Journal of Affective Disorders 302 (2022) 7–14

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

The relationship between text message sentiment and self-reported depression

Tony Liu [a,*], Jonah Meyerhoff [b], Johannes C. Eichstaedt [c], Chris J. Karr [d], Susan M. Kaiser [b], Konrad P. Kording [e], David C. Mohr [b], Lyle H. Ungar [a]

[a] Department of Computer and Information Science, University of Pennsylvania, USA
[b] Center for Behavioral Intervention Technologies (CBITs), Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, USA
[c] Department of Psychology, Stanford University, USA
[d] Audacious Software, USA
[e] Department of Bioengineering, Department of Neuroscience, University of Pennsylvania, USA

ARTICLE INFO

Keywords:
Depression
Machine learning
Language sentiment analysis
Digital phenotyping
Personal sensing

ABSTRACT

Background: Personal sensing has shown promise for detecting behavioral correlates of depression, but there is little work examining personal sensing of cognitive and affective states. Digital language, particularly through personal text messages, is one source that can measure these markers.
Methods: We correlated privacy-preserving sentiment analysis of text messages with self-reported depression symptom severity. We enrolled 219 U.S. adults in a 16 week longitudinal observational study. Participants installed a personal sensing app on their phones, which administered self-report PHQ-8 assessments of their depression severity, collected phone sensor data, and computed anonymized language sentiment scores from their text messages. We also trained machine learning models for predicting end-of-study self-reported depression status using on blocks of phone sensor and text features.
Results: In correlation analyses, we find that degrees of depression, emotional, and personal pronoun language categories correlate most strongly with self-reported depression, validating prior literature. Our classification models which predict binary depression status achieve a leave-one-out AUC of 0.72 when only considering text features and 0.76 when combining text with other networked smartphone sensors.
Limitations: Participants were recruited from a panel that over-represented women, caucasians, and individuals with self-reported depression at baseline. As language use differs across demographic factors, generalizability beyond this population may be limited. The study period also coincided with the initial COVID-19 outbreak in the United States, which may have affected smartphone sensor data quality.
Conclusions: Effective depression prediction through text message sentiment, especially when combined with other personal sensors, could enable comprehensive mental health monitoring and intervention.

1. Introduction

Major depression is a common and debilitating mental health disorder affecting more than 7% of the US population in any given year, resulting in significant impairments in work and social functioning, increased risk of suicide, decreased health, and high costs (Greenberg et al., 2015; Otte et al., 2016). Depression is a complex disorder (Fried and Nesse, 2015), with genetic, biological, environmental, cognitive, and behavioral etiologies (Otte et al., 2016). Due to the heterogeneous symptom combinations that comprise the diagnostic criteria for major depressive disorder, measurement of latent depression is a longstanding challenge in the field of psychology. Most depression assessments, whether self-reports or clinical interviews, relied on retrospectively recalled cognitions, behaviors, and physical symptoms (Fried, 2017). Retrospective reports are subject to a number of biases including the recency and availability heuristic among others that can impede the accurate measurement of depression severity (Tversky and Kahneman, 1973). Furthermore, monitoring depression over time can be burdensome; clinicians report that they do not use symptom measures because they are too time-consuming, and patients often do not complete

* Corresponding author.
E-mail address: liutony@seas.upenn.edu (T. Liu).

https://doi.org/10.1016/j.jad.2021.12.048
Received 22 June 2021; Received in revised form 15 November 2021; Accepted 18 December 2021
Available online 25 December 2021

remotely administered measures (Zimmerman and McGlinchey, 2008).

Passive personal sensing (Mohr et al., 2020) methods that use data from networked sensors in ubiquitous devices such as smartphones open new opportunities to measure cognitive and behavioral constructs, with recent innovations in personal sensing enabling the detection of behavioral correlates of depression passively (physical activity, movement through geographic space, smartphone use) (Insel, 2017; Liao et al., 2019; Saeb et al., 2016; Torous et al., 2015; Zulueta et al., 2018). Sensed data streams can be leveraged to identify trends associated with worsening mental health symptoms, such as reduced movement or activity (Liao et al., 2019; Saeb et al., 2016), or changing rates of keyboard input (Zulueta et al., 2018), and deliver effective, targeted, just-in-time interventions that are personalized based on passively sensed data streams (Torous et al., 2015). Because of unobtrusive data collection and targeted interventions, personal sensing has the potential to fit into the context of individuals' everyday lives (Huckvale et al., 2019b; Insel, 2017, 2018; Marsch, 2018; Onnela and Rauch, 2016; Torous et al., 2015).

One promising passive data stream is digital language, with recent studies demonstrating the successful use of social media language to predict depression (Choudhury et al., 2013; Eichstaedt et al., 2018; Guntuku et al., 2017). However, social media language is sparse with declining usage and user posting frequency (Mavrck, 2017), requiring many months of data for reliable analysis (Merchant et al., 2019). Furthermore, studies that analyze social media language recruit from populations that produce more frequent posts, potentially limiting the generalizability of reported results (Eichstaedt et al., 2018; Merchant et al., 2019). Computer-mediated communication and language is effective for depression prediction, but more generalizable data streams that operate on shorter timescales are needed for effective personal sensing that is responsive to changing language use patterns on the order of days and weeks rather than months.

Text messages are a passive data stream for computer-mediated communication that offers a number of advantages over social media posts. Among the United States adult population, rates of text messaging use (97%) (Smith, 2015) are greater than rates of social media use (72%) (Pew Research Center, 2019). Text messages are personal, often containing information that individuals are not interested in sharing publicly or within their social circles. They are frequent, with public survey polling suggesting that Americans send and receive an average of at least 45 (with a median of 10) text messages daily (Smith, 2011). Short message service (SMS) is the most frequently used function on smartphones, far outpacing social media use (Smith, 2015), and can be monitored passively and continuously (Mohr et al., 2020, 2017). Relative to social media postings, text messages offer a more dense, granular, and more personal data source that may provide more accurate and faster prediction of future depression symptom severity. However, because of the sensitive and personal nature of text messages, data security and privacy are critical when leveraging text-message data. Only with the proper data handling procedures and safeguards in place, such as anonymizing data and on-phone processing of text messages, can privacy concerns be adequately addressed (Jacobson et al., 2020; Onnela, 2021). Although studies have explored SMS message sentiment generally (Andriotis et al., 2014), little work has been done examining the relationship between text message sentiment and psychological outcomes (Glenn et al., 2020). To our knowledge, there has only been one small exploratory study in this space, examining whether SMS language characteristics sentiment were associated with high or low-risk leading up to a suicide attempt.

