User Modeling and User-Adapted Interaction
https://doi.org/10.1007/s11257-020-09276-2

Personality expression and recognition in Chinese language usage

Cuixin Yuan[1] · Ying Hong[2] · Junjie Wu[1,3]

Received: 27 November 2019 / Accepted in revised form: 8 August 2020

Abstract
Personality plays a pivotal role at work. Many scholars have investigated the association between personality and language usage habits in the English corpus. Given that the Chinese language has the largest number of native speakers in the world, it is essential to analyze the pattern of personality expression in Chinese, which has garnered less attention. In this study, we used the TextMind system to examine the correlation between word categories and personality traits based on Chinese Weibo content. We also compared the results with previous studies to demonstrate the similarities and differences of personality expression between English and Chinese. Additionally, this paper established a prediction model based on machine learning methods to recognize personality. Results showed that language features were powerful indicators of personality. Finally, we made recommendations for using personality expression in the recruitment and selection.

Keywords Personality traits · Language use · Personality recognition · Weibo · TextMind · Chinese language · Linear regression

1 Introduction

Personality is an individual difference construct that has been used in the explanation of multiple human attributes and behaviors (Matthews et al. 2003). One of the most commonly used personality models is the Big Five model, which revolves around the traits of openness, conscientiousness, extraversion, agreeableness, and neuroticism (Costa and McCrae 2008), with the acronym "OCEAN." There has been

✉ Junjie Wu
wujj@buaa.edu.cn

[1] School of Economics and Management, Beihang University, No. 37 Xueyuan Road, Haidian District, Beijing 100191, China

[2] Gabelli School of Business, Fordham University, New York, USA

[3] Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China

Published online: 25 August 2020

considerable evidence demonstrating that the Big Five model is a robust predictor of behaviors and can be generalized across cultures (McCrae and Costa 1997). Albeit personality is commonly assessed with self- or other ratings, such as by friends, family members, and colleagues, these approaches are expensive and are subject to social desirability biases (Ellingson et al. 1999). Given that the personality construct itself originated from lexical assumptions (Sanford 1942), i.e., different words are used to describe each personality trait, linguistic features are a natural, alternative tool to assess the theoretical classification of personality traits (Tellegen 1993).

Many researchers have used scientific approaches to analyze the relationship between language usage and personality. The linguistic inquiry and word count (LIWC) (Pennebaker et al. 2007), a notable example of a "closed-vocabulary" psychological dictionary, focuses on parameters such as the frequency of words in the previously defined word categories (e.g., function words, affective processes, social processes, etc.). Using the LIWC dictionary, researchers study the relationship between language usage habits and personality traits (Pennebaker et al. 2003). The results generally support the proposition that one of the most consistent personality observations is through language usage patterns (Holtgraves 2011; Lee et al. 2007; Qiu et al. 2017; Sanford 1942).

While traditionally, researchers relied on historical language samples or publications such as papers and literary fictions to understand how language was used, the recent surge of social media outlets such as blogs and social networking sites presents an unprecedented amount of rich and valuable data, the scientific value of which is greatly under-realized (Park et al. 2015). Language usage exhibited in daily online social activities conveys individual personality even more effectively than well-crafted historical language samples, because the former often represents the unaltered and uninhibited true self (Bargh et al. 2002). Furthermore, online social activities closely pertain to current events and reflect the most leading-edge evolution of language usage habits. In this paper, we aim to analyze online social networking activities to scale and extend the existing research on language usage and personality.

Prior studies have extensively and exclusively focused on how English language impacts personality, and more research is required to gather data on the relationship between other languages and personality. Chinese is the most widely spoken language in the world—almost 1.2 billion people speak Chinese as their native language (Paul et al. 2015). Given that there was no systematic effort to assess the relationship between Chinese language usage and personality, it is important to address this gap in the literature. Particularly, patterns of language usage differ in English and Chinese. For example, the same Chinese word can represent many different lexical categories (e.g., the same word can mean the noun "happiness," the adjective "happy," the adverb "happily," or the stative verb "be happy"). This situation arises in part because Chinese words generally have no grammatical inflections to indicate case, number, gender, tense, or degree (Light et al. 1979; Li and Thompson 1989).

Further, in contrast to English, most Chinese "function" words are optional and are suppressed when they would not lead to ambiguity (Aaronson and Ferres 1987). The specific syntactic and semantic differences between Chinese and English thus lead to differences in expression. For example, Chinese speakers speaking English as a second language use the deliberative function of *I think* in medial or final positions,

while native English speakers do not do so; this may be because the equivalence in Chinese "*wo juede*" can mark the deliberative meaning in medial or final positions. Chinese speakers also use the tentative function of *I think* in initial positions more frequently than the native English speakers; this maybe because they also tend to use *wo juebe* frequently (Liu 2013). This paper, therefore, seeks to investigate how the relationship between Chinese language usage habits and personality differ from that of the English language. Chinese language refers to Mandarin, which is the universal spoken and written language in China. Although people in different regions may have different spoken dialects, these dialects do not have a written language of their own; thus, the written communication online is uniformly Mandarin.

Specifically, we drew our data from Sina Weibo. As the largest social media in China, Weibo has exceeded 516 million monthly active users in the fourth quarter of 2019.[1] Similar to Twitter, Weibo users can publish or republish text, pictures, videos, and other multimedia forms, with text not exceeding 140 Chinese characters. Unlike Facebook or Chinese Wechat, Weibo does not have privacy settings that limit non-followers' access to posts; instead, all posts are publicly accessible. Contrary to Facebook, which uses real names, Weibo uses nicknames. The *open-access* and *anonymous* nature of Weibo may engender more unaltered and uninhibited language patterns than a real-name and closed-loop system (Chan et al. 2012).

In addition to the extraction of language usage features related to personality traits, we also build a prediction model to recognize these traits, which will allow employers to automatically identify the personality traits of individuals using existing online data. Personality prediction has been explored in great depths in the field of computational psychology through the means of machine learning algorithms. Notably, most personality prediction studies in this field are based on varying English corpuses such as emails, blogs, Facebook posts, and essays (e.g., Celli et al. 2014; Farnadi et al. 2013; Golbeck et al. 2011a; Mairesse et al. 2007; Majumder et al. 2017; Nowson and Oberlander 2007). After establishing the personality prediction platform, the personality model of an individual is built through various machine learning methods, such as logistic regressions, Naïve Bayes, recurrent neural networks, support vector machines (SVM), and convolutional neural networks. Using the Big Five personality model, these studies usually construct five binary classification models compared to the median scores. The degree of accuracy in the classification of the five personality dimensions is then computed.

To our knowledge, there have been only a handful of studies which employed Chinese social network data for personality recognition. For example, Bai et al. (2012) used data from RenRen (a popular social network platform in China) and C4.5 decision tree method to identify low, medium, and high observed values of Big Five personality. Peng et al. (2015) analyzed posts from 222 Facebook users with Chinese as their main written language for personality classification (with a focus on the extraversion dimension) via SVM method. It also used the number of friends as an additional feature to improve the accuracy of classification. Although these studies focused on improving the precision rate for the

[1] https://www.sohu.com/a/376087332_114760.

personality recognition process, they did not provide sufficient explanations for the linkage between language usage patterns and personality traits. This paper, therefore, aims to not only reveal the Chinese language usage habits in association with personality, but also build a personality recognition model based on the language features.

2 Theoretical background

2.1 Big Five personality traits

Language is the most common and reliable way for people to translate their internal thoughts and emotions into a form that others can understand. Words and language are then the *sine qua non* of psychology and communication—they become the medium through which social psychologists perceive the cognition, personality, and mentality of human beings (Tausczik and Pennebaker 2010). In speaking of personality, Norman (1967) posited that an important piece of evidence to assess theorized personality attributes into subcomponents is through natural language. In the earliest development of psychology, Freud (1901) provided several compelling examples in his discussion of parapraxes, or slips of the tongue, which suggested that common errors in speech betray people's deeper motives or fears. Jacques Lacan (1968) extended these ideas by suggesting that the unconscious asserts itself through language. In his view, language is the bridge to reality. Nowadays, personality psychologists have long sought to identify individual personality differences based on the lexical hypothesis, which suggests that key features of human personality become a part of the language that we use to describe ourselves (Goldberg 1981; Mairesse and Walker 2010).

