



The Journal of Positive Psychology

Dedicated to furthering research and promoting good practice

ISSN: (Print) (Online) Journal homepage: <https://www.tandfonline.com/loi/rpos20>

The value of social media language for the assessment of wellbeing: a systematic review and meta-analysis

S. Sametoğlu, D.H.M. Pelt, J.C. Eichstaedt, L.H. Ungar & M. Bartels

To cite this article: S. Sametoğlu, D.H.M. Pelt, J.C. Eichstaedt, L.H. Ungar & M. Bartels (2023): The value of social media language for the assessment of wellbeing: a systematic review and meta-analysis, *The Journal of Positive Psychology*, DOI: [10.1080/17439760.2023.2218341](https://doi.org/10.1080/17439760.2023.2218341)

To link to this article: <https://doi.org/10.1080/17439760.2023.2218341>




© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 04 Jun 2023.



[Submit your article to this journal](#) 







[View related articles](#) 



[View Crossmark data](#) 

The value of social media language for the assessment of wellbeing: a systematic review and meta-analysis

S. Sametoğlu ^{a,b}, D.H.M. Pelt ^{a,b}, J.C. Eichstaedt ^{c,d}, L.H. Ungar^{e,f} and M. Bartels ^{a,b}

^aDepartment of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands; ^bAmsterdam Public Health Research Institute, Amsterdam University Medical Centers, Amsterdam, The Netherlands; ^cDepartment of Psychology, Stanford University, Stanford, CA, USA; ^dInstitute for Human-Centered AI, Stanford University, Stanford, CA, USA; ^eDepartment of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA; ^fPositive Psychology Center, University of Pennsylvania, Philadelphia, PA, USA

ABSTRACT

Wellbeing is predominantly measured through self-reports, which is time-consuming and costly. It can also be measured by automatically analysing language expressed on social media platforms, through social media text mining (SMTM). We present a systematic review based on 45 studies, and a meta-analysis of 32 convergent validities from 18 studies reporting correlations between SMTM and survey-based wellbeing. We find that (1) studies were mostly limited to the English language, (2) Twitter was predominantly used for data collection, (3) word-level and data-driven methods were similarly prominent, and (4) life satisfaction was the most common outcome studied. We found that SMTM-based estimates of wellbeing correlated with survey-reported scores across studies at a meta-analytic average of $r = .33$ (95% CI [.25, .40]) for individual-level assessments of wellbeing, and at $r = .54$ (95% CI [.37, .67]) for regional measures of well-being. We provide recommendations for future SMTM wellbeing studies.

ARTICLE HISTORY

Received 22 May 2022
Accepted 20 April 2023

KEYWORDS

Wellbeing; well-being; social media; text mining; validity



There is a growing interest in the concept of wellbeing, given its association with a wide range of positive outcomes. Higher levels of wellbeing are associated with better financial habits and social relations, more altruistic behaviours, higher school grades, and better workplace functioning (Chapman & Guven, 2016; James et al., 2019; Kim et al., 2019; Maccagnan et al., 2019; Okabe-Miyamoto & Lyubomirsky, *in press*; Oswald et al., 2015; Steptoe, 2019; Walsh et al., 2018). Higher levels of wellbeing, supported by governmental policies, may also boost the socio-economic development of nations (Lambert et al., 2020; Santini et al., 2021).

Most of the research on wellbeing relies on self-report questionnaires, which seems justified since by definition wellbeing is centred on the subjective evaluation of one's functioning in life. However, collecting self-report data is time-consuming, expensive, and it can suffer from several biases, such as social desirability (Edwards, 1957) or recollection bias (Shiffman et al., 1997). Furthermore, wellbeing questionnaires are generally static and may therefore not be well suited to capture variance over time. A relatively novel alternative for self-report questionnaires is the automatic analysis of individuals' social-media language (social media text mining; SMTM).

Below, we first discuss how wellbeing is defined in the field and explain in detail how SMTM is conducted and how SMTM estimates are usually evaluated. Next, we review what two existing reviews found for SMTM efficacy in assessing wellbeing. Finally, we present our systematic review, meta-analysis, and conclusions.

Definitions of wellbeing

A wide range of wellbeing definitions, models, and measures exist, but the most common conceptual distinction is made between subjective (or *hedonic*) and psychological (or *eudaimonic*) wellbeing (Deci & Ryan, 2008; Ryff, 1989). Subjective wellbeing (SWB) is defined as the cognitive and affective evaluation of one's life, whereby the cognitive component is often captured by life satisfaction, while the affective component is measured by (the presence of) positive affect and (the absence of) negative affect (Diener et al., 1985). Psychological wellbeing (PWB; Ryff, 1989) is defined as positive functioning in life, consisting of positive relations, autonomy, environmental mastery, personal growth, purpose in life, and

CONTACT S. Sametoğlu  s.sametoglu@vu.nl  Department of Biological Psychology, Faculty of Behavioural and Movement Sciences, Vrije Universiteit Amsterdam, Van der Boechorststraat 7-9, Amsterdam 1081 BT, The Netherlands

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

self-acceptance. Overall, measures of wellbeing correlate moderately to strongly with each other, suggesting an underlying common, broad wellbeing factor (Bartels & Boomsma, 2009; Baselmans & Bartels, 2018; Disabato et al., 2016; Longo et al., 2016).

Social Media Text-Mining (SMTM)

The idea of analysing textual data to infer psychological phenomena can be traced back to the beginning of the 1900s. Freud suggested that mistakes in language use can inform about people's hidden intentions (see for an overview, Tausczik & Pennebaker, 2010). Later on, methods focused on individuals' responses to a pre-determined set of stimuli (e.g., ambiguous inkblots or drawings) as indicators of emotions, thoughts, or motivations (e.g., Holtzman, 1950; McClelland, 1979; Rorschach, 1921). Around the 1950s the use of less stimulus-dependent approaches started to emerge. For instance, Gottschalk et al. (1958) developed a content-analysis protocol to identify Freudian themes in transcriptions of 5-min recordings of patients talking about their thoughts (e.g., Gottschalk et al., 1958, 1969). The first computerized automatic text analysis program, the General Inquirer program (Rosenberg & Tucker, 1979; Stone et al., 1966), appeared in the second half of the 1960s. Today, the most prominent method of analysing text in the social sciences is the Linguistic Inquiry and Word Count software (LIWC; Boyd et al., 2022; Pennebaker et al., 2015).

Although these methods are widely available, collecting human-generated responses to open-ended questions can still be as costly as collecting survey responses. The vast availability of social media data has, however, changed this. It is estimated that more than 3.6 billion people use social media platforms worldwide (Tankovska, 2020) leading to the generation of unprecedented amounts of self-reported textual data every day. These include text-based recordings of thoughts, emotions, and behaviours without the individuals' primary motivation for providing data for research. Social media text data from thousands of subjects can be collected and analysed automatically while offering a less biased, unobtrusive, and more ecologically valid assessment of wellbeing. Collectively, we refer to methods that apply automatic text analyses on data from social media as social media text mining (SMTM; Tay et al., 2020).

Overall, conducting SMTM involves two steps. In the first step, the unstructured language data from individuals' social media accounts is automatically analysed to create language variables or 'features' (e.g., which words are used or the number of times a word is used relatively

to the user's total word count). The methods to build language features can be categorized as closed and open-vocabulary methods (see for an overview, Schwartz & Ungar, 2015).

Closed-vocabulary methods involve dictionaries based on existing psycho-social theories or created through annotations performed by annotators. For example, a dictionary written by experts is the 2022 version of Linguistic Inquiry and Word Count (LIWC-22; Boyd et al., 2022), which includes 337 positive emotion words (e.g., 'love', 'nice', 'sweet') and 612 negative emotion words (e.g., 'hurt', 'ugly', 'nasty'). Examples of annotation-based dictionaries are the Affective Norms for English Words dictionary (ANEW; Bradley & Lang, 1999; also see, Warriner et al., 2013) and the Language Assessment by Mechanical Turk (LabMT; Dodds et al., 2011; also see, Kloumann et al., 2012) providing the average valence scores (between happy and unhappy) for approximately over 10,000 unique words. The relative frequencies of words from the dictionaries can be counted to estimate the positive and negative content in the text.

In open vocabulary methods (Schwartz & Ungar, 2015) language features are 'learned' from the data itself. These methods count on algorithms or decision rules. For example, Latent Dirichlet Allocation (LDA; Blei et al., 2003) groups words in a text that naturally occur together to generate language features called 'topics'. A topic is usually comprised of words that are semantically coherent and meaningful, such as the words 'tonight', 'excited', 'super', and 'stoked' occurring together (Eichstaedt et al., 2021). Similarly, the Pointwise Mutual Information criterion (PMI; Abdi & Williams, 2010) is used to detect two- and three-word sequences that occur at rates that are above chance (e.g., 'have a good day', 'thanks a lot').

In the second step of SMTM, the features can be used for estimating simple correlations (word-level methods; see Jaidka et al., 2020) or building supervised ML-prediction models ('data-driven'). Prediction models may use both open and closed vocabulary features in combination with demographics.

Evaluating the success of an SMTM approach

The success of SMTM can be established through the level of convergence of SMTM estimates and the 'gold standard' or 'ground-truth scores' based on self-reports. Most associations observed between SMTM and survey scores have an upper limit of $r = 0.30$ – 0.40 , which is similar in magnitude to correlations found between for example self-report survey scores and informant-reports of personality (e.g., Park et al.,

2015) and wellbeing (e.g., Schneider & Schimmack, 2009). As alternative validity approaches, researchers compare the words and phrases (or topics) associated with high (and low) survey scores (Kern et al., 2016) or compare the temporal variations (peaks and dips) in SMTM scores (Cao et al., 2018; Dodds et al., 2011; Durahim & Coşkun, 2015; Kramer, 2010; Kristoufek, 2018; Qi et al., 2015).

SMTM to assess wellbeing

The unique SMTM data properties can be leveraged to study complex traits like wellbeing in large samples. SMTM can be applied to measure wellbeing both at the individual and regional levels: akin to survey-based wellbeing assessments, inferences can be made at regional levels by aggregating both the location-stamped language data (e.g., geo-located tweets) and the survey responses within each region (Jaidka et al., 2020).

A review (Luhmann, 2017) and a meta-analysis (Settanni et al., 2018) have been published on the use of SMTM to assess wellbeing. Settanni et al. (2018) found that wellbeing could be estimated accurately through individuals' digital traces (e.g., user demographics, user activity statistics, language), indicated by a meta-analytic correlation of 0.37 (95% CI [0.28–0.45]) between SMTM and survey-based wellbeing scores. The positive association between digital traces and wellbeing was stronger for public social media platforms (Twitter/Sina Weibo, Reddit, and Instagram) than private ones (i.e., Facebook). Luhmann (2017) reported a moderate convergent validity (between $r_s = .20$ and $.40$) of SMTM for wellbeing with the Depression, Anxiety, and Stress Scales (DASS-21; Henry & Crawford, 2005). A weaker convergent validity was observed when the Satisfaction with Life Scale (SWLS; Diener et al., 1985) was used (overall less than $r = .20$), based on which the authors concluded that the validity of SMTM for life satisfaction was limited.

