The Personality Analysis of Characters in Vernacular Novels by SC-LIWC

Yahui Yuan[1], Baobin Li[1(✉)], Dongdong Jiao[2], and Tingshao Zhu[2]

[1] School of Computer and Control, University of Chinese Academy of Sciences, Beijing 100190, China
libb@ucas.ac.cn

[2] Institute of Psychology Chinese Academy of Sciences, Beijing 100101, China
tszhu@psych.ac.cn

Abstract. There are many researches on psychological text analysis, and it has been proved that the words people use can reflect their emotional states. In this paper, we introduce how to analyze the psychology of the characters in vernacular novels automatically. First, we process the dialogs with word segmentation, and analyze the segmented text with SC-LIWC. Then, a vector reflecting the psychology of the character is obtained and we map it to the big five. Finally, taking the dialogs of *the Journey to the West* as corpus, We have got the personalities of four main characters which are verified to be same as some famous comments of *the Journey to the West*, which shows that our work is effective.

Keywords: LIWC · Vernacular · The Journey to the West · The big five
Text analysis

1 Introduction

The use of words in the text can reflect the individual's psychological state and personality [1]. Linguistic Inquiry and Word Count is a tool which we can use to analyze text. The way that the Linguistic Inquiry and Word Count (LIWC) program works is fairly simple. Basically, it reads a given text and counts the percentage of words that reflect different emotions, thinking styles, social concerns, and even parts of speech.

To date, LIWC has been applied to many psychology research. It is often used to examine suicide writings in order to characterize the quantitative linguistic features of suicidal texts, in [2], the authors analyze texts compiled in Marilyn Monroe's Fragments using LIWC, in order to explore the contact between the use of different linguistic categories over the years and her suiside. The result is coincide with different theories of suicide. López-López et al. [3] analyzed the StackOverflow's answers and questions to explore the users' personality traits. They found that the top reputed authors are more extroverted than general users. Moreover, authors who got more votes express significantly less negative emotions than those who got less votes. Markovikj et al. [4] explored

Q. Zu and B. Hu (Eds.): HCC 2017, LNCS 10745, pp. 400–409, 2018.
https://doi.org/10.1007/978-3-319-74521-3_43

the modeling feasibility of user personality based on the features extracted from Facebook. In [5], they collected a sample of 363 participants, including their written self-introductions and final course performance, the result shows that course performance could indeed be predicted by the word usage of linguistic categories.

LIWC for Traditional Chinese, TC-LIWC, is published with the authorization of Pennebaker by Huang et al. (2012). After that, SC-LIWC for Simplified Chinese [6, 7] is published on the basis of TC-LIWC, which lays the foundation for the following research [10, 12].

Recently, some researchers are concerned about automatic personalistic prediction using liwc. Personality is stable in a period of time, so a collected corpus from several months is suitable for this research. Gao [12] selects 1766 participants, first make them fill in a big five inventory for comparison, and then collect their weibo through the API of Sina. 90% of the samples are trained using liwc and the rest act as test set. In the training stage, they compute the Pearson's coefficient between the inventory and the training results, then choose features which behave well in the training. At last, the features are composed to predict personality of test set. They compute the Pearson's coefficient as before, and the results are between 0.3~0.4. While the coefficient between self-rating and rating by observers is about 0.5, hence the method has prediction ability to some extent.

We will seek method to predict the personality of characters in novels written in vernacular. Vernacular is a written language with some artistic processing. It is easy to read, but still have some features of ancient Chinese. Vernacular is generally used for literacy, especially in the novels. Vernacular novels are very popular from the beginning of Ming Dynasty. Three of the four famous Chinese novels were accomplished in Ming Dynasty. After that, vernacular novels were more and more popular. There are many excellent ancient books in China, which created numerous virtual characters, a book named A Dream of Red Mansions only, contains hundreds of characters. We will pay a lot of time to read books, look up in the library, to understand these figures and step into the author's inner world. If we can analyze the characters of the books automatically, it will save us much time and help us follow the books.

In this study, we use LIWC to analyze the personality of the characters in vernacular novels. The process of personality analysis of the characters in the vernacular novels is shown in Fig. 1. The rest of this paper is organized as follows. Section 2 will discuss the preprocessing work for the novel. And then in Sect. 3, we will split the dialogs obtained from Sect. 2, which will be used to analysis the personality of the characters and get the big five of characters in Sect. 4. Finally, we also present the personality change of Sun Wukong before and after the three strikes of White Bone Demon.