One of the most robust and widely applied personality models is the Big Five model (McCrae and Costa 1997). *Openness* is associated with curiosity, creativity, imagination, and an appreciation of new experience. People who manifest high openness have a good sense of aesthetics, appreciate new and unusual ideas, and welcome change. *Conscientiousness* pertains to a preference for an organized approach to activities in life, as opposed to a spontaneous approach. People high in conscientiousness are more likely to be consistent, reliable, and well organized. *Extraversion* is the tendency to seek stimulation in the external world and others' company. Extroverted people are socially active, friendly, and outgoing, as well as energetic and talkative. *Agreeableness* is defined as being cooperative and compassionate, as well as being able to form and maintain positive social relationships. Agreeable people tend to trust others and adapt to others' needs. *Neuroticism*, which is reversely referred to as emotional stability,

measures the tendency to experience mood swings and emotions such as guilt, anger, anxiety, and depression. People who are neurotic are more vulnerable to nervousness and stress in their lives (Farnadi et al. 2016).

2.2 Personality expression in language

The Big Five dimensions have been widely examined in language usage. Many studies have compared personality self- or observers-reports with language usage (Cohen et al. 2008; Fast and Funder 2008; Tausczik and Pennebaker 2010). Specifically, research indicates that individuals with high neuroticism are more likely to use words that express negative emotions ($\beta = .15$), such as anger, suspicion, and revenge, as well as to use the first person singular (Pennebaker and King 1999). Other studies have replicated these findings (Stirman and Pennebaker 2001; Schwartz et al. 2013). Likewise, extraversion is related to positive emotion ($\beta = .13$) and social process words ($\beta = .04$) (Schwartz et al. 2013). In particular, extraversion correlates with family ($\beta = .19$), humans ($\beta = .20$) (Hirsch and Peterson 2009), and other personal words, such as drink ($\beta = .21$), dance ($\beta = .20$), and hostel ($\beta = .21$) (Yarkoni 2010). Also, extraverted people tend to use more verbs in their language (Lee et al. 2007). Agreeableness has a strong correlation with positive emotion words ($\beta = .13$) (Pennebaker and King 1999), while openness is negatively correlated with relativity and time words ($\beta = -.15$) when analyzed through personal essays (Mairesse et al. 2007). Conscientiousness is positively correlated with cognitive processes ($\beta = .33$), expectation ($\beta = .30$), and confidence words ($\beta = .27$), and negatively related to swearing words ($\beta = -.24$) (Lee et al. 2007). In sum, linguistic cues are shown to be powerful indicators of personality.

In particular, many prior studies have examined the association between personality and language usage using the LIWC dictionary. Pennebaker and King (1999) conducted a simple principal components factor analysis based on the words significantly related to personality traits, which generated four factors, namely immediacy, making distinctions, the social past, and rationalization. On this basis, the researchers compared the correlations of the four factors with the personality dimensions. Immediacy is negatively correlated with openness and positively related to neuroticism, while making distinctions is negatively correlated with extraversion and conscientiousness. The social past is correlated with openness, while rationalization is not related to personality traits (Pennebaker and King 1999).

According to research involving other English platforms, profiles on online social networking sites reflect actual personality traits instead of self-idealization (Back et al. 2010). Nonetheless, whether such findings can be generalized in the Chinese context remains less known. Following the approach by Pennebaker and King (1999), Qiu et al. (2017) conducted a factor analysis based on Chinese Weibo data, which, similar

to the English corpus, generated four factors including making a distinction,[2] reflection,[3] objective description,[4] and socialization.[5] Making a distinction is negatively correlated to extraversion and positively correlated to neuroticism. Reflection is negatively correlated with conscientiousness, while objective description is positively correlated with agreeableness and conscientiousness. Lastly, socialization is positively correlated to extraversion (Qiu et al. 2017). Building on Qiu et al. (2017)'s work which used the simplified Chinese version of the LIWC dictionary to examine the relationship between personality and language features, in this research we further extend this work by using language features to build a model to make personality recognition through Weibo.

2.3 The present study

This study aims to investigate the relationship between Chinese language habits and personality traits using the LIWC dictionary. Also, based on the LIWC features related to personality, we will build a personality recognition model through logistic regression. Based on these analyses, we will provide suggestions for human resource management professionals to apply personality recognition through language usage habits in the recruitment and selection of employees.

3 Method

3.1 Participants

We collected two sources of data: First, a survey was administered to the study participants and second, we compiled the Weibo activities of these participants from the website. Students were recruited in a Northern Chinese University by two research assistants; a small financial incentive was provided to the participants for completing the survey. The questionnaires were completed through the WenJuanXing website, an online platform for administering questionnaires in China, which provides services such as questionnaire design, data collection, custom report, and survey results analysis. The survey contained the Big Five personality items, demographic

[2] The first factor included six word categories, three of which were cognitive process categories including tentative, exclusive and discrepancy words, and three were functional categories including impersonal pronouns, adverbs, and conjunctions. This factor was similar to the making distinction factor in English which included tentative, exclusive and discrepancy words (Pennebaker and King 1999). Therefore, this factor was termed making distinction.

[3] The second factor encompassed five word categories, including tense marker, insight, interjunction, assent, and a secondary loading of adverbs. This factor contained thoughtful discussions of the past, present and future, thus was named reflection.

[4] The third factor included eight word categories, namely, time, space, numbers, quantity unit, fewer non-fluencies, fewer positive emotions and a secondary loading of fewer assent. This factor described time, space, and objects, therefore was termed objective description.

[5] The fourth factor contained social words and first- and second-person singular pronouns. It captured social interaction and personal attention and was labeled socialization.

questions such as age, gender, and educational background, as well as one question requesting the respondent's Weibo nickname. We then used their nickname to identify the participants' Weibo profile and collect their online activity data. A total of 204 participants were solicited to complete the questionnaire. Among them, 181 participants provided their Weibo nicknames, who were asked to allow access to their public profile information and microblogs on Sina Weibo. An informed consent document was provided along with the survey questions to assure participants that their data will only be used for research purposes and, once data matching is complete, the identifying information will be removed, which ensures that participants complete the questionnaires voluntarily and confidentially.

After removing the incomplete questionnaires, we obtained a total of 127 Weibo users and their corresponding personality data. Of the final sample, 46 were male, accounting for 36.2% of the sample; 81 were female, representing 63.8% of the sample. These participants were relatively young, but they are representative of the population that uses the online social media platform (Zhang and Pentina 2012). The average age of the participants was 24 years old, with the age group between 21 and 24 years old representing 42.5%. Those between 18 and 20 represented 35.4% of the sample. The remaining participants were between 25 and 30, which were 22.1% of the sample. The majority of the participants were current students, fresh graduates, or graduates who have joined the workplace.

3.2 Measures and procedure

We adopted the 50-item Big Five personality scale from the International Personality Item Pool (IPIP) (Goldberg 1992), which is a well-established and widely used measure of personality. The questionnaires were translated into Chinese following the back-translation procedure (Brislin 1970). The reliability of the scale dimensions and mean and standard deviation scores are exhibited in Table 1.

