Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Using Big Data to study subjective well-being

Maike Luhmann

Subjective well-being comprises emotional experiences and life satisfaction. This article reviews how Big Data can be used to measure, study, and change subjective well-being. Most Big Data approaches measure subjective well-being by analyzing language patterns on Twitter or Facebook. These approaches provide satisfactory accuracy for emotional experiences, but not yet for life satisfaction. Other measurement approaches include the analysis of other digital traces such as Facebook profiles and the analysis of mobile phone usage patterns. Big Data can be used to study subjective well-being on individual levels, regional levels, and across time. Potentials and limitations of using Big Data in studies on subjective well-being are discussed.

Address
Ruhr University Bochum, Department of Psychology, Universitaetsstrasse 150, 44780 Bochum, Germany

Corresponding author: Luhmann, Maike (maike.luhmann@rub.de)

Current Opinion in Behavioral Sciences 2017, **18**:28–33

This review comes from a themed issue on **Big data in the behavioural sciences**

Edited by **Michal Kosinski** and **Tara Behrend**

http://dx.doi.org/10.1016/j.cobeha.2017.07.006

Introduction

The science of subjective well-being (SWB) focuses on the definition, measurement, and correlates of happiness. Subjective well-being encompasses people's emotional experiences (i.e. positive and negative emotions and moods) as well as their evaluations of their lives (i.e. life satisfaction) [1]. SWB is an inherently subjective experience, meaning that each person knows best whether he or she is happy. Consequently, SWB has traditionally been measured with self-reports [2,3]. Self-report measures of SWB exhibit adequate reliability and validity [4] but can be distorted by irrelevant factors such as item order, momentary mood, or motivated self-enhancement [4–6]. Although the effects of these factors tend to be weak [4,7], researchers have nevertheless tried to improve the measurement of SWB through modifications of the research design (e.g. using experience sampling methodology instead of single-occasion surveys) as well as through non-self-report measures (e.g. peer reports or psychophysiological indicators such as cortisol levels and facial expressions). However, studies using these alternative measures remain rare compared to studies relying on self-reports. Moreover, because all of these measures are rather expensive to collect, studies on SWB are often limited by a low temporal and/or geographical resolution and by small and selective samples.

Big Data offer new opportunities to study SWB in ways that circumvent some of these limitations. This paper offers an overview of how Big Data are currently used to measure, study, and change SWB.

Using Big Data to measure SWB

Approaches to measure SWB using Big Data can be distinguished in terms of data source, measurement level, and SWB facet (Table 1).

Data sources

The predominant measurement approach relies on the analysis of so-called digital traces. Digital traces comprise all recorded online activities of an individual that can be accessed through publicly available databases (e.g. Twitter, Google Trends) or by obtaining individuals' permission (e.g. private Facebook profiles). Twitter is particularly popular because it allows researchers to access massive amounts of data quickly, cost-effectively, and without having to obtain informed consent from the users. Tweets are analyzed with respect to language patterns that predict SWB (e.g. [8–11]). This methodology has also been applied to other texts such as Facebook status updates (e.g. [12–14]).

Language patterns are studied using either a closed-vocabulary approach or an open-vocabulary approach [15]. The closed-vocabulary approach scans the text for keywords that are retrieved from a predefined dictionary. Sometimes, these dictionaries contain only a few dozen highly specific words [16,17]. Most dictionaries, however, comprise thousands of words. For example, the LabMT dictionary that is used in a few studies [9,18] contains 10,000 words that were rated by Amazon Mechanical Turk workers on a scale from sad to happy [19]. The most popular dictionary in SWB research, however, is Linguistic Inquiry and Word Count (LIWC) [20] which comprises >6000 words assigned to one or more dimensions. The dimensions most frequently used by SWB researchers are positive and negative emotion words and dimensions related to specific emotions such as sadness or anger. LIWC is easy to use and widely accepted as a validated method to study language patterns. However, as

Table 1

Overview of measurement approaches.

