Smart Health 17 (2020) 100118

Contents lists available at ScienceDirect

Smart Health

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

Moodable: On feasibility of instantaneous depression assessment using machine learning on voice samples with retrospectively harvested smartphone and social media data

Ada Dogrucu [a], Alex Perucic [a], Anabella Isaro [a], Damon Ball [a], Ermal Toto [a], Elke A. Rundensteiner [a], Emmanuel Agu [a,*], Rachel Davis-Martin [b], Edwin Boudreaux [b]

[a] Computer Science Dept., Worcester Polytechnic Institute (WPI), 100 Institute Rd, Worcester, MA, 01609, USA
[b] University of Massachusetts Medical School (UMMS), 55 N Lake Ave, Worcester, MA, 01655, USA

ARTICLE INFO

Keywords:
Smartphone sensing
Sensors
Depression screening
Social media mining
Media
Voice analytics
Machine learning
Retrospective data

ABSTRACT

Depression is a leading cause of disability and is associated with suicide risk. However, a quarter of patients with major depression remain undiagnosed. Prior work has demonstrated that a smartphone user's depression level can be detected by analyzing data gathered from their smartphone's sensors or from their social media posts over a few weeks after enrollment in a user study. These studies typically utilize a prospective study design, which is burdensome as it requires participants smartphone data to be gathered for prolonged periods before their depression level can be assessed. In contrast, we present a feasibility study of our Mood Assessment Capable Framework (Moodable) that facilitates almost instantaneous mood assessment by analyzing instantaneous voice samples provided by the user as well as historical sensor data harvested (scraped) from their smartphone and recent social media posts. Our retrospective, low-burden approach means that Moodable no longer requires study participants to engage with their phone for weeks before a depression score can be inferred. Moodable has the potential to minimize user data collection burden, increase user compliance, avoid study awareness bias and offer a near instantaneous depression screening. To lay a solid foundation for Moodable, we first surveyed 202 volunteer participants about their willingness to share voice samples and various smartphone and social media data types for mental health assessment. Based on these findings, we then developed the Moodable app. Thereafter, we utilized Moodable to collect short voice samples, and a rich array of retrospectively harvested data from users' smartphones (location, browser history, call logs) and social media accounts (instagram, twitter and facebook), with appropriate permissions, of 335 volunteer participants who also responded to 9 depression related questions of the Patient Health Questionaire (PHQ-9). Moodable then used machine learning to build classification models and classify the user's depression and suicidal ideation, for users which scores where unknown to the models. Results of Moodable's screening capability are promising. In particular, for the depression classification task we achieved F1 scores (the harmonic mean of the precision and recall) of 0.766, sensitivity of 0.750, and specificity of 0.792. For the suicidal ideation task we achieved F1 scores of 0.848, sensitivity of 0.864, and specificity of 0.725. This work could significantly increase depression-screening at the population level and opens numerous avenues for further research into

* Corresponding author.
E-mail address: emmanuel@wpi.edu (E. Agu).

https://doi.org/10.1016/j.smhl.2020.100118
Received 30 March 2019; Received in revised form 20 December 2019; Accepted 25 March 2020
Available online 18 April 2020

this newly proposed paradigm of instantaneously screening depression and suicide risk levels from voice samples and retrospective smartphone and social media data.

1. Introduction

Background and Motivation. Background and Motivation. The US Preventive Services Task Force has identified depression as the leading cause of disability in adults and recommends depression screening for anyone over the age of 18 Siu et al. (2016). Due to its crippling effects, depression is predicted to cost US$5.36 trillion between 2011 and 2030 Richards et al. (2016) globally. Additionally, major depressive disorder increases the likelihood of suicide and impairs the patient's ability to manage other health issues, while decreasing patient and family quality of life Witters, Liu, and Agrawal (2015); Siu et al. (2016); Rapaport, Clary, Fayyad, and Endicott (2005).

While screening and detection are the first steps in treating depression Siu et al. (2016), up to a quarter of patients with major depression remain undiagnosed Epstein et al. (2010). This is partly because some patients view traditional depression screening methods, such as the administration of the Patient Health Questionnaire (PHQ-9), as intrusive, cumbersome, and sometimes even feared Weist, Rubin, Moore, Adelsheim, and Wrobel (2007).

State-of-Art in Depression Sensing. Passive methods to detect depression utilizing smartphone have recently emerged Wang et al. (2014); Ma, Yang, Chen, Huang, and Wang (2016). These methods passively gather and analyze smartphone data indicative of depression and make inference as patients live their lives. Such passive screening methods are minimally intrusive as they no longer require the patient to explicitly respond to a questionnaire. Patients responding to health questionnaires (including depression questionnaires) often have trouble accurately recalling their past behaviors, leading to recall bias. Passive smartphone sensing of depression is data-driven, objective and autonomous, and mitigates recall bias.

Several projects have explored smartphone sensing and social media mining for healthcare Saeb, Lattie, Schueller, Kording, and Mohr (2016) Ben-Zeev, Scherer, Wang, Xie, and Campbell (2015) Ma et al. (2016) Valstar et al. (2016) Hasan, Rundensteiner, and Agu (2014b). However, such prior studies have typically utilized a prospective study design, in which participants' smartphone data are gathered continuously for several weeks during the study period, while they simultaneously report their depression status periodically. In such prospective studies, patients' data must be gathered for several weeks before any depression assessment can be done.

The Case for Instantaneous Depression Assessment: In several situations, instantaneous on-the-spot depression assessment without first requiring weeks of depression data gathering is desirable. For instance, patients visiting an Emergency Department (ED) or other healthcare setting like primary care could conveniently be screened for depression, a common co-morbidity of many other ailments. Mental health counselors could generate an accurate depression score for their patients even on their first visit so that evidence-based therapies can be prescribed. At the population level, college students, patients participating in an Accountable Care Organization (ACO), individuals within the military, and the general public could be screened en masse autonomously, with appropriate permissions, facilitating swift interventions that potentially reduce depression and prevents suicides. Finally, this approach can be a useful tool for researchers interested in studying depression treatments. They could use the technology to rapidly screen for depression to help make approach and eligibility interviews more efficient by pruning a given population to those with a signal for depression, averting the expense of having a trained research assistant approach individuals to perform depression screenings.

Challenges: Most prior smartphone sensing work in healthcare require the patient to install a smartphone app, and then to utilize their phone for a period of 2 weeks or longer to enable the longitudinal collection of their behavioral data Wang et al. (2014); Ben-Zeev et al. (2015). This methodology precludes on-the-spot detection of depression. Therefore, to accomplish accurate on-the-spot automated depression screening, several challenges must be overcome:

1. *High study burden:* Studies that rely on prospective, longitudinal data collection require that subjects data be collected using a data gathering app for an extended study period, which imposes a high burden and stringent compliance requirements on subjects.
2. *Unsuitable for on-the-spot depression assessment:* In a prospective study, data collection must be complete before any analysis or inference can be done, which usually takes weeks. As such a prospective study design is likely not amenable for use in on-the-spot diagnosis.
3. *Study awareness bias:* Subjects in a prospective study are typically aware that their data is being gathered and may feel "watched" during the study, which could potentially cause them to modify their behaviors.

Thus, this begs the research question of whether retrospective data that is gathered almost instantaneously could be used instead. If retrospective data collection is possible, additional research questions need to be answered about the suitability of such a dataset for training robust depression prediction models and especially how accurately machine learning depression models trained from such data will predict the depression levels of previously unseen patients (i.e., the generalizability of this technology).

Our Approach: The Mood Assessment Framework (Moodable). In this paper, we explore the feasibility of creating an almost instantaneous depression assessment tool. We propose the Mood Assessment Framework (Moodable) that combines participatory data gathered on the spot through voice samples recorded as patients speak a few sentences with retrospective data harvested from their smartphones, such as location, browser history, and call logs, and social media accounts, including Instagram, Twitter and Facebook with appropriate permissions. Moodable users would need to first download and install it. It uses a machine learning approach to analyze all data gathered.

The Moodable approach could offer several key benefits. First, smartphone users would no longer be required to engage with their phone for the subsequent two weeks. Secondly, Moodable's retrospective data gathering methodology may increase the likelihood of collecting unbiased, digitally manifested depression symptoms, which were generated by the subject before being enrolled in the study or becoming aware that mental health was being assessed. Most importantly, textcolorcyanonce the machine learning models are sufficiently trained, Moodable promises a near instantaneous screening score of the subject's mental state. These trained models no longer need the subjects' PHQ-9 score and thus will present a very low user burden.

