Pre-registration Protocol: Smartphone Sensing Panel Study - Affect Experience in Everyday Language Logged with Smartphones

This pre-registration protocol deals with specific research questions and is completed before the respective analyses have been conducted. Throughout this registration, we will refer to the corresponding basic registration protocol of the panel study. The basic protocol contains information on study procedures and further background information and can be found in the general pre-registration template here: http://dx.doi.org/10.23668/psycharchives.2901.

Working Title
Affect Experience in Everyday Language Logged with Smartphones

Author(s) of the preregistration protocol
Timo Koch, Johannes Eichstaedt, Clemens Stachl

Date
February 14, 2022

Background

Background Information (Optional; Short description of the theoretical background/introduction to research question)
Analyzing language offers a unique window into the inner workings of the human mind. Particularly, the methodological advancements in text processing in recent years and the ubiquity of textual digital trace data have generated opportunities to investigate psychological

constructs, such as affective states, through language in the form of text (Boyd & Schwartz, 2021; Jackson et al., 2021). In this manner, prior studies have predicted affective states from text data, for example social media posts (Eichstaedt & Weidman, 2020). Inferring affective states from language footprints instead of having to administer self-report questionnaires may yield scalable applications, for example, in monitoring the mental well-being of individuals or entire communities (Eichstaedt et al., 2018; Jaidka et al., 2020). Most prior studies relied on human annotations or on establishing sentiment through lexica (e.g., LIWC, VADER) to infer users' affect from social media language samples (e.g., Facebook status updates) – in part because researchers could not assess users' subjective affect experience in the moment they had produced the text. However, affective word usage and human judges' rating conceptually differ from one's subjective affect experience (Kross et al., 2019; Sun et al., 2020). Further, relying on social media text data for timely affect inferences also has drawbacks: First, these social media posts represent only a limited snapshot of one's overall everyday language use. As a consequence, the information stream in social media text has many time gaps (i.e., when people do not post). Second, social media language might be affected by people trying to manage their public image resulting in lower levels of self-disclosure (compared to, for example, private messaging), which possibly make inferences less accurate (Bazarova et al., 2012). Here, phone-level data collection methods offer a promising opportunity to passively log textual data across communication channels (public and private contexts) through the smartphone's keyboard and couple the data with in-situ self-reports on one's affect through experience sampling. Thereby, we hypothesize that we can predict and gain insights into between-person differences in subjective affect experience and respective within-person fluctuations.

Research question(s)

In this work, we want to investigate (in-sample) associations of in-situ self-reported affective states with language features logged with smartphones in everyday life and if these features allow for the (out-of-sample) prediction of between-person differences and within-person fluctuations in affect experience. Further, we want to investigate which language features are most predictive of subjective affect experience. Finally, we want to analyze what the optimal time window is for text analysis (and corresponding amount of text data) around the timestamp of the affective state in question and how affect experience is revealed in different contexts (e.g., public posting vs. private messaging).

Hypotheses
Please provide hypothesis for predicted results. If multiple hypotheses, uniquely number them (e.g. H1, H2a, H2b,) and refer to them the same way at other points in the registration document and in the manuscript.

First, we hypothesize that between-person differences in self-reported affect experience on the dimensions of valence and arousal are predictable beyond chance from everyday language logged with smartphones.
Second, we hypothesize that within-person fluctuations in self-reported affect experience on the dimensions of valence and arousal are predictable beyond chance from everyday language logged with smartphones.
Third, we hypothesize that our predictions (within- and between-person) are more accurate for private communications contexts (e.g., messages in WhatsApp) compared to public contexts (e.g., posts on Facebook).

Variables

Which variables will be used? (see Variables in the basic protocol for an extensive overview of all available variables)
This section shall be used to unambiguously clarify which variables are used to operationalize the specified hypotheses. Please (a) list all variables that will be used in this study and (b) explicitly state the functional role of each variable (i.e., independent variable, dependent variable, covariate, mediator, moderator). It is important to (c) specify for each hypothesis how it is operationalized, i.e., which variables will be used to test the respective hypothesis and how the hypothesis will be operationally defined in terms of these variables. This section is closely related to the statistical models used to test the hypotheses.

