Subjective well-being measurement based on Chinese grassroots blog text sentiment analysis

Jiayin Qi [a,*], Xiangling Fu [b,1], Ge Zhu [a,1]

[a] School of Economic and Management, Beijing University of Posts and Telecommunications, No. 10 Xitucheng Road, Haidian District, Beijing 10086, China
[b] School of Software, Beijing University of Posts and Telecommunications, No. 10 Xitucheng Road, Haidian District, Beijing 10086, China

A R T I C L E   I N F O

A B S T R A C T

In this study, we propose a new method to measure the subjective well-being (SWB) of Chinese people. Based upon the classic framework in psychology, our model constructs a system of multiple weighted emotions in positive and negative affect by applying a text-sentiment analysis. To study SWB in the Chinese context, we also establish and supplement our model with a new lexicon, Ren-CECps-SWB 2.0. Tests on the data of 7 years of grassroots blogs on Sina.com demonstrate the validity of our model. Employing the same data, we find interesting patterns of the SWB of Chinese people on weekly and monthly bases.

© 2015 Published by Elsevier B.V.

1. Introduction

Though self-report scales are the most popular means in psychological studies to measure SWB, self-report scales have disadvantages, such as the limited samples they can assess, their high cost, and their sensitivity to participants' memory, which makes it hard to present the real-time status of respondents [10,15,16,18]. With the rapid development of online social network services (SNS) [25], increasing numbers of people are creating user-generated content (UGC) on the Internet to express their emotions [14]. Because of the rich information in UGC [1,2,22], scholars have attempted to measure SWB through UGC. For example, Dodds and Danforth used Affective Norms for English Words (ANEW) to measure SWB [10,11]. In 2009, Facebook released Facebook Gross National Happiness (FGNH) to measure the aggregate level of SWB [12,13]. Bollen and his colleagues [3] used a similar method with FGNH to calculate SWB. Dodds and Danforth's method is based on economic utility theory and uses valence value to predict SWB. FGNH applies the dualistic classification of negative and positive words to compute SWB.

Despite the insights generated by prior studies, as far as we know, the SWB measurements from text sentiment analysis seldom follow the established methods in psychology. Without harvesting the accumulative insights in psychology for SWB research, we might not be able to stand on the shoulders of giants to gain new understandings of the phenomenon under investigation. Moreover, the extant studies did not offer sufficient reasons why some emotions are selected into the SWB measurement and others are not. In this research, we attempt to overcome these research limitations by following the established psychological measurement for SWB based on multiple weighted emotions through text sentimental analysis.

In addition, although the research that uses sentiment analysis to measure SWB in an English-speaking context is gaining momentum, no research as of today has focused on Chinese text to measure the SWB of Chinese people. However, Chinese semantic analysis and English semantic analysis have striking differences. Applying the results from a sentiment analysis in English contexts to measure Chinese SWB may blind the real SWB of Chinese people because of the huge differences between English and Chinese cultures and languages. Fourth, no present corpus in Chinese can be used to build an SWB measurement model directly. Therefore, we intend to fill the current research gaps in Chinese SWB measurement applying a sentiment analysis technique in the Chinese language. We selected the Positive Affect and Negative Affect Schedule (PANAS) from psychology to construct our SWB

* Corresponding author. Tel.: +86 10 62282039; fax: +86 10 62282039.
E-mail addresses: qijiayin@bupt.edu.cn (J. Qi), fuxiangling@bupt.edu.cn (X. Fu), wwzhuge@gmail.com (G. Zhu).
[1] Tel.: +86 10 62282039; fax: +86 10 62282039.

model. On the basis of Ren-CECps, we constructed a Ren-CECps-SWB 2.0 Chinese lexicon. Integrating localized Chinese PANAS and Ren-CECps-SWB 2.0, we built an SWB measurement model for the Chinese context. We validated our model using data collected from the blog Sina.com from 2008 to 2013.

We selected Sina.com for SWB measurement for three major reasons. First, Sina.com has the largest number of users among all Chinese blog platform services. Sina.com has more than 10 million registered users, and its daily page views exceed three million. Second, Sina.com has the longest operation history in China. For most of the major events that occurred in China in recent years, there are corresponding blogs that can be searched and accessed on the Sina.com blog platform. The data availability facilitated our analysis and tests. Third, bloggers at Sina.com have clear categories, such as entertainment star blogs, intellectual celebrity blogs, and grassroots blogs. Grassroots bloggers normally express their own feelings and emotions in their blogs. They care about the people and events around them. Therefore, grassroots blogs reflect the ups and downs of common people in China. Accordingly, in this research, we chose the blogs of grassroots bloggers as our data source.

