European Journal of Personality, *Eur. J. Pers.* **34**: 845–858 (2020)
Published online 21 May 2020 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/per.2261

Tracking Fluctuations in Psychological States Using Social Media Language: A Case Study of Weekly Emotion

JOHANNES C. EICHSTAEDT[1]* and AARON C. WEIDMAN[2]*

[1]*Stanford University, USA*
[2]*University of Michigan, USA*

Abstract: Personality psychologists are increasingly documenting dynamic, within-person processes. Big data methodologies can augment this endeavour by allowing for the collection of naturalistic and personality-relevant digital traces from online environments. Whereas big data methods have primarily been used to catalogue static personality dimensions, here we present a case study in how they can be used to track dynamic fluctuations in psychological states. We apply a text-based, machine learning prediction model to Facebook status updates to compute weekly trajectories of emotional valence and arousal. We train this model on 2895 human-annotated Facebook statuses and apply the resulting model to 303 575 Facebook statuses posted by 640 US Facebook users who had previously self-reported their Big Five traits, yielding an average of 28 weekly estimates per user. We examine the correlations between model-predicted emotion and self-reported personality, providing a test of the robustness of these links when using weekly aggregated data, rather than momentary data as in prior work. We further present dynamic visualizations of weekly valence and arousal for every user, while making the final data set of 17 937 weeks openly available. We discuss the strengths and drawbacks of this method in the context of personality psychology's evolution into a dynamic science. © 2020 European Association of Personality Psychology

Key words: big data; digital footprints; experience sampling; emotion; personality

Personality psychologists have in recent years shifted their focus towards documenting dynamic, within-person personality processes. This shift can in large part be traced to seminal work by Fleeson (2001), which shed light on the tremendous variability people exhibit in their personality-relevant states. Whole Trait Theory, the most recent instantiation of this work (Fleeson & Jayawickreme, 2015), recognizes that variability represents a meaningful component of personality, above and beyond an individual's typical mode of behaviour. To comprehensively explain personality, therefore, requires examining individual differences in dynamic patterns of microunits such as behaviours, goals, motives, and situational considerations. Personality scientists across numerous domains have followed suit by proposing conceptual models that consider dynamic factors (e.g. Hopwood, 2018) and by designing empirical studies that explicitly examine within-person processes (e.g. Allemand & Hill, 2019; Jayawickreme, Tsukayama, & Kashdan, 2017; Jones, Brown, Serfass, & Sherman, 2017; Sun, Schwartz, Son, Kern, & Vazire, 2019). These studies typically employ protocols that allow for intensive and repeated sampling of individuals' lived experiences, such as experience sampling methodology (ESM; Conner, Tennen, Fleeson, & Barrett, 2009) and intensive longitudinal designs (ILD; Sened, Lazarus, Gleason, Rafaeli, & Fleeson, 2018), in which individuals are assessed frequently over a brief time window, thereby facilitating ideographic analyses of dynamic personality processes.

We believe that personality psychologists' increasing theoretical and methodological focus on dynamic, within-person processes has a natural complement in 'big data' methodologies. This is because digital environments (such as social media) can allow for the collection of naturally occurring *digital traces* that people leave in their online environments and which are indicative of personality (e.g. Tweets and Facebook likes; Harari et al., 2016). Indeed, the widespread use of online environments in recent years has meant that digital traces have become a prominent source of data in personality science. Considerable work has shed light on the validity of digital traces for cataloguing personality-relevant dimensions, including *individual differences in psychological traits* (e.g. stable personality dimensions; Back et al., 2010; Kosinski, Stillwell, & Graepel, 2013; Park et al., 2015) as well as *community-level patterns in psychological traits* (e.g. the aggregate well-being in a county; Dodds, Harris, Kloumann, Bliss, & Danforth, 2011; Golder & Macy, 2011; Schwartz et al., 2013).

*Correspondence to: Johannes C. Eichstaedt, Stanford University, Stanford, CA, USA and Aaron C. Weidman, University of Michigan, Ann Arbor, MI, USA.
E-mail: johannes.stanford@gmail.com; aaron.c.weidman@gmail.com

This article earned Open Materials badge through Open Practices Disclosure from the Center for Open Science: https://osf.io/tvyxz/wiki. The materials are permanently and openly accessible at https://osf.io/pbjer/. Author's disclosure form may also be found at the Supporting Information in the online version.

Handling editor: John Rauthmann
Received 8 May 2019
Revised 25 February 2020, Accepted 30 March 2020

Despite the seeming compatibility of big data methodologies with the increasingly dynamic, process-focused field of personality psychology, these two trends have yet to fully merge. As noted earlier, groundbreaking studies have shown that digital traces can shed light on *static* personality features. At the same time, studies have relied primarily on ESM and ILDs to shed light on within-person personality-relevant processes (e.g. Allemand & Hill, 2019; Jayawickreme et al., 2017; Sun et al., 2019). Yet little work has used big data methods (such as machine learning) to explicitly investigate *dynamic, within-person personality processes*.

The broad goal of this paper is to explicate an initial attempt at bridging this gap between big data analyses and dynamic personality psychology. We aim to provide a case study of how big data analyses of digital traces can be leveraged to track *within-person patterns in psychological states* at scale. To achieve this goal, we use data concerning the fundamental emotion dimensions of valence (i.e., pleasantness) and arousal, which are thought to underlie all emotional experience (Russell & Barrett, 1999). At the same time, in light of the relative novelty of applying natural language processing and machine learning to a within-person research question in psychology, we also endeavour to provide a realistic discussion of challenges associated with this method, including (i) sample non-representativeness; (ii) data sparsity; (iii) the criteria problem; and (iv) privacy concerns.

TRACKING WEEKLY FLUCTUATIONS IN VALENCE AND AROUSAL

The emotional dimensions of valence and arousal are fundamental to personality (Russell & Barrett, 1999). Extensive work has shown that people exhibit stable, trait-like individual differences in both their set point (i.e. typical levels) and variability for valence and arousal (e.g. Kuppens, Van Mechelen, Nezlek, Dossche, & Timmermans, 2007; Kuppens, Oravecz, & Tuerlinckx, 2010; see also Watson & Tellegen, 1985; Diener, Smith, & Fujita, 1995). Furthermore, much like other personality constructs, individual differences in valence and arousal have been shown to have implications for well-being (e.g. Houben, Van Den Noortgate, & Kuppens, 2015; Kuppens et al., 2010; Larsen & Diener, 1985) and to correlate with major personality dimension (e.g. the Big Five; Kuppens et al., 2007; 2010; Yik, Russell, & Steiger, 2011).

Yet modal methods for assessing valence and arousal suffer from several limitations in the context of dynamic, within-person research. Foremost among these is a heavy reliance on self-report, given that simply asking people to introspect about their feelings has the potential to alter those internal states (e.g. Kassam & Mendes, 2013; Lieberman, Inagaki, Tabibnia, & Crockett, 2011). Moreover, substantial burden falls on participants who are asked to self-report their feelings in an intensive, repeated manner; as a result, even studies employing state-of-the-art ESM protocols have a limited temporal scope (e.g. 4–10 assessments per day for 14 days; e.g. Kuppens et al., 2010; Sun et al., 2019). The field's methodological toolbox beyond self-report is equally fraught, often relying on simple word-counting methods to determine the sentiment (i.e. positivity) of speech or written text. These methods were not designed to measure dynamic changes in emotional states and can fail to track momentary feelings because (a) they rely on a relatively small fraction of the vocabulary used by people (often around a few per cent of word occurrences, thus losing statistical power in text samples with low word counts), (b) they disregard word context, and (c) they can be led astray by a small number of highly frequent words (Kring et al., 2019; Sun et al., 2019).

We aimed to move beyond self-report and fixed linguistic categories to examine naturally occurring expression of the fundamental dimensions of valence and arousal via a data-driven, big data analysis of digital traces. Specifically, we applied an unobtrusive method (i.e. we did not require user input) to track fluctuations in valence and arousal in a sample of 640 US Facebook users who post the most frequently within the my Personality Facebook data set of 65 000 users (Kosinski et al., 2013). However, even when considering these highly frequent social media posters, the temporal distribution of their status updates tends to be uneven across days. As a result, day-level time series are relatively sparse, which introduces difficulty in the interpretation of variability, as measurement intervals are spaced unevenly. We obtained much denser time series when aggregating data to the *weekly* level. We thus conceptualized and assessed emotion at this level of temporal aggregation, although feelings of valence and arousal are typically conceptualized and assessed on a *momentary* level (Kuppens et al., 2010; Russell & Barrett, 1999). When we talk about valence and arousal, we are therefore referring to people's *average tendency to feel pleasant (versus unpleasant) or aroused (versus calm) during a given week*. We can colloquially think of weekly valence and arousal as capturing whether a person is having a 'good week' (versus a 'bad week') in the sense that they tend to be feeling positive and/or upbeat (versus negative and/or low energy).