The participants' microblogs and Weibo profiles were extracted through the Sina Weibo API. Information such as comments by others, retweet content, and timestamps was removed. The study gathered Weibo data posted in the 32 months prior to the time of this study, ranging from January 2016 to August 2018. Together, 45,765 posts from Weibo were obtained, with an average of more than a dozen posts published by each participant per month. Specifically, the average number of characters per user per post in the Weibo corpus was 15.73 (s.d. = 17.99, median = 9.81, minimum = 1.00, and maximum = 86.81). The average number of Weibo posts per user was 360.35 (s.d. = 970.18, median = 58, minimum = 1.00, and maximum = 8010). We extracted these posts on Weibo and removed other information such as comments. These texts, which included both full sentences and abbreviations, were then analyzed to calculate the frequencies of the different word categories in TextMind. In addition, participants' geographical locations registered on Weibo suggest that the vast majority of participants came from China (including 19 provinces and 4 municipalities), except for 21 participants who indicated they were in other countries. The top five places were: Beijing (22), Hunan province (16), Sichuan province (8), and Guangdong province (7). The participants covered all regions of China—two

Table 1 Descriptive statistics of Big Five personality traits

	Max	Min	Mean	SD	Cronbach's α	Scale Cronbach's α	1	2	3	4	5
1. Extraversion	44	11	29.11	5.30	.74	.82	1				
2. Agreeableness	47	22	35.40	4.82	.71		.11	1			
3. Conscientiousness	47	19	32.75	4.98	.71		− .02	.33**	1		
4. Neuroticism	48	14	28.86	6.87	.86		.24**	− .00	.23**	1	
5. Openness	47	17	32.99	4.57	.70		.19*	.28**	.29**	.10	1

$N = 127$, max stands for the max value, Min stands for the minimum value. The last five columns are the correlation coefficients among the five big personality traits

* $p < .05$; ** $p < .01$

participants came from the southernmost province of Hainan; 7 participants came from northeast China, including Heilongjiang, Jilin and Liaoning provinces; 3 participants came from the eastern province of Jiangsu; and 7 participants from northwest Shanxi and Ningxia provinces.

Next, we used the TextMind (Gao et al. 2013) system, a Chinese language psychological analysis program, to generate word frequencies. TextMind provides a simplified approach to analyze the preferences and degrees of different categories in text. It was developed based on the dictionary of LIWC2007 and C-LIWC and provides a one-stop solution ranging from automatic Chinese word segmentation to psychological analysis (Gao et al. 2013). The English version of the LIWC has been widely used in language analysis and is known for its robust dictionary. The traditional Chinese version of LIWC dictionary, which is a translation of the LIWC English dictionary, was later released (Gao et al. 2013). Both English LIWC dictionary and traditional Chinese LIWC dictionary were developed for relatively formal text. Weibo, however, is characterized by more current and informal language use. The traditional Chinese LIWC dictionary thus provides less coverage of the popular words that are trending in microblogs, which makes it less feasible for microblog text analysis. As such, in our study of the Weibo data, we use the TextMind system, which added the high-frequency words into the dictionary. The dictionary, text, and punctuation processing in TextMind are optimized for simplified Chinese, and the categories are compatible with the LIWC. TextMind works effectively with high performance.

All tweets were first converted into CSV format aggregating all posts by a person, with one record representing one person's data. The Language Technology Platform (LTP) was then selected to segment the text words, which provided a complete set of bottom-up, rich, and efficient Chinese language processing modules, including six core Chinese processing technologies, such as morphology, syntax, and semantics. Lastly, the frequency of word categories was calculated and extracted through the TextMind system.

4 Results

4.1 Correlations with personality traits

The way language is used in online social network platforms such as Sina Weibo is much more diverse compared to formal text materials (Tausczik and Pennebaker 2010). The TextMind system included the high-frequency usage words (such as those on microblogs) beyond the simplified Chinese linguistic inquiry and word count (SCLIWC). For example, words conveying psychological processes include subcategories such as social processes, affective processes, and cognitive processes. Based on internal reliability and external validity test, the TextMind system works more effectively than the exclusive application of SCLIWC (Gao et al. 2013). Together, TextMind extracted 102 word categories, including special categories in Chinese and English dictionaries. This provides the opportunity to analyze and explain the personality expression patterns in the Chinese language in a way that is more specific and detailed than the study by Qiu et al. (2017).

Several special word categories in Chinese were more suitable for the Chinese context than English. For instance, multifunction words [e.g., 的 (of), 有 (have), 是 (is)] and other words which expressed tense, such as tense makers [e.g., 已经 (already), 之前 (before), 日后 (after)], were better expressed in the Chinese context because there were many similar words used to express time nodes in Chinese that were starkly different from English tense words.[6] Some categories in the TextMind system were similar to the English version of LIWC, which contained summary language variables (e.g., word count, word per sentence, words > 6 characters, rate numeral), linguistic dimensions (e.g., total function words, personal pronouns, I, We, You, verb, adverb, conjunction, etc.), and psychological processes. These factors can also be separated into several subcategories, such as affective processes (e.g., anxiety, negative emotion, positive emotion), social processes (e.g., humans, friends, family), cognitive processes (e.g., tentative, discrepancy, causation, insight), perceptual processes (e.g., feel, hear, see), biological processes (e.g., ingestion, sexual, health, body), drives (e.g., achievement), relativity (e.g., time, space), personal concerns (e.g., death, religion, money, home, work, leisure), and informal language (e.g., non-fluent, accent, swearwords) (Pennebaker et al. 2015). Notably, the TextMind system had a few particular word categories which did not appear in other dictionaries, such as the rate of four-character words, rate of Latin words, mention frequency, amount of emotions, number of hashtags, and number of URLs.

Table 2 illustrates the correlations between word frequencies and participants' personality traits. Comparisons with previous English samples were marked in parenthesis. The correlations indicated relatively stable personality expressions in language usage. Extraversion was positively correlated with friends and humans, suggesting that extraverts enjoy meeting people, seeking excitement from the social world, and being friendly to others. This was consistent with the theoretical definition of extraversion, as well as other studies using English samples (Farnadi et al. 2016; Hirsch and Peterson 2009; Mairesse et al. 2007; Qiu et al. 2012; Scott 2006; Yarkoni 2010). However, in contrast to the perception of extraverts as having positive emotions, extraversion was associated with negative emotion and anger in the Weibo data; this result was consistent with an English conversation study that analyzed the relationship between observer reports of personality score and language usage (Mairesse et al. 2007). This shows that the way extraverts use language likely depends on the method of communication and the language context. In the context of online social media, extraverts are perhaps less reserved from expressing dissatisfaction.

Agreeableness was associated with function words, personal pronouns (PPron), and the first-person singular nominative pronoun "I," replicating the correlations found in the English corpus (Mehl et al. 2006; Pennebaker and King 1999; Yarkoni 2010). These results showed that people with high agreeableness gave more attention to the selves in their language and that they wished to keep good relationships with others. Conscientiousness was positively related to cognitive processes (CogMech);

[6] In Chinese, to convey the time node of "in a moment," one can use multiple expressions, for examples, guoyihui (过一会), denghui (等会), piankehou (片刻后), shaodeng (稍等).

Table 2 Correlations between personality and LIWC word categories through the TextMind program

	Word category	Examples	Mean	SD	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Language variable	Funct	或许 (maybe), 许多 (many), 那些 (those)	.32	.09	.09	**.18*** ([6])	.12	.12	.14
	Pronoun	你 (you), 他们 (they), 它 (it)	.05	.03	.14	.17	.09	**.20*** ([3][6][7])	.08
	PPron	他 (he), 你们 (you), 在下 (I)	.04	.02	.14	**.22*** ([6][10])	.11	**.22*** ([6])	.07
	I	本人 (I), 我 (I)	.03	.02	.11	**.19*** ([5][9][10])	.13	**.25**** ([1][5][6][10][11])	.09
	Quant	一些 (some), 所有 (all), 众多 (many)	.02	.01	−.03	−.12	.02	−.09	**.17***
	MultiFun	的(of), 有(have), 是(is)	.06	.03	−.08	**.18***	.12	.15	**.19***
Social processes	Friend	同伴 (company), 朋友 (friend), 麻吉 (best friend)	.00	.00	**.18*** ([2][6][10])	−.03	−.06	.08	.16
	Humans	成人 (adult), 宝宝 (baby), 男孩 (boy)	.01	.01	**.18*** ([2][4][6][7][8][10])	.09	.03	.03	−.01
Affective processes	NegEmo	担忧 (worried), 丑恶 (ugly), 糟糕 (terrible)	.01	.01	**.18*** ([5][6])	.05	−.04	.05	−.01
Negative emotions	Anger	可恶 (damn), 抱怨 (complain), 破坏 (damage)	.00	.00	**.20*** ([3][6])	.13	−.12	.05	.08
Cognitive processes	CogMech	理解 (understand), 选择 (choose), 质疑 (question)	.14	.04	−.03	.10	**.18*** ([3])	.15	.07
	Insight	了解 (understand), 恍然大悟 (know), 体会 (realize)	.01	.01	−.05	.14	.05	**.23**** ([6])	.12
	Cause	引起 (cause), 使得 (make), 变成 (become)	.01	.00	−.14	.04	.13	**.20*** ([6])	.10