The present study

The existing literature review and meta-analysis provide useful first insights into the potential of applying SMTM to assess wellbeing. However, automatic text analysis methods are rapidly improving, suggesting that an up-to-date systematic review and meta-analysis is needed.

Further, both earlier studies considered self-report-based stress, anxiety, and depression as primary indicators of wellbeing in addition to life satisfaction scores. This approach might have lowered the validity of the results since wellbeing is not equivalent to the absence of psychopathology (Keyes, 2002). To address these points, in this present study, we conduct a systematic review and a meta-analysis with a focus on wellbeing measures specifically. We structure our evaluation across four sections: (1) sample characteristics, (2) design characteristics, (3) validity of the results (using convergent and face validity) based on a qualitative synthesis, and (4) convergent validity assessed via meta-analysis.

Methods

Information source and search strategy

On November 5, 2021, a search was conducted in the bibliographic databases PubMed and Web of Science. The results from both databases were merged. Reference lists of the selected articles were further scrutinized for relevant articles. As our search strategy we used combinations of search terms related to (1) wellbeing (e.g., 'Wellbeing', 'Well-being', 'Life satisfaction'), (2) social media platforms (e.g., 'Social-media', 'Facebook', 'Twitter'), (3) language message type ('Language', 'Post', 'Updates', 'Status'), and (4) language analysis methods (e.g., LDA – "Latent Dirichlet Allocation, 'LIWC – Linguistic Inquiry and Word Count') (see Table 1 for detailed information), and the Boolean search operators 'AND' and 'OR'. Initially, we conducted a comprehensive search that involved all possible combinations of the four search term categories. Subsequently, we narrowed down our search to include only the combinations of three search term categories

Table 1. Search terms.

Search Term 1	Search Term 2	Search Term 3	Search Term 4
Well-being	Social media	Language	GI – The General Inquirer
Wellbeing	Social-media	Posts	DICTION
Quality of Life	Twitter	Updates	LIWC – Linguistic Inquiry and Word Count (most prominent)
Satisfaction with Life	Instagram	Status	LSA – Latent Semantic Analysis
Life satisfaction	Forum		LDA – Latent Dirichlet Allocation
Positive affect	Blog		DLA – Differential Language Analysis
Happiness	Online		Word Embeddings
	Facebook		Vector space semantics
	Reddit		

The database search was done by using combinations of the terms above. The Boolean search operators AND (horizontal) and OR (vertical) were used to combine the 4 columns, after that, first 3, then first 2.

(social media platforms, language message types, and language analysis methods). This was followed by using only two search term categories (language message types and language analysis methods).

Study selection, eligibility criteria, and data extraction

Following the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) guidelines (Moher et al., 2009), a flow diagram of our study selection

process is presented in Figure 1. The titles and abstracts of all identified articles were screened after exact duplicates were removed. The screening was performed by the first author. Uncertain cases were resolved through discussions among the authors. The title and abstract of articles were screened according to the following eligibility criteria: (1) A study must utilize 'social media language' to investigate 'wellbeing', (2) must use a 'quantitative approach', (3) must be published in a peer-reviewed journal, (4) is not a meta-analysis or a review paper (5) and is written in English. We

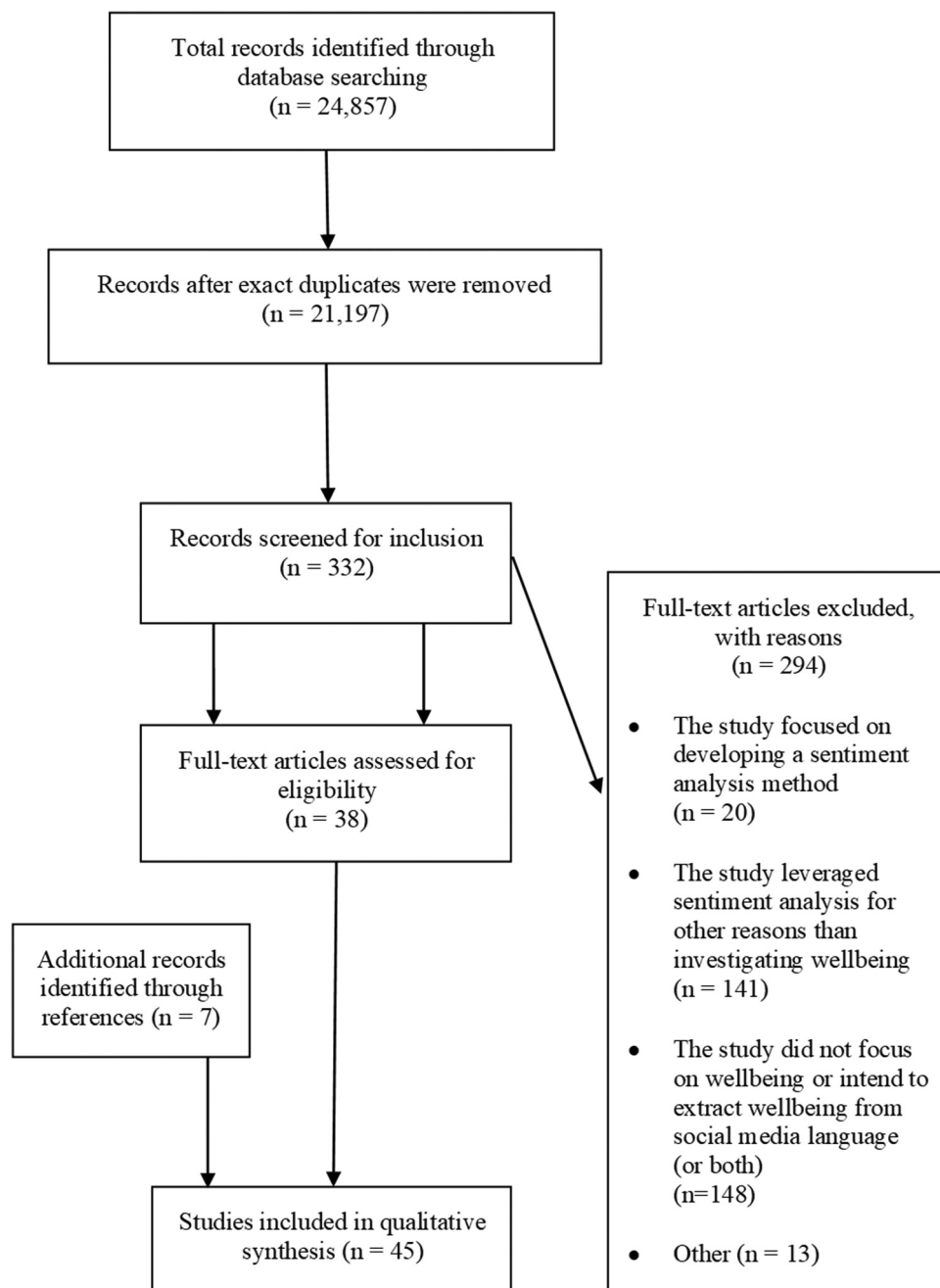


Figure 1. PRISMA Flow Diagram of the included studies

additionally included articles by scanning the references of the two aforementioned review studies (Luhmann, 2017; Settanni et al., 2018).

Qualitative synthesis

Sample characteristics

In the current review, we investigated the type of language (e.g., English, Chinese), platform (e.g., Facebook, Twitter), and sample size used in each study.

Design characteristics

We reported whether a ground-truth measure (i.e., self-report) was included, and if so, which wellbeing measures were used, whether the main focus was on individual, subnational, or national wellbeing levels, whether closed and/or open vocabulary methods were used for textual data, and whether data-driven or word-level methods were used. We use the term 'subnational' to refer to assessments made for location, but not country-level studies, e.g., states, counties, or neighbourhoods.

Validity of results

Convergent validity. To assess SMTM's convergent validity for wellbeing, we report, meta-analyse, and evaluate the correlations between SMTM and ground-truth scores.

Face validity. To assess SMTM's face validity for wellbeing, we compare the evidence from the SMTM and the self-report-based wellbeing literature. If similar conclusions can be drawn, this can provide evidence for the face validity of SMTM in the wellbeing context. To facilitate a meaningful comparison, four recurring topics from the wellbeing literature were chosen: (1) the characteristics of happy individuals, (2) temporal trends in wellbeing, (3) the relation between positive affect (PA) and negative affect (NA), and (4) demographics (age and sex differences).

Quantitative synthesis: meta-analysis and publication bias

A meta-analysis was conducted using the Metafor package in R (R Core Team, 2021; Viechtbauer, 2010) to obtain a meta-analytic estimate for converging validity across studies. We identified a total of 32 effect sizes (correlation coefficients) from 18 studies. Two studies were excluded to ensure homogeneity. The excluded studies involved a considerably large time gap between the SMTM and survey-based wellbeing measurements or did not use self-reports but face-to-face interviews.

For sake of comparability with previous meta-analyses (e.g., Settanni et al., 2018) the effect sizes (i.e., correlations) used from each study were based on the best-performing language machine learning prediction models or language features (while being able to generalize to newer datasets – i.e. not overfitting)

Correlation coefficients were converted to standardized z-values using Fisher transformation. After conducting the meta-analysis on the transformed effect sizes, the meta-analytic estimate was converted back to a correlation coefficient to allow interpretation in the original metric.

We applied a random effects model (Rubio-Aparicio et al., 2020) with robust variance estimation (Hedges et al., 2010) using a restricted maximum likelihood estimator (REML; Kenward & Roger, 1997). Cochran's Q-test (Hedges & Olkin, 2014) was applied to assess whether the null hypothesis that the true heterogeneity (τ^2) between the effect sizes is equal to 0. Further, the extent of how much of the heterogeneity was attributable to true heterogeneity can be assessed through the I^2 statistic, with values around 25, 50, and 75% categorized as 'low', 'medium', and 'high' (Higgins et al., 2003). Higher levels of I^2 values can be considered legitimate grounds for including potential moderators. We included two possible categorical moderator variables. The first moderator variable indicated if an effect size was estimated at individual level or location level (reference category), and the second moderator indicated whether the methods for estimating wellbeing were through data-driven or word-level methods (reference category). A third interaction term between our two moderators was not included because the number of effect sizes for the location-level word-level method group was far lower than for the other groups, 2 compared to $k = 8, 9,$ and 13 (individual/word, individual/data, location/data, respectively), increasing the risk for type-2 errors.

To assess the risk of publication bias, we visually created and inspected a funnel plot, and applied Egger's test to statistically assess its asymmetry (Egger et al., 1997). We also estimated the number of (potentially) missing studies on the left side of the funnel through the trim and fill method (Duval & Tweedie, 2000). Observing a symmetrical funnel plot, obtaining a statistically non-significant Egger's test, and finding no studies missing based on the trim-and-fill method would provide no evidence for publication bias.