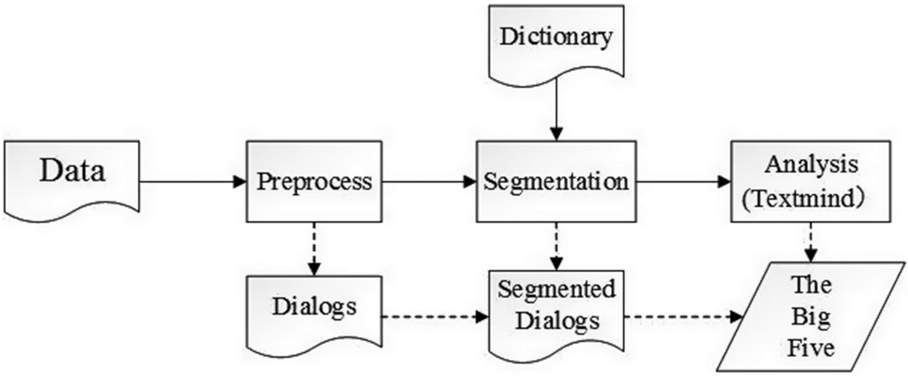

Fig. 1. The flow chart of automatic personality analysis of vernacular novels.

2 Data and Data Preprocessing

The Journey to the West is the first ancient Chinese romantic novel. The book deeply depicts the social reality of the time, mainly describes the origin of Tang priest, Sun Wukong, Zhu Bajie, Sha monk, and together with the story of pilgrimage to the west. After the spread of centuries, *the Journey to the West* has been translated into many languages, and a number of relevant research monographs have been published, which made a high evaluation of the novel. *The Journey to the West* is known as one of the four famous Chinese classics. There are four main characters in *the Journey to the West*. They are Sun Wukong, Sha monk, Zhu Bajie and their master, Tang priest.

James W. Pennebaker proved that words used in their daily lives could contains important information of psychological information [8]. Especially, he proved that not only nouns and verbs serve as markers of emotional state, social identity, and cognitive styles, particles, serve as the glue that holds nouns and regular verbs together, can also do the same things. That means we can study the particles instead of more complex methods. On the basis of his work, we decide to use the dialogs of the characters to study their personality. We select the dialogues and their inner monologues for each character respectively, and put them into 5 different files. There are a lot of descriptive verses in the text, we should delete them because they would interface the process of participation, and these verses are often said by other people, not the roles themselves, so we believe that the descriptive verse have little influence on the personality. Notice that it shouldn't include quotes on the end of the sentence. But other interval will be retained for the next step. At last we get four files.

3 Segmentation

In order to get the particles for analysis, we should first do the text segmentation. As we know, the current methods of Chinese word segmentation can be divided into three kinds: the method based on lexicons, the method based on statistics and one based on

semantics. Also there are methods mixing two or three of them in order to improve the accurate. There have been many kinds of word segmentation systems for modern Chinese. For example, LTP-CLOUD, NLPIR, jieba, and so on. These systems all have good results in Chinese word segmentation. Among these tools, LTP and NLPIR are systematic tools, while jieba only contains segmentation function. Moreover, LTP and jieba are open source tools, but NLPIR is not.

Few studies are involved in ancient Chinese segmentation. First, there is little corpus for ancient Chinese. As we all know, segmentation need a lot of corpus which has been marked manual to improve the accurate, but no one have done the work for ancient Chinese. Second, it seems that we cannot obtain any economic benefits from it. However, Hou, et al. have studied ancient Chinese segmentation [9], but the corpus is still very small, which couldn't be generalized easily.

Vernacular has the features of both ancient Chinese and modern Chinese. Therefore, we can refer to the methods for modern Chinese segmentation.

The punctuation and function words are not changed much over the years. They are also used in ancient Chinese. Besides, there are a lot of words that still exist in ancient Chinese. In addition, the Chinese word segmentation methods are able to identify new words according to statistical methods.

In order to simplify the segmentation procession, we make a simple test on the segmentation of vernacular and find, most words are segmented correctly by applying the LTP directly. But there are still many mistakes; notional words are not distinguished from others, idioms are segmented wrong, and there are other mistakes generated in the algorithm. That is mainly because of the difference between vernacular and modern Chinese. There are many words which are not used now, especially those appear only several times in the text.

A simpler and more efficient method we used to solve above questions is add a manual dictionary to the LTP. The dictionary includes notional words and some idioms. Notional words include place name, monster name, Tang priest and his three apprentices' name and nickname, gods' name and their nickname, the particular items and some words about the emperor and the dynasty.