THE CURRENT CASE STUDY

We implemented a big data method to track weekly fluctuations in valence and arousal as outlined in Figure 1. First, we trained a predictive model using a *calibration sample*, following the steps developed in previous work (Preotiuc-Pietro et al., 2016). Specifically, in this previous work, two trained research assistants annotated 2895 public Facebook posts for valence and arousal on 9-point ordinal scales drawn from an age and gender-stratified sample of 2786 Facebook users (with a maximum of two statuses from each user). The annotations from both raters were averaged to yield a final estimate. The text of these Facebook statuses was then encoded as distributions of relative word, phrase, and topic frequencies using methods of natural language processing [specifically, emoticon-aware tokenization, phrase detection using a pointwise-mutual information criterion, and the extraction of 2000 previously modelled latent Dirichlet allocation topics and Linguistic Inquiry and Word Count (LIWC) 2007 dictionaries]. Dimensionality reduction (using principal component

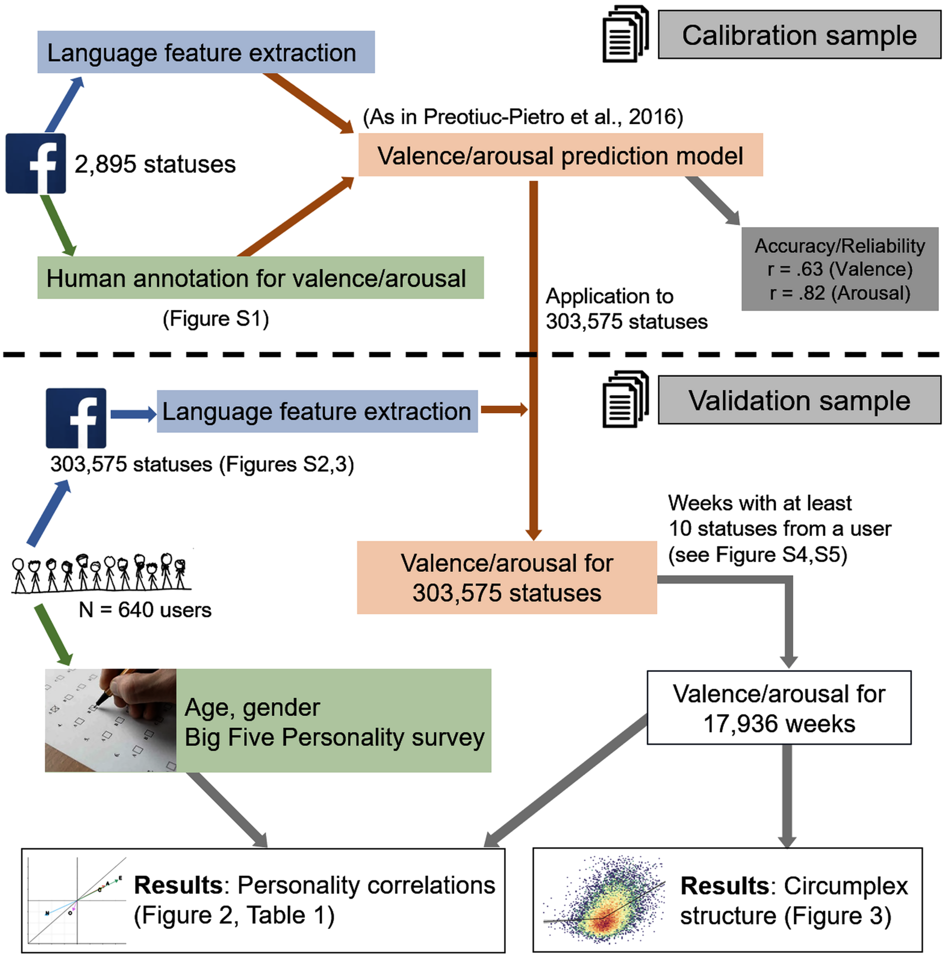


Figure 1. Analytic strategy and use of data across the two samples comprising this study. [Colour figure can be viewed at wileyonlinelibrary.com]

analysis) was then applied to these feature spaces, and penalized regression was used to create text-based prediction models for valence and arousal, which were cross-validated out of sample (see Preotiuc-Pietro et al., 2016, for additional detail on the model training process).

This approach and other approaches that use the whole vocabulary (e.g. ANEW or LabMT; Bradley & Lang, 1999; Mitchell, Frank, Harris, Dodds, & Danforth, 2013) have the advantage over approaches that count the words that occur in a single, limited dictionary (such as LIWC’s Positive Emotion dictionary). This is because these approaches use a majority of the observed vocabulary to estimate the valence and arousal of a Facebook status. This in turn increases the stability of the estimates and reduces the number of words needed to derive a meaningful estimate. For comparison, LIWC’s Positive Emotion dictionary matches roughly 5% of word occurrences, based on our analyses.

Second, we applied and evaluated this prediction model in a separate *validation sample*. We used data from the MyPersonality application, which contains status updates and survey-based personality scores of consenting Facebook users (Kosinski et al., 2013). We selected the 640 users (equal numbers of male and female) who had the highest word counts across their Facebook statuses (M = 14 492 [SD = 7585] words per user; in the succeeding text, we discuss the effect of this sampling strategy on the generalizability of our findings). We used the prediction model constructed in the *calibration sample* to derive valence and arousal estimates for the 303 575 Facebook statuses posted by users in our *validation sample*. That is, in the *validation sample*, we extracted the same word, phrase, topic, and dictionary features described earlier and applied the regression weights learned on the *calibration sample* to derive valence and arousal estimates for each of the 303 575 Facebook statuses. We averaged the resulting valence and arousal estimates of the Facebook statuses within weeks for each user, to yield weekly estimates of valence and arousal. Sensitivity analyses suggested that 10 messages per week and 14 weeks per user were needed to yield stable weekly and user-level valence and arousal estimates (both for means and standard deviations). These cut-offs yielded an average of 28 weeks per user, containing an average of nearly 17 messages per week (see Figure S3).

Third, we leveraged our predictive model to better understand the link between valence, arousal, and the Big Five. Prior work using ESM to estimate set point and variability in valence and arousal has shown that individuals with high extraversion, agreeableness, and conscientiousness tend to experience relatively pleasant and aroused affect, whereas individuals with high neuroticism tend to experience relatively

DOI: 10.1002/per

unpleasant affect and heightened variability in affect. This work has also shown that openness to experience tends to be only weakly related to valence and arousal (Kuppens et al., 2007; 2010; Yik et al., 2011). We examined whether the links between valence, arousal, and the Big Five would replicate or deviate from those observed in prior self-report studies, in the hope of shedding light on the robustness of these links across different levels of analysis and different data sources. Note that we did not preregister the exact hypothesized links between the Big Five and emotional dimensions.

METHODS

Calibration sample

Annotations

In previous work (Preotiuc-Pietro et al., 2016), a total of 2895 public Facebook posts were collected from 2786 unique users (maximum of two statuses per user; for the distribution of word frequencies, see Figure S1, and for a scatter plot of the ratings, see Figure S2). These posts were annotated for valence and arousal by two trained research assistants on 9-point ordinal scales [valence: 1 = 'negative' and 9 = 'positive'; arousal (which the prompt called 'intensity'): 1 = 'low' and 9 = 'high']. The research assistants received training with examples of posts that were high and low on these dimensions and annotated 120 training statuses that were checked against the annotations of a senior psychologist. We found the annotation quality to be adequate. Across the 2895 posts, the research assistants reached agreement correlations of r = .77 for valence and r = .83 for arousal. The two raters' annotations were averaged to yield final estimates. The statuses and their annotations can be obtained in anonymized form from the project's OSF repository (https://osf.io/pbjer).