To research, develop and validate the Moodable Framework, we conducted two crowdsourced studies. First, to establish what data subjects are willing to permit Moodable to collect, we surveyed 202 participants for their willingness to provide voice samples, and to allow usage of various smartphone and social media data as sources for mental health screening. Secondly, we gathered PHQ-9 depression scores, voice samples, and harvested smartphone and social media data from 335 volunteer subjects, from which we synthesized machine learning classifiers to infer subjects' depression (PHQ-9) and suicidal ideation (PHQ-9, Item 9) scores on-the-spot. Lastly, our best performing classifiers were programmed into a robust working application that generated subjects' depression and suicide ideation scores from their voice samples, smartphone and social media data.

In this work, we demonstrate that the proposed methodology is promising, and the collected data holds informative signals. In particular, for the depression classification task we achieve F1 measures of 0.766, sensitivity of 0.750, and specificity of 0.792. In addition, by investigating Question 9 of the PHQ-9, we find that Moodable could also be leveraged to infer suicidal thoughts, achieving an F1 score of 0.848, sensitivity of 0.864, and specificity of 0.725. The F1 score is a measure of classification accuracy for unbalanced datasets These baseline results of Moodable are encouraging; as they offer the promise for enabling a potential paradigm shift to on-the-spot depression screening using machine learning on instant, low-burden biometrics and retrospective smartphone and social media data. This work opens numerous opportunities for follow-on research from the development of additional feature engineering, the addition of additional modes, the exploration of more advanced machine learning models, the design of more comprehensive case studies, to conducting additional evaluations of the Moodable technology on targeted subpopulations for validation.

2. Related work

Active vs participatory data collection from smartphones: Smartphone data collected for the purpose of mental health assessment can either involve active user participation (or participatory) data or passive (no user involvement or opportunistically gathered) data. Examples of active data include users taking surveys, reporting their food consumption or indicating their mood by selecting an image that reflects their current mood from a palette of images. Essentially, the human smartphone user is in the data collection loop and needs to be incentivized especially for high-burden data gathering Burke et al. (2006). In passive data collection, data such as GPS locations, accelerometer data, communication logs from voice calls and text messages are gathered autonomously without the user's involvement Torous, Staples, and Onnela (2015). Data collection is automatic and generates a large amount of data quickly without burdening the smartphone user Lane et al. (2010). However, the data tends to be noisy with missing values and some errors.

Mental health assessment using opportunistic data: A number of recent studies Ben-Zeev et al. (2015); Saeb et al. (2016); Wang et al. (2014) have investigated the correlation between opportunistic smartphone data and mental health. Saeb et al. showed that Saeb et al. (2016) GPS features, including location variance, entropy, and circadian movement, were significantly correlated with PHQ-9 scores (r's ranging from −0.43 to −0.46, p-values ¡ 0.05). Similar correlations have also been observed in features extracted from other opportunistic data types such as call frequency, application usage, twitter posts and text message frequency Ben-Zeev et al. (2015); Wang et al. (2014); Hasan et al. (2014b, a, 2017, 2018).

Mental health assessment using participatory data: Voice features such as pitch, intensity, speaking rate and pauses while talking have also been demonstrated to be good predictors of depression. Gathering data for machine learning studies to infer subjects' depression levels from voice typically involves gathering voice samples from subjects (e.g. while they read a passage) and simultaneously requiring them to fill out clinically validated measures of depression such as the PHQ-9 Ma et al. (2016); Valstar et al. (2016). Subsequently, machine learning methods are used to classify the patients' depression levels (PHQ-9 scores) using features extracted from the voice data. Other prior work Wang et al. (2014) has required participants to also log other wellness information such as their eating and sleeping patterns.

Social media mining of depression: Increasing evidence has shown that mental health issues and especially depression can be detected by analyzing social media messages of subjects Munmun De Choudhury, Gamon, and Horvitz (2013); M Park and Cha (2012). Moreno et al. found that 25% of the Facebook (social media) status updates of 200 college students displayed depressive symptoms based on the DSM depression evaluation criteria used in clinics Moreno et al. (2011). DSM criteria include depressed mood, hopelessness, and sleep difficulties. Their methods involved manually retrieving and post-analyzing social media posts by a human. Park et al. found that Twitter users openly disclosed their depressive state M Park and Cha (2012). Choudhury et al. M De Choudhury and Counts (2013) developed a statistical model to forecast postpartum depression in new mothers by analyzing prenatal social media messages, relationships, and linguistic style. Kotikalapudi discovered web browsing patterns of students that could signal depression R Katikalapaudi, Montgomery, Wunsch, Lutzen, and Chellappan (2012). Park et al. S Park, Kwak, Cha, Lee, and Jeong (2013) found that Facebook profiles could indicate its owner's depressive states. Reece and Danforth Reece and Danforth (2017a) also established that certain colors, metadata components and the number of faces of Instragram photos were good predictors of depressed Instragram users. These works all utilized manual data gathering and analysis methods.

Duration of data collection. Current state of the art studies in automated PHQ-9 estimation Saeb et al. (2016) Ben-Zeev et al. (2015) Ma et al. (2016) Valstar et al. (2016) Hasan et al. (2014b) base their data collection window on the DSM-V American Psychiatric Association (2013) definition of depression, which analyzes the persistence of symptoms in the prior 2-week period. Consequently, much of prior

work gathers longitudinal patient data via a mobile app for a minimum of two weeks. However, this approach requires that the data gathering is complete before analysis can be done, which is not suitable for instantaneous depression screening.

Moodable On-the-Spot depression Assessment Framework. In contrast, our research focuses on almost instantaneous data collection and machine learning analysis with the goal of on-the-spot depression detection. We combine participatory collection of low-burden voice data with a retrospective harvesting of subjects' smartphone and social media data. Consistent with the DSM-V requirements that symptoms persist for a minimum of two weeks, retrospective data is collected (scraped) from the prior two-week window. While the data harvested is not as clean (e.g. missing items) as in the case of longitudinal prospective studies, our retrospective data harvesting approach is lower burden, lower cost data, and yields larger datasets that are amenable to on-the-spot depression detection.

3. Study 1: investigating users' willingness-to-disclose voice, smartphone and social media data

A major constraint that researchers face when collecting mental health data, is the willingness of the study participants to allow their data to be collected or to disclose sensitive health information (e.g PHQ-9 scores). To ensure that participants were indeed willing to give the information we needed to detect depression, we first generated a list of data types and modalities that were relevant to mental health prediction, which could be retrieved from a smartphone or social media retrospectively Torous et al. (2015); Lane et al. (2010). Next, we designed the willingness to disclose Study, a survey to determine which of these data modalities the participants were willing to allow us to collect from them. The modalities included in this survey include: 1) *GPS location traces*, which have been shown in prior work to effectively predict depression Canzian and Musolesi (2015); Saeb et al. (2015); Wang et al. (2014). 2) *Sensor Data*, including gyroscope recordings Saeb et al. (2015). 3) *Text*, in various forms that prior work has shown to be predictive of depression, including Facebook posts Rodrigues, das Dores, Camilo-Junior, and Rosa (2016), Twitter posts Park, Cha, and Cha (2012); De Choudhury et al. (2013); Hasan et al. (2014b, a, 2017, 2018), and text messages. 4) *Images*, in the form of Instagram posts Reece and Danforth (2017b), and self-taken facial images Valstar et al. (2016), which have also been shown to be predictive of depression. 5) *Voice Samples*, that have also been shown to be predictive of depression Valstar et al. (2016). We asked participants about their willingness to record a voice clip. 6) *Other Logs*, finally we asked participants about their willingness to allow collection of smartphone logs, such as call logs, and browser history. While these modalities are relevant, their efficacy in retroactively collected data has not been demonstrated.

Due to the retrospective nature of the study, opportunistic data presents an additional challenge as some data types are not routinely logged without the user having installed a specialized smartphone application prior to data collection. For example, accelerometer data are not typically gathered and stored by the smartphones operating system. Thus, in addition to subjects' willingness to disclose a data type, its availability on the smartphone also needed to be factored into our decision on whether to include a given data type in our final study.

