupational class through twitter content,” in Proceedings of the 53rd ACL (Volume 1: Long Papers), vol. 1, pp. 1754–1764, 2015.

[12] D. Nguyen, D. Trieschnigg, A. S. Do˘gru¨oz, R. Gravel, M. Theune, T. Meder, and F. De Jong, “Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment,” in Proceedings of COLING 2014, pp. 1950–1961, 2014.

[13] F. Hsieh, R. Dias, and I. Paraboni, “Author profiling from facebook corpora,” in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), 2018.

[14] Alekseev, Anton, and Sergey Nikolenko. "Word embeddings for user profiling in online social networks." Computación y Sistemas 21.2 (2017): 203-226.

[15] I. Cantador, I. Fernandez-Tobıas, and A. Bellogın, “Relating personality types with user preferences in multiple entertainment domains,” in CEUR Workshop Proceedings. Shlomo Berkovsky, 2013.

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

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

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

[19] 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.

[20] Pennebaker, James W., et al. The development and psychometric properties of LIWC2015. 2015.

[21] 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.

[22] G. Sidorov, A. Gelbukh, H. Gómez-Adorno H, 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.