Table 2 (continued)

	Word category	Examples	Mean	SD	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Current concerns	Sexual	上床 (sex), 性欲 (lust), 裸体 (nudity)	.00	.00	.01	**.20*** ([10])	−.02	−.00	.15
	Money	富有 (wealth), 年薪 (salary), 折扣 (discounts)	.02	.02	−.10	−.04	**.24****	.07	.02
Informal language	Nonfl	呃 (uh), 然后 (then), 那 (that)	.00	.01	.01	**.19***	.07	.03	.07
	Filler	就 (just), 像是 (like) 话说 (say)	.01	.01	−.09	−.03	.00	−**.20***	−.05
Positive emotion	Love	爱 (love), 友好 (nice), 甜蜜 (sweet)	.00	.00	.13	.03	.07	.10	**.18***
Relativity	tNow	今天 (today), 现在 (now)	.00	.00	**.24**** ([8])	.10	.04	.00	.15
Punctuation	Comma	, (comma)	.08	.06	.003	−.09	.04	**.18***	−**.25****
	QMark	?(question mark)	.00	.01	.02	−.07	−.11	−**.21*** ([3])	−.09
	Quote	“”(quotation marks)	.01	.02	−**.18***	.02	.13	.12	.03
	Apostrophe	e.g., (Abbreviation symbol)	.00	.00	−.12	.13	**.18*** ([3])	.13	−.01
	OtherP	;(Other punctuation)	.04	.03	**.20***	−.04	−.01	−.01	**.18***

Table 2 (continued)

	Word category	Examples	Mean	SD	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Summary language variables	RateDicCover	(Dictionary coverage rate)	.79	.06	**.18***	.03	.09	−.04	−.07
	RateNumeral	(Number ratio)	.03	.03	−.12	.05	.08	**.22**** ([3])	.01
	RateSixLtrWord	(Ratio of the length of words longer than or equal to 6)	.02	.03	−.07	.02	.05	**.28**** ([3][6])	−.01
	RateFourCharWord	(Ratio of the length of words longer than or equal to 4)	.04	.04	−.08	.04	.05	**.24****	.03
	RateLatinWord	拉丁词比率(ratio of Latin words)	.05	.05	−.08	.06	.04	**.17***	.08

Bold values indicate that the correlations were significant

Mean value represents the percentage of the frequencies of each word categories. SD stands for stander deviation. Only features that correlate significantly with at least one trait are shown. Parenthesis indicates findings in prior English corpus, and the corresponding numbers are marked in the references

*$p < .05$; **$p < .01$

[1] Gill and Oberlander (2003)

[2] Hirsch and Peterson (2009

[3] Mairesse et al. (2007)

[4] Majumder et al. (2017)

[5] Pennebaker and King (1999)

[6] Qiu et al. (2012)

[7] Qiu et al. (2017)

[8] Scott (2006)

[9] Tellegen (1993)

[10] Yarkoni (2010)

[11] Yee et al. (2011)

highly conscientious individuals tend to use words such as understanding, choices, and questions. This also illustrated that people with high conscientiousness were more likely to be planful, reliable, and well-organized (Lee et al. 2007; Mairesse et al. 2007). Neuroticism correlated with pronouns, personal pronouns, and the first-person singular nominative pronoun "I," suggesting that neurotic individuals tend to view issues from the self-perspective. This finding was consistent with English corpus research (Gill and Oberlander 2003; Mairesse et al. 2007; Pennebaker and King 1999; Yarkoni 2010). Neuroticism was also positively related to insight and cause, suggesting that neurotic people were more sensitive to external stimuli, which was consistent with earlier discoveries (Yarkoni 2010). Also, neuroticism was negatively related to question marks (Qmark), which echoed the result of Mairesse et al. (2007). Openness was positively related to "quantification" and love, suggesting that open people were more curious about other things and loved new things.

Besides the consistent findings with previous research based on the English corpus, there were some unique findings in the Chinese language usage context. As a special word category in the Chinese dictionary, multifunction (MultiFun) was positively correlated with agreeableness and openness, indicating that agreeable or open people used Chinese words such as 的 (of), 有 (have), and 是 (is) more frequently. Also, two distinctive categories in TextMind, namely four-character words and Latin words, were significantly related to neuroticism. Furthermore, neuroticism was related to the rate of numeral- and six-character words (Mairesse et al. 2007). This reveals that highly neurotic individuals are more likely to use long words, numbers, and characters in their language.

Furthermore, the differences between personality expression in the English and Chinese languages were also reflected by the use of punctuation. Extraversion was negatively related to the quotation mark, but positively related to other punctuation; conscientiousness was correlated with the use of an apostrophe, which was also found in the study by Yarkoni (2010); neuroticism was positively related to the use of a comma; openness was negatively related to a comma and positively related to other punctuation. Associations between different personality traits and punctuation habits were not significant in English samples, suggesting that punctuation in Chinese was more complex.[7] For the category of informal language, agreeableness was associated with non-fluent language, suggesting agreeable people use more euphemistic, repeated, and redundant words. Neuroticism was negatively related to filler words, which reflects language expression that is uncertain and casual. As for the category of current concern, agreeableness was positively related to sexual words, consistent with a previous study (Yarkoni 2010); conscientiousness was significantly related to money, reflecting rigorousness and carefulness.

[7] Some punctuation marks in Chinese are not found in English, such as "、" plays a role of separating the juxtaposition in a sentence in Chinese, which does not exist in English; "《》" is used as title marks in Chinese; "." is used between the month and date, and between the translated given name and family name in Chinese; "……", or solid dots underneath the text, are used to indicate emphasis.

Table 3 Stepwise linear regression results of language features on Big Five dimensions

Dimension	Model	R^2	Unstandardized coefficients		Standard coefficients	Collinearity statistics	
			B	Standard error	Beta	Tolerance	VIF
Extraversion	Constant		−5.56	1.58			
	tNow	.06	74.40	31.51	.19*	.94	1.06
	OtherP	.11	12.45	3.68	.28**	.92	1.08
	RateDicCover	.15	5.85	1.85	.28**	.79	1.27
	Work	.18	9.80	3.90	.21*	.89	1.12
Agreeableness	Constant		−0.16	0.25			
	PPron	.05	14.47	4.35	.28**	.89	1.12
	Space	.10	−15.21	4.33	−.32**	.78	1.29
	Psychology	.13	−16.89	6.84	−.20*	.94	1.07
	MultiFun	.17	8.87	3.56	.20*	.95	1.05
	Ingest	.20	15.33	7.09	.19*	.79	1.26
Conscientiousness	Constant		−0.26	0.11			
	Money	.06	13.80	4.35	.26**	.99	1.01
	Apostrophe	.10	69.64	28.15	.21*	.99	1.01
Neuroticism	Constant		−1.14	0.23			
	RateSixLtrWord	.08	9.18	2.92	.25**	.96	1.04
	Insight	.12	22.85	8.01	.24**	.91	1.10
	Pronoun	.15	10.05	3.69	.22**	.95	1.06
	Religion	.18	44.63	17.84	.20*	.97	1.03
	Money	.21	9.37	4.40	.18*	.87	1.15
Openness	Constant		0.43	0.30			
	Comma	.06	−10.85	3.58	−.25**	.95	1.05
	Love	.09	97.57	31.30	.27**	.86	1.17
	MultiFun	.13	13.55	3.92	.31**	.80	1.25
	Affect	.16	−10.02	4.57	−.20*	.80	1.25
	Filler	.18	−42.72	21.47	−.17*	.89	1.13

Only features that enter the model and correlate significantly with the personality trait are shown

*$p < .05$; **$p < .01$

4.2 Personality recognition model

Given that language usage is associated with personality traits, we can use features of word frequency to predict personality scores. In this paper, the prediction model was conducted based on the TextMind system outputs using a linear regression algorithm (following the method by Kosinski et al. (2013)). All of the word categories' frequency values were inputted as the independent variables and the personality scores as the dependent variables. All the variables had been normalized before analysis, and five linear regression models were built to predict the Big Five personality traits separately. Forward stepwise regression was used given the large number

of independent variables. Only independent variables with significant coefficients were entered into the model. The results are shown in Table 3.