Here we examine text message sentiment as a digital marker of depression status in a population of 219 U.S. adults enrolled in a 16 week longitudinal study (Fig. 1). Participants completed baseline assessments of their mental wellbeing and installed a personal sensing app on their phones, which administered regular self-report PHQ-8 assessments of their depression severity (Kroenke et al., 2009) and passively collected sensor data: GPS location, application usage, and communication metadata. The app also computed anonymized sentiment scores on participant's devices, which allows for text sentiment analysis without having to store sensitive raw text message data. Participants were included in our final analysis sample if they sent at least 100 text messages throughout the entire study to meet reliable thresholds of data for language analysis (Merchant et al., 2019). We explore the associations

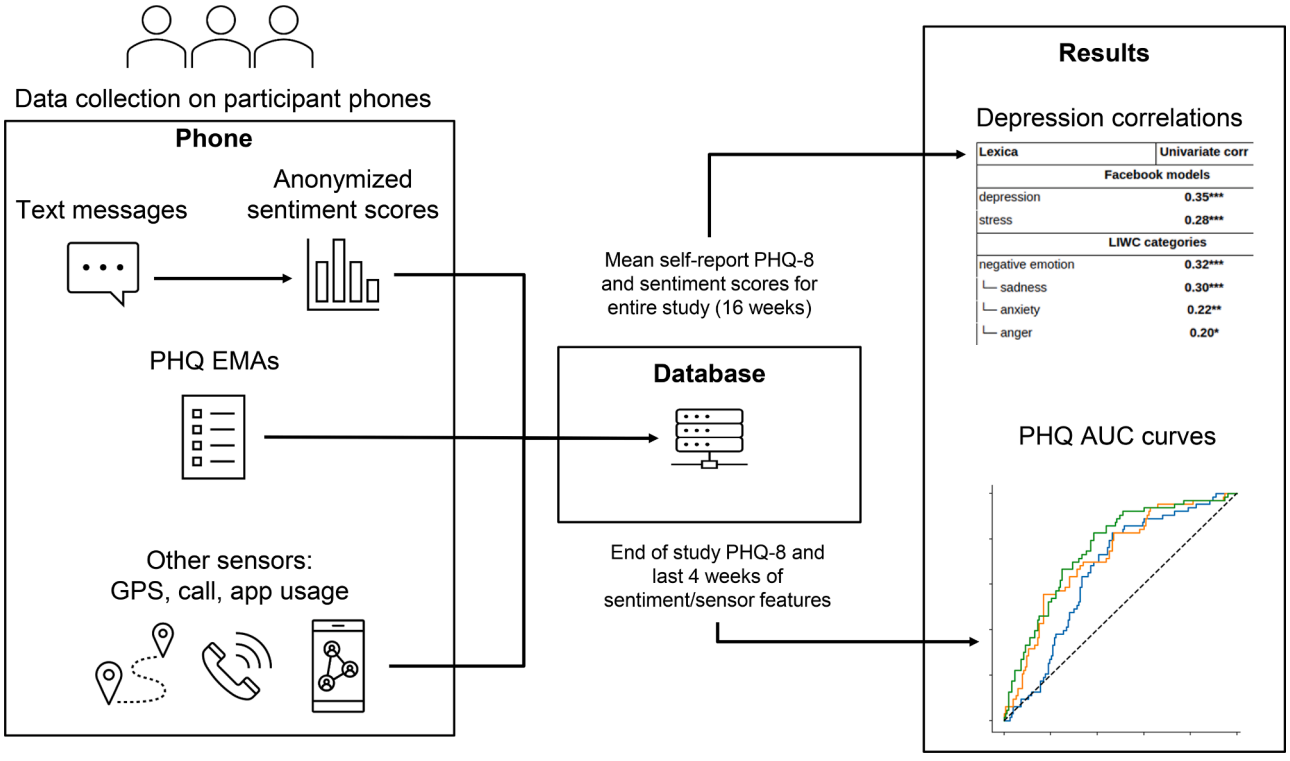

Fig. 1. Data collection and analysis strategy of our study.

between existing psychologically validated lexica and depression symptom severity. We also examine the ability of text message sentiment to classify future depression both as a standalone feature and in conjunction with other networked smartphone sensors, using predictions made using baseline depression as a benchmark for performance.

2. Methods

2.1. Participants

Participants were recruited across the United States through Focus Pointe Global, a national research panel, and enrolled in two periods: February 3, 2020 to February 7, 2020, and April 6, 2020 to April 10, 2020. The study was advertised as a study on depression, deliberately over-sampling depressed individuals so that at least 50% of participants experienced at least moderate depression symptoms according to a baseline PHQ-8 self-report. We included participants who were at least 18 years old, used Android smartphones, and did not have any self-reported co-morbid severe mental illness: bipolar disorder, schizophrenia, or other psychotic disorders. iPhone users were excluded due to sensor collection restrictions on the iOS operating system, and participants could decline to consent or withdraw if they had any privacy concerns with the study. Participants were compensated for completing regularly scheduled self-report online assessments as well as for completing ecological momentary assessments (EMAs), with a maximum total compensation of $142 for completing all assessments across the entire study. Study protocols and procedures were approved by the Northwestern University Institutional Review Board (IRB) and written electronic informed consent from all participants was obtained before the beginning of the study.

2.2. Data collection and procedures

Participation in the 16 week study began after the end of the enrollment period, with 370 participants combined across both groups of study participants enrolled. Participants completed online surveys through the REDCap platform (Harris et al., 2019, 2009) administered at baseline as well as once every three weeks during study. Passive sensing phone data, including GPS location, communication metadata (i.e. text message count, call count, call duration), and application usage, were collected through the Passive Data Kit (PDK) mobile app (Audacious Software, 2018). The PDK app also administered on-phone PHQ-8 surveys at the beginning and end of every third week in the study, with evaluations in weeks 1, 4, 7, 10, 13, and 16. Text message sentiment scores per text message sent were computed on-device as a weighted word count sum for every lexica category in LIWC 2015 (Pennebaker et al., 2015), the degrees of depression lexica (Schwartz et al., 2014), and the stress lexica (Guntuku et al., 2019). No raw text message content was stored as part of the study, as only the aggregated sentiment scores were transmitted from participants' phone.