Statuses on average were rated as expressing moderately pleasant mood (M = 5.26, SD = 1.19) and moderate levels of arousal (M = 3.35, SD = 1.98). Following Kuppens, Tuerlinckx, Russell, and Barrett (2013), we fit a series of models to the data, each of which represented a distinct possible link between valence and arousal. Specifically, these models differed based on (i) whether they represented arousal as orthogonal to valence, as a linear function of valence, or as a *v*-shaped function of valence; (ii) whether they included a positivity offset (i.e. different intercepts for positive and negative affect); and (iii) whether they included a positivity bias (i.e. different slopes for positive and negative affect; see the Supporting Information for model fitting details). We observed a *v*-shaped relation between arousal and valence, such that arousal increased as valence became both more positive and, to a lesser extent, more negative (see Tables S1 and S2, Figure S2, and the Supporting Information for full model fitting details). This relation showed a positivity bias, meaning that arousal increased more rapidly with increases in positive valence, b = 1.15, t(2891) = 24.60, $p < .001$, compared with negative valence, b = .56, t(2891) = 5.71, $p < .001$; interaction testing the difference: b = .59, t(2891) = 5.42, $p < .001$. The parameter testing for a negativity/positivity offset was not significant (b = −.23, t(2891) = 1.62, p = .11; see Table S1). This asymmetric, *v*-shaped relation between valence and arousal is similar to what has been previously observed in prior studies examining momentary emotion assessed through self-report (Kuppens et al., 2013). This concordance in the valence–arousal structural link across methods lends confidence to the validity of our annotations (although note that we did not preregister a hypothesis with respect to these structural links).

Model creation

Following the steps developed in previous work (Preotiuc-Pietro et al., 2016), for the 2895 human-annotated Facebook statuses in the calibration sample, we used the Differential Language Analysis ToolKit (DLATK; Schwartz et al., 2017; see dlatk.wwbp.org) to extract three sets of linguistic features: (i) the relative frequency of occurrences of words and phrases; (ii) 2000 latent Dirichlet allocation topics derived in previous work from 18 million Facebook status updates using the MALLET package (Schwartz et al., 2013[1]); and (iii) LIWC dictionaries (LIWC 2007; Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007). DLATK implements emoticon-aware tokenization (splitting of statuses into 'words').

As in previous work (Schwartz et al., 2013), we used a Pointwise-Mutual Information criterion to identify phrases, with a threshold of pointwise-mutual information >3 (i.e. we retained only phrases [e.g. 'happy birthday'] that were more than three times as likely to occur than the frequency of their underlying tokens ['happy' and 'birthday'] would entail by chance). We reduced the size of the combined word, phrase, topic, and LIWC dictionary feature space through a dimensionality reduction procedure, which combined univariate feature selection and principal component analysis, leaving a total of 1439 components for the prediction of valence and 675 for arousal, respectively. We next trained a ridge regression model to predict valence and arousal annotations based on the entire set of linguistic features and evaluated it using 10-fold cross-validation (i.e. the model is built on 90% of the data and then evaluated on the remaining 10% —which itself is not considered during the model building phase—to avoid overfitting). The cross-validated out-of-sample prediction accuracies of the model were r = .63 for valence and r = .82 for arousal, expressed as correlations between annotated and model-predicted valence and arousal values, and can be interpreted as reliabilities. In previous work, the model was compared against other standard measures of sentiment (such as ANEW, the Affective Norms for English Words) across the annotated statuses and found to provide more accurate estimates than these alternative measures (Preotiuc-Pietro et al., 2016).

Validation sample

Participants

[1] Available at a http://wwbp.org/downloads/public_data/wwbpFBtopics_freq.csv.

DOI: 10.1002/per

From the 65 000+ consenting Facebook users in the MyPersonality data set (Kosinski et al., 2013), we selected the 640 US Facebook users who had the largest word count across Facebook statuses (*M* = 14 492 words per user, *SD* = 7585), sampling an equal number of men and women (self-reported age: $M_{\text{age}} = 22.27$, *SD* = 5.86[2]). We determined the sample size such that (i) male and female users were balanced, (ii) the sample was as large as possible, and (iii) we had at least 14 weeks per user with at least 10 statuses for a given week (see data thresholds in the succeeding text). None of the participants who met these criteria were excluded from the subsequent analysis. Participants in our sample were younger on average than participants in the MyPersonality data set who were not included in our sample (excluded participants: $M_{\text{age}} = 26.28$, *SD* = 11.42; *d* = −.18, confidence interval = [−.66, .28][3]). This likely reflects the fact that we selected for inclusion participants who posted more, and these people tended to be younger. Participants included in our sample also scored somewhat higher on openness to experience (*d* = .25, confidence interval, CI = [.17, .33]) and neuroticism (*d* = .22, CI = [.14, .30]), somewhat lower on conscientiousness (*d* = −.22, CI = [−.30, −.14]), and slightly lower on extraversion (*d* = −.07, CI = [−.15, .00]) and agreeableness (*d* = −.06, CI = [−.14, −.01]). These standardized mean differences are in the ‘small’ range but nevertheless constitute a sample bias, a point to which we will return in the succeeding text. Finally, reflecting the fact that we selected participants with the aim of having a gender-balanced sample, the proportion of women in our sample (50%) was lower than the proportion of women in the MyPersonality data set (56%, $\chi^2(1) = 9.63$, $p < .01$).

Model application

We extracted the same language features as in the calibration sample (i.e. words and phrases, topics, and LIWC dictionaries) for the 303 575 Facebook statuses posted between 2 January 2009 and 24 November 2011, in our MyPersonality validation sample (see Figure S3 for temporal distribution of statuses). We then applied the same feature reduction steps as in the calibration sample and the same predictive model built on the calibration sample. This procedure yielded a predicted valence and arousal score for every Facebook status. We retained weeks that had a sufficient number of statuses to ensure reliable estimation of weekly emotion (see data thresholds in the succeeding text). We computed average valence and arousal within a given week to derive week-level set point estimates. User-level set point estimates of valence and arousal were similarly derived by averaging valence and arousal across weeks for a given user (again using a reliability threshold; see the succeeding text), and user-level variability estimates for valence and arousal were derived by computing the standard deviation of weekly valence and arousal across weeks for a given user. Predicted weekly valence was moderately positive on average (*M* = 5.11, *SD* = 0.32), as was predicted weekly arousal (*M* = 3.11, *SD* = 0.76), in line with prior work using ESM (Kuppens et al., 2007, 2010).

Of note, we chose to aggregate valence and arousal at the weekly level for pragmatic purposes. Weeks were the shortest possible unit of time that would ensure most observations to be consecutive—that is, of the 17 937 weekly data points included in this study, 12 705 (71%) are from consecutive weeks. Shorter units of temporal aggregation (e.g. days) or no aggregation (i.e. analysing each individual status update) would have resulted in widely differing windows between observations (ranging from minutes to multiple days), rendering it impossible to interpret variability in emotion in a theoretically meaningful manner.

Data thresholds/sensitivity analyses

We determined two thresholds to ensure reliable estimation of emotion while maximizing sample size: (i) the minimum *number of messages per week* to include a given week in the analysis and (ii) the minimum *number of weeks per user* to include a given user in the analysis. When setting a threshold for messages per week, too high of a threshold would result in fewer weeks to be included per user, truncating the length of the time series and yielding a noisier estimate for the overall user mean, whereas too low a threshold would result in unreliable estimates for a given week. Similarly, requiring too high a number of weeks per user would reduce the sample size, while too low a requirement would result in unstable estimates of the overall user mean and standard deviation. We applied a bootstrapping procedure to determine these thresholds, and we retained all data points which met these thresholds.

Messages per week. In the main study data set, we first limited the sample to statuses from 1489 weeks during which at least 30 messages were available, because we anticipated that 30 messages per week would yield a reliable estimate of weekly valence and arousal (Glass & Hopkins, 1984). We then averaged messages within each week to yield stable estimates for these weeks (i.e. a ‘ground truth’). We then randomly selected separate samples of size *k* = 2 to *k* = 29 statuses from these weeks (i.e. we selected a random sample of 2, a separate random sample of 3, and a separate random sample of 4). We averaged the valence and arousal estimates for all statuses, separately for each sample size (2 to 29), and correlated the resulting average valence and arousal estimates in each sample with the 30-message, ground truth weekly estimates (i.e. this yielded 28 correlations per user for both valence and arousal). We repeated the procedure 100 times, yielding one valence and one arousal average correlation for each weekly sample size as a measure of reliability (see Figure S4). We aimed to determine the minimum number of messages, such that both the valence and arousal average correlations exceeded .80 with the ground truth 30+ message estimates. This standard was met for *k* = 10 messages a week, which we determined to be our threshold for including a given week in the analysis.