Firstly, the variance inflation factor (VIF) values of the statistics were less than 2; this indicated that the independent variables in the constructed regression model had no collinearity. For each dimension of the five personality traits, the coefficients of the LIWC word categories entered into the model were significant at $p < 0.05$. As can be seen from Table 3, the predictors of extraversion included tense (tNow, $r = .19$), other punctuations (OtherP, $r = .28$), rate of dictionary cover (RateDicCover, $r = .28$), and work (Work, $r = .21$). The predictors of agreeableness included personal pronouns (PPron, $r = .28$), space (Space, $r = .-32$), psychology (Psychology, $r = -.20$), multifunction (MultiFun, $r = .20$), and ingest (Ingest, $r = .19$). The remaining three dimensions of the predictors are shown in Table 3. All coefficients are significant.

Based on the features that were related to personality significantly, we performed personality score predictions. All the results were averaged over a fivefold cross-validation, and the root mean-squared error (RMSE) was used to evaluate the results, which measured the difference between the predicted values by the regression model and the self-report values. The values of the RMSEs of prediction of each personality trait were .94 (extraversion), .97 (agreeableness), 1.01 (conscientiousness), .97 (neuroticism), and .97 (openness). The majority of the previous studies about personality prediction were based on English corpus, which usually conducted binary classification for personality instead of continuous scores. As such, we compared our results with Farnadi et al. (2016), which also predicted personality scores through regression using RMSE as the evaluation indicator, albeit with datasets from Facebook and YouTube content. The results showed that extraversion prediction in our study had lower RMSE value than Farnadi et al. (2016), which obtained a RMSE of .98 based on the YouTube dataset. The prediction accuracy of other personality dimensions was not significantly different. In order to further demonstrate the prediction performance of the language features, we also constructed a binary recognition model to predict personality.

4.3 Binary classification model

As aforementioned, much personality recognition research has been conducted using machine learning algorithms. To extend previous research in this study, a classification model was created to further recognize personality. For each dimension, if the personality score was higher than the mean score, it was labeled as 1. Personality scores lower than the mean score were labeled as 0. As such, we divided personality scores into two categories, which stand for a prominent or a not prominent personality trait. Then, we built the personality recognition model through the logistic regression method. The feature extractions were from the word categories significantly related to personality traits, as shown in Table 2. Lastly, all the results were averaged over a fivefold cross-validation, and the accuracy and precision values were calculated to evaluate the performance.

Table 4 Classification accuracy results as compared to prior essays corpus

Dataset	Method	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Weibo	LR-LIWC	**0.70**	**0.62**	**0.66**	**0.62**	**0.69**
Essays	NB-LIWC	0.53	0.54	0.54	0.56	0.59
Essays	SMO-LIWC	0.53	0.56	0.56	0.58	0.63

Bold values indicate the highest accuracy value of classification for each personality dimension

Table 5 Classification precision results compared to the Facebook corpus

Dataset	Method	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Weibo	LR-LIWC	0.67	0.58	**0.65**	**0.63**	**0.69**
Weibo	SVM	0.52	**0.62**	0.40	0.52	0.50
Facebook	SVM-LIWC	0.58	0.47	0.55	0.48	0.60
Facebook	SVM	**0.71**	0.60	0.45	0.36	0.50

Bold values indicate the highest precision value of classification for each personality dimension

We conclude that the features of word frequency extracted from the TextMind system are significant cues for personality recognition. As a comparison, we conducted another experiment through the support vector machine (SVM) algorithm, which is an effective classifier (Mairesse et al. 2007) and can extract features from content automatically. In the model, the features were extracted using means of term frequency-inverse document frequency (TF-IDF), which is widely used in the text pre-processing and feature engineering. Finally, we evaluated the precision. Both prediction results are presented in Tables 4 and 5.

Furthermore, we compared our results with previous studies that used the two common English corpora: Facebook and Essays. For instance, Alam et al. (2013) built a prediction model through the LIWC dictionary and used the NB and SMO algorithms to predict personality in essays. As Table 4 shows, the prediction accuracy of each personality trait in our result outperforms the other methods with essays. Extraversion had the highest accuracy of up to .70, which illustrates that extraversion expression in language is quite visible. The other four personality dimensions' recognition accuracy was all above .60, suggesting an excellent performance in identifying personality in the Weibo dataset. Also, as shown in Table 5, the prediction precision of conscientiousness, neuroticism, and openness through logistic regression with word frequency features was higher than that through SVM with automatic feature extraction. The precision rates were all greater than .60, with openness having the best prediction precision of .69.

Moreover, we compared our results with Corff and Toupin (2010), which analyzed the Facebook corpus through SVM and was effective in recognizing extraversion and agreeableness, with corresponding precision values of .71 and .62. Nevertheless, our results of conscientiousness, neuroticism, and openness outperformed Corff and Toupin (2010). In addition, Peng et al. (2015) conducted personality

Table 6 Correlation between Weibo seniority and LIWC word categories

LIWC word categories	Weibo seniority
Swear	.22*
Humans	.20*
NegEmo	.17*
Anger	.21*
Exclusive	.23**
Filler	.18*
tnow	.21*
Comma	− .29**
WordCount	.36**
RateDicCover	.22*
RateNumeral	− .31**
Ratesixltrword	− .27**
RateFourcharword	− .22*
NumEmotion	.21*
Extraversion	.07
Agreeableness	− .02
Conscientiousness	− .06
Neuroticism	− .08
Openness	.09

*$p < .05$; **$p < .01$

recognition of extraversion based on Chinese textual data published on Facebook. Using SVM as the classifier, this study achieved a best average prediction accuracy of .70 through the x^2-test for feature selection. This result is slightly better than ours in the extraversion dimension, perhaps because of a fewer number of selected features. Taken together, personality can be recognized accurately using linguistic features. For extraversion and agreeableness, SVM has a good prediction performance for both Chinese Weibo data and the English Facebook data. However, word frequency features extracted from the TextMind system have better performance on average.

4.4 The seniority of Weibo users and language usage habits

Given that Weibo documents each user's extent of activity, which is described as the Weibo seniority, it would be informative to examine whether Weibo seniority has any relationship to language usage patterns. Weibo seniority was determined by the accumulated experience of each individual on the website—each day, when users complete certain tasks, they can earn experience values and improve their level of seniority. These tasks include posting microblogs (creating or forwarding a post, five experience points), making continuous logins, adding followers (e.g., if the number of followers reaches 35, one will get 30 experience points), and following others.

Based on this information, we calculated the correlations between Weibo seniority, the Big Five personality model, and word categories (Table 6). The relationship between Weibo seniority and personality was not significant. However, results show that Weibo seniority was related to many word categories, as shown in Table 6. Specifically, Weibo seniority was positively related to swear ($r = .22$, $p = .01$), humans ($r = .20$, $p = .02$), negative emotion ($r = .17$, $p = .04$), anger ($r = .21$, $p = .01$), and exclusive ($r = .23$, $p = .00$). They were all significant at the level of $p < .05$. This indicated that the higher the Weibo seniority, the more likely people tend to curse, express negative emotions, and vent their moods on social media. Words such as "moron," "worry," "suspicion," "revenge," "hateful," "complain," and so on appear in their language more frequently. Besides, Weibo seniority was also correlated with filler, now tense (tnow), word count, rate of dictionary cover (RateDicCover), and the number of emotions (numemotion). This suggested that users with more extensive Weibo usage preferred to use more emoticons, present tense, and filler words in their language expression. However, Weibo seniority was negatively related to comma, rate of number (ratenumeral), rate of six-character words (ratesixltrword), and rate of four-character words (ratefourcharword). This further indicated that these people were accustomed to expressing their emotions directly, concisely, and sharply.