Publication	General approach	Data source	SWB facet	Measurement level
Algan *et al.* [47]	Frequency of specific search terms on Google	Google Trends	Life satisfaction Emotional well-being	Longitudinal trends within one nation (United States)
Carlquist *et al.* [16]	Closed vocabulary (self-constructed lexicon)	Newspaper articles	Emotional well-being	Longitudinal trends within one nation (Norway)
Collins *et al.* [12]	Online activity; closed vocabulary (LIWC)	Facebook status updates and likes	Life satisfaction	Individual
Curini *et al.* [45]	Open vocabulary	Twitter	Emotional well-being	Italian provinces
Doré *et al.* [46••]	Closed vocabulary (LIWC)	Twitter	Emotional well-being	Local with exact coordinates
Durahim & Coskun [44]	Closed vocabulary (SentiStrength, Turkish version)	Twitter	Emotional well-being	Turkish provinces and national
Hao *et al.* [24]	Open vocabulary	Weibo posts	Emotional well-being	Individual
Hung *et al.* [34]	Mobile phone usage	Calling states, app usage	Emotional well-being	Individual
Jones *et al.* [41]	Closed vocabulary (LIWC)	Twitter	Emotional well-being	Regional
Kosinski *et al.* [26]	Online activity	Facebook likes	Life satisfaction	Individual
Kramer [40]	Closed vocabulary (LIWC)	Facebook status updates	Life satisfaction	Regional
Lee *et al.* [27]	Online activity	Daily activity on Facebook and Twitter	Emotional well-being	Individual
LiKamWa *et al.* [39]	Mobile phone usage	Call and messaging logs, app usage, location	Emotional well-being	Individual
Liu *et al.* [13]	Closed vocabulary (LIWC)	Facebook status updates	Life satisfaction	Individual
MacKerron & Mourato [33]	Mobile phone app	Location data as basis for data type of location, weather	Emotional well-being	Individual
Mitchell *et al.* [18]	Closed vocabulary (LabMT)	Twitter	Emotional well-being	Neighborhood (U.S. cities)
Miura *et al.* [17]	Closed vocabulary (self-constructed lexicon)	Twitter	Emotional well-being	Japanese regions
Nguyen *et al.* [23]	Open vocabulary	Twitter	Emotional well-being	U.S. ZIP codes
Nguyen *et al.* [9]	Closed vocabulary (LabMT)	Twitter	Emotional well-being	Neighborhoood (3 U.S. cities)
Prata *et al.* [43]	Open vocabulary	Twitter	Emotional well-being	Precise location in Brazil
Rickard *et al.* [38•]	Combination of language analysis, other online activities	Facebook and Twitter activity and posts	Emotional well-being	Individual
Saeb *et al.* [36]	Mobile phone usage	Location, usage patterns	Emotional well-being	Individual
Schwartz *et al.* [8]	Combination of closed vocabulary (LIWC) and open vocabulary	Twitter	Life satisfaction	U.S. counties
Schwartz *et al.* [14]	Open vocabulary	Facebook status updates	Life satisfaction	Individual
Settanni & Marengo [51]	Closed vocabulary (LIWC)	Facebook status updates	Emotional well-being	Individual
Volkova & Bachrach [22]	Open vocabulary	Twitter	Emotional well-being	Individual
Volkova *et al.* [42]	Open vocabulary	Twitter	Emotional well-being	Neighborhood (U.S. universities)
Wang *et al.* [52]	Closed vocabulary (LIWC)	Facebook status updates	Life satisfaction	Individual
Wang *et al.* [11]	Closed vocabulary (LIWC)	Twitter	Emotional well-being	National
Wojcik *et al.* [56•]	Closed vocabulary (LIWC) for language analysis	Twitter, LinkedIn, etc.	Emotional well-being	Individual
Yang & Srinivasan [10]	Closed vocabulary (LIWC)	Twitter	Life satisfaction	Individual
Yu & Wang [54]	Closed vocabulary (NRC Word-Emotion Association lexicon)	Twitter	Emotional well-being	National

all dictionaries, it comprises a limited number of words and does not work well for colloquial language, leading some to question its validity for the analysis of social media language [21••].

Open-vocabulary approaches, in contrast, do not rely on an a priori defined list of words but rather identify relevant words, topics (i.e. clusters of words used frequently together), and other linguistic features using a data-driven approach [8,14,22–24]. Open-vocabulary-based predictions of SWB tend to exhibit greater predictive validity than closed-vocabulary-based predictions. In one exemplary study, life satisfaction correlated weaker with words identified using a closed-vocabulary approach