[2] Across *N* = 615 users who self-reported a plausible age between 15 and 60. The age of users who reported ages above or below these thresholds were set to these thresholds yielding $M_{\text{age\_15 to 60}} = 22.48$ and $SD_{\text{age\_15 to 60}} = 7.08$ across all 640 users.

[3] Difference computed over all users who reported a plausible age between 15 and 60.

DOI: 10.1002/per

Weeks per user. Similarly, we sought to determine how many weeks were needed to generate a stable estimate of a user's overall valence and arousal average and across-week standard deviation. We again postulated 30 weeks with 10 messages per week (as determined in the previous step) to yield stable estimates (i.e. a 'ground truth') and, through an analogous bootstrapping procedure as earlier, determined 14 weeks per user to yield reliable estimates that were correlated at least .80 with the 'ground truth' estimates of valence and arousal averages and variability.

These procedures allowed us to choose thresholds that promised reasonable stability for our affect estimates. We included in our data set only users who had at least 14 weeks with at least 10 messages in them, yielding a data set with an average of 28 weeks per user, for a total of 17 937 weekly estimates, based on an average of 16.9 messages per week (see Figures S5 and S6 for the distribution of statuses and weeks across users).

Regression analyses

In the main analyses, when determining associations with Big Five traits, we adjusted for age and gender by entering them as covariates, and we report standardized regression coefficients (βs). We report associations without this adjustment in the Supporting Information (see Table S3; the results are largely unchanged). When reporting associations with variability, we control for mean levels, as variables with larger means can be expected to have larger variances. Mean levels, standard deviations, and intercorrelations for all variables are given in Table S4.

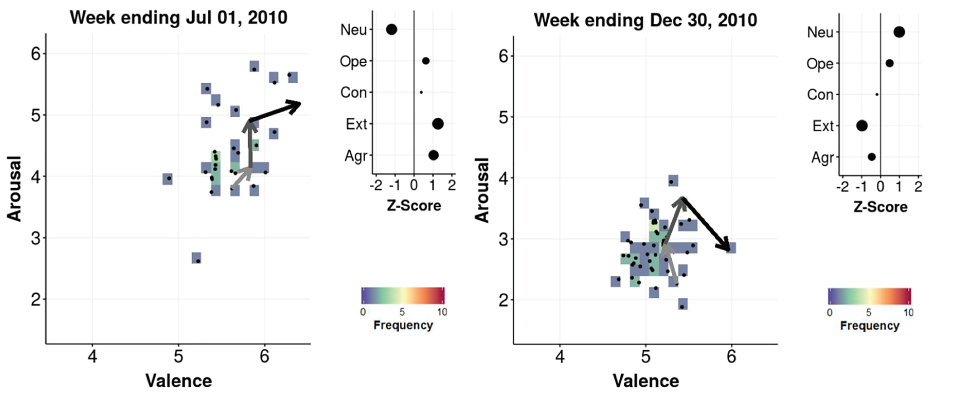
Data sharing

The de-identified data set of 17 937 weekly estimates of valence and arousal of 640 users and their demographic and personality information is publicly available on the Open Science Framework (https://osf.io/pbjer). Readers interested in learning more about the original MyPersonality data set can visit https://sites.google.com/michalkosinski.com/mypersonality. To our knowledge, the current *validation sample* is the largest data set of its kind describing within-person emotional trajectories (in terms of total number of temporal observations). We also share dynamic week-by-week animations for each user, as well as user-specific plots with regression lines across the negative and positive valence domains, to visualize positivity biases. This repository also contains syntax used to run our primary analyses, allowing others to reproduce our findings if they desire. We also reshare the annotated and anonymized 2895 Facebook posts from the *calibration sample* in the same repository.

RESULTS

Visualizing fluctuations in weekly emotion

We created dynamic visualizations of each user's model-predicted weekly fluctuations in emotional valence and arousal across the entire duration of the study (see https://osf.io/pbjer). For illustrative purposes, Figure 2 depicts two such weekly fluctuations in valence and arousal for both a woman (left) and a man (right), shown along with each user's Big Five personality profiles (see the Supporting Information for animations). These visualizations yield several apparent contrasts in each user's emotional experience. The user on the left shows fluctuations largely involving high-arousal, highly pleasant affect and rarely experiences a week with below-average arousal or below-neutral valence. In contrast, the user on the right shows fluctuations largely around average arousal and neutral valence and rarely experiences a week with high-arousal or highly pleasant emotion.



Note: Each visualization represents one user in our validation sample. A valence of 5 represents neutral valence on the 1-9 valence scale. One new point appears for each week indicated at the top of the plot; arrows connect points for adjacent weeks. Frequency refers to the number of weeks during which the user showed a given combination of valence and arousal. Each user's personality profile is shown in the top right. *Z*-scores for each trait are computed relative to the entire sample. *Left*: 21-year-old woman with elevated agreeableness, conscientiousness and extraversion. *Right*: 24-year-old man with elevated neuroticism and relatively low agreeableness, conscientiousness and extraversion. Animations for these users can be viewed in the supplementary materials and animations for all 640 users can be found at https://osf.io/pbjer.

Figure 2. Visualizations of weekly fluctuations in valence and arousal.

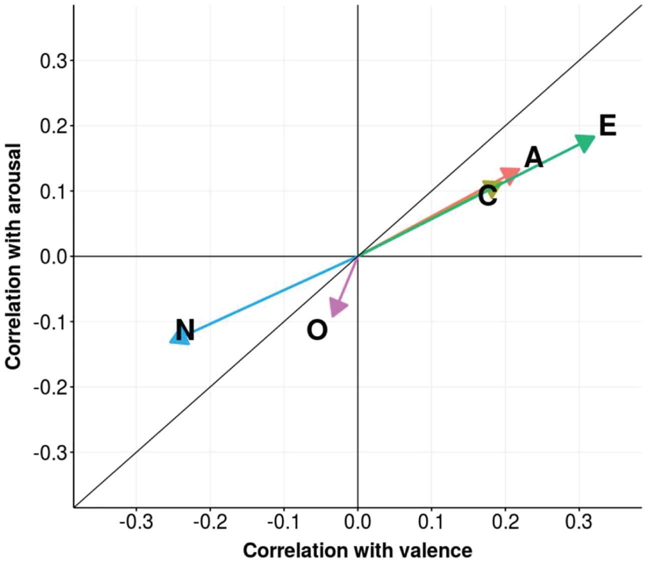
Big Five correlates of weekly valence and arousal

We next examined the relationships between individual differences in valence and arousal and the Big Five personality traits, with an eye towards making a comparison between these links and those that have emerged in prior ESM studies relying on self-report (Kuppens et al., 2007, 2010; Yik et al., 2011). Following prior work, we computed associations between both set point (i.e. average level) and variability (i.e. standard deviation) in users' weekly valence and arousal (computed across weeks for each user) and self-reported Big Five personality traits (completed as part of the MyPersonality application; see Table 2 and Figure 3). When examining links with valence and arousal variability, we controlled for each user's average level of valence and arousal, given that mean levels on any measure are confounded with that measure's variability (Baird, Le, & Lucas, 2006). Given the association of gender with valence and arousal mean levels and variability (*r*s = .14 to .22; see Table S4), we adjusted all estimates for both gender and age.

Individual differences in model-predicted weekly valence and arousal showed broadly consistent links with Big Five traits as compared to the links observed in prior work measuring momentary valence and arousal, particularly for set points (see Table 1 for a guide to the relationships we might predict). Extraversion, agreeableness, and conscientiousness showed moderate, positive associations with set points for weekly valence (*β*s = .19–.32, *p*s < .001), openness was relatively orthogonal to valence set point (*β* = −.03, *p* = .38), and neuroticism showed a negative link with valence set point (*β* = −.25, *p* < .001). Also, consistent with prior work, extraversion, agreeableness, and conscientiousness showed positive links with weekly arousal set point (*β*s = .11–.18, *p*s < .003). The link between arousal set point and openness was stronger than has been reported in prior studies but was still relatively weak (*r* = −.09, *p* = .018). Similarly, neuroticism was negatively associated with the arousal set point (*β* = −.13, *p* = .001), whereas this link has tended to be positive in most prior studies (although see Study 1 in Kuppens et al., 2007).