Data collection for this work was part of a six-month panel study using the PhoneStudy research app at Ludwig-Maximilian-Universität München (LMU) from May until November 2020 (for more details see http://dx.doi.org/10.23668/psycharchives.2901). All data collection procedures were approved by the ethics board at LMU.
The study also comprised two two-week experience sampling phases (July 27, 2020 to August 09, 2020; September 21, 2020 to October 04, 2020) during which participants received two to four short questionnaires per day. Here, self-reported experience of affective states was assessed. We assessed affective states based on the Circumplex Model of Affect, which suggests that affective states can be mapped onto a space with the two dimensions of valence (i.e., pleasure) and arousal (i.e., physical and psychological activation) (Russell et al., 1989). Valence and arousal were assessed in two separate items on six-point Likert scales among other psychological properties.
We applied a privacy-respecting language logging method to collect data of participants' everyday language use. That is, our research app accessed the smartphone's Android

accessibility services to log participants' text input. The captured keyboard input was directly summarized on the participant's device (that is, words were matched against dictionaries and only the dictionary frequency stored and transmitted – but not the words themselves; emoji and emoticons were logged in raw format) and only the summary data were sent to our research server. Further, for each text input, we stored metadata containing a timestamp, the app name, and the hint text of the input field provided by the app (e.g., "message" on WhatsApp).
At the moment of pre-registration, the data has already been collected and accessed. We have run quality checks to ensure correct data storage was achieved and started to process the data, but no descriptive or predictive analyses for this project have been conducted yet.

Analysis Plan

Preprocessing

Inclusion criteria (e.g., criteria for including (1) participants (e.g., Do you only use a subsample?, (2) study days (e.g., only weekdays, certain number of study days), (3) any other criteria concerning data quality (e.g., only days with at least x% of logging data) etc. If you cannot specify these aspects now, please state why.

We will exclude experience samples, where participants did not provide information on valence and arousal. Further, we will exclude participants with less than five days with at least one completed experience sampling questionnaire and, if there is reason to believe that participants did not fill out the affect items thoroughly (e.g., no variance in affect responses across all questionnaires indicating straightlining). Further, we exclude data from experience sampling instances if no text input had been made in the respective time window (e.g., 1 hour, 3 hours) around the experience sampling instance.

Definition of variables based on smartphone sensing. Please specify your degrees of freedom in variable extraction procedures, e.g.,

- *time information (e.g., what does night, daily, weekend exactly mean?)*
- *Aggregation measures (e.g., measures of central tendency/dispersion).*

If you cannot specify these aspects now, please state why.

We will extract four feature groups from the privacy-respectfully logged language data and aggregate features across different time windows (e.g., 1 hour, …, 3 hours, …) around the time stamp of the respective experience sampling instance:

- ***Meta-data***: e.g., number of text inputs across apps and contexts, number of words & characters & emoji/emoticons in total and avg. per text input and their standard deviation, avg. length per app, share of words that had been added/ changed/ removed
- ***Word dictionaries (LIWC)***: Percentage of LIWC2015 word dictionaries from all logged words in that time frame
- ***Word sentiment + emoji sentiment***: Sentiment score from text using SentiWS (Remus et al., 2010) and emoji sentiment score (Kralj Novak et al., 2015) from emoji
- ***Preferences for specific emoji and emoticons***: e.g., share of specific emoji and emoticons from all emoji/ emoticons used in that time frame

Further preprocessing steps (e.g., transformation of data, handling of missing data/outliers etc.)

-

Data Analysis

Statistical models
Please specify the statistical model (e.g. t-test, ANOVA, LMM) or algorithms that will be used to test each of your hypotheses. Give all necessary information about model specification (e.g., variables, interactions, planned contrasts) and follow-up analyses. Include model selection criteria (e.g., fit indices), corrections for multiple testing, and tests for statistical violations, if applicable. Please also indicate Inference Criteria (e.g., p-values, effect sizes, performance measures etc.).