We then filtered participants based on the density of their text message data. 303 participants remained in the study to its conclusion. Of these participants, 264 reported text message sensor data. Using the 500 word threshold established in other literature for stable language analysis (Merchant et al., 2019), we included participants with at least 100 outgoing text messages as a heuristic given the median text message length in our sample (35 characters) and the average length of a word in English text (~5 characters) (Miller et al., 1958). A total of 219 participants met this texting criteria. Demographic characteristics of these participants are shown in Table 2.

2.3. Lexica correlations

To study the relationship between language and depression severity, we correlated participant's reported mean PHQ-8 across the study with the relative frequencies of lexica categories. We use significance thresholds corrected with the Benjamini-Hochberg procedure to account for multiple comparisons (Benjamini and Hochberg, 1995).

2.4. End-of-study depression prediction target outcome

We binarized participant's week 16 PHQ-8 score according to the established cutoff score of 10 (Kroenke et al., 2009) in order to evaluate the predictive performance of personal sensing features for future depression status classification. After binarization, 30% of our participants were above the PHQ-8 cutoff. We use receiver operating characteristic (ROC) curves and the accompanying area under the ROC curve (AUC), which are appropriate measures of classifier performance in the presence of imbalanced labels, to evaluate our models.

2.5. Personal sensing feature extraction

To build end-of-study depression status classification models, we need to derive features from our collected personal sensing data. The relative frequencies of selected lexica based on categories shown in previous literature to be correlated with depression (Eichstaedt et al., 2018; Schwartz et al., 2014) were used as features: the degrees of depression and stress lexica which are pre-trained weighted lexica dictionaries fitted on social media data, and the LIWC categories negative emotion, sadness, anxiety, discrepancy, pronouns, personal pronouns, first person singular, feeling, and health. We chose not to use data-driven methods to select for lexica features, as the small sample size relative to the number of lexica categories could lead to overfitting due to statistical artifacts. These features comprised the "text-only" model results shown in Fig. 1.

We also considered phone sensing markers in addition to text sentiment features. We computed high-level GPS-based movement and location features based on the methodology presented in Saeb et al. (2016): location variance, circadian movement, velocity, location clusters, raw and normalized location entropy, home stay, and total transitions between locations. Communication metadata features included daily means of incoming/outgoing text message counts, call counts, and total call durations. Application usage features were computed as mean daily screen-on durations of aggregated categories informed by the communication field (Bayer et al., 2020): browser, social media, message, and email applications. A complete list of application names searched for in each category can be found in Appendix Table A.2. We also include time spent on the app "launcher" (home screen) as another measure of general smartphone usage. These GPS, communication, and application features comprise our "sensor-only" model results.

2.6. Data window size selection

When determining the timeframe for aggregating sensor data into features for end-of-study prediction, we wanted to balance having enough data for stable language analysis with temporal proximity to the end-of-study measurement. To do so, we examined time windows of 1, 2, 4, 8, and 15 weeks from the end of study measurement. We found that a time window of 4 weeks produced the best performance of our preliminary language-only models (Fig. A.3). 216 of our participants met the 100 outgoing messages threshold when considering data within 4 weeks of the final PHQ-8 assessment, and we use text and sensor data collected in this timeframe for all of our subsequent model creation.

2.7. Depression prediction model creation

We considered two models for our prediction task: logistic regression with L2 regularization and histogram-based gradient boosted trees, both implemented in scikit-learn (Pedregosa et al., 2011). We use leave-one-out cross validation to evaluate out-of-sample performance due to the relatively small sample of participants. The hyperparameters

of all models were tuned using stratified cross validation within the training set.

For our text-only model, we found that logistic regression performed the best, consistent with previous literature using language models for depression prediction (Eichstaedt et al., 2018), while the gradient-boosted model performed best when using the sensor-only feature block. To combine the features, we trained a gradient-boosted model that uses the sensor features to predict the residuals of our text-only model in a "stage-wise" fashion (Hastie et al., 2009), adding the predictions of the two models to produce final output. We used ROC curves to evaluate performance, and DeLong's test (DeLong et al., 1988) is used for significance testing between the ROC curves.

3. Results

3.1. Texting language correlates of depression

We sought to understand the relationship between text message language used and severity of depression. To do so, we correlated participant-level mean Patient Health Questionnaire (PHQ-8) scores with relative frequencies of words from three existing sentiment lexica dictionaries: the weighted (1) *degree of depression* (Schwartz et al., 2014) and (2) *stress* (Guntuku et al., 2019) lexica trained on social media language, as well as (3) the unweighted Linguistic Inquiry and Word Count (LIWC) lexica (Pennebaker et al., 2015) commonly used in psychological linguistics research (Table 1, univariate correlations). 24 lexica categories were significantly correlated with mean PHQ-8 at $\alpha \leq 0.05$ after corrections for multiple comparisons (Benjamini and Hochberg, 1995). To provide context on the composition of some of these categories, we show the top ten words in selected lexica collected in a prior study in Appendix A.

The pre-trained social media lexica (Guntuku et al., 2017), degrees of depression ($r = 0.35, p < 0.001$) and stress ($r = 0.28, p < 0.001$) yielded strong associations with depression symptom severity when using text messaging language.

Emotional LIWC lexica, such as overall negative emotion ($r = 0.32, p < 0.001$), which includes sadness ($r = 0.30$, $p < 0.001$), anxiety ($r = 0.22, p < 0.01$), and anger ($r = 0.2, p < 0.05$) words also showed positive associations with depression severity. Negative emotion and sadness words both serve as indicators of low mood — a hallmark symptom of depression. Anger and anxiety words, while not measured in PHQ-8, are indicators of emotional states that are frequently comorbid with depression (Painuly et al., 2005; Sartorius et al., 1996). We also found that the swear lexica ($r = 0.17, p < 0.05$) and sexual lexica ($r = 0.18, p < 0.05$), which are also dominated by swear words were associated with depressive symptoms, consistent with previous findings that hostility is a marker for depression (Eichstaedt et al., 2018) (Appendix A.1 and A.2).