With respect to variability, although neuroticism showed a positive link with valence (but not arousal) variability when not controlling for age and gender (β = .08, p = .029), this link disappeared when controlling for age and gender (*β* = .01, *p* = .834). None of the other Big Five traits showed significant associations with valence or arousal variability (*β*s = −.03 to .04, *p*s > .313); prior work has shown a negative link between agreeableness and conscientiousness (but not extraversion or openness) and arousal variability (Kuppens et al., 2007). The overall lack of a relationship between variability in weekly valence and arousal and the Big Five traits in this study—particularly when controlling for age and gender—was surprising in light of prior findings and theoretical links between neuroticism and emotional variability; additional research is needed to investigate this issue further.

These results indicate that our algorithm portrayed extraverted, agreeable, and conscientious users as typically experiencing pleasant, moderately aroused weekly emotion and users high in openness as not showing a particularly distinctive emotional profile. In contrast, neurotic users were portrayed as typically experiencing unpleasant affect, yet at the same time, they tended to show greater fluctuations in how pleasant they felt from week to week. Although some of these relations were small compared with



Note: Vectors represent standardized coefficients between self-reported personality traits and user-level estimates of valence and arousal (aggregated across weeks for each user), adjusted for age and gender. Horizontal and vertical components of each vector convey the corresponding *β* associations with valence and arousal. See Table 2 for corresponding data with CIs.

Figure 3. Correlations of Big Five personality traits with weekly valence and arousal. [Colour figure can be viewed at wileyonlinelibrary.com]

DOI: 10.1002/per

Table 1. Expected relations between demographics, personality, and weekly valence and arousal based on prior work using experience sampling methodologies

	Mean levels		Variability	
	Valence	Arousal	Valence	Arousal
Demographics				
Age	+	0	−	0
Gender	++	+	+	0
Personality				
Agreeableness	++	+	−	−
Extraversion	++	+	0	0
Conscientiousness	+	+	−	−
Openness	0	0	0	0
Neuroticism	−	+	++	+

Note: 0 = relationship expected to be near zero. −/+ = relationship expected to be weak (<.20 and negative or positive, respectively). −−/++ = relationnship expected to be moderate (>.20 and negative or positive, respectively). For gender, '+' indicates higher values for women than men. Predictions based on Kuppens et al. (2007, 2010), Yik et al. (2011), Charles, Reynolds, and Gatz (2001), Gard and Kring (2007), Grossman and Wood (1993), Kring and Gordon (1998), LaFrance, Hecht, and Paluk (2003), and Röcke, Li, and Smith, 2009.

links observed in prior work using ESM (e.g. the association between neuroticism and valence variability was only β = .08 when not controlling for age and gender), the pattern of relations across the entire Big Five domain was largely in line with that found in prior work (Kuppens et al., 2007, 2010; Yik et al., 2011). Establishing the robustness of the links between emotional and Big Five dimensions across both self-report survey data (as in prior ESM studies) and naturally occurring linguistic data (as in the present study) is critical given that ecologically valid data are often viewed as a gold standard for personality psychology but in practice are relatively underutilized (Furr, 2009). For descriptive purposes, we also examined the link between weekly valence and arousal and both gender and age, reported as part of the MyPersonality data (see Table 2 and the Supporting Information).

Model validity check: Structure of valence and arousal

We next ran a validity check on the weekly valence and arousal estimates produced by our predictive model. Prior work has shown that the within-person relationship between self-reported, momentary valence and arousal typically emerges such that arousal increases sharply as valence becomes more positive and, to a far less extent, more negative. The result is typically an asymmetric, *v*-shaped link between valence and arousal or, in some cases, a positive, linear relation with a steeper slope for positive (versus negative) valence data points (Kuppens et al., 2013). We therefore tested whether the weekly valence and arousal estimates produced by our model exhibited a similar internal structure as has been previously observed in studies relying on ESM. Note that we conducted these analyses without preregistering a specific hypothesis as to the nature of the valence–arousal structural link.

Specifically, following Kuppens et al. (2013), and as in the calibration sample, we fit a series of models to the data, each of which represented a distinct hypothesized relation between valence and arousal. Model 1 represented arousal as orthogonal to valence, Model 2 represented arousal as a symmetric positive linear function of valence, Model 3 represented arousal as a symmetric *v*-shaped function of valence (i.e. by predicting arousal from the absolute value of valence), and Models 4–6 represented arousal as an asymmetric *v*-shaped function of valence. Specifically, Model 4 included a parameter allowing for the positive and negative valence slopes to have different intercepts (i.e. a positivity/negativity offset), Model 5 included a parameter allowing for the positive and negative slopes to have different steepness (i.e. a positivity/negativity bias), and Model 6 included both an offset and a bias parameter. Finally, Model 7 represented arousal as an asymmetric, positive linear

Table 2. Relations between demographics, personality, and weekly valence and arousal

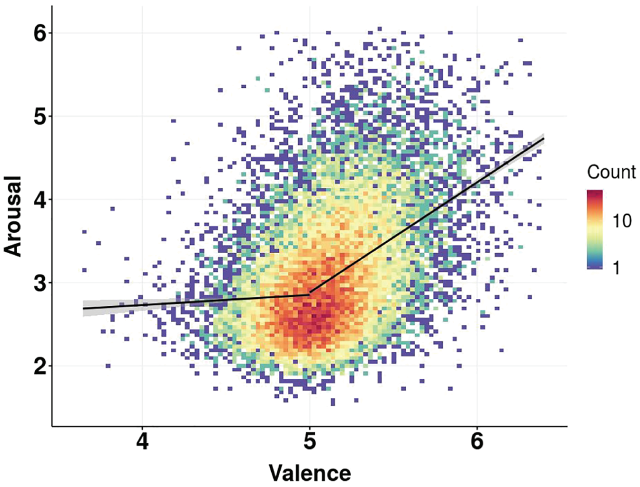
	Mean levels				Variability (controlled for mean levels)			
	Valence		Arousal		Valence		Arousal	
	β [95% CI]	p	β [95% CI]	p	β [95% CI]	p	β [95% CI]	p
Demographics								
Age	.07 [.07, .07]	.074	.04 [.03, .04]	.360	−.14 [−.14, −.14]	.000	−.05 [−.05, −.04]	.115
Gender	.22 [.19, .25]	.000	.23 [.14, .32]	.000	.19 [.18, 20]	.000	−.01 [−.04, .01]	.648
Personality								
Agreeableness	.22 [.19, .24]	.000	.13 [.06, .21]	.001	.04 [.03, .05]	.333	.03 [.01, .05]	.346
Extraversion	.32 [.30, .34]]	.000	.18 [.13, .24]	.000	.04 [.03, .05]	.313	−.02 [−.03, .00]	.571
Conscientiousness	.19 [.17, .22]	.000	.11 [.04. .18]	.003	−.02 [−.03, −.01]	.644	−.03 [−.04, −.01]	.380
Openness	−.03 [−.07, .00]	.383	−.09 [−.18, .00]	.018	−.03 [−.04, −.02]	.380	−.01 [−.03, .02]	.831
Neuroticism	−.25 [.27, −.23]	.000	−.13 [−.19, −.07]	.001	.01 [.00, .02]	.834	−.01 [−.01, .02]	.781

Note: Valence and arousal standardized regression coefficients for users' mean levels across weeks or standard deviation across weeks (variability). Personality regressions are adjusted for age and gender, the age regression for gender, and the gender regression for age. Coefficients for variability are also adjusted for mean levels.

DOI: 10.1002/per

function of valence, in that it modeled a linear relation but also included parameters for both positivity/negativity offset and positivity/negativity bias (see the Supporting Information for full model fitting details).

Our analyses provided evidence for an asymmetric, positive, linear relation (Figure 4; see Table S2 for full model fit details). A positive, linear relation emerged on average between valence and arousal (b = .53, z = 22.89, p < .001), yet this relation showed two asymmetries. First, predicted arousal increased rapidly with *increases* in positive valence (b = .63, z = 17.90, p < .001) and also increased slightly with *decreases* in negative valence (b = .28, z = 8.14, p < .001; interaction testing the difference: b = .35, z = 8.14, p < .001. The parameter testing for a negativity/positivity offset was small (b = .04, z = 3.18, p = .002). These relations are similar to those typically found in ESM studies that assess momentary valence and arousal via self-report (Kuppens et al., 2013), lending confidence to the validity of our predictive algorithm. Note, however, that the structural relation between valence and arousal merits a distinct interpretation at the weekly level compared with the momentary level. Our findings indicate that people tend to feel more active and aroused during more pleasant (versus unpleasant) weeks. The positivity bias indicates that during unpleasant weeks, this link is weak, meaning that a highly unpleasant week likely only involves slightly less active/aroused feelings than a less unpleasant week. In contrast, during pleasant weeks, this link is strong, meaning that a highly pleasant week is likely to involve much more active/aroused feelings than a slightly pleasant week.[4]



Note: Each point represents an estimate of valence and arousal, predicted by our language-based model, for one week for one user, across a total of 17,937 weeks.