5 Discussion

Although many studies have explored the correlations between English language usage and personality traits, this study further extends personality expression research based on the Chinese Weibo language. There are clear traces of evidence on how English and Chinese are used for different personality traits (Chen and Bond 2010). Also, it has been revealed that personality expression in any language is influenced by culture and context (Triandis and Suh 2002). Therefore, in the investigation of other languages and personality, influential factors such as culture and context must be considered.

5.1 Personality expression

Given that language is one of the most natural and cost-effective means to understand personality, many previous studies have investigated the relationship between personality traits and language patterns and created prediction models to identify personality using various corpuses in English social networking platforms (Golbeck et al. 2011b; Iacobelli and Culotta 2013; Skowron et al. 2016). This study extends the existing research by examining the associations between Chinese Weibo content and personality traits through the TextMind system. The results showed that Sina microblogs contained valid linguistic cues to predict personality. These linguistic personality associations are largely consistent with previous studies, although there are some differences between the English and Chinese corpora on language usage habits. In particular, we found that extraversion is positively correlated with friends

and humans, which belong to the social process word category. Agreeableness is positively related to function words, personal pronouns, and sexual words. Conscientiousness correlates with cognitive processes and an apostrophe. Neuroticism is significantly related to pronoun and first-person pronouns. Openness is positively correlated with quant and positive emotion words.

Furthermore, this study identifies some unique associations between Chinese social media activities and personality, which are different from their relationships in the English corpus. For instance, agreeableness and openness are positively correlated with multifunction words, which is a special word category in the Chinese dictionary of the TextMind system. Agreeableness and neuroticism are correlated with informal languages, such as non-fluent and filler words. Conscientiousness is positively related to current concerns, such as money.

Punctuation use habits are also diverse in Chinese and other languages. Extraversion is correlated with quotation marks, while neuroticism is related to the comma. These findings reflect the fact that Weibo offers a platform that encourages individuals to talk, share their daily lives, and express emotions freely. Although sharing some similarities with English social media, the activities on Weibo are also subject to the context of Chinese language and therefore exhibit complicated grammar. Different punctuation and sentence length can also reflect different emotions. This information is uniquely captured by studying the Weibo content, which supplements the previous research that was primarily conducted in the English context.

Interestingly, we found that language usage habits are also a factor of the Weibo user's seniority. The correlation analysis between word frequency features and Weibo seniority suggests that Weibo seniority is positively related to swearing, anger, negative emotion, and exclusive word categories. This appears to be in line with the belief that online social networking sites are a platform for people to demonstrate their ideal self (Back et al. 2010). The anonymous and open-access nature of Weibo creates a virtual world in which the social display rules do not constrain individuals in their virtual lives. In this virtual life, users participate in the discussion of hot topics, celebrity gossip, or social problems, and often insult and attack others, in a so-called cyber violence phenomenon (Hanewald 2008), to vent their negative emotions.

Given that many people must conform to certain behavioral norms in their personal and work lives, most individuals are often required to externally exhibit positive and friendly attitudes and behaviors in their real-life transactions. However, when they are internally disgruntled or dissatisfied with their work or life, it creates a level of emotional dissonance (Morris and Feldman 1996), which is defined as the discrepancy between ones' experienced emotions and displayed emotions. Nonetheless, there is a limited opportunity for people to release such negativity in their daily lives except through social networking platforms like Weibo. Unlike WeChat or Facebook, Weibo is *anonymous*, which makes the consequences of one's negative behaviors significant to the recipient but negligible to the benefactor. As such, dumping negative emotions on online social networking sites can help these individuals relieve their emotional dissonance. This echoes the argument that people who experience loneliness or social frustration in real life are likely to choose sites that impose less social pressure to relieve their emotions (Juan et al. 2018).

We found that the most seasoned Weibo users tend to express their opinions online without reservations. It would be informative for future research to replicate this finding in the Western context. Perhaps there is a more considerable gap between Chinese internet users' true self and ideal self than Westerners due to the common Chinese cultural practice of preserving "face" to appear appropriate and hide true feelings in front of others (Lim et al. 2015). As such, Chinese internet users may experience more emotional dissonance, which creates a greater need for them to release such tension on an anonymous online platform or a platform in which they do not need to face their counterparts directly.

5.2 Personality recognition

Corroborating the findings of previous research, this study shows that language usage habits and word choices can reflect individual personality traits. Employing a split-sample approach, we used the word frequency features to build a personality prediction model. We adopted the linear regression function to predict the continuous personality score. To test the effect of the prediction, we compared our results with the commonly used English corpus using RMSE. Results show that the prediction using the word frequency features has great performance. This illustrates that personality traits can be effectively identified through people's language content in online social networking activities on Weibo.

Furthermore, with the machine learning algorithm development, we developed five binary classification models for personality prediction through logistic regression. Compared with previous studies, the average precision value of this prediction model was higher, especially for conscientiousness, neuroticism, and openness. Both results support the argument that word frequency features extracted by the TextMind were a better predictor for personality traits than other feature sets (Farnadi et al. 2013, 2016). This prediction model enables faster and more convenient personality prediction, which has a profound implication for human resource managers to use social networking activities on Weibo as a basis to predict individual personality.

5.3 Practical contributions

Our personality prediction model has many practical applications in the modern world. As aforementioned, personality traits are associated with an individual's behavior and psychology activity. Given the significance of personality recognition, human resource managers in companies have utilized personality questionnaire-based assessments as part of the recruitment and selection processes. However, with self-report, many applicants do not complete these personality surveys based on the true self but provide socially desirable results (Crowne and Marlowe 1960). Therefore, it is essential and meaningful to find a quick, effective, and objective means to identify an employee' s personality traits. From the study results, HR managers can identify an employee's personality by browsing their microblog to identify their language usage habits. Alternatively, employees can also be asked to write an essay. Based on these language activities, human resource managers can easily capture the candidate's word categories

and frequency through the TextMind system using the prediction model. This would provide a better assessment for their personality traits compared to the traditional self-report approach, and HR managers can select targeted candidates according to the requirement of specific positions. For example, people with high scores of extraversion and openness were shown to be good candidates for sales positions (Turnbull 1976). If the organization intends to recruit employees with high extraversion, based on the findings of word categories correlation with certain personality traits, the recruitment team can determine whether the employee's language usage contains words such as friends, humans, and social process. This approach is not only more cost-effective but also provides accurate insights regarding the personality traits of the candidate. More notably, this method facilitates massive recruitment, hence improving the process efficiency.

It is worth noting that organizations need to adhere to proper ethical guidelines for collecting data from sources such as Weibo. First, human resource managers need to present an informed consent document, which explains what data will be collected and how the organization will use the network data and analysis. Second, there needs to be an option that allows the candidates to provide their social network account voluntarily. Finally, HR needs to ensure confidentiality once data are collected, e.g., the nickname or id number can be replaced by other codes such that no one else would know the identifying information except the key personnel (Borgatti and Molina 2005).

Besides human resource management, our model also has other practical implications in consumer products businesses, especially e-commerce. Personality recognition helps to first provide target services and products to a specific market niche of customers, which could be categorized by different personality traits. Furthermore, personality trait recognition is used in the detection of deception, recognition of criminals, and opinion mining. With the fast development of machine learning and artificial intelligence, the demand for computational psychology application is expected to continuously increase; hence, personality trait identification and prediction is expected to be applied more extensively.