Lastly, some of the effect sizes in the present meta-analysis were non-independent: some studies provided multiple effect sizes based on different wellbeing measures (e.g., life satisfaction, PA/NA or eudaimonic wellbeing dimensions) for the same samples. To solve the non-independence issue, we applied robust variance

estimation (RVE; Hedges et al., 2010; Moeyaert et al., 2017) to our meta-analytic estimate. RVE ensures that the studies with more effect sizes are assigned a smaller weight to obtain unbiased standard errors (Hedges et al., 2010; Moeyaert et al., 2017).

Results

Our initial search resulted in 28,857 papers. After removing duplicates 21,197 articles remained. By screening titles and abstracts, 332 articles were found potentially relevant. Based on full-text readings, 38 articles were included in the present study. Seven articles were added from external references. The final systematic review included 45 articles (See Figure 1).

In Table 2, we list all studies referred to in the sample/design characteristics section of the results. In Table 3, we provide results acquired from studies to judge the face validity of SMTM for wellbeing. In Table 4, correlations reported between SMTM-based wellbeing and ground-truth scores are provided. In all tables, studies are referred to with an identification number, which is indicated in the reference list as well.

Sample characteristics

Type of language

In the majority of the studies, analyses were based on single language data ($k = 42$) with English datasets being most common ($k = 30$), followed by Chinese ($k = 6$), Italian ($k = 4$), Russian ($k = 1$), and Turkish ($k = 1$). Only a few studies used data from multiple languages ($k = 3$).

Platform

Mostly, the data were collected through Twitter ($k = 28$) and its Chinese equivalent Sina Weibo ($k = 6$). Facebook was used in less than one-fourth of all studies ($k = 10$). A single study used datasets from both Twitter and Facebook.

Sample size

Most of the studies ($k = 31$) reported information on sample size, yet a smaller number of studies ($k = 14$) did not. Reported sample sizes varied between 133 and around 100 million, which could be categorized into 4 categories: studies involving less than 1,000 participants ($k = 4$), between 1,000 and 10,000 ($k = 9$), between

Table 2. Sample and design characteristics of the 45 reviewed studies.

	Corresponding article number(s)
Language	
English	2, 4–8, 10, 11, 13–17, 19–22, 24, 27, 29, 31, 32, 37–39, 41–45
Chinese	18, 23, 25, 26, 36, 40
Italian	9, 12, 28, 35
Russian	3
Turkish	30
Multilingual	11, 33, 34
Platform	
Twitter	1, 2, 4, 5, 7, 8, 9, 10, 13–17, 19, 22, 24, 27–30, 32, 33, 38, 40, 43, 44, 46
Sina Weibo	18, 23, 25, 26, 37, 41
Facebook	3, 6, 11, 12, 20, 21, 35, 38, 41, 44
Twitter and Facebook	31
Sample Size	
Less than 1,000 participants	12, 18, 35, 42
Between 1,000 and 10,000 participants	3, 6, 11, 20, 22, 31, 36, 40
Between 10,000 and a million participants	5, 15–17, 19, 21, 27, 30, 34, 39, 43–45
More than a million participants	4, 7, 23, 24, 29, 31, 41
Multiple values ranging between 171 and 86,073	38
Individual vs. Regional	
Individual	2, 3, 4, 6, 7, 10, 11, 12, 20–22, 24, 27, 29, 35, 36, 38, 39–41, 43–45
Subnational	1, 5, 8, 9, 13–19, 23, 25, 26, 28, 30, 32, 34, 37, 42
National	18, 33, 34
Individual and subnational	31
Closed vs. Open Vocabulary	
Closed vocabulary	1, 2, 4, 7, 9–19, 21–30, 32–36, 38–41, 43–45
Open vocabulary	5, 42
Open and closed vocabulary	3, 6, 8, 20, 31, 37
Word-level vs. Data-driven	
Word-level	1, 2, 7, 10, 11, 13, 14, 21, 22, 23, 24, 25, 26, 29, 35, 36, 39, 40, 41, 45
Data-driven	3, 4, 5, 6, 8, 9, 12, 15, 16, 17, 18, 19, 20, 27, 28, 30, 33, 32, 34, 37, 38, 42, 44
Word-level and data-driven	31, 43
Inclusion of Ground-Truth Measures	
Ground-truth wellbeing measure	3, 6, 9, 11, 12, 14, 15, 20, 21, 28, 30, 31, 34–38, 40, 41

Table 3. Results from the 45 reviewed studies.

	Corresponding article number(s)
Language of high levels of wellbeing	
Enjoying weekend, being happy on Sunday, romance, friends and family gatherings, birthdays, parties at night, life and living, and sharing photos	8
Professional occupation, engagement, family-friends, communal engagement (similar to involvement, dedication, and organizational citizenship behavior)	20
Exercise, altruism/donation, business skills, amazement	37
Money/achievements, positive word categories, first person plural nouns (e.g., 'we', 'our')	37
Money, work, achievements, religion	24
Higher exclamation mark use	24
Language of low levels of wellbeing	
Disengagement, swearing, boredom	20, 37
Impersonal predicates (e.g., 'must', 'should not'), negative emotion words	3
Negative emotion/affect, swear and anger, first person singular nouns (e.g., 'I', 'me'), present tense, disengagement, negative relationships, lack of meaning and achievement	12, 24, 37
(Less) positive emotion and (more) negative emotions use	12
Temporal trends	
SMTM-WB increases and decreases in line with regional or worldwide events	5, 10, 18, 29, 30,41
SMTM-WB is stable against regional and worldwide events	24
SMTM-WB changes in hours	
SMTM-WB changes in days	7, 26, 28
SMTM-PA and NA (combined) peaks at early morning, decreases drastically until midday followed by a less steep decrease until midnight.	29
SMTM-PA peaks twice a day (early morning and near midnight), SMTM-NA dips in the mornings increases stable until a night-time peak	7
SMTM-WB is highest overall during weekends	5, 7, 8, 18, 28, 29
SMTM-WB is the lowest on Wednesdays	18, 29
SMTM-WB is the highest on Tuesdays	28
SMTM-WB is the lowest on Tuesdays	29
SMTM-WB is highest in July and lowest in February	28
SMTM-PA decreases with shorter day length but not SMTM-NA	7
SMTM-WB increases as the temperature increases in winter and spring, no association after 30 degrees Celsius in summer	28
SMTM-WB decreases with rain, but not with snowfall	28
Decreases in SMTM-WB related to air pollution increases with bad weather	26
Relation between PA and NA	
SMTM-PA and SMTM-NA shows independent within-person level trajectories	7
SMTM-PA and SMTM-NA are independent at province level	23
SMTM-PA is more present than SMTM-NA on social media	6, 11, 19, 30, 32, 39
SMTM-PA and SMTM-NA have different 'affect dynamics'	43, 44
SMTM-PA peaks quicker, SMTM-NA builds up slowly but dissipates faster	43
Demographics (age & sex)	
SMTM-PA is more present for females (but not SMTM-NA)	39, 43
SMTM-PA increases faster for females and dissipates slower	43
SMTM-WB decreases in relation to air pollution more for females	26
SMTM-WB is higher for older individuals (more joyful and less sad)	28, 39
SMTM-WB is higher in regions with older populations	32

SMTM-WB, SMTM-PA, SMTM-NA = overall wellbeing, positive affect, and negative affect based on social media text mining.

10,000 and a million ($k = 12$), or more than a million participants ($k = 7$). One study mentioned multiple sample sizes for their main analyses that ranged between 171 and 86,073.

Design characteristics

Inclusion of ground-truth measures

Less than half of the studies employed a wellbeing related ground-truth measure ($k = 19$).

Individual vs. regional focus

The majority of the studies investigated either individual-level wellbeing ($k = 23$) or subnational (e.g., county

or state) wellbeing ($k = 18$), while a few studies investigated nation-level wellbeing ($k = 3$). A single study investigated both individual and subnational level wellbeing.

Closed vs. open vocabulary methods

Most of the studies applied closed vocabulary methods, such as Linguistic inquiry and word count (LIWC; Tausczik & Pennebaker, 2010), affective norms for English words (ANEW; Bradley & Lang, 1999), or the Language assessment by Mechanical Turk (LabMT; Dodds et al., 2011) ($k = 36$), while the rest used a combination of open vocabulary methods such as Latent Dirichlet Allocation (LDA; Blei et al., 2003),

Table 4. Overview of the highest achieved correlations reported for convergent validity.

Nr	Study	Sample size	Scale	Construct	<i>r</i>	Measurement level
[3]	Bogolyubova et al. (2020)	1,972	WHO-5	WB	0.08	Individual
[6]	Chen et al. (2017)	2,612	SWLS	LS	0.36	Individual
[11]	Liu et al. (2015)	1,124	SWLS	LS	0.15	Individual
[12]	Marengo et al. (2021)	603	QoL	QoL	0.43	Individual
[20]	Schwartz et al. (2016)	2,198	SWLS	LS	0.33	Individual
[21]	N. Wang et al. (2014)	24,193	SWLS	LS	0.72	Individual
[35]	Settanni and Marengo (2015)	201	DASS-21	Negative Affect	0.32	Individual
[36]	Bai et al. (2014)	2,018	URRSAQ-LS	LS	0.53	Individual
[31]	Jaidka et al. (2020)	2,321	Gallup WB-LS	LS	0.26	Individual
[31]	Jaidka et al. (2020)	2,321	Gallup WB-H	Happiness	0.21	Individual
[31]	Jaidka et al. (2020)	2,321	Gallup WB-W	Worry	0.15	Individual
[31]	Jaidka et al. (2020)	2,321	Gallup WB-S	Sadness	0.15	Individual
[38]	Collins et al. (2015)	3,505	SWLS	LS	0.16	Individual
[40]	Hao et al. (2014)	1,785	PANAS	Affect	0.45	Individual
[40]	Hao et al. (2014)	1,785	PWBS	Eudaimonic WB	0.45	Individual
[41]	Kramer (2010)	1,341	SWLS	LS	0.17	Individual
[9]	Iacus et al. (2020)	19	ISTAT	LS	0.45	Location
[14]	Mitchell et al. (2013)	50	BRFSS-LS	LS	0.25	Location
[14]	Mitchell et al. (2013)	50	Gallup WB-6D	WB	0.51	Location
[28]	Curini et al. (2015)	110	QoL (by Il Sole 24 Ore)	QoL	0.19	Location
[31]	Jaidka et al. (2020)	1,208	Gallup WB-LS	LS	0.62	Location
[31]	Jaidka et al. (2020)	1,208	Gallup WB-H	Happiness	0.51	Location
[31]	Jaidka et al. (2020)	1,208	Gallup WB-W	Worry	0.52	Location
[31]	Jaidka et al. (2020)	1,208	Gallup WB-S	Sadness	0.64	Location
[34]	Coşkun and Ozturan (2018)	110,062	Cantril Ladder	LS	0.85	Location
[37]	Schwartz et al. (2013)	1,293	SWLS	LS	0.54	Location

3 studies that reported null results were not included in the table (15, 30).