We observed a relationship between the LIWC lexica of cognitive processes and depressive symptom severity ($r = 0.28$, $p < 0.001$). The LIWC cognitive processes lexica comprises smaller dictionaries that we also found to be significantly associated with depression severity. These included discrepancy ($r = 0.25, p < 0.01$) tentative ($r = 0.20, p < 0.05$) differentiation ($r = 0.23$, $p < 0.01$) and causation($r = 0.19$, $p < 0.05$) lexica. These findings could be markers for ruminative processes, as has been found in other studies on language use in depression (Eichstaedt et al., 2018).

The use of first person singular words ($r = 0.24$, $p < 0.01$) is also correlated with depressive symptom severity. This finding is consistent with the literature that first-person singular pronouns are robust language markers for depression and negative affectivity more generally (Edwards and Holtzman, 2017; Tackman et al., 2019). As a result, lexica categories that are supersets of first person singular words are also significantly correlated with depression severity, including personal pronouns ($r = 0.25, p < 0.01$), pronouns ($r = 0.25, p < 0.01$ and function words ($r = 0.23$, $p < 0.01$).

Table 1
Demographics and baseline characteristics for included participants.

Variable	Total $n = 219$
Age, mean (sd)	43.4 (12)
Sex (assigned at birth), n (%)	
Female	169 (77.2%)
Male	50 (22.8%)
Gender identity, n (%)	
Woman	168 (76.7%)
Man	50 (22.8%)
Non-binary	1 (0.5%)
Race, n (%)	
White	175 (79.9%)
Black/African American	32 (14.6%)
Asian	2 (0.9%)
Native American/Alaskan Native	1 (0.5%)
More than one Race	8 (3.7%)
Prefer not to answer	1 (0.5%)
Ethnicity, n (%)	
Hispanic/Latinx	12 (5.5%)
Non-Hispanic/Non-Latinx	207 (94.5%)
Highest level education completed, n (%)	
Some high school, no diploma	3 (1.4%)
High school/GED	16 (7.3%)
Some college, no degree	43 (19.6%)
Associate's degree	53 (24.2%)
Bachelor's degree	68 (31.1%)
Graduate degree	36 (16.4%)
Marital status, n (%)	
Single/never married	64 (29.4%)
Domestic partnership	2 (0.9%)
Married	75 (34.4%)
Separated	7 (3.2%)
Divorced	39 (17.9%)
Unknown/Prefer not to answer	3 (1.4%)
Household income, n (%)	
<$10,000	17 (7.8%)
$10,000–19,999	23 (10.5%)
$20,000–39,999	42 (19.2%)
$40,000–59,999	46 (21.0%)
$60,000–99,999	49 (22.4%)
>$100,000	39 (17.8%)
Unknown/Prefer not to answer	3 (1.4%)
Employment, n (%)	
Employed	143 (65.3%)
Unemployed	26 (11.9%)
Disability	21 (9.6%)
Retired	9 (4.1%)
Other	19 (8.7%)
Prefer not to answer	1 (0.5%)
Baseline PHQ-8, mean (sd)	9.73 (6.51)

In addition to first person singular word usage, we found that various parts of speech were related to depression severity. The use of adverbs ($r = 0.28, p < 0.001$), (common) verbs ($r = 0.20, p < 0.05$), auxiliary verbs ($r = 0.19$, $p < 0.05$), suggest the adoption of an informal, passive voice (Tausczik and Pennebaker, 2010), as well as conjunctions ($r = 0.26, p < 0.01$) and quantifiers ($r = 0.17$, $p < 0.05$).

Our results also reveal links between temporal features of language and depressive symptoms. Notably, we found associations between LIWC lexica that contain words focused on the past ($r = 0.16, p < 0.05$) as well as lexica that contain words focused on the present ($r = 0.17, p < 0.05$). These findings suggest an orientation on the here and now and on the past, rather than future-minded goal-orientation (Tausczik and Pennebaker, 2010).

We also wanted to evaluate the relationship between language markers and depressive symptoms after adjusting for baseline depression severity (Table 1, partial correlations). We find that degree of depression ($r = 0.25, p < 0.01$), negative emotion ($r = 0.23, p < 0.05$), sadness ($r = 0.21$, $p < 0.05$), stress ($r = 0.19$, $p < 0.05$), personal pronouns ($r = 0.21$, $p < 0.05$), and sexual language ($r = 0.23$, $p < 0.01$) partial correlations remain significant after correcting for multiple comparisons. Our overall findings still hold when controlling for

baseline PHQ-8 scores, emphasizing the robustness of language as a marker for depression severity.

3.2. Depression status classification prediction using text sentiment features

To use personal sensing markers in depression assessment, we need to evaluate the effectiveness of text sentiment as a predictor of future depression status. To this end, we built a predictive model of end-of-study depression status classification using the participants' last four weeks of language lexica data. We chose to use the last four weeks of data to balance between temporal recency and a sufficient amount of data to make reliable language estimates, as our preliminary analyses showed that window size produced the best out-of-sample performance. End-of-study depression status was converted into a binary indicator of PHQ-8 $\geq$ 10, a well-studied cutoff for current depression (Kroenke et al., 2009). 29% of participants in our sample had an end-of-study PHQ-8 above the cutoff. Because of the limited sample size, we use lexica previously established to predict depression status as features for our model (Eichstaedt et al., 2018; Schwartz et al., 2014) and use leave-one-out cross validation for out of sample prediction. We compare our language model to models using GPS and application usage sensor features, which have been previously studied to predict depression severity (Saeb et al., 2016). We also train a model that combines both language and sensor modalities to contextualize the predictive performance across all of our models (Fig. 2).

We find that using only text sentiment features produces fair out-of-sample performance (AUC=0.72). This is greater than the sensor-only performance in our sample (AUC=0.66), though not a significant difference ($p = 0.25$). The combined model using both language and sensor features (AUC=0.76) produces a significant increase ($p = 0.01$) in performance over the sensor-only model, suggesting that these data modalities are complementary. As text communication can be continually monitored while PHQ questionnaires may not be administered on a regular basis, these results provide a meaningful interpretation of the predictive power of text sentiment and highlight its potential as a marker for future depression status, especially in the context of other personal sensing features.