Novelty of Willingness to Disclose Study: Prior work has not determined what modalities subjects were willing to give in order to receive an assessment of their depression levels. Determining this was important as if subjects are unwilling to give certain required modalities, Moodable could not work. The contribution of the willingness to disclose study is determining what specific modalities participants were willing to allow Moodadble collect.

3.1. Design of the willingness to disclose study

We deployed our willingness to disclose study as an anonymous survey on Mechanical Turk, a crowdsourcing platform where volunteers participate in various user studies for a fee. We paid subjects in this study $0.05. Mechanical Turk reaches a wide variety of demographics Ipeirotis (2010); Mason and Suri (2012), and gathered samples are representative of the general population. Participants had to be over 18 years of age. Initially participants were given an introduction that motivated the need for a better depression detection method along with a description of a possible use case. They were thanked for helping to develop applications that could improve depression screening. Next, a series of demographic questions asked for the participants' gender, age, and employment. Participants were also asked if they had ever been treated in an emergency department. In the remaining questions, participants were asked to label their willingness to disclose different data modalities and PHQ-9 scores on a Likert scale ranging from Completely Unwilling, Somewhat Unwilling, Unsure, Somewhat Willing, to Completely Willing.

In the survey, participants were first asked for their willingness to disclose data from social media accounts to medical personnel. These types of data included their Twitter username, Twitter tweets, Facebook Posts, and messages on applications such as GroupeMe, Discord, and WhatsApp.

The second group of questions surveyed the willingness to disclose retrospective data stored on a typical smartphone such as GPS, gyroscope, accelerometer, browser history, call logs, and app usage logs. Finally, we surveyed their willingness to record a phrase and capture an image of their face.

In order to check for automated robot responses, the question about the participant's age was duplicated at the end of the survey, followed by an optional feedback textbox. Results of a chi test of the data and a list of the data modalities surveyed in the Willingness to Disclose Study are given in Table 1 and Table 2 respectively. The format of the survey is illustrated in Fig. A.8.

3.2. Findings of the willingness to disclose study for moodable

A total of 202 participants completed the Willingness to Disclose survey. Their answers are summarized in Table 2 and in Fig. A.7, sorted from most likely to share (top) to least likely to share (bottom). These results indicate that participants are most likely to share participatory data such as voice clips and self portraits. They are somewhat less likely to share certain opportunistic data such as their

Table 1
Chi Square Test. *** p-value 0.001, ** p-value 0.01, * p-value 0.05

	Voice Clips	Images of Face	Twitter Username	GPS Data	App Usage	Twiter Tweets	Gyroscope	Call Logs	Facebook Posts	Text Chat App	Browser History
Voice Clips		4.3	38.894***	28.413***	40.505***	42.301***	32.014***	44.357***	42.5***	54.862***	61.882***
Images of Face			25.795***	13.847**	24***	24.609***	15.278**	27.669***	24.069***	34.582***	39.874***
Twitter Username				4.717	1.594	3.094	8.379	2.031	5.805	5.291	8.935
GPS Data					2.407	2.696	1.008	3.376	3.431	6.887	8.047
App Usage						0.303	3.866	0.549	1.648	1.769	4.266
Twiter Tweets							3.077	0.637	0.576	1.078	3.184
Gyroscope								4.398	2.413	6.756	6.823
Call Logs									1.697	1.781	2.932
Facebook Posts										1.659	3.394
Text Chat App											1.995
Browser History											

Table 2
Willingness to share data with a medical professional. CU = Number of participants that responded Completely Unwilling, SU = Somewhat Unwilling, U = Unsure, SW = Somewhat Willing, CU = Completely Willing.

Mode of Acquisition	Data Modality	CU	SU	U	SW	CW
Participatory	Voice Clips	19	21	31	77	54
	Image of Face	29	27	29	63	54
Opportunistic	Twitter Username	70	15	16	59	42
	GPS Data	59	26	21	55	41
	App Usage	71	20	18	50	43
	Twiter Tweets	72	21	20	46	43
	Gyroscope	60	29	25	48	40
	Call Logs	75	19	20	50	38
	Facebook Posts	71	22	24	42	43
	Text Chat App	81	20	18	40	43
	Browser History	85	25	19	40	33

browser history and text chats, roughly 2 to 3 times more likely to share voice clips than their browser history.

3.3. Moodable feature engineering

Based on the results of our willingness to disclose study (Section 3.2) and the availability of retrospective data from subjects, we selected a set of data modalities to be collected by the Moodable smartphone App. We limited participatory data, which requires active participation from user to audio samples that could be collected with minimal effort from the participant. While sensor data such as accelerometer data would have been useful for determining the user's activity levels (which are indicative of depression levels), only certain data modalities are routinely stored by most phones and are available retrospectively. Specifically, while the smartphone users' call history and messages are routinely stored, their accelerometer data is not. A list of data modalities is given in Table 3.

To facilitate the deployment of machine learning classification models and ensure anonymity, the Featurizer module extracts 1967 features from the raw sensor data on the Smartphone. Below we give a description of the different features.

3.3.1. Phone usage features

- Number of Contacts in the phonebook.
- Call frequency for the previous 14 days. There are 14 call frequency features cf_i, where for each day, a different feature is computed as the average frequency of the previous 14 days (Equation (1)).

Table 3
A list of data types that were collected and the number of features extracted from each modality.

Mode of Acquisition	Data Type	Features Extracted
Opportunistic	Contacts	1
	Twitter Posts	4
	Text Messages	73
	GPS Data	5
	Call Logs	14
	Instagram Posts	18
Participatory	Audio Sample	1583

$$cf_i = \frac{\sum_{i=1}^{14} \text{Call Frequency at } t_{0-i}}{14}. \quad (1)$$

Text Features include features extracted from both text messaging applications and Twitter. Current evidence indicates that text in general (including usage of language, frequency of some words, user responsiveness and expression of certain affect) and especially text in social media can be a predictor of depression Rodrigues et al. (2016); De Choudhury et al. (2013); Hasan et al. (2014b, a, 2017, 2018).

3.3.2. Text message features

- Incoming text sentiment score moving average for the past 14 days. There are 14 text sentiment features ts_i, where for each day, a different feature is computed as the average sentiment of the previous 5 days (Equation (2)). Sentiment analysis was performed using TextBlob.[1] TextBlob is a python API for common Natural Language Processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification and translation. Textblob returned the polarity (−1 to 1) of and subjectivity (0–1) of input text.

$$ts_i = \frac{\sum_{i=1}^{5} \text{Incoming Text Sentiment at } t_{0-i}}{5}. \quad (2)$$

- Incoming text frequency moving average for the past 14 days. There are 14 text frequency features tf_i, where for each day, a different feature is computed as the average frequency of the previous 5 days (Equation (3)).

$$tf_i = \frac{\sum_{i=1}^{5} \text{Incoming Text Frequency at } t_{0-i}}{5}. \quad (3)$$

- A list of 45 features counting the number of words that fall into the 45 parts of speech defined by the Penn Treebank syntactic tree Taylor, Marcus, and Santorini (2003), which splits the English language into 45 types of words.

3.3.3. Twitter features

- Total Number of Twitter users that the participant follows.
- Total Number of Twitter users following the participant.
- Average daily Twitter likes on the participants posts during previous 14 days.
- Average daily Twitter re-tweets on the participant posts during the previous 14 days.

Instagram Features. Images in the form of Instagram posts Reece and Danforth (2017b), and self taken facial images which are often posted on Instagram, are shown to be predictive of depression Valstar et al. (2016). Based on this prior research we extracted the following features:

- Total Number of Instagram users that the participant follows.
- Total Number of Instagram users following the participant.
- Average filter usage frequency (how often does the participant uses any filter) during the previous 14 days.
- Average daily usage ratio of the following 8 filters: Amaro, Crema, Hefe, Inkwell, Rise, Valencia, Willow, X-Pro II for the previous 14 days. There is a feature for each filter, for a total of 8 features.
- Average daily Instagram Likes that the participant has received during the previous 14 days.
- Average daily Instagram Comments that the participant has received during the previous 14 days.
- Average daily Instagram Posts during the previous 14 days.
- Average daily pixel wise average values for Hue, Saturation and Value for all Instagram posts for the previous 14 days.
- Average face count in all Instagram posts (by the participant) for the previous 14 days. We do not analyze facial expressions, which have also been found to be predictive of depression.