Figure 4. Internal structure of weekly valence and arousal in validation sample, as predicted by Facebook language.

GENERAL DISCUSSION

In the present work, we present a case study in how big data methodologies, using natural language processing and machine learning-based estimation applied to digital traces, have the potential to help track *within-person fluctuations in dynamic personality states* unobtrusively across time and at scale. This is exactly the type of ideographic analysis that has increasingly become a primary focus of personality scientists (e.g. Fleeson & Jayawickreme, 2015; Hopwood, 2018; Jones et al., 2017; Sun et al., 2019). We deployed a feature extraction, reduction, and prediction model pipeline (developed in previous work; Preotiuc-Pietro et al., 2016) to track 640 Facebook users' weekly fluctuations in valence and arousal—or colloquially whether someone has been feeling pleasant and/or upbeat (versus unhappy and/or lethargic) during a given week—across an average of 28 weeks per user (17 937 weeks total). We conducted this analysis solely based on the text of these individuals' posted statuses and without reliance on self-report assessments. We then explored the links between model-predicted weekly emotion and self-reported Big Five traits, observing a similar pattern of links as has been observed in prior studies assessing momentary emotion via self-report, particularly for individuals' set points (i.e. mean levels) of valence and arousal. These findings lend confidence to the notion that big data analysis of digital traces can be used to gain valid insight into people's emotional experience, while also shedding light on the relations between aggregated weekly feelings and major personality trait dimensions.

Importantly, although we applied this method using a data set of affective personality states, the method itself could *in theory* be utilized with the goal of assessing nearly any personality-relevant state in a naturalistic, ecologically valid manner and with a large scope. First, with respect to ecological validity, big data analyses of digital footprints rely on data that are generated in the real-world through people's normal, everyday course of life. Capturing digital traces is therefore tantamount to capturing life *in vivo*, as it unfolds, and most importantly, without interference (i.e. no participant reported how they were feeling in the present study). This is a particularly important advantage when one wishes to assess subjective feelings, because the act of pausing one's day-to-day activities to introspect can alter how people report feeling or behaving (e.g. Kassam & Mendes, 2013; Lieberman et al., 2011). In the case of more behavioural traits, one could easily imagine that being forced to reflect on and report one's momentary personality (e.g. state extraversion or conscientiousness), particularly if done repeatedly as part of an ESM protocol, could cause people to change how they are behaving or at the very least could change their self-perceptions (e.g. Baird & Lucas, 2011).

Second, with respect to scope, big data methodologies can be employed on large samples and across many platforms, given that they are relatively cheap to apply. In theory, the method presented here could be used to track the emotion of any frequent Facebook poster, a population that includes billions across the world. More broadly, the method can be easily extended to track fluctuations in emotion or other transient psychological states across other text-based social media platforms (e.g. Twitter; Dodds et al., 2011; Eichstaedt

[4] The positivity offset indicates that people tend to feel marginally more aroused during weeks that are only slightly pleasant (versus slightly unpleasant), but we hesitate to overinterpret this finding given its small magnitude.

DOI: 10.1002/per

et al., 2015). In comparison, most traditional methods such as ESM and ILD require individual participants to be enrolled, compensated with credit or payment, and tracked over the following days or weeks, all of which increase the difficulty of recruiting extremely large samples.

Challenges and considerations associated with big data analyses of digital footprints

Although we have detailed what we view as the potential of big data analyses for personality psychologists, in practice, these methods present a set of logistical challenges and considerations quite distinct from those associated with other methods of choice for cataloguing dynamic, within-person personality processes via highly granular data (e.g. ESM and ILD). We discuss these in the succeeding text in the context of dynamic, within-person analyses relevant to personality psychologists.

Data representativeness

The high and increasing frequency with which users share autobiographical text on social media platforms means that over time, a tremendous number of data points may become available for any given user enrolled in a study. It is common that a data collection authorization provided by a consenting user (through a mechanism such as a Facebook app, as in the case of the myPersonality data set) generally allows for the retroactive collection of user content spanning years. For example, in the present study, for a 21-year-old woman, we derived estimates of valence and arousal for 81 weeks drawing on 1875 statuses. However, we drew on a sample of 640 high-volume social media users, meaning that the depth of data we observed for these individuals is not likely to be available for the majority of individuals in a given social media population, calling into question the representativeness of our sample.[5] On the one hand, we were encouraged to find only small differences between our sample and the entire MyPersonality sample on major personality characteristics (e.g. our participants were slightly more neurotic than those in the general population). On the other hand, the individuals who used the MyPersonality application—and therefore were included in the population from which we selected our current sample—may still differ from the general population (e.g. users tend to be younger; see MyPersonality.org). It is therefore important in the application of these big data samples to determine user demographics and in turn assess sample representativeness.

In contrast to big data methods, traditional methods such as ESM and ILD will never yield such a large amount of data for individual participants as we collected for our high-volume users (e.g. a highly intensive ESM protocol might sample participants 4–10 times per day for two weeks; e.g. Kuppens et al., 2010; Sun et al., 2019). Yet traditional methods typically yield a consistently high number of data points across the majority of participants in a given sample and, in theory, in a target population. Big data methods might therefore promise tremendous data depth for a select few individuals, whereas more traditional methods might promise adequate data depth for a large and more representatively sampled set of individuals.

Data sparsity

Social media users typically generate insufficiently dense digital traces to allow for estimation of within-person fluctuations in psychological states over short periods of time. To estimate how people feel over a given time period on the content of social media posts, one would need to accumulate a large number of posts, corresponding to all of the occasions at which the researcher hoped to assess a participant's feelings (e.g. to assess hourly fluctuations in emotion, one would likely need social media posts generated by each participant every hour across a time period of interest). For the vast majority of people, it would be unrealistic to expect this frequency of posting, particularly if it had to occur on a set schedule. Even a frequent poster who updates her Facebook status at breakfast (e.g. 7:45am), lunch (e.g. 12:30pm), and dinner (e.g. 7:00pm) will leave long gaps during which no digital trace is available. In the present work, this issue manifested even at the weekly level: The average user had sufficient Facebook posts to yield a reliable estimate of weekly valence and arousal in just over half of the weeks between his or her initial week and final week in the sample ($M = 52\%$, $SD = 21\%$).

In contrast, gold standard methods such as ESM and ILD typically yield a very dense set of data, albeit over a briefer window (e.g. 1–2 weeks). As a result, these methods would be better suited for assessing valence and arousal on a moment-to-moment basis (as is typically done; e.g. Kuppens et al., 2010; Russell & Barrett, 1999), whereas in the present work, we were only able to obtain reliable estimates of emotion at the weekly level. Conventional methods might therefore be more appropriate when researchers wish to examine dynamic shifts in personality states across short time intervals (e.g. minutes and hours), whereas big data analyses may be best suited for capturing people's tendency to enact a specific personality state across a longer time period over which otherwise sparse data can be aggregated.

The criterion problem

A major strength of big data analyses is that they often do not require temporally concurrent self-report assessments to be gathered alongside digital traces. Yet this also presents a unique challenge, in that researchers do not always have a clear 'ground truth' value for the feeling or behaviour meant to correspond to digital traces. In the present study, we did not have self-reports of weekly emotion against which to compare our algorithm's predictions. Instead, we relied on human-annotated ratings of the emotion expressed in a large set of Facebook status. We therefore had to assume that the words expressed on Facebook convey emotion, whereas they could have been motivated in part by self-presentational concerns that are relatively unrelated to current feelings of valence and arousal. Yet seminal work in personality

[5] However, in the years since the my Personality data set has been collected, we have observed in subsequent data collections that the volume of text shared on Facebook and Twitter (and recoverable through app-based data collections) has increased to the order of thousands of words for the average user.

psychology has demonstrated that social media profiles in fact provide reasonably accurate depictions of how people typically think, feel, and act, more so that they depict people's idealized self-views (e.g. Back et al., 2010). Based on these and other similar findings (Park et al., 2015), we feel confident that language expressed on social media in large part reflects *how people feel* more so than *how people wish others think they feel*.