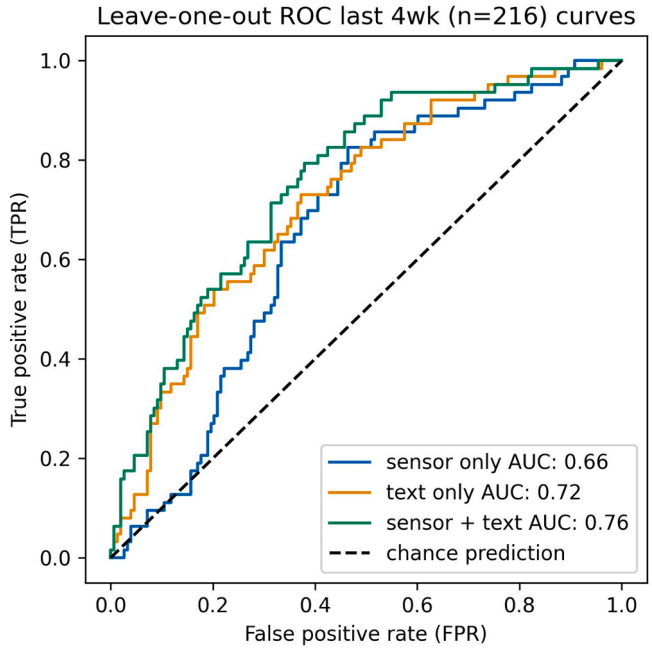

Fig. 2. Leave-one-out ROC curves for depression status classification. The difference between the sensor-only and combined sensor + text performance is statistically significant ($p = 0.01$).

4. Discussion

We show strong predictive performance for in-the-wild depression status classification using text message sentiment in combination with other personal sensing features. Text sentiment alone as a predictor performed well, and combining language features with established phone sensor data modalities significantly improved performance. Our results highlight the potential value of text messages as a component of a larger personal sensing suite that also integrates data from networked smartphone sensors to predict future depression status continuously and passively.

We also found that many sentiment lexica correlate strongly with mean study PHQ-8 consistent with other psychological work on language sentiment. Our results replicate findings in the literature, that certain language use patterns on social media are related to depression, and extend these findings to private text messages. Specifically, we find that a degrees of depression language model, trained on a large corpus of Facebook language (Schwartz et al., 2014), is highly associated with PHQ-8 despite different sources of digital language (semi-public social media vs. private conversations) and different outcome measures (personality depression facet vs. PHQ-8). We also replicated findings regarding the association of emotional language, cognitive rumination, and personal pronoun usage in social media with depressive symptoms (Eichstaedt et al., 2018; Guntuku et al., 2017). While it is not entirely surprising that language markers of depression transfer to text messages, these results establish text message language as another digital feature (Insel, 2017) that can be measured through personal sensing.

In fact, text messaging presents a number of distinctions over social media as a data stream for digital communication. Text messaging is more commonly used than social media usage, thus our results could be applicable to a wider population (Pew Research Center, 2019; Smith, 2015). Additionally, people often generate a greater volume of text message language compared to social media language (Smith, 2011), allowing for sentiment analysis on shorter timescales. We meet established word count thresholds and achieve good classification performance using only four weeks of text messages, while studies using social media may require many more months or even years of data. The ubiquity of text message usage and density of text message language make it a promising personal sensing data source for measuring depression status and severity.

In practice, text messages can be used together with GPS mobility, app use data, activity levels, and other sensed behaviors to predict and identify periods of risk or periods of worsening symptom severity. While other sensed features provide passive data about an individuals' behaviors, text message data, like social media data (Bathina et al., 2021), provides a richer data source indicative of an individual's cognitions, feelings, and sociality and may be especially well suited for capturing cognitive and affective aspects of depression, allowing for more comprehensive assessments of depressive symptoms.

4.1. Privacy and safety

We need to carefully consider and address data privacy and safety when using personal sensing technologies for mental health monitoring and prediction. First, presenting the plausible risks to participants about how their data may reveal sensitive information at the point of obtaining informed consent is an important element of ethical research practices. For many participants, it may not be obvious that their digital language or behavioral patterns can reveal information about their mental health status (Nicholas et al., 2019), so informing participants how text language or other sensed data can reveal sensitive health information (e.g. through a machine learning algorithm) is crucial in personal sensing study design. Though we do not record outcomes regarding participants' perceived trust and privacy in this study, it is imperative that future work examine how different degrees of personal information shared affect a participant's comfort and trust level with medical professionals.

It is also necessary to minimize unintended harm to participants by protecting their personal data. Though data anonymization is a best practice, it does not guarantee research participants' privacy, as it is possible to re-identify individuals from collected datasets (Rocher et al., 2019). Thus, we also need to minimize the amount of sensed data collected, and, if possible, ensure that particularly sensitive data is never actually collected by only storing anonymous data summaries (Rawassizadeh et al., 2018). We take this approach when analyzing participants' text messages by computing language sentiment on their devices, and only collecting these anonymous metrics as opposed to the raw text message content. While there is great promise for the application of personal sensing and digital language to innovate mental health care delivery, privacy and safe data handling processes must be prioritized to minimize the risk of unintended harm and increase patients' control over their data.

We also need to consider the ethical implications of developing and deploying digital phenotyping systems in the real world beyond a controlled research setting. Although university-based research universally goes through IRB approval that ensures ethical management of data, app companies are far from uniform in having careful privacy policies (Huckvale et al., 2019a). Thus, while passive text sentiment analysis methods such as those developed in this paper could potentially be useful in the emerging digital mental health industry, companies would have to establish and adhere to clear and transparent privacy policies that prevent inappropriate use or the sale of data to third parties. When looking forward to a future where digital text analysis methods could be deployed in the wild, privacy and data sharing agreements must be scrutinized with an appropriate level of rigor to ensure user safety.