GPS Features are extracted retrospective GPS data gathered from the participant's Google location services. Each participant had to log in to their Google account to give access to this GPS data. The raw data consists of a series of timestamps paired with longitude and latitude coordinates as well as placemarks identified by Google, and auto-detected activities such as running, walking, or biking. From this data, based on current literature Canzian and Musolesi (2015) we extracted the following features.

- Number of different places visited in the time span of two weeks. This was achieved by counting the placemarks.
- The total distance d covered during the timespan of 2 weeks. This distance is the sum of all distances between i ordered coordinate pairs as expressed in Equation (4).

[1] https://textblob.readthedocs.io/en/dev/.

$$d = \sum \sqrt{(latitude_i - latitude_{i-1})^2 + (longitude_i - longitude_{i-1})^2}. \quad (4)$$

- The maximum distance between two locations L_i and L_j. It corresponds to the distance between the two farthest places from the set of all places L a participant visited during the previous 2 weeks.

$$d_{max} = \underset{i,j \in L}{\operatorname{argmax}} \sqrt{(latitude_i - latitude_j)^2 + (longitude_i - longitude_j)^2}. \quad (5)$$

- Number of times an activity occurs, that is, the frequency of activities such as running, walking and biking during the previous 2 weeks. Due to the sparsity of the data, running, walking and biking are grouped together. Physical activity (quantity and intensity) has previously been shown to be negatively correlated with the incidence of depression Ströhle (2009).
- Total running distance of the participants in the previous 2 weeks.

Audio Features are extracted from participant's recordings using openSMILE[2] Eyben, Weninger, Gross, and Schuller (2013). SMILE performs audio signal processing including windowing, fast-fourier transform and conversion to Mel Frequency Cepstral Coefficients. It then extracts speech-related features such as signal energy, loudness, MFCC features and pitch. SMILE also computes statisical properties of the audio including its moments, percentiles, peaks, durations, DCT coefficients and zero-crossings. SMILE extracts a total of 1583 features. In order to reduce complexity and the high dimensionality of the SMILE features, we utilize a subset of 40 of the SMILE features.

3.4. Development of moodable classifier using machine learning

Model Training: The Machine Learning Model Generation Module trains depression and suicide indicator classification models using classical machine learning algorithms and feature extracted from the raw data. PHQ-9 scores are used as ground truth for the classification labels. For this task we compare several popular machine learning algorithms, namely, k-Nearest Neighbors (kNN), Support Vector Machine (SVM), and Random Forest (RF), commonly utilized for classification.

Model Deployment. These machine learning models are then deployed to the Moodable Predictor application and tested using features extracted from new participants. In order to measure the performance of machine learning classification models during the model evaluation phase, the depression labels predicted by the classification models were compared with the ground truth labels from the PHQ-9 survey. After deployment of Moodable, future usage will predict subjects' depression and suicidal ideation labels without having to reference their PHQ-9 scores or require them to pre-fill out PHQ-9 questionnaires. Below is a brief description of the machine learning algorithms that are used in this paper.

Support *Vector Machine (SVM)* Hearst, Dumais, Osuna, Platt, and Scholkopf (1998); Fu et al. (2008)*:* is a supervised machine learning algorithm that is commonly used for classification. More specifically, a SVM model is a representation of the data instances as points in space, mapped in such a way that instances belonging to different classes are divided by a hyperplane. The hyperplane is chosen so that its associated gap zone (margin) between classes is maximized. Feature vectors on the margin are also called support vectors. SVM can be tuned by applying different kernel functions that transform the points to support a nonlinear separation. After a model is trained, new instances are then mapped into that same feature space and their class is predicted (i.e. depressed or not depressed) based on which side of the gap they fall.

Random forest: is an ensemble algorithm that combines the predictions from multiple decision trees, with each tree trained using a random subset of the features and/or the data. The individual decision trees iteratively split the data set into separate subsets based on an attribute split-criteria on a so-called tree branch to generate homogeneous groups of points at each leaf, i.e., all mostly depressed or mostly non-depressed examples.

K-Nearest Neighbors algorithm (kNN): is a non-parametric method that maps data into a multi-dimensional feature space. It then classifies the new instance based on the class label of its *k* closest training examples in this feature space. The instance is classified by a majority vote of its neighbors, and it is assigned the most common class among its k nearest neighbors.

4. Study 2: evaluation study of Moodable's depression and suicidal ideation prediction

To evaluate the performance and how accurately the Moodable framework predicts depression and suicide ideation, we deployed the Moodable mobile application on Mechanical Turk and collected data from 335 participants who accepted to run the application. For this data collection, we enrolled only participants who had previously successfully completed at least 50 other tasks on Mechanical Turk. Additionally, Mechanical Turk participants are required to be at least 18 years or older. The design of this study is described below.

Depression Assessment using the PHQ-9: For the assessment of their depression state, we collected their answers to questions on the Patient Health Questionnaire (PHQ-9). The PHQ-9 is a well-validated measure of depression widely used in both research and clinical settings. While it cannot be used as a standalone assessment to formally diagnose clinical depression like lengthy Structured Clinical

[2] http://audeering.com/technology/opensmile/.

Interviews conducted by trained clinicians, this simple instrument can be self-administered to identify depressive symptoms and is the appropriate instrument for an online design study such as this one. The PHQ-9 requires subjects to respond on a 0–3 scale to 9 psychophysical questions that are indicative of depression. The questions ask users about their psychophysical state in the prior two weeks including feeling down, appetite, feeling tired, slowness in movement and about thoughts of suicide in question Q9. PHQ-9 was designed as a convenient assessment tool to identify depression. Since prior studies have shown that high PHQ-9 scores are indeed strongly correlated with clinical depression Kroenke, Kurt, and Williams (2001), we adopt it here for our analysis. We used question Q9 about thoughts of suicide on the PHQ-9 as our ground truth label for suicide ideation prediction.

As per Table 4, the diagnostic interpretation of PHQ-9 scores with higher scores are associated with more severe levels of depression. Subjects with scores in the 0–9 range are generally considered healthy, scores in the 10–19 range are considered to have mild depression, and subjects with scores 20 and above are considered to be severely depressed. Given that we work with regression scores, we investigated selecting different cut-off thresholds for the PHQ-9 score such that subjects with scores above that threshold were considered depressed, while those below considered not depressed Manea, Gilbody, and McMillan (2012).

4.1. Data collection methodology for study 2

As discussed in Section A.1, the Moodable mobile app is designed for ease-of-use, and minimal burden on the participants. The following workflow that was utilized in the study observes those principles.

Step 1: The permission of participants is requested (Fig. A.5b) to collect retrospective data in the background, while the rest of the experience was completed. Users could give permission to one, many or no data modalities. Completion of this step was not required for the study to proceed.

Step 2: Next, participants were required to fill out the PHQ-9 survey which is used to label the data (Fig. A.5d). Completion of this step was required for the study to proceed.

Step 3: Participants were asked to optionally record a voice sample (Fig. A.5c), while reading the phrase "The quick brown fox jumps over the lazy dog". Most participants completed this step, which was consistent with the willingness to disclose study.

Step 4: In addition, participants were asked to optionally sign into their social media accounts and grant access to their data. This step was designed to collect anonymous features and was chosen to be collected at the end due to findings by the first study that they would be less willing to provide this type of information.

Step 5: Finally, participants were given a completion code necessary to verify that they completed the study. This code was used to track participants and pay them.

A small monetary compensation was given as an incentive to participate in this study. To account for differences we observed in our willingness to disclose study, we compensated participants for giving different modalities at varying rates. The incentive structure was such that participants earned more money as they permitted the collection of data from more modalities. A full list of incentives (by modality) provided to the participants for permitting collection of this data, is given in Table 5.

4.2. Evaluation metrics

This section describes machine learning metrics that we use in the evaluation of our models. These metrics are based on the notation given in Table 6.

Sensitivity (Equation (6)), also called True Positive Rate (TPR) or *Recall*, refers to the ability of the medical diagnosis (test) to correctly identify depressed subjects.

$$Sensitivity(or, TPR) = \frac{TP}{TP + FN} = \frac{TP}{P}. \tag{6}$$

Table 4
Ranges of PHQ-9 scores and levels of depression severity.

PHQ-9 Score	Diagnosis
0–4	Minimal depression
5–9	Mild depression
10–14	Moderate depression
15–19	Moderately severe depression
20–27	Severe depression

Table 5
Participant compensation based on modality.