In addition, given our reliance of annotations made by observers in our *calibration sample* training data, our models may have capitalized on linguistic cues that *are perceived as indicative of emotion by observers*, rather than detecting linguistic cues that *are indicative of emotion itself*. Our predictive model is therefore predicated on the reliability of these human annotations. This is why we took several steps to confirm the sensibility of the human-annotated emotion estimates, such as comparing the internal structure of model-predicted valence and arousal with the structure found in prior ESM studies (e.g. Kuppens et al., 2013). Unfortunately, this type of criteria problem would apply in all studies designed to track momentary personality states across time via digital traces without relying on self-report. Even if personality psychologists wished to use an alternative criteria in a big data context, such as observer reports of personality states, this presents challenges both logistically (e.g. collecting observer reports at scale would require considerable time and effort) and conceptually (e.g. observers do not always have accurate insight into people's intrapsychic states; e.g. Vazire, 2010).[6] Personality researchers hoping to conduct big data analyses of digital traces must at some level put their faith in the validity of the naturalistic traces themselves, unless future 'big data' study designs include an ESM component for at least a subset of the sample, to validate the unobtrusive methods against self-report. Of course, establishing predictive links between digital traces and other self-reported criteria (such as the Big Five analyses presented in this paper) helps strengthen our confidence in the validity of the digital traces themselves.

Privacy concerns

Big data analyses typically involve unobtrusively monitoring or collecting personal information (e.g. social media posts and smartphone data such as geolocation). It is therefore critical to ensure that participants have the explicit opportunity to provide informed consent after reading and understanding the degree to which researchers will access their personal information. Standard ethical guidelines for research with human subjects are typically followed by researchers conducting big data studies, including in the case of the myPersonality data set (Kosinski et al., 2013). However, in real-world applications, it may not always be clear to participants exactly what kind of sensitive information could be derived from the seemingly benign data they are sharing (e.g. Facebook statuses). Furthermore, outside of research contexts, participants may not be aware of whom their data may be sold to for marketing or other commercial purposes that may be disadvantageous to the user. Given the sensitive nature of big data analyses and the consent process, researchers still at times face understandable backlash over the potential of these methods to derive sensitive personal information (e.g. Kramer, Guillory, & Hancock, 2014; Wang & Kosinski, 2018). Sharing of Facebook data in particular has recently come under increased scrutiny in light of the potential data breach involving Cambridge Analytica (Granville, 2019). Of note, this event did not involve any myPersonality data, and to our knowledge, myPersonality data have not been exploited for non-research-related purposes.

Constraints on generality and future directions

The present work employed a sample of 640 heavily active Facebook users living in the United States. These 'super-users' were selected because they provided sufficiently frequent status updates to support the present work, but as a result, they are unlikely to be representative of all Americans. Furthermore, our use of a Western, individualistic cultural sample raises the question of whether the present findings would generalize to users from more collectivistic cultural contexts (Markus & Kitayama, 1991). Facebook users in more individualistic (versus collectivistic) cultures are known to engage in more self-oriented activities on Facebook, such as frequently broadcasting their present 'status updates' to people with whom they share close connections (Hong & Na, 2018; Na, Kosinski, & Stillwell, 2015). This type of individualistic Facebook use could engender emotional expression and/or disclosure, increasing the likelihood that Facebook status updates contain meaningful emotional information that is amenable to big data analyses. Furthermore, individualistic (versus collectivistic) cultures tend to place a higher value on feeling highly arousing, pleasant emotion, which can manifest in linguistic expressions (e.g. frequent use of the word 'great'; Tsai, 2007). These types of overt emotional expressions provide the very basis for inferring feelings of valence and arousal via Facebook posts. These issues raise the possibility that norms in collectivistic cultures could curtail outward emotional expression via social media in a manner that hinders our ability to detect emotional feelings via big data analyses. This possibility would be fascinating to test in future work. More broadly, this concern is a special case of the general consideration that machine learning algorithms tend to encode in their prediction models the presentation and sample biases present in their training data.

Another worthwhile avenue for future work would be to examine whether the present findings obtained via Facebook would replicate on different digital platforms, most notably Twitter, which has been the subject of much psychological inquiry in recent years (e.g. Dodds et al., 2011; Eichstaedt et al., 2015). Twitter is commonly thought to differ from Facebook on two key dimensions: (i) data density and (ii) disclosure intimacy. On one hand, Twitter is likely to provide more dense data, in the sense that the rapid-fire nature of Twitter conversations is likely to yield a larger number of messages that represent

[6] One could argue that, in the present study, we have used machine learning to scale observer report (albeit of statuses, not of people) to a large sample.

DOI: 10.1002/per

potentially usable data points when estimating personality states. In contrast, as noted throughout this paper, the present data set was very sparse, with hours or even days separating a user's Facebook posts. On the other hand, Twitter is commonly thought to provide less intimate disclosures than Facebook, given that people often use it for more information sharing as opposed to socializing (e.g. academic psychologists frequently share news related to the field on Twitter). However, we have seen across data collections from 2015 onward that younger cohorts may use Twitter as a messaging platform for highly personal information. Whether Twitter posts prove more or less useful for estimating users' personality or emotional states compared with more sparse but potentially more revealing Facebook statuses remains an open question for future research.

A third potential avenue for future work would be to harness additional digital traces from the present data set to more deeply explore markers of valence and arousal via social media. We focused exclusively on the words people used when estimating weekly emotion, but other meta-linguistic digital traces could prove useful. For example, we might expect that the time elapsed between Facebook posts could be indicative of a user's mood; in line with the notion that behaving extraverted engenders positive mood (e.g. Fleeson, Malanos, & Achille, 2002), users may be feeling more upbeat and/or aroused during times of high social media activity. In contrast, in the present data set, the types of prolonged periods of absence from social media that we treated as sparse/missing weeks may in fact meaningfully indicate that an individual is going through a period of low or unhappy affect. A time-series analysis that uses additional data streams (such as self-report) during times during which users 'go dark' on social media may shed light on these questions.

Conclusion

Harnessing technology to capture dynamic, within-person fluctuations in psychological states is increasingly a goal across personality science (e.g. Fleeson & Jayawickreme, 2015; Hopwood, 2018; Vazire & Sherman, 2017). With this goal in mind, we have presented a case study in a type of methodology that could help further this endeavour: Researchers could harness big data analyses to track the ups and downs of thousands of individuals across many weeks *in vivo*, using algorithms to analyse semi-public social media posts, authorized with a figurative click of the mouse or tap of the finger. We hope that the current research helps personality psychologists better understand and, ultimately, apply these methods in a manner that promotes a personality science grounded to a greater extent in naturally occurring digital traces.

AUTHOR CONTRIBUTIONS

J. C. E. and A. C. W. developed the study concept. J. C. E. performed data analysis, with assistance from A. C. W. Both authors wrote portions of the manuscript, contributed critical revisions, and approved the manuscript for submission.

ACKNOWLEDGEMENTS

The authors would like to thank Ryan L. Boyd, Peter Kuppens, and Lyle. H. Ungar for their constructive feedback on prior versions of this manuscript.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1: Model fit statistics for relation between valence and arousal in calibration sample
Table S2: Model fit statistics for relation between valence and arousal in validation sample
Figure S1: Histogram of words per Facebook status in the calibration sample.
Figure S2: Valence as a function of arousal in the calibration sample.
Figure S3: Temporal distribution of weeks included in the validation sample.
Figure S4: Stability of valence and arousal mean and standard deviations as a function of different thresholds.
Figure S5: Histogram of Facebook statuses per user included in the validation sample.
Figure S6: Histogram of weeks per user included in the validation sample.
Figure S7. User-level time series in validation data set.
Table S3: Relations between demographics, personality and weekly valence and arousal with no age or gender controls.
Table S4: Descriptive statistics for primary variables in validation sample.
Table S5. Average autocorrelations for lags of 1 to 7 weeks for valence and arousal in validation data set.
Table S6. Associations of lag 1 autocorrelation coefficients across users in validation data set.

REFERENCES

Allemand, M., & Hill, P. L. (2019). Future time perspective and gratitude in daily life: A micro-longitudinal study. *European Journal of Personality*, *33*, 385–399.