- Place name. There are a lot of places made up by the author, such as the monster's cave, the god's mountain, the monkey's birthplace, and so on. Take these words into dictionary will make sure they are segmented correctly.
- Monster's name. Tang priest and his four apprentices meet many big monsters on *the Journey to the West*, and each of them get a nickname, even some of small monster under them get one, too. To split them correctly, we had better put them into the dictionary.
- Four characters' name and their nickname. Though only four people, each of them have many nickname. Only Tang priest has more than 5 nicknames, for example, priest, Tang priest, Xuanzang, elder, Tang elder, master, and so on. Especially they often emerge in the dialogs. Therefore, split them from other words are important.
- Gods' name and their nickname. When walking through the long way, the four meet a lot of gods. Each of them have several nickname, especially Guanyin, his nicknames even catch up with Tang priest.

- The particular items. There are many particular things in *the Journey to the West*, such as kinds of weapons, treasures, and so on. Much of them are not usually used in modesty Chinese.
- Some idioms. Idioms are fixed phrases in the ancient Chinese.
- Some words about the emperor and the dynasty.

We select 1000 characters randomly from the segmentation results for each of the four, five of our workmates check it respectively. When finished all, they vote on the contradicted ones. Table 1 shows that the error rate is about 2%~3%, in other words, the accurate rate is about 97%~98%. Though the punctuation is count into the words, the error rate will not exceed 1.2 times of the existing data. As we can see, this method is quite effective.

Table 1. The error rate of word segmentation.

	Total	Wrong	Error Rate
Tang priest	850	27	0.0318
Sun Wukong	766	21	0.0274
Zhu Bajie	1048	25	0.0239
Sha monk	753	30	0.0398

In [12], five of six feature classes are properties of Weibo, so in our work, we choose only the features of liwc. Due to the differences between ancient Chinese and modern Chinese, liwc dictionary will have corresponding change, so will the features selected based on liwc. To solve it, we get rid of the features which are not consistent with ancient Chinese. The rest features will serve as input of the model which have been trained in [12]. In order to inspect the personality more intuitive, we adopt a commonly used quantitative method in psychology—the big five personality traits [11]. In the big five traits model, the user's personality is abstracted into five dimensions, which are shown in Tables 2 and 3. The big five score of Tang priest and his apprentice is shown in Table 4.

Table 2. The big five 1. Each of them has 6 facets.

Agreeableness	Conscientiousness	Extraversion
Trust	Competence	Warmth
Straightforwardness	Order	Gregariousness
Altruism	Dutifulness	Assertiveness
Compliance	Achivement striving	Activity
Modesty	Self-discipline	Excitement seeking
Tender mindedness	Deliberation	Positive emotion

Table 3. The big five 2.

Openness to experience	Neuroticism
Fantasy	Anxiety
Aesthetics	Hostility
Feelings	Depression
Actions	Self-consciousness
Ideas	Impulsiveness
Values	Vulnerability to stress

Table 4. The big five of Tang priest and his three apprentice.

	Agree.	Cons.	Extra.	Open.	Neur.
Tang priest	13.06	11.26	7.94	3.50	25.42
Sun Wukong	9.63	4.97	1.34	0.92	1.93
Zhu Bajie	9.25	5.02	18.54	15.08	18.27
Sha monk	15.24	6.54	14.15	1.10	26.31

4 Personality Analysis of Characters

4.1 Agreeableness

As an eminent monk from Tang dynasty, Tang priest is very kind and compassionate [13]. On the Journey to the West, he tries to help others, though when he is in danger. He is modest and subject to authority, such as emperor of Tang Dynasty, and all the gods they meet on the Journey to the West. But sometimes he is egoistic, and often shifts responsibility. His agreeableness is relatively high.

Sha monk is very careful and slavish since he was surrendered by Sun Wukong. He has never been egoistic, doing his best to serve the master and help his brothers [14]. Tang priest and Sun Wukong all have deep trust on him. The agreeableness of him is highest.

Sun Wukong is capable, but his master does not trust him. He has helped many people, but that does not mean he is willing to sacrifice. If someone harms his interests, he will not hesitate to teach him a lesson. Sun Wukong is also an arrogance role, never understanding what modesty is. He is not gentle too. In conclude, his agreeableness is lowest.

4.2 Conscientiousness

Tang priest has no ability to protect himself, and has no experience to deal with monsters. He strictly abides by the doctrine, trying his best to protect it. And he has a strong will, which makes him go to the west firmly to obtain Mahayana Buddhism [13]. All in all, his conscientiousness is relatively high.