Data Modality	**Payout**
Text Logs, Call logs, Contacts	$0.40 (Base payout)
Twitter	$0.10
Google GPS Data	$0.30
Instagram	$0.10
Voice Recording	$0.10

Table 6
Classification notation.

Notation	Description
P	Number of actually positive samples, i.e., depressed.
N	Number of actually negative samples, i.e., non-depressed.
TP	Number of true positives, i.e., depressed correctly classified as depressed).
TN	Number of true negatives, i.e., non-depressed correctly classified as non-depressed.
FP	Number of false positives, i.e., non-depressed wrongly classified as depressed (false alarm, Type I error).
FN	Number of false negatives, i.e., depressed wrongly classified as non-depressed (Type II error).
TPR	True positive rate, sensitivity, or recall (Equivalent terms).
TNR	True negative rate, specificity.
PPV	Positive predictive value, precision.

Specificity (Equation (7)), also called the True Negative Rate (*TNR*), measures the proportion of actual negatives that are correctly identified as such. That is, the percentage of healthy people who are correctly identified as not having the depression condition.

$$Specificity(or, TNR) = \frac{TN}{N} = \frac{TN}{TN + FP}. \tag{7}$$

Precision (Equation (8)), also called Positive Predictive Value (*PPV*), is the fraction of relevant (positively labeled) instances among the predicted instances.

$$Precision(PPV) = \frac{TP}{TP + FP}. \tag{8}$$

F1 score (Equation (9)), is the harmonic mean of precision and sensitivity, combining these scores. Hence, *F*1 is a commonly accepted measure to evaluate the performance of learning methods especially when working with unbalanced data.

$$F1 = \frac{PPV \times TPR}{PPV + TPR}. \tag{9}$$

Accuracy (ACC) and the related Error Rate (*ERR*) (Equation (10)) capture the overall performance of the models, without taking in account sensitivity or specificity.

$$Accuracy(ACC) = \frac{TP + TN}{P + N}, ErrorRate(ERR) = 1 - ACC. \tag{10}$$

4.3. Descriptive analytics of The Moodable study

A total of 335 participants contributed one or more data modalities during the Moodable study. Their PHQ-9 scores varied from 0 to 27, and exhibited a quasi-bimodal distribution illustrated in Fig. 1g. In terms of data modalities, 266 participants gave access to their contact data (number of contacts), 74 to Twitter posts, 240 to text messages, 147 to GPS data, 36 to call logs, 27 to Instagram posts, and 266 recorded a voice sample.

Participants could refuse to contribute any data modality, and at the minimum, only the results of the PHQ-9 survey were required to complete the study. While the majority of the participants contributed at least one data modality, a small subset contributed all the modalities.

Fig. 1g shows the distribution of PHQ-9 scores for participants with *at least one* data modality, while Fig. 1h shows the distribution of PHQ-9 scores for participants with *all* the data modalities. The rest of the histograms in Fig. 1 show the distribution of all the data modalities and their PHQ-9 scores. Interesting patterns in willingness to give can be observed, including the fact that participants with higher PHQ-9 scores were generally more willing to disclose their data.

4.4. Predicting depression using moodable classifiers

For the task of depression detection, retrospective data shows great promise, with measured F1 scores of up to 0.766, sensitivity of 0.750, and specificity of 0.792 (Table 7).

In addition, we demonstrate that these results can vary between PHQ-9 classification cutoffs. We also experiment with different machine learning algorithms and demonstrate that some of the ML algorithms are better suited for this task. Full results of classifying PHQ-9 scores using different algorithms and cutoff points are summarized in Fig. 2. These charts capture the performance of the PHQ-9 classification models as described by Accuracy, Sensitivity, Specificy, Precision and F1 score. Each metric is coded in the point shape and color as described in the common legend. The Y-Axis gives the metric value, which has a common range of 0–1 for all the measured metrics. The X-Axis corresponds to the cutoff chosen for the depression classification. Fig. 2a, b, and 2c show the performance of different Moodable classifiers trained with different machine learning algorithms, kNN, RF and SVM respectively. Overall the RF algorithm shows a robust prediction, that appropriately balances sensitivity and specificity. Down-sampling was used to balance the dataset, resulting in a reduction in the number of instances of the majority class.

Results generated by the RF algorithm at commonly accepted cutoff points as per Manea et al. (2012) are summarized in Table 7. In

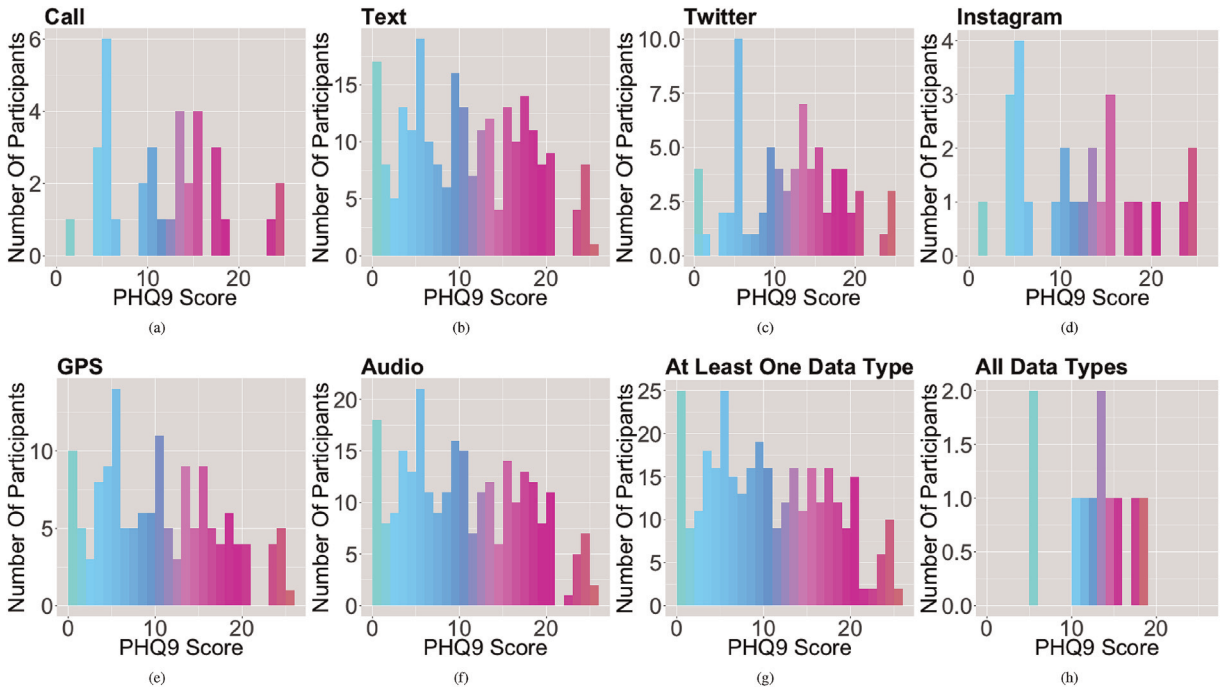


Fig. 1. Comparing the total PHQ9 Scores of different subsets of participants based on the availability of data modalities. 1a Call, 1b text message data, 1c Twitter data, 1d Instagram data, 1e GPS Data, 1f audio recording, 1g at least one data modality, and all data modalities. 1h.

Table 7
Classification results for predicting TOTAL score of PHQ-9 using ALL features with the dataset balanced using down sampling. Commonly used cutoff points are selected. Note that as the cut-off point changes, the number of features in each condition changes, therefore balancing produces datasets with different number of instances. TP, FP, TN, FN stand for True Positive, False Positive, True Negative and False Negative respectively.