5.4 Limitations and future directions

Personality traits are imperative in many aspects of life. While this paper focuses only on language usage as a key predictor of personality traits, future studies should explore other aspects of individual activities such as behaviors (Lepri et al. 2016). This would provide multi-dimensional insights in revealing personality traits; hence, it would enhance the effectiveness as well as accuracy in personality prediction. Also, one fundamental premise of this study was that it focused on the words categorized by TextMind from a linguistic point of view. As such, future research should encompass more information on social media, such as comments, retweets, number of friends, likes, and followers. We would like to recognize that given that so many analyses were performed in this study, there was a chance of Type 1 error occurring, i.e., some relationships may be significant by chance. Future research can replicate this study using a larger sample size to reduce the probability of this error, and to obtain more accurate and reliable insights regarding personality traits and personality prediction.

6 Conclusion

This study applies the recent development in Chinese Weibo content to examine personality expression in language usage habits. The study corroborates previous correlations between personality and word categories through the use of the TextMind system and also identifies new relationships that have not been previously documented in the English context. This points to the fact that there exists a difference between English and Chinese language usage concerning personality traits. Furthermore, we build a robust regression function that facilitated personality prediction through word frequency features. From the results obtained, it was evident that the prediction model can be used as a reliable and effective tool to identify personality traits using the Chinese corpus. Lastly, the paper provides HRM recommendations based on personality identification and prediction results, which are aimed to enhance the effectiveness of recruitment and selection, as well as to inform other divisions of management. There is a growing interest in the use of personality modeling for adaptive systems, as can be seen in the UMUAI 2016 special issue on "Personality in Personalized Systems" (Tkalčič et al. 2016). Personality profiles have been used to investigate adaptations in persuasive technology, intelligent tutoring systems, recommender systems, etc. (Smith et al. 2019). The model of personality recognition we constructed is essentially an adaptive system. Human resource management can select suitable employees according to their personality profiles and conduct target training. Based on the personality identification, this allows the adaptability of human resource management according to candidates' personality.

Acknowledgements Dr. Junjie Wu's work was partially supported by the National Key R&D Program of China (2019YFB2101804), and the National Natural Science Foundation of China (71725002, 71531001, U1636210).

Authors' contributions All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Cuixin Yuan and Ying Hong. The first draft of the manuscript was written by Cuixin Yuan. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Aaronson, D., Ferres, S.: The impact of language differences on language processing: an example from Chinese–English bilingualism. In: Homet, P., Palij, M., Aaronson, D. (eds.) Childhood Bilingualism: Aspects of Linguistic, Cognitive, and Social Development, pp. 75–119. Erlbaum, Hillsdale (1987)

Alam, F., Stepanov, E.A., Riccardi, G.: Personality traits recognition on social network-facebook. In: Proceedings of Workshop on Computational Personality Recognition, pp. 6–9. AAAI Press, Melon Park, CA (2013)

Back, M.D., Stopfer, J.M., Vazire, S., Gaddis, S., Schmukle, S.C., Egloff, B., Gosling, S.D.: Facebook profiles reflect actual personality, not self-idealization. Psychol. Sci. **21**(3), 372–374 (2010)
Bai, S., Zhu, T., Cheng, L.: Big-five personality prediction based on user behaviors at social network sites. Comput. Sci. **8**(2), e2682–e2682 (2012)
Bargh, J.A., McKenna, K.Y., Fitzsimons, G.M.: Can you see the real me? Activation and expression of the ''true self'' on the Internet. J. Soc. Issues **58**(1), 33–48 (2002)
Borgatti, S.P., Molina, J.: Toward ethical guidelines for network research in organizations. Soc. Netw. **27**(2), 107–117 (2005)
Brislin, R.W.: Back-translation for cross-cultural research. J. Cross Cult. Psychol. **1**(3), 185–216 (1970)
Celli, F., Lepri, B., Biel, J.I., Gatica-Perez, D., Riccardi, G., Pianesi, F.: The workshop on computational personality recognition 2014. In: Proceedings of the ACM International Conference on Multimedia—MM'14, pp. 1245–1246 (2014)
Chan, M., Wu, X., Hao, Y.Q., Xi, R., Jin, T.: Microblogging, online expression, and political efficacy among young Chinese citizens: the moderating role of information and entertainment needs in the use of Weibo. Cyberpsychol. Behav. Soc. Netw. **15**(7), 345–349 (2012)
Chen, S.X., Bond, M.H.: Two languages, two personalities? Examining language effects on the expression of personality in a bilingual context. Pers. Soc. Psychol. Bull. **36**(11), 1514–1528 (2010)
Cohen, A.S., Minor, K.S., Baillie, L.E., Dahir, A.M.: Clarifying the linguistic signature: measuring personality from natural speech. J. Pers. Assess. **90**, 559–563 (2008)
Corff, Y.L., Toupin, J.: The five-factor model of personality at the facet level: association with antisocial personality disorder symptoms and prediction of antisocial behavior. J. Psychopathol. Behav. Asses. **32**(4), 586–594 (2010)
Costa, P.T., McCrae, R.: The revised NEO personality inventory (NEO-PI-R). In: Boyles, G., Matthews, G., Saklofske, D. (eds.) The SAGE Handbook of Personality Theory and Assessment: Volume 2-Personality Measurement and Testing, pp. 179–199. Sage, London (2008)
Crowne, D.P., Marlowe, D.: A new scale of social desirability independent of psychopathology. J. Consult. Psychol. **24**(4), 349–354 (1960)
Ellingson, J.E., Sackett, P.R., Hough, L.M.: Social desirability corrections in personality measurement: issues of applicant comparison and construct validity. J. Appl. Psychol. **84**(2), 155–166 (1999)
Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., De Cock, M.: Computational personality recognition in social media. User Model. User Adapt. Inter. **26**(2–3), 109–142 (2016)
Farnadi, G., Zoghbi, S., Moens, M.-F., De Cock, M.: Recognising personality traits using facebook status updates. In: Proceedings of Workshop on Computational Personality Recognition, pp. 14–18. AAAI Press, Melon Park, CA (2013)
Fast, L.A., Funder, D.C.: Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior. J. Pers. Soc. Psychol. **94**, 334–346 (2008)
Freud, S.: Psychopathology of Everyday Life. Basic Books, New York (1901)
Gao, R., Hao, B., Li, H., Gao, Y., Zhu, T.: Developing simplified Chinese psychological linguistic analysis dictionary for microblog. In: International Conference on Brain and Health Informatics, pp. 359–368. Maebashi, Japan (2013)
Gill, A., Oberlander, J.: Perception of email personality at zero-acquaintance: extraversion takes care of itself; neuroticism is a worry. In: Proceedings of the 25th Annual Conference of the Cognitive Science Society, pp. 456–461. Boston, MA (2003)
Golbeck, J., Robles, C., Turner, K.: Predicting personality with social media. In: Proceeding of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA'11, pp. 253–262. New York, USA (2011a)
Golbeck, J., Robles, C., Edmondson, M., Turner, T.: Predicting personality from twitter. In: Proceedings of the 3rd IEEE International Conference on Social Computing, pp. 149–156 (2011b)
Goldberg, L.R.: The development of markers for the big-five factor structure. Psychol. Assess. **4**(1), 26–42 (1992)
Goldberg, L.R.: Language and individual differences: the search for universals in personality lexicons. In: Wheeler, L. (ed.) Review of Personality and Social Psychology, 2(1), pp. 141–165. Sage, Beverly Hills (1981)
Hanewald, R.: Confronting the pedagogical challenge of cyber safety. Aust. J. Teach. Edu. **33**(3), 1–16 (2008)
Holtgraves, T.: Text messaging, personality, and the social context. J. Res. Pers. **45**(1), 92–99 (2011)