Though exiled from heaven just because knocked over a glass made of colored glaze, Sha monk not only gets no angry, but also accepts the destiny to atone for his sin. In this point, he is something like Tang priest [14]. Compared with his brothers, he seems plain, but he is better than his brothers in human nature. So his conscientiousness is relatively higher.

Sun Wukong is strongest among his brothers. He pursues the idea that the stronger should hold the power, never yielding to authority [15]. Therefore, many gods serve him as a servant. He tries his best to protect the master, not for any benefits, but for returning Tang priest's salvation. He could not restrain his aggressive instincts, and that makes Tang priest most unhappy. It is not strange that his conscientiousness is lower than his brothers.

Zhu Bajie is similar with Sun Wukong in agreeableness and conscientiousness. There is also difference: Zhu Bajie gets more score in tender mindedness, while Sun Wukong gets more in altruism [16, 17]. Zhu Bajie obeys the laws. Sun Wukong is a king of monkey before he follows Tang priest, so he has no knowledge of it. Thus in order, Zhu Bajie gets more score. But he always declares to go back to Gao Village when he encounters danger. Sun Wukong loves battle while Zhu Bajie loves women. In total, they go head in head with each other in agreeableness and conscientiousness.

4.3 Extraversion

In the point of extraversion, Zhu Bajie gets first without question. He is very lustful, showing great enthusiasm for women. He loves to eat, too [18]. Each time when they arrive to a new town, he is happiest because he can eat a lot. Zhu Bajie is outgoing compared with other people.

Sha monk is upright and honest [19]. He talks little, but what he said is very useful. When there is contradiction among the group, Sha monk is the one who tries to solve it.

Tang priest is kind and friendly when dealing with people, but not enthusiastic. He prefers quiet than noisy. So his extraversion is low.

In Table 4, Sun Wukong is lowest in extraversion. Though his warmth is not as good as his brothers, but the rest features should be better than others. The possible reason may be the dictionary we use may not be so fit with his dialogs.

4.4 Openness

Openness is an indicator of the level of intelligence. From Table 4 we can observe that Zhu Bajie gets the highest score. Zhu Bajie is always being called "fool", but that is not the case [18]. He is fond of eating and sleeping, and good at flattering in front of master, so Tang priest trusts him very much. In general, he always makes the best decision for himself.

As a leader of the group, Tang priest is well learned and behaved. But he is often cheated by monsters and confused by Zhu Bajie, and getting rid of Sun Wukong several times, who tries to protect him.

Sha monk actually is a servant of Tang priest. He is responsible for all trifles, but never complaining about it [20].

Sun Wukong is powerful and good at dealing with enemies, however, he is often fooled by Zhu Bajie. So in fact, Zhu Bajie is the most intelligent person among the group.

4.5 Neuroticism

In the term of neuroticism, Tang priest and Sha monk are similar to each other. When facing the danger, Tang priest is anxious and scared, and often bursts into tears [13]. Sha monk is puzzled, and 'What should we do?' is his pet phrase. Zhu Bajie helps Sun Wukong a lot in fighting, but once failed the first thing he thinks of is escaping. Sun Wukong never gives up, even if he was alone, he will fights until success.

5 The Personality Change of Sun Wukong

We also did a research on the personality change of Sun Wukong before and after the three strikes of White Bone Demon with this model. The results are shown in Table 5.

Table 5. The big five of Sun Wukong before and after his beating the White Bone Demon for three times.

	Agree.	Cons.	Extra.	Open.	Neur.
Before	6.28	3.63	2.83	19.27	6.62
After	10.56	4.99	0.39	1.03	1.19

Beating the White Bone Demon for three times was a turning point of Sun Wukong. The author explained the mind change of Sun Wukong by the words of Zhu Bajie.

Before that, Sun Wukong was very irritable, but under the constraint of Tang priest, he changed gradually. He became no more impulsive at all. It is corresponding to the decrease of neuroticism.

In the early stage, Sun Wukong despised the authority, refused to obey the discipline, and showed a clear sense of rebellion. While later he was influenced by Tang priest, no longer having a strong sense of resistance. As we can see in Table 5, his openness is greatly changed before and after the three beats of White Bone Demon.

The agreeableness is higher after the three strikes. In the respect of getting along with people, Sun Wukong changed obviously, especially to Bodhisattva, Buddha, and some others who is venerable. He was more and more courteous and no longer as arrogant as before.

In the early stage, the responsibility of Sun Wukong is not clear, he even tried killing Tang priest and attacked the Bodhisattva at first. However, he accepted his duty in the late, becoming a qualified defender.