Target	Cut.	ML	Inst.	TP	FP	TN	FN	Acc.	Err.	Prec.	Sens.	Spec.	F1
Total	5	RF	126	43	16	47	20	0.714	0.286	0.729	0.683	0.746	0.705
Total	15	RF	228	64	32	82	50	0.64	0.36	0.667	0.561	0.719	0.61
Total	10	RF	294	82	55	91	66	0.588	0.412	0.599	0.554	0.623	0.575
Total	20	RF	96	36	10	38	12	0.771	0.229	0.783	0.75	0.792	0.766

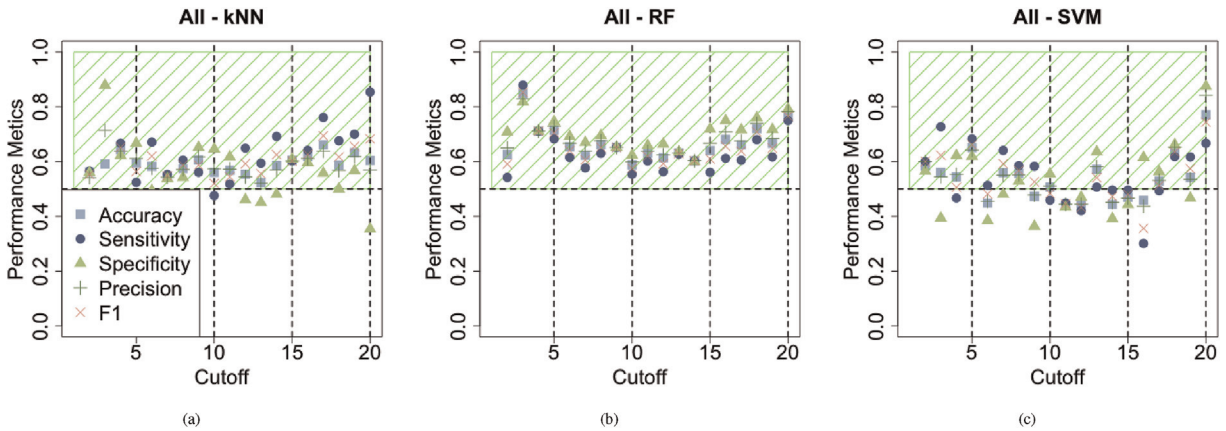


Fig. 2. The charts in Fig. 2 capture the performance of the PHQ-9 classification models as described by Accuracy, Sensitivity, Specificy, Precision and F1 score. Each metric is coded in the point shape and color as described in the common legend. The Y-Axis gives the metric value, which has a common range of 0–1 for all the measured metrics. The X-Axis corresponds to the cutoff chosen for the depression classification. Fig. 2a, b, and 2c, show the performance of the kNN, RF and SVM algorithms respectively.

this table, for full transparency, in addition to Accuracy, Error, Precision, Sensitivity, Specificity and F1 scores, we also report the number of True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) predictions. The number of instances listed in the table (column 4) reflects instance reduction due to downsampling.

We also analyze the effectiveness of predicting the PHQ-9 depression score using data from only a single data modality such as GPS, voice, or text. We note that this study is preliminary, as it our results depend on the amount of data gathered for each modality. However, it provides initial insights into the feasibility and promise of each modality to serve as foundation for depression prediction.

The charts in Fig. 3 capture the performance of the PHQ-9 classification models as described by Accuracy, Sensitivity, Specificity, Precision and F1 score for different modalities. Each metric is coded in the point, shape and color as described in the common legend. The Y-Axis gives the value of the metric, which has a range of 0–1 for all the measured metrics. The X-Axis corresponds to the cutoff

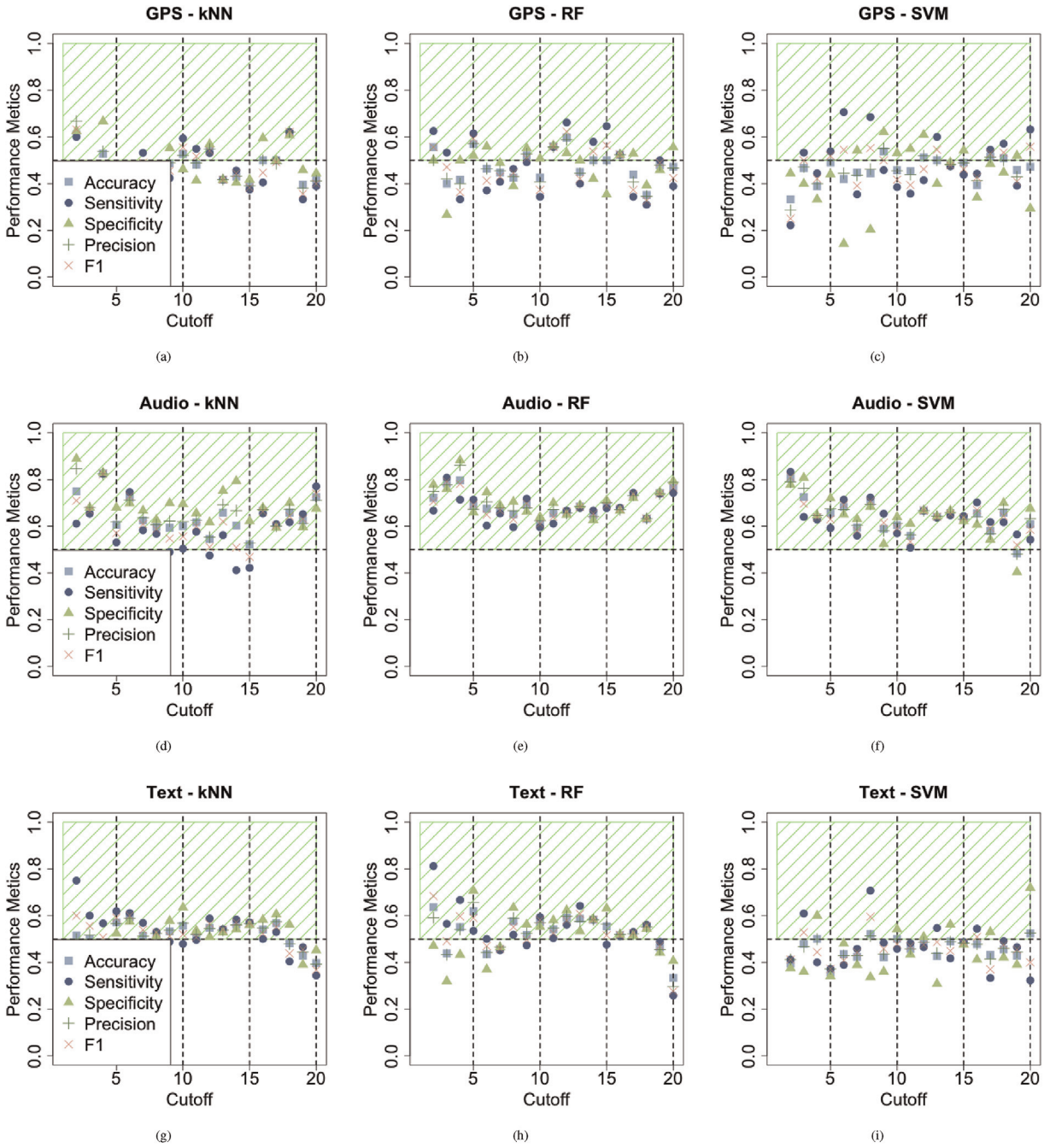


Fig. 3. Results of Predicting Total PHQ-9 Score Classification using the GPS (Fig. 3a for kNN, 3b for RF, and 3c for SVM results), Audio (Fig. 3d for kNN, 3e for RF, and 3f for SVM results) and Text (Fig. 3g for kNN, 3h for RF, and 3i for SVM results) features.

selected for the depression classification.

Some modalities are not included due to lack of sufficient instances at different cut-off points after downsampling. That is, when building modality-specific models, only participants that contributed data for that particular modality are being utilized. In some cases, these subsets of participants are considerably smaller than the total number of participants; which then is further reduced after downsampling.

4.5. Predicting suicidal ideation using moodable classifiers

We also evaluated the performance and accuracy of Moodable when Question 9 of the PHQ-9 (about suicidal thoughts) at different cutoffs. For the suicidal ideation task we achieve F1 measures of 0.848, sensitivity of 0.864, and specificity of 0.725, again showing promise in utilizing the combination of retrospective and participatory data for this task. Full results are given in Fig. 4 and are plotted based on Sensitivity and Specificity. Each data type is coded in the point shape and color as described in the legend. The Y-Axis corresponds to sensitivity, which has a range from 0 to 1. The X-Axis corresponds to specificity which also varies from 0 to 1. The upper right quadrant covered in slanted lines, contains modality-cutoff combinations with good sensitivity and specificity. The cutoff for the classification is included as a label next to each data pointQ9. Results of predicting Q9 Scores using the RF algorithm are summarized in Table 8.