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmuckle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological Science*, *21*, 372–374. https://doi.org/10.1177/0956797609360756.

Baird, B. M., Le, K., & Lucas, R. E. (2006). On the nature of intraindividual personality variability: Reliability, validity, and associations with well-being. *Journal of Personality and Social Psychology*, *90*, 512–527. https://doi.org/10.1037/0022-3514.90.3.512.

Baird, B. M., & Lucas, R. E. (2011). "... And how about now?": Effects of item redundancy on contextualized self-reports of personality. *Journal of Personality*, *79*, 1081–1112. https://doi.org/10.1111/j.1467-6494.2011.00716.x.

Bradley, M. M., & Lang, P. J. (1999). *Affective Norms for English Words (ANEW): Instruction manual and affective ratings* (Vol. *30*, no. 1, 25-36). Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Charles, S. T., Reynolds, C. A., & Gatz, M. (2001). Age-related differences and change in positive and negative affect over 23 years. *Journal of Personality and Social Psychology*, *80*, 136-151.

Conner, T. S., Tennen, H., Fleeson, W., & Barrett, L. F. (2009). Experience sampling methods: A modern idiographic approach to personality research. *Social and Personality Psychology Compass*, *3*, 292–313. https://doi.org/10.1111/j.1751-9004.2009.00170.x.

Diener, E., Smith, H., & Fujita, F. (1995). The personality structure of affect. *Journal of Personality and Social Psychology*, *69*, 130–141.

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*, e26752. https://doi.org/10.1371/journal.pone.0026752.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., … Seligman, M. E. P. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, *26*, 159–169. https://doi.org/10.1177/0956797614557867.

Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*, 1011–1027. https://doi.org/10.1037/0022-3514.80.6.1011.

Fleeson, W., & Jayawickreme, E. (2015). Whole trait theory. *Journal of Research in Personality*, *56*, 82–92. https://doi.org/10.1016/j.jrp.2014.10.009.

Fleeson, W., Malanos, A. B., & Achille, N. M. (2002). An intraindividual process approach to the relationship between extraversion and positive affect: Is acting extraverted as "good" as being extraverted? *Journal of Personality and Social Psychology*, *83*, 1409–1422.

Furr, R. M. (2009). Personality psychology as a truly behavioral science. *European Journal of Personality*, *23*, 369–401.

Gard, M. G., & Kring, A. M. (2007). Sex differences in the time course of emotion. *Emotion*, *7*, 429-43.

Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in education and psychology* (3rd ed.). New York, NY: Allyn & Bacon.

Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across several cultures. *Science*, *333*, 1878–1881. https://doi.org/10.1126/science.1202775.

Granville, K. (2019). *Facebook and Cambridge Analytica: What you need to know as the fallout widens*. *New York Times*. Retrieved from https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html.

Grossman, M., & Wood, W. (1993). Sex differences in intensity of emotional experience: A social role interpretation. *Journal of Personality and Social Psychology*, *65*, 1010-1022.

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science*, *11*, 838–854. https://doi.org/10.1177/1745691616650285.

Hong, S., & Na, J. (2018). How Facebook is perceived and used by people across cultures: The implications of cultural differences in the use of Facebook. *Social Psychological and Personality Science*, *9*, 435–443.

Hopwood, C. J. (2018). Interpersonal dynamics in personality and personality disorders. *European Journal of Personality*, *32*, 499–524.

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*, *141*, 901–930. https://doi.org/10.1037/a0038822.

Jayawickreme, E., Tsukayama, E., & Kashdan, T. B. (2017). Examining the effect of affect on life satisfaction judgments: A within-person perspective. *Journal of Research in Personality*, *68*, 32–37.

Jones, A. B., Brown, N. A., Serfass, D. G., & Sherman, R. A. (2017). Personality and density distributions of behavior, emotions, and situations. *Journal of Research in Personality*, *69*, 225–236.

Kassam, K. S., & Mendes, W. B. (2013). The effects of measuring emotion: Physiological reactions to emotional situations depend on whether someone is asking. *PLoS ONE*, *8*, e64959. https://doi.org/10.1371/journal.pone.0064959.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 5802–5805.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*, 8788–8790.

Kring, A. M. & Gordon, A. H. (1998). Sex differences in emotion: Expression, experience, and physiology. *Journal of Personality and Social Psychology*, *74*, 686-703.

Kring, A. M., Kross, E., Verduyn, P., Boyer, M., Drake, B., Gainsburg, I., … Jonides, J. (2019). *Does counting emotion words on online social networks provide a window into people's subjective experience of emotion?* A case study on Facebook. Emotion.

Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: Accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology*, *99*, 1042–1060. https://doi.org/10.1037/a0020962.

Kuppens, P., Tuerlinckx, F., Russell, J. A., & Barrett, L. F. (2013). The relation between valence and arousal in subjective experience. *Psychological Bulletin*, *139*, 917–940. https://doi.org/10.1037/a0030811.

Kuppens, P., Van Mechelen, I., Nezlek, J. B., Dossche, D., & Timmermans, T. (2007). Individual differences in core affect variability and their relationship to personality and psychological adjustment. *Emotion*, *7*, 262–274. https://doi.org/10.1037/1528-3542.7.2.262.

LaFrance, M., Hecht, M. A., & Paluck, E. L. (2003). The contingent smile: a meta-lanalysis of sex differences in smiling. *Psychological Bulletin*, *129*, 305-334.

Larsen, R. J., & Diener, E. (1985). A multitrait–multimethod examination of affect structure: Hedonic level and emotional intensity. *Personality and Individual Differences*, *6*, 631–636.

Lieberman, M. D., Inagaki, T. K., Tabibnia, G., & Crockett, M. J. (2011). Subjective responses to emotional stimuli during labeling, reappraisal, and distraction. *Emotion*, *11*, 468–480. https://doi.org/10.1037/a0023503.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*, 224–253.

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, *8*, e64417. https://doi.org/10.1371/journal.pone.0064417.

Na, J., Kosinski, M., & Stillwell, D. J. (2015). When a new tool is introduced in different cultural contexts: Individualism–collectivism and social network on Facebook. *Journal of Cross-Cultural Psychology*, *46*, 355–370.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., et al. (2015). Accurate personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*, 934–952.

DOI: 10.1002/per

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007*. Austin, TX: LIWC.net.

Preotiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., & Shulman, E. P. (2016). *Modelling valence and arousal in Facebook posts*. Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA), NAACL.

Röcke, C., Li, S. C., & Smith, J. (2009). Intraindividual variability in positive and negative affect over 45 days: Do older adults fluctuate less than young adults?. *Psychology and Aging*, *24*, 863.

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called *emotion*: Dissecting the elephant. *Journal of Personality and Social Psychology*, *76*, 805–819. https://doi.org/10.1037//0022-3514.76.5.805.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Agrawal, M., Park, G. J., … & Lucas, R. E. (2013). *Characterizing geographic variation in well-being using tweets*. *In Seventh International AAAI Conference on Weblogs and Social Media (ICWSM)*. Boston, MA.

Schwartz, H. A., Giorgi, S., Sap, M., Crutchley, P., Ungar, L., & Eichstaedt, J. (2017). *DLATK: Differential Language Analysis ToolKit*. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 55-60.

Sened, H., Lazarus, G., Gleason, M. E., Rafaeli, E., & Fleeson, W. (2018). The use of intensive longitudinal methods in explanatory personality research. *European Journal of Personality*, *32*, 269–285.

Sun, J., Schwartz, H. A., Son, Y., Kern, M. L., & Vazire, S. (2019). The language of well-being: Tracking fluctuations in emotion experience through everyday speech. *Journal of Personality and Social Psychology*, *118*, 364–387.

Tsai, J. L. (2007). Ideal affect: Cultural causes and behavioral consequences. *Perspectives on Psychological Science*, *2*, 242–259. https://doi.org/10.1111/j.1745-6916.2007.00043.x.

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*, 281-300.

Vazire, S., & Sherman, R. A. (2017). Introduction to the special issue on within-person variability in personality. *Journal of Research in Personality*, *69*, 1–3.

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, *114*, 246–257. https://doi.org/10.1037/pspa0000098.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin*, *98*, 219–235. https://doi.org/10.1037//0033-2909.98.2.219.

Yik, M., Russell, J. A., & Steiger, J. H. (2011). A 12-point circumplex structure of core affect. *Emotion*, *11*, 705–731. https://doi.org/10.1037/a0023980.

DOI: 10.1002/per