SWLS = Satisfaction with Life Scale, PWBS = Ryff's Psychological Well-being scales, PANAS = Positive and Negative Affect Schedule, WHO-5 = World Health Organization Well-being Index, URRSAQ-LS = Urban and Rural Residents Social Attitudes Questionnaire – Life satisfaction, DASS-21 = The Depression, Anxiety, and Stress Scales, BRFSS-LS = The Behavioral Risk Factor Surveillance System – Life satisfaction, ISTAT = Italian National Institute of Statistics. QoL = Quality of Life (not a previously validated scale).

Word2Vec (Rong, 2014) ($k = 2$), or a combination of closed and open vocabulary methods ($k = 7$).

Word-level vs. data-driven methods

The majority used data-driven methods ($k = 23$), and the remaining studies used word-level methods ($k = 20$) except for two studies ($k = 2$) using both.

Inclusion of ground-truth measures

In the 19 studies that included a ground truth measure, most studies used measures for satisfaction with life ($k = 12$). The remaining studies used measures for affect ($k = 1$), general wellbeing ($k = 1$), quality of life ($k = 2$), affect and satisfaction with life ($k = 1$), affect and psychological wellbeing ($k = 1$). Only a single study used a eudaimonic wellbeing measure.

Convergent validity of results

Overall, studies indicated convergent validity for SMTM ($k = 18$) with correlation coefficients on average $r = 0.39$, $SD = 0.19$, ranging between $r = 0.08$ and 0.85 . Only a few studies found unexpected results ($k = 3$); One study found life (dis)satisfaction was not associated with wellbeing at location level, while another study found no association between SMTM wellbeing and ground-truth measures across 81 provinces in Turkey. The last study

found that negative word use was associated with wellbeing but in a positive direction. A summary estimate for the convergent validities across the studies will be provided in the meta-analysis section of the present study.

Face validity of results

Language of high and low wellbeing

The language of individuals who score higher on survey-based wellbeing ($k = 6$) included words related to topics, such as leisure time, achievements, exercise, altruism, amazement, religion, and first person plural nouns (e.g., 'We', 'us'). The language of individuals who scored lower on survey-based wellbeing included words related to swearing, disengagement, lack of meaning, problems with relationships, first-person singular nouns (e.g., 'I, me') and impersonal predicates (e.g., 'must').

Temporal trends in wellbeing

Fifteen studies have investigated the temporal trends in SMTM-based wellbeing. Six studies have examined the changes in SMTM-based wellbeing as a response to an emotionally valent worldwide and nation-scale event (e.g., festivals, disasters, economic crisis) ($k = 6$). Despite the overall results, one study has suggested wellbeing fluctuations cannot be found in social media text data ($k = 1$). Nine studies reported changes in

SMTM-based wellbeing in reoccurring/cyclical fashion (hours for working and resting during the day, weekends, and seasons with good weather). These changes have occurred in a range of time resolutions such as hours ($k = 3$), days ($k = 5$), and months/seasons ($k = 1$). Some studies have also reported changes in SMTM-based wellbeing related to changes in temperature or weather conditions ($k = 3$).

The two studies that investigated hourly changes in SMTM wellbeing reported different results depending on whether both PA and NA were included in the SMTM wellbeing measure. The study with a composite measure of PA and NA ($k = 1$) found that SMTM wellbeing peaked early in the morning (e.g., when waking up, commuting, and starting to work) and decreased drastically until mid-day, which was followed by a less strong decay until midnight (resting and eventually going to sleep). The other study that focused independently on SMTM-based PA and NA found that PA peaked twice a day (early morning and near midnight) while NA was lowest in the mornings and had a stable increase until its single peak at night-time.

The studies focusing on the daily fluctuations of SMTM wellbeing have revealed that wellbeing was consistently highest at the weekends ($k = 6$). Nonetheless, it was less clear on which weekday lowest wellbeing levels are found. In two studies, Wednesday was suggested ($k = 2$), while Tuesday was also considered as the day with the highest wellbeing ($k = 1$) but also the lowest wellbeing ($k = 1$).

SMTM wellbeing is found to decrease with lower temperatures, less light, and rougher weather conditions (e.g., wind, rain) ($k = 3$). For instance, SMTM wellbeing was the highest in July and lowest in February in Italy which is located in the Northern Hemisphere ($k = 1$), and it rose in parallel with increasing temperatures during winter and spring but was limited until 30 degrees Celsius in summer ($k = 1$). Shorter day length – which changes in accordance with seasons – was also associated with decreases in PA (but not associated with an increase in NA) in both Northern and Southern Hemisphere countries (e.g., the United States and Canada, India, and Australia) ($k = 1$). In addition, SMTM wellbeing decreased if it rained (but not if it snowed) ($k = 1$) and the effect of air pollution on SMTM wellbeing increased when air conditions were rough (i.e., if there is too much rain, wind and too many clouds) ($k = 1$).

Relation between positive and negative affect

Only a few studies assessed and compared SMTM-based PA and NA, and the results from these studies on the differences between the two constructs are in line with evidence-based on self-reports. PA and NA showed independent within-person level trajectories over time ($k = 1$),

and PA and NA were not significantly associated with each other at province level ($k = 1$). PA was more prevalent than NA in the language of individuals on social media ($k = 6$) and was also characterized by different 'affect dynamics' ($k = 2$). For instance, whenever users on Twitter state that they are feeling either 'good' or 'bad' (i.e., express their current emotions), the emotional content of their following tweets first peaks, and eventually returns to baseline levels. SMTM-based PA tends to peak quicker, while NA builds up more slowly and eventually returns to baseline levels even faster than PA ($k = 1$).

Demographics

Sex differences were found in SMTM-based wellbeing ($k = 4$). Females used more positive and negative language use than males, yet the findings for the negative word use were not consistent across studies ($k = 2$). Female users' trajectories show that positive word use frequencies increased faster but dissipated slower than in trajectories for males ($k = 1$). The negative effects of air pollution were more visible in the SMTM wellbeing levels of females than males ($k = 1$).

Concerning age, at the individual level older people used more positive language ($k = 2$). In line with individual-level results, at the regional level, happier tweets were observed in neighbourhoods with older populations ($k = 1$).

Meta-analysis and publication bias

In total, we retrieved 32 effect sizes from 18 studies and the number of outcomes per study ranged from 1 to 8

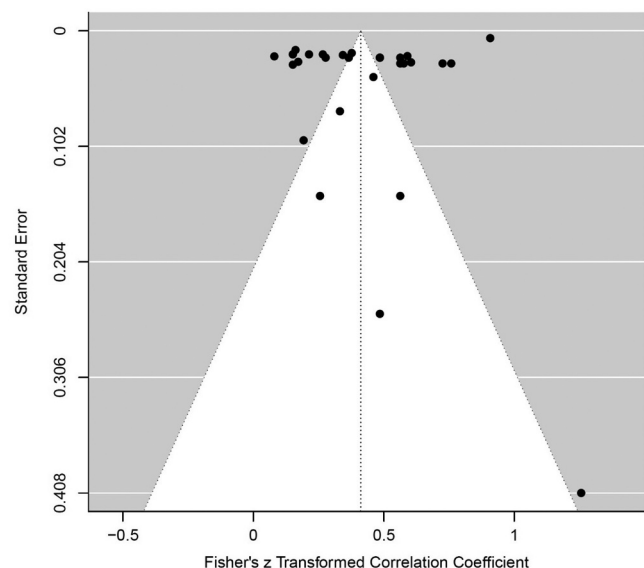


Figure 2. Funnel plot based on the 32 included effect sizes in the meta-analysis

(mean = 1.78, median = 1). To assess the risk of publication bias, we visually inspected our funnel plot and concluded that it was mostly symmetrical (see Figure 2). The Egger’s test results showed no statistical evidence for a funnel plot asymmetry ($z = 1.02, p = .31$). The trim and fill algorithm estimated the number of missing/non-reported effect sizes/studies on the left side of the mean effect of our funnel plot as 0 ($SE = 3.42$). Based on an initial random-effects meta-analysis model without any moderator variables, we found that the estimated true heterogeneity ($\hat{\tau}^2$) between the effect sizes was significant, as indicated by Cochran’s Q -test (Hedges & Olkin, 2014), $Q (df = 31) = 6204.30, p < .0001$. Nonetheless, the estimated true heterogeneity was small ($\hat{\tau}^2 = .04, SE = .01$). Most of the variability between the effect sizes was attributable to true heterogeneity ($I^2 = 98.81\%$). Thus, we proceeded with including our categorical moderator variables (individual vs location level and data-driven versus word level) in our meta-analytic model.

The results of the moderation analysis revealed that individual vs location level moderator was significant ($\beta = -.26 (SE = .01), p < .05, 95\% CI [.03, -.50]$), while the second moderator, data-driven versus word-level study, was not ($\beta = -.13 (SE = .08), p > .05, 95\% CI [-.31,$

.05]). Overall, including both moderators accounted for 25.71% of the heterogeneity in the initial model without the moderator variables. After applying the Robust Variance Estimator to address dependence across the effect sizes and converting z -values back to Pearson correlation coefficients, we found that the meta-analytic correlation was .54 (95% CI .37, .67) for location level studies and .33 (95% CI .25, .40) for individual-level studies (See Figure 3).

Discussion

Following the PRISMA guidelines, we presented a systematic review based on 45 studies that use SMTM to assess wellbeing and conducted a (1) qualitative synthesis and (2) a meta-analysis based on 32 effect sizes from a subset of eighteen studies reporting correlations between SMTM and survey-based wellbeing.