5. Discussion

Due to the longitudinal nature of depression, most current research has focused on data collection over several weeks using a prospective study design (Ben-Zeev et al. (2015); Saeb et al. (2016); Wang et al. (2014); Ma et al. (2016); Valstar et al. (2016). While controlled data collection can produce quality datasets, prospective study designs are costly, hard to scale and unsuitable for on the spot mental health assessment (e.g. PHQ-9). In our work, we have studied the feasibility of using instantaneously collected data for this task using a mostly retrospective approach. That is, our data consists of retrospective data from the participants' smartphone usage, and low-burden participatory data that can be contributed within a few seconds to a minute. Retrospective data collection consisted of text messages, call logs, Twitter data, Instagram data, and GPS traces. Participatory data in our study was kept to a minimum, namely, a brief voice sample collection.

Preliminary results show that our proposed Moodable approach is promising, and opens up the opportunity for further in-depth

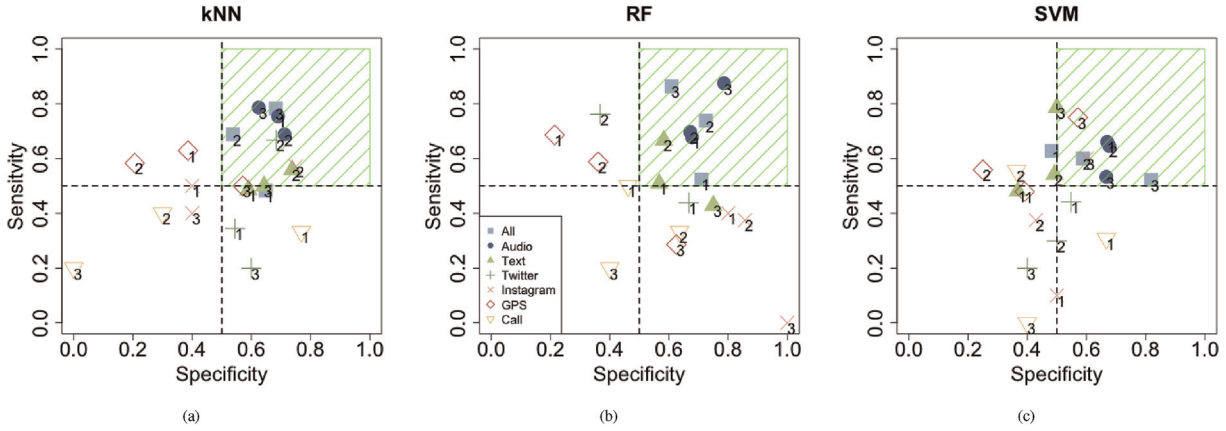


Fig. 4. Specificity and Sensitivity chart of the classification results for Q9 of the PHQ-9 Questionnaire using different Feature types. The results of kNN, RF and SVM algorithms, are given in Fig. 4a, b, and 4c respectively. The charts in Fig. 4 capture the performance of the Question 9 classification models as described by both Sensitivity and Specificity. Each data type is coded in the point shape and color as described in the legend. The Y-Axis corresponds to sensitivity, which has a range from 0 to 1. The X-Axis corresponds to Specificity which also varies from 0 to 1. The upper right quadrant covered in slanted lines, contains modality-cutoff combinations with good sensitivity and specificity. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 8
Classification results for predicting Q9 score of PHQ-9 using ALL features with the dataset balanced using down sampling. Commonly used cutoff points are selected. Note that as the cut-o_ point changes, the number of features in each condition changes, therefore balancing produces datasets with different number of instances. TP, FP, TN, FN stand for True Positive, False Positive, True Negative and False Negative respectively.

Target	Cut.	ML	Inst.	TP	FP	TN	FN	Acc.	Err.	Prec.	Sens.	Spec.	F1
Q9	1	RF	290	76	42	103	69	0.617	0.383	0.644	0.524	0.71	0.578
Q9	2	RF	160	59	22	58	21	0.731	0.269	0.728	0.738	0.725	0.733
Q9	3	RF	45	19	9	14	3	0.733	0.267	0.679	0.864	0.609	0.76

research into the development of a reliable instantaneous Mental Health screening instrument. In particular, the short duration, minimal interaction approach to data collection makes this framework ideal for on-the-spot depression assessment. Another important finding was that people were quite willing to share access to smartphone and social media data, which would likely make the use of this technology feasible in other populations.

Our sensitivity and specificity results are decent, but not perfect. This has implications on the utility of Moodable. If applied to large enough population, some false positive depression cases would be found and some true depression cases would be missed. However, we believe that having mostly accurate depression inferences albeit with a few errors is probably better than not reaching underserved populations at all due to resource constraints and/or due to those individuals simply not coming in contact with medical staff who can administer more reliable testing.

Moodable in its present design is a general instrument for quick screening that we show has reasonable accuracy. As with all screening methods, subjects who are flagged will be referred to a medical professional, who will then make the final assessment. It can be considered relevant for individualized depression detection because Moodable uses the data from an individual to compute their PhQ-9 score.

This work thus opens numerous avenues for further research, moving beyond subjects recruited from Amazon mechanical turk and including the evaluation of the Moodable approach on targeted sub-populations such as healthcare facilities and various patients of various demographics. In particular, the Moodable application could be utilized to screen patients for depression upon arrival at emergency rooms or as they wait in the doctors' office. This lightweight depression screening approach could be quite useful for screening patients who are being treated for ailments that are co-morbid to depression. Such a screen using Moodable could provide an initial indication of depression risk. Subsequently, a mental health therapist can be invited to additionally screen the patient if Moodable predicts a high PHQ-9 score indicative of depression. This proposed use would require further validation of Moodable among a medically ill population as some PHQ9 items map onto physical ailments, which could artificially inflate depression scores. Also, it is important to note that multiple ailments may have similar smartphone-sensed patterns. For instance, subjects voices are affected by depression, but also exhaustion, alcohol and many other conditions. This confounds depression inference and Moodable is best viewed as an initial screen that flags potentially depressed patients, which a therapist can then focus detailed examinations on.

As our techniques are refined further, we expect the inference models to become further refined. Lessons learned from this study could be incorporated into a future version of this Moodable app to increase its reliability as a depression screening instrument. More advanced feature extraction and generation strategies could be designed for each of the modalities to increase the quality of Moodable classification. For instance, for text mining, extracting additional features may be beneficial for extracting sentiment and other higher-order signals from the text data.

6. Conclusion

In this study we conducted a feasibility study, researched, designed and developed Moodable, a practical mobile application for near instantaneous smartphone and social media data collection for mental health screening. The core objective met by our work, is to develop a mental health assessment methodology that no longer requires compliance with, and utilization of a smartphone app for a long study periods. To evaluate the efficacy of this method, we use the Moodable app to collect data from 335 participants.

Our initial results are promising, in particular, for the depression classification task using out-of-the box machine learning models we achieve F1 measures up to 0.766, sensitivity up to 0.750, and specificity up to 0.792. In addition, Question 9 of the PHQ-9, which relates to suicidal ideation, was investigated. For this task we achieve F1 measures up to 0.848, sensitivity up to 0.864, and specificity up to 0.725. This result is encouraging and indicates that our Moodable approach of instantaneous depression assessment is viable.

Acknowledgements

The authors would like to thank ML Tlachac for reviewing the paper. Thanks also to the WPI computer science department for funding the Amazon mechnical turk studies associated with the Moodable survey and deployment studies.