Hirsch, J.B., Peterson, J.B.: Personality and language use in self-narratives. J. Res. Pers. **43**, 524–527 (2009)
Iacobelli, F., Culotta, A.: Too neurotic, not too friendly: structured personality classification on textual data. In: Proceeding of the Workshop on Computational Personality Recognition, pp. 19–22. AAAI Press, Melon Park, CA (2013)
Juan, H., Yamikani, N., Xuefei, P., Shuangyi, C., Fei, X., Xiaochu, Z.: Weibo or wechat? Assessing preference for social networking sites and role of personality traits and psychological factors. Front. Psychol. **9**, 545 (2018)
Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proc. Natl. Acad. Sci. U.S.A. **110**(15), 5802–5805 (2013)
Lacan, J.: The Language of the Self: The Function of Language in Psychoanalysis. Johns Hopkins Press, Baltimore (1968)
Lee, C.H., Kim, K., Seo, Y.S., Chung, C.K.: The relations between personality and language use. J. Gen. Psychol. **134**(4), 405–413 (2007)
Lepri, B., Staiano, J., Shmueli, E., Pianesi, F., Pentland, A.: The role of personality in shaping social networks and mediating behavioral change. User Model. User Adapt. Inter. **26**(2), 143–175 (2016)
Li, C.N., Thompson, S.A.: Mandarin Chinese: A Functional Reference Grammar. University of California Press, London (1989)
Light, T., Henne, H., Rongen, O.B., Hansen, L.J.: A handbook on Chinese language structure. J. Asian Stud. **38**(2), 376 (1979)
Lim, J.S., Nicholson, J., Yang, S.U., Kim, H.K.: Online authenticity, popularity, and the "real me" in a microblogging environment. Comput. Hum. Behav. **52**, 132–143 (2015)
Liu, B.: Effect of first language on the use of English discourse markers by L1 Chinese speakers of English. J. Pragmat. **45**(1), 149–172 (2013)
Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using linguistic cues for the automatic recognition of personality in conversation and text. J. Artif. Intell. Res. **30**, 457–500 (2007)
Mairesse, F., Walker, M.A.: Towards personality-based user adaptation: psychologically informed stylistic language generation. User Model. User Adapt. Inter. **20**(3), 227–278 (2010)
Majumder, N., Poria, S., Gelbukh, A., Cambria, E.: Deep learning-based document modeling for personality detection from text. IEEE Intell. Syst. **32**(2), 74–79 (2017)
Matthews, G., Deary, I.J., Whiteman, M.C.: Personality Traits, 2nd edn. Cambridge University Press, Cambridge Books Online (2003)
McCrae, R.R., Costa Jr., P.T.: Personality trait structure as a human universal. Am. Psychol. **52**, 509–516 (1997)
Mehl, M.R., Gosling, S.D., Pennebaker, J.W.: Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. J. Pers. Soc. Psychol. **90**(5), 862–877 (2006)
Morris, J.A., Feldman, D.C.: The dimensions, antecedents, and consequences of emotional labor. Acad. Mana. Rev. **21**(4), 986–1010 (1996)
Norman, W.T.: 2800 Personality Trait Descriptors: Normative Operating Characteristics for a University Population. Department of Psychology, University of Michigan, Ann Arbor (1967)
Nowson, S., & Oberlander, J.: Identifying more bloggers: towards large scale personality classification of personal weblogs. In: Proceedings of the International Conference Personality Traits Recognition on Social Non Weblogs and Social Media (2007)
Park, G., Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Kosinski, M., Stillwell, D.J., Seligman, M.E.P.: Automatic personality assessment through social media language. J. Pers. Soc. Psychol. **108**(6), 934–952 (2015)
Paul, L.M., Simons, G.F., Fennig, C.D.: Ethnologue: Languages of the World. SIL International, Dallas (2015)
Peng, K.H., Liou, L.H., Chang, C.S., Lee, D.S.: Predicting personality traits of Chinese users based on Facebook wall posts. In: 2015 24th Wireless and Optical Communication Conference (WOCC), Taipei, 2015, pp. 9–14 (2015)
Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G.: Psychological aspects of natural language use: our words, our selves. Annu. Rev. Psychol. **54**(1), 547–577 (2003)
Pennebaker, J.W., Booth, R.J., Francis, M.E.: Linguistic Inquiry and Word Count: LIWC 2007. LIWC, Austin (2007)
Pennebaker, J.W., King, L.A.: Linguistic styles: language use as an individual difference. J. Pers. Soc. Psychol. **77**(6), 1296–1312 (1999)

Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The Development and Psychometric Properties of LIWC2015. University of Texas at Austin, Austin (2015)
Qiu, L., Lin, H., Ramsay, J., Yang, F.: You are what you tweet: personality expression and perception on twitter. J. Res. Pers. **46**(6), 710–718 (2012)
Qiu, L., Lu, J., Ramsay, J., Yang, S., Qu, W., Zhu, T.: Personality expression in Chinese language use. Int. J. Psychol. **52**(6), 463–472 (2017)
Sanford, F.H.: Speech and personality. Psychol. Bull. **39**(10), 811–845 (1942)
Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Ungar, L.H.: Personality, gender, and age in the language of social media: the open vocabulary approach. PLoS ONE **8**, 773–791 (2013)
Scott, N.: The language of weblogs: a study of genre and individual differences. PhD thesis, University of Edinburgh (2006)
Skowron, M., Tkalčič, M., Ferwerda, B., Schedl, M.: Fusing social media cues. In: Proceedings of the 25th International Conference Companion on World Wide Web—WWW '16 Companion, pp. 107–108 (2016)
Smith, K.A., Dennis, M., Masthoff, J., Tintarev, N.: A methodology for creating and validating psychological stories for conveying and measuring psychological traits. User Model. User Adapt. Interact. **29**, 573–618 (2019)
Stirman, S.W., Pennebaker, J.W.: Word use in the poetry of suicidal and nonsuicidal poets. Psychosom. Med. **63**, 517–522 (2001)
Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: lIWC and computerized text analysis methods. J. Lang. Soc. Psychol. **29**(1), 24–54 (2010)
Tellegen, A.: Folk concepts and psychological concepts of personality and personality disorder. Psychol. Inq. **4**, 122–130 (1993)
Tkalčič, M., Quercia, D., Graf, S.: Preface to the special issue on personality in personalized systems. User Model. User Adapt. Inter. **26**(2–3), 103–107 (2016)
Triandis, H.C., Suh, E.M.: Cultural influences on personality. Annu. Rev. Psychol. **53**, 133–160 (2002)
Allen, A.: Selling and the salesman: prediction of success and personality change. Psychol. Rep. **38**(3_suppl), 1175–1180 (1976)
Yarkoni, T.: Personality in 100,000 words: a large-scale analysis of personality and word use among bloggers. J. Res. Pers. **44**, 363–373 (2010)
Yee, N., Harris, H., Jabon, M., Bailenson, J.N.: The expression of personality in virtual worlds. Soc. Psychol. Personal. Sci. **2**(1), 5–12 (2011)
Zhang, L., Pentina, I.: Motivations and usage patterns of Weibo. Cyberpsychol. Behav. Soc. Netw. **15**(6), 312–317 (2012)

Cuixin Yuan is a PhD student at School of Economics and Management, Beihang University. Her main research focus on organization behavior, personality recognition, motivation, machine learning in computational psychology.

Ying Hong is an associate professor in the Leading People and Organizations area at the Gabelli School of Business, Fordham University. Prior to this appointment, she was an assistant professor in the DeGroote School of Business at McMaster University. She received her PhD in industrial relations/human resources from Rutgers University. She specializes in research on the strategic role of human resource management, behavior, and psychology. Her work has appeared in peer-reviewed journals such as the Academy of Management Journal and the Journal of Applied Psychology, and she received the Scholarly Achievement Award from the human resources division of the Academy of Management and Dean's Research Awards from the Gabelli School of Business.

Junjie Wu received the BE degree in civil engineering and the PhD degree in management science and engineering from Tsinghua University. He is currently a full professor with the Information Systems Department, School of Economics and Management, Beihang University, the director of the Research

Center for Data Intelligence (DIG), and the vice director of the Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations. His general area of research is data mining and complex networks, with special interests include social, urban, and financial computing. He is the recipient of various national awards in China, including NSFC Distinguished Young Scholars, MOE Changjiang Young Scholars, and MOE Excellent Doctoral Dissertation.