6 Conclusion

In this paper, we make a automatic personality analysis of characters *in the Journey to the West* using LIWC, and map the feature vector into the big five to make it easy to

observe. We compare the results with many famous reviews, and compare the results between characters, and compare the results back and forward. These comparisons all show that automatic personality analysis of characters with LIWC is feasible.

References

1. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. **29**, 24–54 (2010)
2. Fernándezcabana, M., Garcíacaballero, A., Alvespérez, M.T., Garcíagarcía, M.J., Mateos, R.: Suicidal traits in Marilyn Monroe's fragments: an LIWC analysis. Crisis **34**, 124–130 (2013)
3. López-López, E., del Pilar Salas-Zárate, M., Almela, Á., Rodríguez-García, M.Á., Valencia-García, R., Alor-Hernández, G.: LIWC-based sentiment analysis in Spanish product reviews. In: Omatu, S., Bersini, H., Corchado, J.M., Rodríguez, S., Pawlewski, P., Bucciarelli, E. (eds.) Distributed Computing and Artificial Intelligence, 11th International Conference. AISC, vol. 290, pp. 379–386. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07593-8_44
4. Markovikj, D., Gievska, S., Kosinski, M., Stillwell, D.J.: Mining Facebook data for predictive personality modeling. In: AAAI International Conference on Weblogs and Social Media (2013)
5. Robinson, R.L., Navea, R., Ickes, W.: Predicting final course performance from students' written self-introductions: a LIWC analysis. J. Lang. Soc. Psychol. **32**, 469–479 (2013)
6. Gao, R., Hao, B., Li, H., Gao, Y., Zhu, T.: Developing simplified Chinese psychological linguistic analysis dictionary for microblog. In: Imamura, K., Usui, S., Shirao, T., Kasamatsu, T., Schwabe, L., Zhong, N. (eds.) BHI 2013. LNCS (LNAI), vol. 8211, pp. 359–368. Springer, Cham (2013). https://doi.org/10.1007/978-3-319-02753-1_36
7. Zhao, N., Jiao, D.D., Bai, S.T., Zhu, T.S.: Evaluating the Validity of Simplified Chinese Version of LIWC in Detecting Psychological Expressions in Short Texts on Social Network Services, vol. 11, p. e0157947 (2016)
8. Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G.: Psychological aspects of natural language use: our words, our selves. Annu. Rev. Psychol. **54**, 547–577 (2003)
9. Zeng, Y., Hou, H.: Research on the extraction of ancient texts. J. China Soc. Sci. Tech. Inf. **30**, 132–135 (2008)
10. Gao, R., Hao, B.B., Li, L., Bai, S., Zhu, T.S.: The establishment of the software system of Chinese language psychological analysis. In: Psychology and the Promotion of Innovation Ability: the Sixteenth National Conference on Psychology, p. 3 (2013)
11. de Raad, B.: The Big Five Personality Factors: The psycholexical Approach to Personality, pp. 309–311. Hogrefe & Huber Publishers, Ashland (2000)
12. Gao, R.: The Research and Application of the Psychological Analysis Technology of Weibo Content, University of the Chinese Academy of Sciences, vol. Master. University of the Chinese Academy of Sciences (2014)
13. Cao, B.J.: The reflection and critique of pure confucianism personality: a new comment of Tang priest. Acad. J. Zhongzhou **4**, 110–112 (1999)
14. Jiang, K., Huang., L.J.: Compromise and secularization: an image evolution of Sha monk. J. Zibo Normal Coll. **4**, 58–62 (2007)
15. Zhou, X.S.: The time spirit and cultural connotation of Sun Wukong. J. Southeast Univ. **8**, 63–73 (2006). (Philosophy and Social Science)
16. Cao, B.: Secularized Comic Image and Civil Personae 1: A New Reading of Ba Jie in A Journey to the West. J. Huaihai Inst. Technol. **5**, 23–27 (2007). (Social Science Edition)

17. Cao, B.: Secularized Comic Image and Civil Personae 2: A New Reading of Ba Jie in A Journey to the West. J. Huaihai Inst. Technol. **5**, 27–30 (2007). (Social Science Edition)
18. Liu, W.L., Qiu, H.C.: The Empiricism of Zhu Bajie. J. Keshan Teachers Coll. **2**, 42–46 (2002)
19. Cai, S.: Discussion on Sha monk in "Journey to the West". J. Educ. Inst. Jilin Province **30**, 114–115 (2014)
20. Lu, L.: Industrious, duty, firm and persistent: a comment of Sha monk. South J. **11**, 98–99, 110 (2014)