This research makes three contributions to the SWB research. First, we build a novel SWB measurement model based on multiple weighted emotions from PANAS. Second, we construct a lexicon, Ren-CECps-SWB 2.0, used specifically to measure the SWB implied by Chinese texts based on the Ren-CECps corpus. Third, we provide an SWB measurement model including five basic emotions in Chinese.

The remainder of this paper is organized as follows. We first provide a literature review. Then, we develop an SWB measurement model based on PANAS. Third, we modify Ren-CECps corpus and construct the Chinese lexicon Ren-CECps-SWB 2.0. Fourth, we build a novel SWB measurement model for Chinese and validate our measurement model. Finally, we draw implications for research and practice.

2. Literature review

2.1. Definition of SWB

Though SWB has been well studied, there is no consensus definition for it. In this research, following Ed Diener, one of the leading researchers in the field of SWB, we define SWB as the way a person evaluates his/her own life, including emotional experiences of pleasure versus pain in response to specific events and cognitive evaluations of what a person considers a good life [7,6]. Based on this definition, SWB consists of cognitive well-being and affective well-being [8]. In this paper, we restrict our investigation on SWB to affective well-being merely due to data availability. (As of today, there is no valid method available for us to mine measurements for cognitive wellbeing from text sentiment analysis.) This study conceives of certain types of pleasant experiences (viz. pleasant moods) as more valuable than others, e.g., "transient pleasant sensations" [4]. In our study, therefore, SWB is a summarization of individuals' emotional experiences of the continuous events in their daily life and work, including both positive emotions, such as love and happiness, and negative emotions, such as sorrow and anxiety.

2.2. Prior research on SWB measurement through text sentiment analysis

Several methods have been proposed to measure SWB through text sentiment analysis. One method is used by Facebook to build its Facebook Gross National Happiness (FGNH) index. When constructing the index, the company first uses the number of positive (negative) words in users' status updates to proxy positivity/negativity, and the FGNH index is the standardized difference between positivity and negativity. Another method is proposed by Dodds and Danforth [10]. They used Affective Norms for English Words (ANEW) to measure the implied SWB by estimating the overall valence score for a text. Though the two basic methods have their merits, they also have limitations. For example, in FGNH, only genetic positive emotions and negative emotions are involved. The dualistic classification and equal weights of general positive emotion and negative emotion in SWB are over simplified. Dodds and Danforth integrated good-bad (valence), but psychologists insist that good and bad emotions are independent scales for SWB [9]. Moreover, SWB involves multiple dimensions of emotion, and each emotion may make a different contribution to SWB [17]. To overcome the limitations of prior methods for measuring SWB in text sentiment analysis, as discussed in the next section, this paper constructs a new SWB method using automated UGC sentiment analysis. Our approach is based on a more delicate classification of emotions, and each emotion has a specified weight in the SWB analysis.

3. SWB measurement model and its specification in Chinese context

In this section, we propose a new SWB model specific to the Chinese context based on UGC. Our model extends the PANAS framework by constructing and measuring its key components, positive affect and negative affect, using the online UGC and text sentiment analysis techniques. To construct SWB based on Chinese UGC, we also introduce a new sentiment lexicon by extending the Ren-CECps lexicon.

3.1. SWB measurement model based on PANAS

The Positive and Negative Affect Schedule (PANAS) is one of the most widely used scales to measure mood or emotion [24]. This brief scale is comprised of 20 items, with 10 to measure positive affect (PA, e.g., excited, inspired) and the other 10 to measure negative affect (NA, e.g., upset, afraid). Each item is rated on a five-point Likert scale. PANAS provides a classification for both PA and NA. Schmukle et al. [21] provide solid evidence that the PA and the NA are unrelated, which suggests that they can be used as independent indicators for measuring SWB [21].

We propose an SWB measurement model based on PANAS. The key difference between our model and PANAS scale is that in PANAS,[1] PA and NA are measured by a self-report survey. In our model, however, we use the UGC of online grassroots blogs to conduct text sentiment analysis to measure PA, NA, and SWB. In our model, first we calculate the proportion of each sentiment word in a text. The more frequently a word appears in an online text conversation, the more representative the word is in the conversation, and the more weight will be assigned to this word in measuring SWB. Then, we calculate the value of each emotion vector from PA and NA in a text by summing up each emotion vector value from each sentiment word in a text. Finally, we obtain the weighted sum of all emotion vectors in a text to gauge SWB in this text. When we average all SWB from those blogs in a period of time, we obtain the aggregate SWB in this period of time. On

---

[1] Because we want to measure Chinese people's well-being, the PANAS framework from Watson, Clark and Tellegen cannot be used directly because of huge differences between English and Chinese. We have to establish a Chinese version of the PANAS framework to measure Chinese SWB. Based on the original PANAS from Watson, Clark and Tellegen, Qiu, Zheng and Wang developed a Chinese localized PANAS, which includes nine emotions in PA and NA, respectively [20]. Our model is based on this Chinese-localized PANAS.