Qualitative synthesis

The systematic review and qualitative synthesis resulted in the following overall observations. Across the 45 studies, 70% of all studies were based on English-speaking

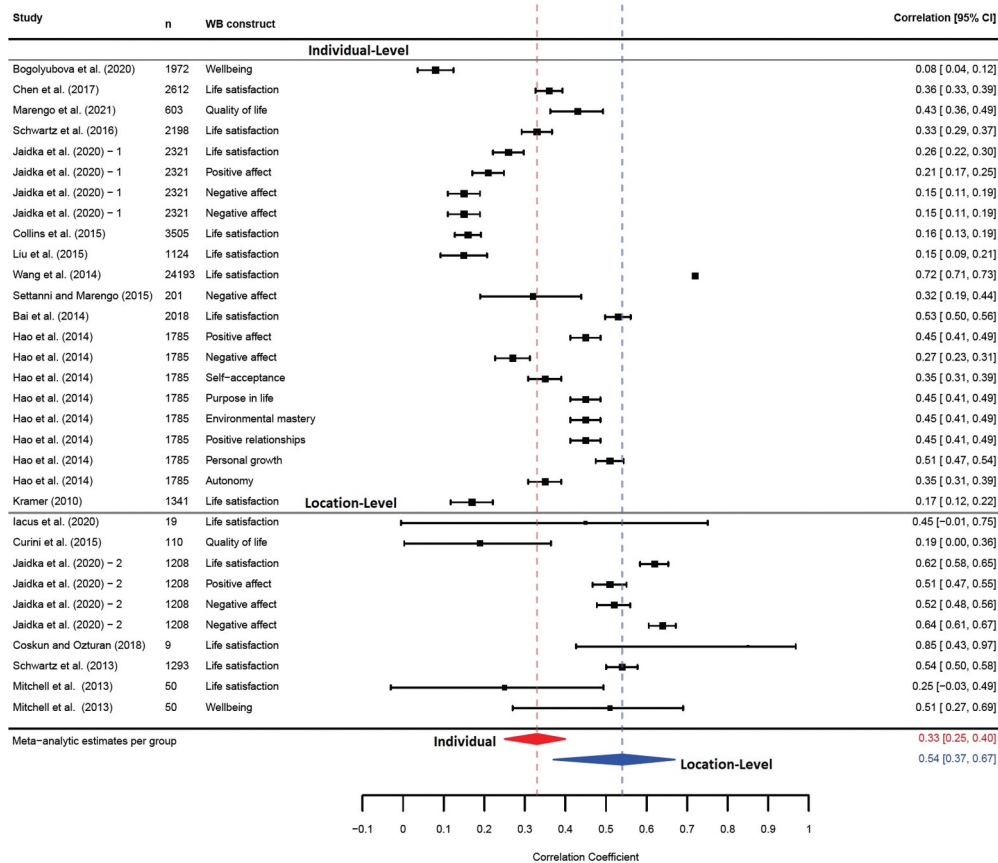


Figure 3. Meta-analytic estimates of the correlations observed between Social Media Text Mining (SMTM) and survey-based wellbeing assessments at both individual and location levels.

samples and Twitter was the most popular platform (60% of the studies). In general, large sample sizes were used (between 10,000 and a million individuals), though with a wide range (the smallest samples had less than 1000 individuals while the largest samples included even more than a million individuals). Half of the studies focused on individual-level wellbeing and the other half focused on regional wellbeing. Half of the studies used closed vocabulary methods (such as LIWC dictionaries), and the remaining studies either used open vocabulary (such as LDA topic models) or combined closed and open vocabulary methods. Word-level and data-driven methods were used equally. Satisfaction with life was the most used ground-truth measure (25% of the studies with ground-truth measures), while eudaimonic wellbeing (such as autonomy, personal growth, and environmental mastery) was only assessed in a single study.

Our results showed a clear majority (70% of all studies using English-speaking samples). This limits the applicability of the current evidence to non-English speaking populations. Existing studies have already suggested differences in the expression and conceptualization of happiness/wellbeing in different cultures. For instance, a study has found that European Americans (EA) and Asian Americans (AA) valued high-arousal positive affect (reflecting 'excitement') more than the Hong Kong Chinese people (CH), whereas AA and CH participants valued low-arousal positive affect (reflecting 'calmness') more than EA participants (Tsai et al., 2006). Therefore, a particular set of language features (e.g., words, topics) or a model used in SMTM may assess wellbeing in one culture well but can fail in another. Currently, SMTM might thus only be a reliable option to assess wellbeing in English, thereby excluding large parts of the world population. Nonetheless, the recent development of open-access language models based on massive multilingual data such as Multilingual BERT (M-BERT; Pires et al., 2019), XLM (Cross-lingual Language Model; Lample & Conneau, 2019), and XLM-R (XLM-RoBERTa; Conneau et al., 2020) may help to alleviate these representativeness problems. Such models, once trained, can be applied in a wide variety of natural language processing tasks in a wide range of languages not limited to English, allowing for the detection of wellbeing of individuals from different populations.

The qualitative synthesis, furthermore, indicated that most studies use large sample sizes, highlighting one of the advantages of social media data use. The combination of large-scale social media data and computerized text analysis methods allow for assessing wellbeing in a complementary, and perhaps an alternative way, to traditional survey-based self-reports. The unique characteristics of large language data used in SMTM (e.g., the

longitudinal prospective structure, the ecologically valid setting, large reach, anonymous large-scale data collection) provide an efficient way to assess wellbeing at different units of analysis (e.g., individual or location). Wellbeing assessed through SMTM at both the individual and location level can aid in developing improved personalized interventions to increase happiness while it can also inform better policies in neighbourhoods, cities, and countries. Both practitioners and policy-makers can use SMTM and aim to increase the already known positive outcomes related to higher levels of wellbeing such as better financial habits, social relations, more altruistic behaviours, higher school grades, and better workplace functioning (Chapman & Guven, 2016; James et al., 2019; Kim et al., 2019; Maccagnan et al., 2019; Okabe-Miyamoto & Lyubomirsky, *in press*; Oswald et al., 2015; Steptoe, 2019; Walsh et al., 2018), as well as increased socio-economic development of regions (Lambert et al., 2020; Santini et al., 2021).

Among the reviewed studies, there was an equal preference for word-level and data-driven methods. In addition, open-vocabulary methods were applied in 20% of the studies, less than the closed-vocabulary methods. However, data-driven and open vocabulary methods may better leverage larger datasets. Used with large social media datasets, these methods allow for finding previously unknown associations and help generate new hypotheses. Given the vast availability of language data on social media will likely expand further, the increasing preference for these computational methods is understandable. Nonetheless, open-vocabulary methods also have potential shortcomings, such as that study variables are generally not comparable across studies, while closed-vocabulary methods (dictionaries), for example, remain constant across studies. Overall, open vocabulary methods require more expertise to implement, need larger datasets, and are less easy to use than closed-vocabulary methods (for a full discussion, see Eichstaedt et al., 2021).

Half of the studies included at least one survey-based ground-truth measure ensuring a valuable source to test for the validity of social media data. With recent developments of data-driven methods, though, ground-truth measures are becoming less essential as these are only necessary when new models for SMTM are developed or when models are adapted for new populations. The novel contextual word embeddings are pre-trained on high-quality large samples (e.g., Devlin et al., 2018; Sanh et al., 2020; Z. Yang et al., 2019), which may increasingly liberate researchers from the burden of collecting large amounts of ground-truth measures. In principle, pre-trained models can be directly used to assess wellbeing in independent text data. Researchers should, however,

ensure there are no large differences between development and target samples.

There was a considerable amount of variety of wellbeing ground-truth measures to validate SMTM wellbeing scores. Most studies (around 60%) used self-reported life satisfaction scores as the ground truth. The remaining studies used other wellbeing measures capturing subjective wellbeing, mental wellbeing, or quality of life. These inconsistencies between measures make it difficult to readily compare results across studies. For instance, some of these measures included items for objective wellbeing (e.g., income, access to basic services), while others used items for the affective or cognitive components of wellbeing. At the same time, studies have shown that different wellbeing measures correlate with each other at moderate to high levels (e.g., Bartels & Boomsma, 2009; Busseri, 2018), implying a general wellbeing factor (Longo et al., 2016). Based on this, results informed by different wellbeing measures may still be comparable. Overall, both clarity on conceptualization of wellbeing and caution when inferring results from different wellbeing measures are needed. In the future, it is also commendable to include eudaimonic wellbeing measures (reported only in one study) to obtain a complete picture of SMTM wellbeing.

Our review indicated that SMTM-based wellbeing had mostly similar qualities as survey-based wellbeing. For instance, language expressed by individuals scoring high and low on wellbeing were largely in line with self-report-based results (e.g., Diener et al., 2018). People with higher wellbeing levels talked about social gatherings, leisure time, engagement, and enjoyment, whereas the language of people with lower wellbeing included words related to low levels of motivation, lack of meaning, impersonal predicates, and swearing. This aligns with previous survey studies showing that individuals who were more engaged and socially connected reported higher wellbeing (Keyes, 2010; Ryff, 1989; Seligman, 2018). In addition, our review showed that SMTM-based wellbeing both increase and decrease in response to positive and negative events similar to result acquired through survey-based wellbeing studies as well (Luhmann et al., 2012). Both one-time impactful (e.g., earthquakes or economic crisis) and cyclically occurring events (e.g., work vs. leisure hours) resulted in changes in SMTM wellbeing, hence providing additional evidence for the face validity of using SMTM capture wellbeing and its fluctuations.

A minority of studies have reported unexpected results concerning convergence between SMTM and survey-based wellbeing. One study found SMTM

wellbeing was not associated with life (dis)satisfaction at the location level (LabMT applied to 232 zip codes in Utah, Nguyen, Kath, et al., 2016). The reason for the null associations between life (dis)satisfaction and wellbeing at location level may be due to the time gap between the assessment of the ground-truth measure (between 2009 and 2010) and when the language data was collected (between 2009 and 2014). Another study reported no association between SMTM wellbeing and ground-truth wellbeing measures across 81 provinces in Turkey (Durahim & Coşkun, 2015). The second study's null results may be explained by the fact that the ground-truth measure for wellbeing consisted of scores given by government officials, which differed from the other studies' methods, where social desirability and anonymity concerns may have played a role. Finally, the last study found negative word use on social media to be positively associated with survey-based wellbeing (N. Wang et al., 2014). The unexpected results from the study finding negative word use being positively related to wellbeing may be due to the use of LIWC negative words in the reverse context or sarcasm (e.g., I feel 'terribly' happy). Such problems with word-level methods (like LIWC) might have been particularly alleviated in this study where the language data from individuals were aggregated at different time windows like days, weeks, and months. Despite these few unexpected results, we have observed higher convergent validities for regional studies compared to individual-level were observed. This may be due to the higher prevalence of data-driven methods and larger sample sizes observed in the former group of studies.

Overall, the qualitative synthesis has shown the value of large-scale data collection and wellbeing assessment via social media language and provided a detailed picture of the current state of the field. Large sample sizes, frequent use of ground-truth measures, as well as widely used data-driven methods ensures the quality of the studies to improve further in the future as well. The multilingual versions of newer data-driven methods may allow for better assessment of wellbeing in non-English speaking populations.

Quantitative synthesis: meta-analysis and publication bias

A meta-analysis was conducted to obtain a meta-analytic estimate for converging validities across studies. We meta-analysed 32 effect sizes acquired from 18 studies that reported convergent validities of SMTM wellbeing and seemed, based on Egger's test (Egger et al., 1997) and the trim and fill method (Duval & Tweedie,

2000), unaffected by publication bias. The results based on a meta-analysis including effect sizes of optimal models (thus signifying the upper limit) indicated moderate convergence between SMTM and survey-based wellbeing (meta-analytic $r=0.40$, 95% CI [0.33–0.47]). This correlation is largely similar to the results of an earlier meta-analysis (meta-analytic $r=0.37$, CI 95% [0.28–0.45]) (Settanni et al., 2018) and are similar to convergent validity coefficients achieved by other methods (e.g., peer reports with self-report surveys) for personality and wellbeing which typically range between $r=0.30$ – 0.40 (e.g., Park et al., 2015; Schneider & Schimmack, 2009). The correlation is, however, higher than the values in previous literature reviews which reported correlations being between 0.20 and 0.40 for affect, and smaller than 0.20 for life satisfaction (e.g., Bellet & Frijters, 2019; Luhmann, 2017). These differences may be explained by these studies being literature reviews and not systematic reviews, which may have resulted in not including all of the available evidence in the field.