4.2. Outlook

These findings suggest the potential of text messages for personal sensing, as (1) it is frequently and widely used, (2) sheds light on an individuals' thoughts, feelings, and social relationships, and (3) has good predictive ability to identify worsening depression severity. In a research context, text messaging can be paired with other networked sensors to help us understand the relationship between an individuals' behaviors and environment (e.g., places they visit and frequent, people they communicate with, how active they are, their sleep habits, etc.) and an individual's thoughts, feelings, and subsequent symptom severity. Additionally, using text messages as a passive data source opens up new possibilities for just-in-time-adaptive-interventions (Nahum-Shani et al., 2016, 2014) and micro-interventions (Baumel et al., 2020). Using an individual's sensed cognitive or affective state could allow for context-aware and in-the-moment digital interventions. For example, communication and social skills problems are common in individuals with depression and social functioning is an outcome that they value highly (Chevance et al., 2020). Monitoring messages for correlates associated with social difficulties and isolation could be used to deliver useful communication skills training in the moment. Finally, text messaging raises the possibility of truly passive depression screening systems that have the potential to proactively address emerging depressive episodes before they become more severe and prolonged. While text messaging alone is unlikely to provide sufficient signal for this task, when paired with other sensed data that can be monitored continuously, a future in which passive monitoring to detect individuals who may benefit from intervention may be feasible.

4.3. Limitations

A number of important limitations must be considered for our study. First, the sample is likely not wholly representative of people with depression. While rates of depression prevalence are greater in women than men (Salk et al., 2017), participants were recruited from a research panel that over-represented women (76.7%) and individuals with self-reported baseline depression (mean PHQ-8 = 9.76) (Table 2). Furthermore, the sample excluded due to not reaching the minimum texting threshold contained significantly more males than the sample that met the texting threshold (Table A.4), further increasing gender differences. The PHQ-8 scores we use in the study represent self-reported symptom severity, and diagnostic criteria were not confirmed via clinical interview. Subsequently, individuals with PHQ-8 scores >= 10 have a likely diagnosis of Major Depressive Disorder, but it is conceivable that other conditions may account for the reported symptoms and their severity, which limits the scope of interpretation to self-reported symptoms. Our sample is also predominantly white (79.9%) and tends towards middle-aged individuals (mean age = 43.4 years). The composition of our sample could influence the results of our study, as it is well known that language use differs across many demographic factors, most prominently gender and age (Fast and Funder, 2010; Schwartz et al., 2013), so we caution against interpreting the results beyond the population considered here.

Another data limitation is that our study period coincided with the initial COVID-19 outbreak in the United States, where much of the country was under social-distancing and stay-at-home restrictions. These circumstances may have altered participant behavior as well as the personal sensing data we collected, particularly GPS locations. Consequently, our results may have also been influenced by these circumstances, such as the relatively poor performance of sensor-only prediction of future depression severity. Thus, there may be limits in the generalizability of our results, and we look forward to future work exploring the relationship between text sentiment, phone sensors, and depression severity in more varied populations and circumstances.

5. Conclusion

Text messages are a ubiquitous form of communication that show promise as a digital marker of mental health that, by extracting sentiment scores, can be collected unobtrusively while preserving privacy. Our findings highlight the potential utility of text message language, especially when combined with other sources of data, to predict

Table 2
Significant correlations between lexica and mean study PHQ.

Lexica	Univariate corr	Partial corr
Pre-trained semantic features from pre-trained models		
depression	0.35***	0.25**
stress	0.28***	0.19*
LIWC categories		
negative emotion	0.32***	0.23*
└ sadness	0.30***	0.21*
└ anxiety	0.22**	0.15
└ anger	0.20*	0.15
sexual	0.18*	0.25**
swear	0.17*	0.14
cognitive processes	0.28***	0.16
└ discrepancy	0.25**	0.13
└ tentative	0.20*	0.11
└ differentiation	0.23**	0.13
└ causation	0.19*	0.20*
function words	0.23**	0.14
└ pronouns	0.23**	0.18
└ personal pronouns	0.25**	0.21*
└ first person singular	0.24**	0.20*
└ adverbs	0.28***	0.17
└ common verbs	0.20*	0.13
└ auxiliary verbs	0.19*	0.10
└ conjunctions	0.26**	0.16
quantifiers	0.17*	0.10
past focus	0.16*	0.09
present focus	0.17*	0.13

All correlations are tested for significance with a BH correction for multiple comparisons. * < 0.05, ** < 0.01, *** < 0.001.

depression status and passively monitor depressive symptom severity changes.

Funding

This work was supported by a grant from the National Institute of Mental Health (5R01MH111610) to David C. Mohr and Konrad Kording. Jonah Meyerhoff is supported by a grant from the National Institute of Mental Health (T32 MH115882). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Contributor

TL performed literature search, data analysis, and lead manuscript writing. JM and JCE also performed literature searches and wrote the manuscript. CJK developed the data collection software, led mobile sensing platform development and deployment, and assisted in manuscript drafting. SMK managed all aspects of data collection and drafted key portions of the methods section. KPK, DCM, and LHU oversaw data analyses and helped write the manuscript. All authors discussed the results throughout the development of the manuscript and have approved the final manuscript.

Data statement

The phone sensor data are not publicly available due to the presence of potentially identifying information that could compromise participant privacy and consent, but the de-identified self-report data (PHQ and EMAs) will be made available through the NIMH Data Archive at the conclusion of the study.

Declaration of Competing Interest

David C. Mohr has accepted consulting fees from Apple Inc, Otsuka Pharmaceuticals, and the One Mind Foundation. He also has an ownership interest in Adaptive Health, Inc. None of the other authors have conflicts of interest to declare.

Acknowledgements

We thank Nameyeh Alam and Zara Mir for help with study management as well as data collection and management.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jad.2021.12.048.

References

Andriotis, P., Takasu, A., Tryfonas, T., 2014. Smartphone message sentiment analysis. In: Peterson, G., Shenoi, S. (Eds.), Advances in Digital Forensics X, IFIP Advances in Information and Communication Technology. Springer, Berlin, Heidelberg, pp. 253–265. https://doi.org/10.1007/978-3-662-44952-3_17.

Audacious Software, 2018. Passive Data Kit.

Bathina, K.C., ten Thij, M., Lorenzo-Luaces, L., Rutter, L.A., Bollen, J., 2021. Individuals with depression express more distorted thinking on social media. Nat. Hum. Behav. 1–9. https://doi.org/10.1038/s41562-021-01050-7.