After reporting descriptive correlations of language features and self-reports on affect experience, we will train multiple machine learning models for the prediction of raw self-reported valence and arousal (between-person differences) and fluctuations from participant's individual affect baseline (within-person differences; affect baseline is the median valence/ arousal score across experience samples in the study period for a given participant). We will

use language features extracted from keyboard logs as predictor sets to train separate machine learning models and compare their predictive performance within one single benchmark experiment for each outcome variable (i.e., valence and arousal).
We plan to compare the predictive performance of multiple regression (and potentially also classification) algorithms, for example, Elastic Net regularized regression models (Zou & Hastie, 2005), non-linear tree-based Random Forest models (Breiman, 2001), and a baseline model, which would predict the mean value from the training set for all cases in the test set.
We will tune model hyperparameters in a nested cross-validation scheme and evaluate the predictive performance of our models. To prevent overlaps between training and test data, we will block participants in the resampling procedure ensuring that for one train/test set pair all data points of a given participant are either in the training set or in the test set. Further, we plan to run separate analyses on data from private contexts (e.g., messaging on WhatsApp) and public contexts (e.g., posting on social media) to compare their predictive value.
Our prediction models will be evaluated based on how accurate new (unseen) samples can be predicted in comparison to a baseline model. Regression model fit will be evaluated based on multiple statistical parameters, for example, Pearson/Spearman correlation and mean absolute error (*MAE*).
Further, we plan to run variance-corrected significance tests to determine if we can predict valence and arousal significantly above baseline levels (for example, from Nadeau & Bengio, 2003).
Further, we want to understand our prediction models. Therefore, for (regularized) regression models, we will investigate features' regression weights. For Random Forest models, we aim to compute feature importance measures for features in order to investigate which language features are predictive of the experience of affective states and, for example, accumulated local effects (ALE) plots and/ or partial dependence plots (PDP) in order to get insights into the direction of feature effects. Finally, we will analyze residuals in different value areas of valence and arousal, testing if the models are more accurate for certain values. Further, we will run sensitivity analyses investigating the impact of the temporal distance of the text inputs from the experience sampling instance and the corresponding amount of text data available (e.g., number of words) on prediction performance of our models.

Planned exploratory analysis (Optional)

-

References

Bazarova, N. N., Taft, J. G., Choi, Y. H., & Cosley, D. (2013). Managing impressions and relationships on Facebook: Self-presentational and relational concerns revealed through the analysis of language style. *Journal of Language and Social Psychology*, *32*(2), 121-141.

Boyd, R. L., & Schwartz, H. A. (2021). Natural Language Analysis and the Psychology of Verbal Behavior: The Past, Present, and Future States of the Field. *Journal of Language and Social Psychology, 40*(1), 21–41.

Breiman, L. (2001). Random forests. *Machine learning, 45(1)*, 5-32.

Eichstaedt, J. C., & Weidman, A. (2020). Tracking fluctuations in psychological states: a case study of weekly emotion using social media language. *European Journal of Personality.*

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., ... & Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, *115*(44), 11203-11208.

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.

Jackson, J. C., Watts, J., List, J.-M., Puryear, C., Drabble, R., & Lindquist, K. A. (2021). From Text to Thought: How Analyzing Language Can Advance Psychological Science. *Perspectives on Psychological Science*. https://doi.org/10.1177/17456916211004899

Kralj Novak, P., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of emojis. *PloS one*, *10*(12), e0144296.

Kross, E., Verduyn, P., Boyer, M., Drake, B., Gainsburg, I., Vickers, B., ... & 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, 19*(1), 97.

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine learning*, *52*(3), 239-281.

Remus, R., Quasthoff, U., & Heyer, G. (2010, May). SentiWS-A Publicly Available German-language Resource for Sentiment Analysis. In *LREC*.

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect grid: a single-item scale of pleasure and arousal. *Journal of personality and social psychology*, *57*(3), 493.

Sun, J., Schwartz, H. A., Son, Y., Kern, M. L., & Vazire, S. (2020). The language of well-being: Tracking fluctuations in emotion experience through everyday speech. *Journal of Personality and Social Psychology, 118(2)*, 364.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology), 67(2)*, 301-320