The results showed higher convergent validities for location level studies (between 0.37 and 0.67) compared to individual-level studies (between 0.25 and 0.40), as previously reported in the World Happiness Report in 2019 (Bellet & Frijters, 2019). At the individual level, word-level methods performed better than data-driven methods (average $r=0.38$ and 0.26 respectively; $SD=0.22$ and 0.13). At the location level, however, data-driven methods performed better than word-level methods (average $r=0.52$ and 0.38, respectively; $SD=0.24$, SD for the second mean cannot be calculated as it is based on a single score). Nevertheless, it should be noted that the latter result was based on a single study, limiting the interpretability of this result. Overall, this pattern of findings may be explained by the fact that word-level methods (e.g., LIWC) were developed with a particular focus on interpreting individuals' psychological states and traits rather than regions. On the contrary, data-driven methods are thought to capture the nuances and differences in the language of different geographical regions and achieve higher accuracies, but require bigger datasets, as observed for regional studies (Jaidka et al., 2020).

Limitations of the study and social media data use in general

Our review and meta-analysis should be interpreted in light of the following limitations. In the present review, pre-prints were not included which may have caused missing very recent developments in the field, but it guarantees that the included studies were peer-reviewed. In the meta-analysis section, the models

achieving the highest convergence in each study were included for the calculation of the effect sizes. Thus, the meta-analytic convergence between SMTM and survey-based wellbeing reflects the upper limit. Among the search terms used, we did not include eudaimonic wellbeing explicitly. However, the other search terms we used for general wellbeing (e.g., wellbeing, well being, or well-being) probably would have captured eudaimonic wellbeing studies if they existed. Clearly, eudaimonic wellbeing is researched less than hedonic wellbeing (e.g., life satisfaction, positive or negative effects), perhaps due to the length and scope of the self-report measures dedicated to this construct (Ryff, 1989). Therefore, we estimated that we covered most of the studies available in the field – if not all of them.

More in general, although SMTM appears to be a valid method for assessing wellbeing, it is important to acknowledge the limitations inherent to social media language data. Social media language data are often 'noisier' than survey data. There is temporal variation in terms of the amount of text data produced between individuals (some write more, and more frequently), and within individuals (text production per time unit varies). In addition, social media data are not representative of the population concerning age, sex, income levels, educational levels, and ethnicities (e.g., Blank & Lutz, 2017; Hargittai & Dobransky, 2017; Hargittai, 2020; Mislove et al., 2011). For example, a recent report mentioned fewer global female Facebook users compared to males, although female users were more active (e.g., frequency of post likes, comments, or the number of clicks on advertisements) (Kemp, 2021). In addition, language features (e.g., words) show a Zipfian distribution, i.e., most words only occur a few times (Eichstaedt et al., 2021), leading to sparse data for most words. People use a very large number of different words or topics – with low base rates and uneven distributions across the population – each with small effects on the phenotype of interest (e.g., wellbeing). In order to find these small effects, large sample sizes are needed. At the same time, while a single word or collection of words may be an imperfect measure of wellbeing, given the sample sizes that can be achieved, language may provide a valid measure of wellbeing when aggregating all the small effects across all words. Overall, researchers must acknowledge the potential limitations of social media language datasets and correct for those, if possible.

Recommendations and future studies

Based on the results of our systematic review and meta-analysis, we have the following recommendations for the field of SMTM:

Expand to other languages and populations

The results of this review indicated that most studies focused on English language datasets. However, as we foresee an increase in the use of SMTM, it may become more important to reassure that all the voices on social media are being heard, especially if policymakers and researchers aim to infer the wellbeing levels of a particular region to aid (social) policies. A substantial number of people in a specific region may use a different language to express their happiness and worries instead of a dominantly spoken language/dialect of a country. Similar issues have been observed in other research areas, such as in genetics, where inferences are predominantly based on European ancestry samples, worsening the existing health disparities between over and underrepresented groups (Martin et al., 2019). In line with large-scale worldwide initiatives in the field of genetics, for instance, creating methods that are compatible with other ancestry populations-datasets (e.g., Multi ancestry Meta-Analysis; Turley et al., 2021), or compiling datasets representing diverse populations (e.g., 23andMe, All of Us, China Kadoorie Biobank), applications of multilingual SMTM and collecting multilingual datasets may become important to make SMTM more inclusive (Hsu et al., 2021).

Combine data from different social media platforms

We observed that most datasets were acquired from Twitter, and only a single study has combined data from two platforms (Facebook and Twitter) (Jaidka et al., 2020). Making general population-level inferences based on single platform data may result in misleading conclusions, given the presence of potential platform-specific sample selection mechanisms (e.g., females and younger individuals use Facebook more than males and older individuals; Blank & Lutz, 2017). The issue of non-representativeness can be relieved by collecting data from multiple social media platforms (to acquire a more comprehensive picture), and applying existing de-biasing techniques (e.g., Giorgi et al., 2019, 2021; Z. Wang et al., 2019) if population-level conclusions are being made.

Collect social media data in existing cohorts

To further control for biases in social media data, researchers can request access to social media data from individuals who already participate in large-scale population-based studies such as the CDC's Behavioral Risk Factor Surveillance System (BRFSS; Johnson et al., 2014; Nelson et al., 2001), the UK Biobank (UKB; Sudlow et al., 2015), the Midlife in the United States national survey (MIDUS; Brim et al., 2004), the Health and Retirement Study (HRS; Juster & Suzman, 1995), and

the National Longitudinal Study of Adolescent Health (Add Health; Harris, 2013). It is typical for these studies to already collect a wide array of (demographic) information from their participants (for instance, yearly). These population-based samples may be more representative of the general population, and by collecting social media data from such samples, potential sampling biases can be reduced.

Such large-scale datasets can be used for other purposes as well. For instance, the existing survey scores collected in the past can be easily augmented with the social media language of the same individuals from the same time point in the past. By doing so, the convergent validity of SMTM for various traits including but not limited to wellbeing (e.g., depression, personality) can be investigated for multiple time points. In addition to that, combining multiple types of data from the same individuals in a continuous fashion (e.g., survey, SMTM) makes a real-time assessment of wellbeing (and other traits) possible. Social media language features and survey scores can also be combined in the same model to increase prediction model performances.

It should be noted, however, linking these data, as well as collecting the text-based social media data of individuals requires caution for privacy concerns. Researchers must respect the individuals' rights to privacy and reassure that all ethical requirements are sufficiently met.

Use open vocabulary and data-driven approaches

Most of the papers we examined use closed-vocabulary methods to extract language features. We recommend that researchers also use open vocabulary methods and apply data-driven approaches which can offer improved predictions compared to when only closed vocabulary methods or word-level approaches are used.

Conclusion

The present qualitative synthesis and meta-analysis supported the value of SMTM to cost-efficiently assess wellbeing both at the individual and regional levels. SMTM can be used to assess past and present wellbeing. Application of SMTM to assess wellbeing – in real-time – can eventually help develop personalized interventions to increase wellbeing, or aid policymakers to adjust their decisions to maximize the wellbeing of (inhabitants of) neighbourhoods, countries, and cities. The use of SMTM for assessing wellbeing may also provide new opportunities for researchers. For instance, individuals' wellbeing levels and their

variation over time can be analysed in combination with other existing datasets including surveys, physiological, and laboratory measures.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by a European Research Council Consolidator Grant (WELL-BEING 771057, PI Bartels)

ORCID

S. Sametoğlu  <http://orcid.org/0000-0002-5447-5410>
 D.H.M. Pelt  <http://orcid.org/0000-0001-5926-5247>
 J.C. Eichstaedt  <http://orcid.org/0000-0002-3220-2972>
 M. Bartels  <http://orcid.org/0000-0002-9667-7555>

Data availability statement

The data that support the findings of this study are available from the corresponding author, S. Sametoğlu, upon reasonable request.