Baumel, A., Fleming, T., Schueller, S.M., 2020. Digital micro interventions for behavioral and mental health gains: core components and conceptualization of digital micro intervention care. J. Med. Internet Res. 22, e20631. https://doi.org/10.2196/20631.

Bayer, J.B., Triệu, P., Ellison, N.B., 2020. Social media elements, ecologies, and effects. Annu. Rev. Psychol. 71, 471–497. https://doi.org/10.1146/annurev-psych-010419-050944.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. 57, 289–300.

Chevance, A., Ravaud, P., Tomlinson, A., Le Berre, C., Teufer, B., Touboul, S., Fried, E.I., Gartlehner, G., Cipriani, A., Tran, V.T., 2020. Identifying outcomes for depression that matter to patients, informal caregivers, and health-care professionals: qualitative content analysis of a large international online survey. Lancet Psychiatry 7, 692–702. https://doi.org/10.1016/S2215-0366(20)30191-7.

Choudhury, M.D., Gamon, M., Counts, S., Horvitz, E., 2013. Predicting depression via social media 10.

DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837. https://doi.org/10.2307/2531595.

Edwards, T., Holtzman, N.S., 2017. A meta-analysis of correlations between depression and first person singular pronoun use. J. Res. Personal. 68, 63–68. https://doi.org/10.1016/j.jrp.2017.02.005.

Eichstaedt, J.C., Smith, R.J., Merchant, R.M., Ungar, L.H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D.A., Schwartz, H.A., 2018. Facebook language predicts depression in medical records. Proc. Natl. Acad. Sci. 115, 11203–11208. https://doi.org/10.1073/pnas.1802331115.

Fast, L.A., Funder, D.C., 2010. Gender differences in the correlates of self-referent word use: authority, entitlement, and depressive symptoms. J. Pers. 78, 313–338. https://doi.org/10.1111/j.1467-6494.2009.00617.x.

Fried, E.I., 2017. The 52 symptoms of major depression: lack of content overlap among seven common depression scales. J. Affect. Disord. 208, 191–197. https://doi.org/10.1016/j.jad.2016.10.019.

Fried, E.I., Nesse, R.M., 2015. Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR*D study. J. Affect. Disord. 172, 96–102. https://doi.org/10.1016/j.jad.2014.10.010.

Glenn, J.J., Nobles, A.L., Barnes, L.E., Teachman, B.A., 2020. Can text messages identify suicide risk in real time? A within-subjects pilot examination of temporally sensitive markers of suicide risk. Clin. Psychol. Sci. 8, 704–722. https://doi.org/10.1177/2167702620906146.

Greenberg, P.E., Fournier, A.-.A., Sisitsky, T., Pike, C.T., Kessler, R.C., 2015. The economic burden of adults with major depressive disorder in the United States (2005 and 2010). J. Clin. Psychiatry 76, 155–162. https://doi.org/10.4088/JCP.14m09298.

Guntuku, S.C., Buffone, A., Jaidka, K., Eichstaedt, J.C., Ungar, L.H., 2019. Understanding and measuring psychological stress using social media. In: Proceedings of the International AAAI Conference on Web and Social Media 13, pp. 214–225.

Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H., Eichstaedt, J.C., 2017. Detecting depression and mental illness on social media: an integrative review. Curr. Opin. Behav. Sci. 18, 43–49. https://doi.org/10.1016/j.cobeha.2017.07.005. Big data in the behavioural sciences.

Harris, P.A., Taylor, R., Minor, B.L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., Duda, S.N., 2019. The REDCap consortium: building an international community of software platform partners. J. Biomed. Inform. 95, 103208 https://doi.org/10.1016/j.jbi.2019.103208.

Harris, P.A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., Conde, J.G., 2009. Research electronic data capture (REDCap)-a metadata-driven methodology and workflow process for providing translational research informatics support. J. Biomed. Inform. 42, 377–381. https://doi.org/10.1016/j.jbi.2008.08.010.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning. Springer Series in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/978-0-387-84858-7.

Huckvale, K., Torous, J., Larsen, M.E., 2019a. Assessment of the data sharing and privacy practices of smartphone apps for depression and smoking cessation. JAMA Netw. Open 2, e192542. https://doi.org/10.1001/jamanetworkopen.2019.2542.

Huckvale, K., Venkatesh, S., Christensen, H., 2019b. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. Npj Digit. Med. 2, 1–11. https://doi.org/10.1038/s41746-019-0166-1.

Insel, T.R., 2018. Digital phenotyping: a global tool for psychiatry. World Psychiatry 17, 276–277. https://doi.org/10.1002/wps.20550.

Insel, T.R., 2017. Digital phenotyping: technology for a new science of behavior. JAMA 318, 1215–1216. https://doi.org/10.1001/jama.2017.11295.

Jacobson, N.C., Bentley, K.H., Walton, A., Wang, S.B., Fortgang, R.G., Millner, A.J., Coombs, G., Rodman, A.M., Coppersmith, D.D.L., 2020. Ethical dilemmas posed by mobile health and machine learning in psychiatry research. Bull. World Health Organ. 98, 270–276. https://doi.org/10.2471/BLT.19.237107.

Kroenke, K., Strine, T.W., Spitzer, R.L., Williams, J.B., Berry, J.T., Mokdad, A.H., 2009. The PHQ-8 as a measure of current depression in the general population. J. Affect. Disord. 114, 163–173.

Liao, P., Greenewald, K., Klasnja, P., Murphy, S., 2019. Personalized HeartSteps: a reinforcement learning algorithm for optimizing physical activity. ArXiv190903539 Cs.

Marsch, L.A., 2018. Opportunities and needs in digital phenotyping. Neuropsychopharmacology 43, 1637–1638. https://doi.org/10.1038/s41386-018-0051-7.

Mavrck, 2017. 2017 Facebook User-generated content (UGC) benchmark report [WWW Document]. URL https://info.mavrck.co/hubfs/Anchor%20Content/Ebooks,%20White%20Papers/%5BFINAL%5D%20Facebook_UGC_Benchmark_Report_Mavrck_2017-Special_Edition.pdf?hsLang=en (accessed 3.7.21).

Merchant, R.M., Asch, D.A., Crutchley, P., Ungar, L.H., Guntuku, S.C., Eichstaedt, J.C., Hill, S., Padrez, K., Smith, R.J., Schwartz, H.A., 2019. Evaluating the predictability of medical conditions from social media posts. PLoS ONE 14, e0215476. https://doi.org/10.1371/journal.pone.0215476.