References

*Studies included in the meta-analysis

- Liu, P., Tov, W., Kosinski, M., Stillwell, D. J., & Qiu, L. (2015). Do Facebook Status Updates Reflect Subjective Well-Being? *Cyberpsychology, Behavior and Social Networking*, 18(7), 373–379. <https://doi.org/10.1089/cyber.2015.0022>
- Marengo, D., Azucar, D., Longobardi, C., & Settanni, M. (2021). Mining Facebook data for quality of life assessment. *Behaviour & Information Technology*, 40(6), 597–607. <https://doi.org/10.1080/0144929X.2019.1711454>
- Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., Danforth, C. M., & Sánchez, A. (2013). The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *Plos One*, 8(5), e64417. <https://doi.org/10.1371/journal.pone.0064417>
- Qi, J., Fu, X., & Zhu, G. (2015). Subjective well-being measurement based on Chinese grassroots blog text sentiment analysis. *Information & Management*, 52(7), 859–869. <https://doi.org/10.1016/j.im.2015.06.002>
- Schwartz, H. A., Sap, M., Kern, M. L., Eichstaedt, J. C., Kapelner, A., Agrawal, M., Blanco, E., Dziurzynski, L., Park, G., Stillwell, D., Kosinski, M., Seligman, M. E. P., & Ungar, L. H. (2016). Predicting individual well-being through the language of social media. *Biocomputing*, 2016, 516–527. <https://doi.org/10.1142/97898147494110047>
- Wang, N., Kosinski, M., Stillwell, D. J., & Rust, J. (2014). Can well-being be measured using Facebook status updates? Validation of Facebook's gross national happiness index. *Social Indicators Research*, 115(1), 483–491. <https://doi.org/10.1007/s11205-012-9996-9>
- Curini, L., Iacus, S., & Canova, L. (2015). Measuring idiosyncratic happiness through the analysis of Twitter: An application to the Italian case. *Social Indicators Research*, 121(2), 525–542. <https://doi.org/10.1007/s11205-014-0646-2>
- Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS One*, 6(12), e26752. <https://doi.org/10.1371/journal.pone.0026752>
- Jaidka, K., Giorgi, S., Schwartz, H. A., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2020). Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*, 117(19), 10165–10171. <https://doi.org/10.1073/pnas.1906364117>
- Nguyen, Q. C., Kath, S., Meng, H.-W., Li, D., Smith, K. R., VanDerslice, J. A., Wen, M., & Li, F. (2016). Leveraging geo-tagged Twitter data to examine neighborhood happiness, diet, and physical activity. *Applied Geography*, 73, 77–88. <https://doi.org/10.1016/j.apgeog.2016.06.003>
- Coşkun, M., & Ozturan, M. (2018). #europehappinessmap: A framework for multi-lingual sentiment analysis via social media big data (a Twitter case study). *Information*, 9(5), 102. <https://doi.org/10.3390/info9050102>
- Settanni, M., & Marengo, D. (2015). Sharing feelings online: Studying emotional well-being via automated text analysis of Facebook posts. *Frontiers in Psychology*, 6, 6. <https://doi.org/10.3389/fpsyg.2015.01045>
- Bai, S., Gao, R., Hao, B., Yuan, S., & Zhu, T. (2014). Identifying social satisfaction from social media. ArXiv: 14073552 [Physics]. <http://arxiv.org/abs/1407.3552>
- Schwartz, H. A., Eichstaedt, J., Kern, M., Dziurzynski, L., Lucas, R., Agrawal, M., Park, G., Lakshminanth, S., Jha, S., Seligman, M., & Ungar, L. (2013). Characterizing geographic variation in well-being using tweets. *Proceedings of the International AAAI Conference on Web & Social Media*, 7(1), 583–591. <https://doi.org/10.1609/icwsm.v7i1.14442>
- Collins, S., Sun, Y., Kosinski, M., Stillwell, D., & Markuzon, N. (2015). Are you satisfied with life?: Predicting satisfaction with life from Facebook. In N. Agarwal, K. Xu, & N. Osgood (Eds.), *Social computing, behavioral-cultural modeling, and prediction* (Vol. 9021, pp. 24–33). Springer International Publishing. https://doi.org/10.1007/978-3-319-16268-3_3
- Bogolyubova, O., Panicheva, P., Ledovaya, Y., Tikhonov, R., & Yaminov, B. (2020). The language of positive mental health: Findings from a sample of Russian Facebook users. *SAGE Open*, 10(2), 215824402092437. <https://doi.org/10.1177/2158244020924370>
- Hao, B., Li, L., Gao, R., Li, A., & Zhu, T. (2014). Sensing subjective well-being from social media. In D. Ślęzak, G. Schaefer, S. T. Vuong, & Y.-S. Kim (Eds.), *Active media technology* (pp. 324–335). Springer International Publishing. https://doi.org/10.1007/978-3-319-09912-5_27
- Kramer, A. D. I. (2010). An unobtrusive behavioral model of 'gross national happiness'. *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI April 10-15, 2010 Atlanta, Georgia, USA '10*. New York, NY, United States: Association for Computing Machinery, 287–290. <https://doi.org/10.1145/1753326.1753369>
- Cao, X., MacNaughton, P., Deng, Z., Yin, J., Zhang, X., & Allen, J. (2018). Using Twitter to better understand the

- spatiotemporal patterns of public sentiment: A case study in Massachusetts, USA. *International Journal of Environmental Research and Public Health*, 15(2), 250. <https://doi.org/10.3390/ijerph15020250>
- Chen, L., Gong, T., Kosinski, M., Stillwell, D., Davidson, R. L., & Xia, F. (2017). Building a profile of subjective well-being for social media users. *PLoS One*, 12(11), e018727. <https://doi.org/10.1371/journal.pone.0187278>
- Iacus, S. M., Porro, G., Salini, S., & Siletti, E. (2020). An Italian composite subjective well-being index: The voice of Twitter Users from 2012 to 2017. *Social Indicators Research*, 161(2–3), 471–489. <https://doi.org/10.1007/s11205-020-02319-6>
- Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews, Computational Statistics*, 2(4), 433–459. <https://doi.org/10.1002/wics.101>
- Bartels, M., & Boomsma, D. I. (2009). Born to be happy? The etiology of subjective well-being. *Behavior Genetics*, 39(6), 605–615. <https://doi.org/10.1007/s10519-009-9294-8>
- Baselmans, B. M., & Bartels, M. (2018). A genetic perspective on the relationship between eudaimonic and hedonic well-being. *Scientific Reports*, 8(1), 1–10. <https://doi.org/10.1038/s41598-018-32638-1>
- Bellet, C., & Frijters, P. (2019, March 20). *Big data and well-being*. World Happiness Report 2019 (New York: Sustainable Development Solutions Network.). <https://worldhappiness.report/ed/2019/big-data-and-well-being/>
- Blank, G., & Lutz, C. (2017). Representativeness of social media in great britain: Investigating Facebook, LinkedIn, Twitter, Pinterest, Google+, and Instagram. *The American Behavioral Scientist*, 61(7), 741–756. <https://doi.org/10.1177/0002764217717559>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022 <https://dl.acm.org/doi/10.5555/944919.944937>.
- Boyd, R., Ashokkumar, A., Seraj, S., & Pennebaker, J. (2022). *The development and psychometric properties of LIWC-22*. University of Texas at Austin. <https://doi.org/10.13140/RG.2.2.23890.43205>
- Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings (Tech. Report C-1) Vol. 30. Gainesville: University of Florida, Center for Research in Psychophysiology.
- Brim, O. G., Ryff, C. D., & Kessler, R. C. (2004) The MIDUS National Survey: An Overview. In Brim, O. G., Ryff, C. D., Kessler, R. C. (Eds.) *How healthy are we?: A national study of well-being at midlife* (pp. 1-34). Chicago, IL: The University of Chicago Press.
- Busseri, M. A. (2018). Examining the structure of subjective well-being through meta-analysis of the associations among positive affect, negative affect, and life satisfaction. *Personality & Individual Differences*, 122, 68–71. <https://doi.org/10.1016/j.paid.2017.10.003>
- Chapman, B., & Guven, C. (2016). Revisiting the relationship between marriage and wellbeing: Does marriage quality matter? *Journal of Happiness Studies*, 17(2), 533–551. <https://doi.org/10.1007/s10902-014-9607-3>
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. ArXiv:191102116 [Cs]. arxiv.org/abs/1911.02116
- Deci, E. L., & Ryan, R. M. (2008). Hedonia, eudaimonia, and well-being: An introduction. *Journal of Happiness Studies*, 9(1), 1–11. <https://doi.org/10.1007/s10902-006-9018-1>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv:181004805 [Cs]. <https://arxiv.org/abs/1810.04805>
- Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, 49(1), 71–75. https://doi.org/10.1207/s15327752jpa4901_13
- Diener, E., Seligman, M. E. P., Choi, H., & Oishi, S. (2018). Happiest people revisited. *Perspectives on Psychological Science*, 13(2), 176–184. <https://doi.org/10.1177/1745691617697077>
- Disabato, D. J., Goodman, F. R., Kashdan, T. B., Short, J. L., & Jarden, A. (2016). Different types of well-being? A cross-cultural examination of hedonic and eudaimonic well-being. *Psychological Assessment*, 28(5), 471. <https://doi.org/10.1037/pas0000209>
- Durahim, A. O., & Coşkun, M. (2015). #iamhappybecause: Gross national happiness through Twitter analysis and big data. *Technological Forecasting & Social Change*, 99, 92–105. <https://doi.org/10.1016/j.techfore.2015.06.035>
- Duval, S., & Tweedie, R. (2000). A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, 95(449), 89–98. <https://doi.org/10.1080/01621459.2000.10473905>
- Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. Dryden Press.
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629–634. <https://doi.org/10.1136/bmj.315.7109.629>
- Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., Hagan, C. A., Tobolsky, V. A., Smith, L. K., Buffone, A., Iwry, J., Seligman, M. E. P., & Ungar, L. H. (2021). Closed-and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, 26(4), 398. <https://doi.org/10.1037/met0000349>
- Giorgi, S., Lynn, V., Gupta, K., Ahmed, F., Matz, S., Ungar, L., & Schwartz, H. A. (2019). Correcting sociodemographic selection biases for population prediction from social media arXiv:1911.03855 [Cs]. <https://arxiv.org/abs/1911.03855>
- Giorgi, S., Nguyen, K. L., Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Kosinski, M., Seligman, M. E., Ungar, L. H., Schwartz, H. A., & Park, G. (2021). Regional personality assessment through social media language. *Journal of Personality*, 90(3), 405–425. <https://doi.org/10.1111/jopy.12674>
- Gottschalk, L. A., Gleser, G. C., Daniels, R. S., & Block, S. (1958). The speech patterns of schizophrenic patients: A method of assessing relative degree of personal disorganization and social alienation. *Journal of Nervous & Mental Disease*, 127(2), 153–166. <https://doi.org/10.1097/00005053-195808000-00008>
- Gottschalk, L. A., Gleser, G. C., & Winget, C. N. (1969). *Manual of instructions for using the Gottschalk-Gleser content analysis scales: Anxiety, hostility, and social alienation-personal disorganization*. Univ of California Press. <https://doi.org/10.1525/9780520318816>
- Hargittai, E. (2020). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review*, 38(1), 10–24. <https://doi.org/10.1177/0894439318788322>

- Hargittai, E., & Dobransky, K. (2017). Old dogs, new clicks: Digital inequality in skills and uses among older adults. *Canadian Journal of Communication*, 42(2). <https://doi.org/10.22230/cjc.2017v42n2a3176>
- Harris, K. M. (2013). *The add health study: Design and accomplishments* (Vol. 1). Carolina Population Center, University of North Carolina at Chapel Hill.
- Hedges, L. V., & Olkin, I. (2014). *Statistical methods for meta-analysis*. Academic Press.
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1(1), 39–65. <https://doi.org/10.1002/jrsm.5>
- Henry, J. D., & Crawford, J. R. (2005). The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, 44(2), 227–239. <https://doi.org/10.1348/014466505X29657>
- Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, 327(7414), 557–560. <https://doi.org/10.1136/bmj.327.7414.557>
- Holtzman, W. H. (1950). Validation studies of the Rorschach test: Shyness and gregariousness in the normal superior adult. *Journal of Clinical Psychology*, 6(4), 343–347. [https://doi.org/10.1002/1097-4679\(195010\)6:4<343:AID-JCLP2270060407>3.0.CO;2-B](https://doi.org/10.1002/1097-4679(195010)6:4<343:AID-JCLP2270060407>3.0.CO;2-B)
- Hsu, T. W., Niiya, Y., Thelwall, M., Ko, M., Knutson, B., & Tsai, J. L. (2021). Social media users produce more affect that supports cultural values, but are more influenced by affect that violates cultural values. *Journal of Personality & Social Psychology*, 121(5), 969–983. <https://doi.org/10.1037/pspa0000282>
- James, P., Kim, E. S., Kubzansky, L. D., Zevon, E. S., Trudel-Fitzgerald, C., & Grodstein, F. (2019). Optimism and healthy aging in women. *American Journal of Preventive Medicine*, 56(1), 116–124. <https://doi.org/10.1016/j.amepre.2018.07.037>
- Johnson, N. B., Hayes, L. D., Brown, K., Hoo, E. C., & Ethier, K. A. (2014). US: Centers for Disease Control and Prevention. CDC national health report: Leading causes of morbidity and mortality and associated behavioral risk and protective factors—United States, 2005–2013.
- Juster, F. T., & Suzman, R. (1995). An overview of the health and retirement study. *The Journal of Human Resources*, 30, S7–S56. <https://doi.org/10.2307/146277>
- Kemp, S. (2021, January 27). Digital 2021: Global overview report. Datareportal-Kepios Pte. Ltd., We Are Social Ltd. and Hootsuite Inc. <https://datareportal.com/reports/digital-2021-global-overview-report>
- Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53(3), 983–997. <https://doi.org/10.2307/2533558>
- Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological Methods*, 21(4), 507. <https://doi.org/10.1037/met0000091>
- Keyes, C. L. M. (2002). The mental health continuum: From languishing to flourishing in life. *Journal of Health & Social Behavior*, 43(2), 207–222. <https://doi.org/10.2307/3090197>
- Keyes, C. L. M. (2010). Flourishing. In I. B. Weiner & W. E. Craighead (Eds.), *The corsini encyclopedia of psychology* (p. corpsy 363). John Wiley & Sons, Inc. <https://doi.org/10.1002/9780470479216.corpsy0363>
- Kim, E. S., James, P., Zevon, E. S., Trudel-Fitzgerald, C., Kubzansky, L. D., & Grodstein, F. (2019). Optimism and healthy aging in women and men. *American Journal of Epidemiology*, 188(6), 1084–1091. <https://doi.org/10.1093/aje/kwz056>
- Kloumann, I. M., Danforth, C. M., Harris, K. D., Bliss, C. A., & Dodds, P. S. (2012). Positivity of the English language. *PLoS One*, 7(1), e29484. <https://doi.org/10.1371/journal.pone.0029484>
- Kristoufek, L. (2018). Does solar activity affect human happiness? *Physica A Statistical Mechanics & Its Applications*, 493, 47–53. <https://doi.org/10.1016/j.physa.2017.10.031>
- Lambert, L., Lomas, T., van de Weijer, M. P., Passmore, H. A., Joshanloo, M., Harter, J., Ishikawa, Y., Lai, A., Kitagawa, T., Chen, D., Kawakami, T., Miyata, H., & Diener, E. (2020). Towards a greater global understanding of wellbeing: A proposal for a more inclusive measure. *International Journal of Wellbeing*, 10(2), 1–18. <https://doi.org/10.5502/ijw.v10i2.1037>
- Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. *ArXiv: 190107291 [Cs]*. arxiv.org/abs/1901.07291
- Longo, Y., Coyne, I., Joseph, S., & Gustavsson, P. (2016). Support for a general factor of well-being. *Personality & Individual Differences*, 100, 68–72. <https://doi.org/10.1016/j.paid.2016.03.082>
- Luhmann, M. (2017). Using big data to study subjective well-being. *Current Opinion in Behavioral Sciences*, 18, 28–33. <https://doi.org/10.1016/j.cobeha.2017.07.006>
- Luhmann, M., Hofmann, W., Eid, M., & Lucas, R. E. (2012). Subjective well-being and adaptation to life events: A meta-analysis. *Journal of Personality & Social Psychology*, 102(3), 592–615. <https://doi.org/10.1037/a0025948>
- Maccagnan, A., Wren Lewis, S., Brown, H., & Taylor, T. (2019). Wellbeing and society: Towards Quantification of the co-benefits of wellbeing. *Social Indicators Research*, 141(1), 217–243. <https://doi.org/10.1007/s11205-017-1826-7>
- Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4), 584–591. <https://doi.org/10.1038/s41588-019-0379-x>
- McClelland, D. C. (1979). Inhibited power motivation and high blood pressure in men. *Journal of Abnormal Psychology*, 88(2), 182. <https://doi.org/10.1037/0021-843X.88.2.182>
- Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. (2011). Understanding the demographics of Twitter users. *Proceedings of the International AAAI Conference on Web & Social Media*, 5(1), 554–557. <https://doi.org/10.1609/icwsm.v5i1.14168>
- Moeyaert, M., Ugille, M., Natasha Beretvas, S., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2017). Methods for dealing with multiple outcomes in meta-analysis: A comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology*, 20(6), 559–572. <https://doi.org/10.1080/13645579.2016.1252189>
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, T. P. (2009). Preferred reporting items for systematic reviews and

- meta-analyses: The PRISMA statement. *PLoS Medicine*, 6(7), e1000097. <https://doi.org/10.1371/journal.pmed.1000097>
- Nelson, D. E., Holtzman, D., Bolen, J., Stanwyck, C. A., & Mack, K. A. (2001). Reliability and validity of measures from the Behavioral Risk Factor Surveillance System (BRFSS). *Sozial-Und Praventivmedizin*, 46, Suppl 1, S3–42.
- Okabe-Miyamoto, K., & Lyubomirsky, S. (in press). Happiness shapes and is shaped by social cognition and social connection. In D. Carlston, K. Johnson, & K. Hugenberg (Eds.), *The Oxford handbook of social cognition* (2nd ed.). Oxford University Press. <https://sonjalyubomirsky.com/papers-publications/>
- Oswald, A. J., Proto, E., & Sgroi, D. (2015). Happiness and productivity. *Journal of Labor Economics*, 33(4), 789–822. <https://doi.org/10.1086/681096>
- Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality & Social Psychology*, 108(6), 934–952. <https://doi.org/10.1037/pspp0000020>
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. Austin, TX: University of Texas at Austin.
- Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is Multilingual BERT? *ArXiv: 190601502 [Cs]*. <https://doi.org/10.1101/190601502>
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rong, X. (2014). Word2vec parameter learning explained. *arXiv:1411.2738 [Cs]*. <https://arxiv.org/abs/1411.2738>
- Rorschach, H. (1921). *Psychodiagnostik*. Ernst Bircher Verlag.
- Rosenberg, S. D., & Tucker, G. J. (1979). Verbal behavior and schizophrenia: The semantic dimension. *Archives of General Psychiatry*, 36(12), 1331–1337. <https://doi.org/10.1001/archpsyc.1979.01780120061008>
- Rubio-Aparicio, M., López-López, J. A., Viechtbauer, W., Marín-Martínez, F., Botella, J., & Sánchez-Meca, J. (2020). Testing categorical moderators in mixed-effects meta-analysis in the presence of heteroscedasticity. *The Journal of Experimental Education*, 88(2), 288–310. <https://doi.org/10.1080/00220973.2018.1561404>
- Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality & Social Psychology*, 57(6), 1069. <https://doi.org/10.1037/0022-3514.57.6.1069>
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *ArXiv: 191001108 [Cs]*. <http://arxiv.org/abs/1910.01108>
- Santini, Z. I., Becher, H., Jørgensen, M. B., Davidsen, M., Nielsen, L., Hinrichsen, C., Madsen, K. R., Meilstrup, C., Koyanagi, A., Stewart-Brown, S., McDaid, D., & Koushede, V. (2021). Economics of mental well-being: A prospective study estimating associated health care costs and sickness benefit transfers in Denmark. *The European Journal of Health Economics*, 22(7), 1053–1065. <https://doi.org/10.1007/s10198-021-01305-0>
- Schneider, L., & Schimmack, U. (2009). Self-informant agreement in well-being ratings: A meta-analysis. *Social Indicators Research*, 94(3), 363–376. <https://doi.org/10.1007/s11205-009-9440-y>
- Schwartz, H. A., & Ungar, L. H. (2015). Data-driven content analysis of social media: A systematic overview of automated methods. *The Annals of the American Academy of Political and Social Science*, 659(1), 78–94. <https://doi.org/10.1177/0002716215569197>
- Seligman, M. (2018). PERMA and the building blocks of well-being. *The Journal of Positive Psychology*, 13(4), 333–335. <https://doi.org/10.1080/17439760.2018.1437466>
- Settanni, M., Azucar, D., & Marengo, D. (2018). Predicting individual characteristics from digital traces on social media: A meta-analysis. *Cyberpsychology, Behavior and Social Networking*, 21(4), 217–228. <https://doi.org/10.1089/cyber.2017.0384>
- Shiffman, S., Hufford, M., Hickcox, M., Paty, J. A., Gnys, M., & Kassel, J. D. (1997). Remember that? A comparison of real-time versus retrospective recall of smoking lapses. *Journal of Consulting & Clinical Psychology*, 65(2), 292. <https://doi.org/10.1037/0022-006X.65.2.292.a>
- Steptoe, A. (2019). Happiness and health. *Annual Review of Public Health*, 40, 339–359. <https://doi.org/10.1146/annurev-publhealth-040218-044150>
- Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). *The General Inquirer: A Computer Approach to Content Analysis* (Cambridge, Massachusetts: M. I. T. Press).
- Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3), e1001779. <https://doi.org/10.1371/journal.pmed.1001779>
- Tankovska, H. (2020). *Number of social network users worldwide from 2017 to 2025*. Statista. <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/>
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language & Social Psychology*, 29(1), 24–54. <https://doi.org/10.1177/0261927X09351676>
- Tay, L., Woo, S. E., Hickman, L., & Saef, R. M. (2020). Psychometric and validity issues in machine learning approaches to personality assessment: A focus on social media text mining. *European Journal of Personality*, 34(5), 826–844. <https://doi.org/10.1002/per.2290>
- Tsai, J. L., Knutson, B., & Fung, H. H. (2006). Cultural variation in affect valuation. *Journal of Personality & Social Psychology*, 90(2), 288–307. <https://doi.org/10.1037/0022-3514.90.2.288>
- Turley, P., Martin, A. R., Goldman, G., Li, H., Kanai, M., Walters, R. K., Jala, J. B., Lin, K., Millwood, I. Y., Carey, C. E., Palmer, D. S., Zacher, E. G., Chen, Z., Li, L., Akiyama, M., Okada, Y., Kamatani, Y., Walters, R. G., Callier, S., Laibson, D., ... Neale, B. M. (2021). Multi-ancestry meta-analysis yields novel genetic discoveries and ancestry-specific associations. *bioRxiv 2021.04.23.441003*. <https://doi.org/10.1101/2021.04.23.441003>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3). <https://doi.org/10.18637/jss.v036.i03>
- Walsh, L. C., Boehm, J. K., & Lyubomirsky, S. (2018). Does happiness promote career success? Revisiting the evidence. *Journal of Career Assessment*, 26(2), 199–219. <https://doi.org/10.1177/1069072717751441>

Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flöck, F., & Jurgens, D. Demographic inference and representative population estimates from multilingual social media data (2019). <https://arxiv.org/abs/1905.05961> Proceedings of the 2019 World Wide Web Conference (WWW '19) arXiv:1905.05961 [Cs]. 2056–2067.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas.

Behavior Research Methods, 45(4), 1191–1207. <https://doi.org/10.3758/s13428-012-0314-x>

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems Vancouver, Canada. 5753–5763. <https://dl.acm.org/doi/10.5555/3454287.3454804>