Miller, G., Newman, E., Friedman, E., 1958. Length-frequency statistics for written English. Inf. Control 1, 370–389. https://doi.org/10.1016/S0019-9958(58)90229-8.

Mohr, D.C., Shilton, K., Hotopf, M., 2020. Digital phenotyping, behavioral sensing, or personal sensing: names and transparency in the digital age. Npj Digit. Med. 3, 45. https://doi.org/10.1038/s41746-020-0251-5.

Mohr, D.C., Zhang, M., Schueller, S.M., 2017. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu. Rev. Clin. Psychol. 13, 23–47.
Nahum-Shani, I., Smith, S.N., Spring, B.J., Collins, L.M., Witkiewitz, K., Tewari, A., Murphy, S.A., 2016. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann. Behav. Med. 52, 446–462. https://doi.org/10.1007/s12160-016-9830-8.
Nahum-Shani, I., Smith, S.N., Tewari, A., Witkiewitz, K., Collins, L.M., Spring, B., Murphy, S., 2014. Just in time adaptive interventions (jitais): an organizing framework for ongoing health behavior support. Methodol. Cent. Tech. Rep. 2014, 14–126.
Nicholas, J., Shilton, K., Schueller, S.M., Gray, E.L., Kwasny, M.J., Mohr, D.C., 2019. The role of data type and recipient in individuals' perspectives on sharing passively collected smartphone data for mental health: cross-sectional questionnaire study. JMIR MHealth UHealth 7, e12578. https://doi.org/10.2196/12578.
Onnela, J.P., 2021. Opportunities and challenges in the collection and analysis of digital phenotyping data. Neuropsychopharmacology 46, 45–54. https://doi.org/10.1038/s41386-020-0771-3.
Onnela, J.P., Rauch, S.L., 2016. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology 41, 1691–1696. https://doi.org/10.1038/npp.2016.7.
Otte, C., Gold, S.M., Penninx, B.W., Pariante, C.M., Etkin, A., Fava, M., Mohr, D.C., Schatzberg, A.F., 2016. Major depressive disorder. Nat. Rev. Dis. Primer 2, 16065. https://doi.org/10.1038/nrdp.2016.65.
Painuly, N., Sharan, P., Mattoo, S.K., 2005. Relationship of anger and anger attacks with depression. Eur. Arch. Psychiatry Clin. Neurosci. 255, 215–222. https://doi.org/10.1007/s00406-004-0539-5.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Python 6.
Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K., 2015. The development and psychometric properties of LIWC2015.
Pew Research Center, 2019. Social Media Fact Sheet: Demographics of Social Media Users and Adoption in the United States. Pew Research Center.
Rawassizadeh, R., Pierson, T.J., Peterson, R., Kotz, D., 2018. NoCloud: exploring network disconnection through on-device data analysis. IEEE Pervasive Comput. 17, 64–74. https://doi.org/10.1109/MPRV.2018.011591063.
Rocher, L., Hendrickx, J.M., de Montjoye, YA., 2019. Estimating the success of re-identifications in incomplete datasets using generative models. Nat. Commun. 10, 3069. https://doi.org/10.1038/s41467-019-10933-3.
Saeb, S., Lattie, E.G., Schueller, S.M., Kording, K.P., Mohr, D.C., 2016. The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ 4, e2537.
Salk, R.H., Hyde, J.S., Abramson, L.Y., 2017. Gender differences in depression in representative national samples: meta-analyses of diagnoses and symptoms. Psychol. Bull. 143, 783–822. https://doi.org/10.1037/bul0000102.
Sartorius, N., Ustün, T.B., Lecrubier, Y., Wittchen, H.U., 1996. Depression comorbid with anxiety: results from the WHO study on psychological disorders in primary health care. Br. J. Psychiatry. 38–43. Suppl.
Schwartz, H.A., Eichstaedt, J., Kern, M.L., Park, G., Sap, M., Stillwell, D., Kosinski, M., Ungar, L., 2014. Towards assessing changes in degree of depression through Facebook. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. Association for Computational Linguistics, Baltimore, Maryland, USA, pp. 118–125. https://doi.org/10.3115/v1/W14-3214. Presented at the Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality.
Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E.P., Ungar, L.H., 2013. Personality, gender, and age in the language of social media: the open-vocabulary approach. PLoS ONE 8, e73791. https://doi.org/10.1371/journal.pone.0073791.
Smith, A., 2015. U.S. Smartphone Use in 2015. Pew Research Center.
Smith, A., 2011. Americans and Text Messaging. Pew Research Center.
Tackman, A.M., Sbarra, D.A., Carey, A.L., Donnellan, M.B., Horn, A.B., Holtzman, N.S., Edwards, T.S., Pennebaker, J.W., Mehl, M.R., 2019. Depression, negative emotionality, and self-referential language: a multi-lab, multi-measure, and multi-language-task research synthesis. J. Pers. Soc. Psychol. 116, 817–834. https://doi.org/10.1037/pspp0000187.
Tausczik, Y.R., Pennebaker, J.W., 2010. The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. 29, 24–54. https://doi.org/10.1177/0261927X09351676.
Torous, J., Staples, P., Onnela, J.P., 2015. Realizing the potential of mobile mental health: new methods for new data in psychiatry. Curr. Psychiatry Rep. 17, 61. https://doi.org/10.1007/s11920-015-0602-0.
Tversky, A., Kahneman, D., 1973. Availability: a heuristic for judging frequency and probability. Cognit. Psychol. 5, 207–232.
Zimmerman, M., McGlinchey, J.B., 2008. Why don't psychiatrists use scales to measure outcome when treating depressed patients? J. Clin. Psychiatry 69, 1916–1919. https://doi.org/10.4088/jcp.v69n1209.
Zulueta, J., Piscitello, A., Rasic, M., Easter, R., Babu, P., Langenecker, S.A., McInnis, M., Ajilore, O., Nelson, P.C., Ryan, K., Leow, A., 2018. Predicting mood disturbance severity with mobile phone keystroke metadata: a BiAffect digital phenotyping study. J. Med. Internet Res. 20, e241. https://doi.org/10.2196/jmir.9775.