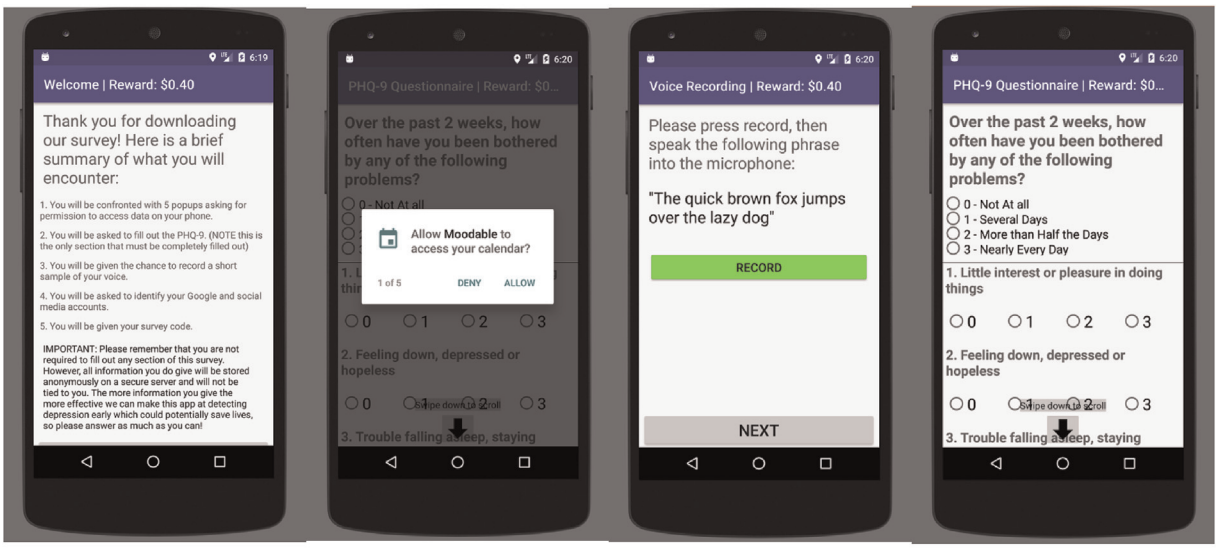
Appendix A. Design of the Moodable Application for Mental Health Data Sensing

The Moodable depression detection framework is implemented using a client-server paradigm. The Moodable Android smartphone app gathers retrospective and participatory smartphone data as well as social media data from subjects, which is then transmitted to the Moodable server where data analysis, machine learning and depression inference occurs. The Moodable application design and architecture are described below.

Appendix A.1. Overall Workflow of the Moodable Smartphone Application

In the design of the Moodable Smartphone application (app) with screens shown in Fig. A.5, the *user experience* was a key consideration. User permissions for gathering retrospective data (Fig. A.5b) are requested early on. Thereafter, data scraping occcurs in the background while participatory data (voice samples) is collected in the foreground (Fig. A.5c). Permissions for more crucial data such as the PHQ-9 (which are target labels for machine learning) are collected first in case the user decides to leave the study before giving all requested data. If a user decides to leave the study, they are given the option of deleting all data they have contributed to the study up till that point.

The application then follows a linear work flow, scraping retrospective participatory smartphone data, social media data and PHQ-9 scores. Sample permission request and user survey screens of the Moodable smartphone application are shown in Fig. 3. The Moodable Framework/Architecture and data collection mechanisms are described in Section A.2.



(a) (b) (c) (d)

Fig. A.5. Figure A.5a displays the Moodable introduction screen, explaining what data collection will be collected. In Figure A.5b the participant is asked for permission to access retrospective data, Figure A.5c displays the voice sample collection screen, and Figure A.5d is a screen shot from the PHQ-9 questionnaire.

Appendix A.2. Moodable Framework

The Moodable framework illustrated in Fig. A.6 has a client-server architecture.that is comprised of the *Moodable Data Collector*, *Data POST Processor*, *Machine Learning Model Generation* and *Predictor* modules described below.

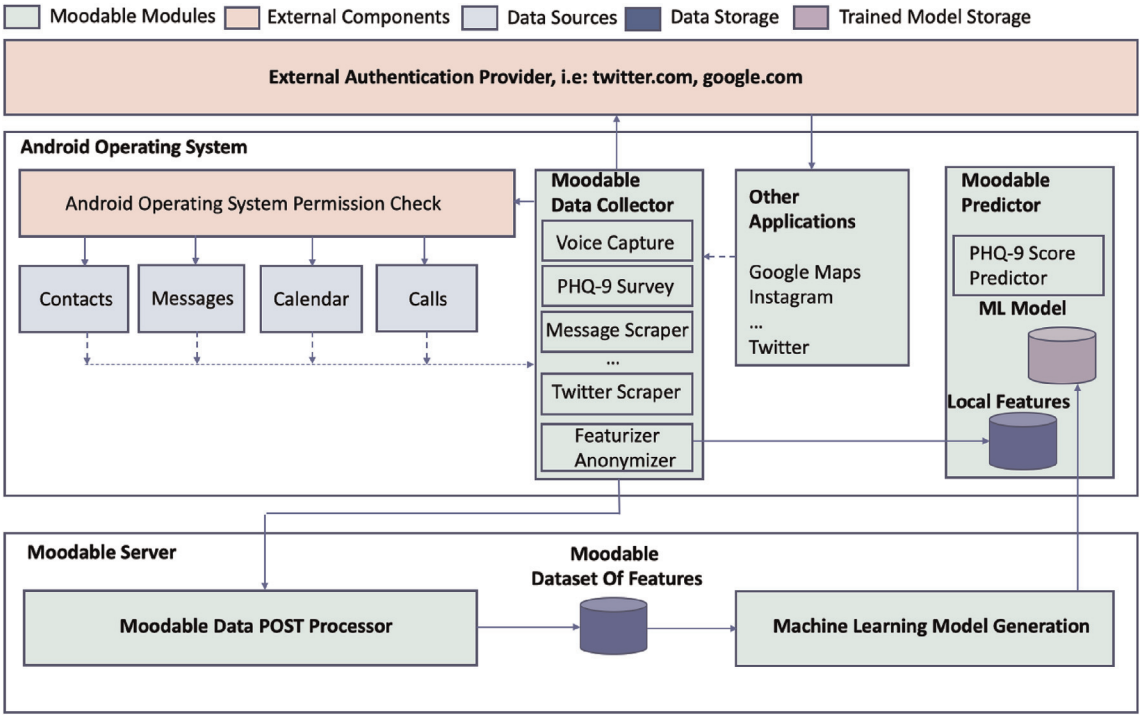


Fig. A.6. Moodable Framework/Architecture.

Moodable Data Collector is the main retrospective data scraper, and is implemented as a native Android Application that has multiple data collection units. This module utilizes the Android OS Permission System to gain access to local smartphone data, and external remote authentication systems such as Google[3] and Twitter[4] to gain access to social media and GPS data. All data is processed with user consent and is converted into 1697 features, which ensures user anonymity. A list of features extracted by the Featurizer sub-module, which is part of the Moodable smartphone application, is described in detail in Section 3.3.

Moodable Data POST Processor is the Moodable server API that receives data from the Moodable Data Collector and stores it as a feature dataset.

Machine Learning Model Generation utilizes machine learning algorithms to generate prediction models from the Moodable feature data. Different algorithms utilized in this module are described in further detail in Section 3.4.

Moodable Predictor is an Android application that uses Moodable Data features and previously trained machine learning models to make near-instantaneous assessments of the user's depression level, based on their retrospective data and voice samples. This module is part of the Moodable smartphone application and is further discussed in Section 3.4.

[3] https://google.com.
[4] https://twitter.com.

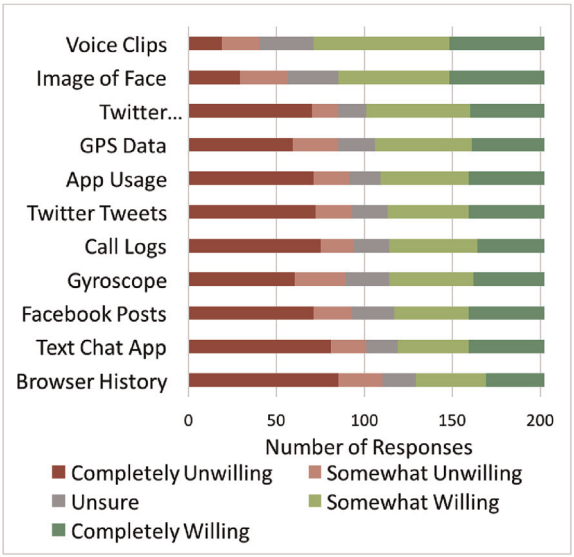


Fig. A.7. Willingness to share data with a Medical Professional, sorted by data type from most likely to share (top) to least likely (bottom).

How willing would you be to allow a member of the medical staff to run software that retrieves the following information stored on your phone and feeds it into a program that estimates your mental health status:

	Completely Unwilling (1)	Somewhat Unwilling (2)	Unsure (3)	Somewhat Willing (4)	Completely Willing (5)
Your phone's GPS data (1)	○	○	○	○	○
Your phone's gyroscope/accelerometer data (2)	○	○	○	○	○
Your phone's browser history (3)	○	○	○	○	○
Your phone's call logs (4)	○	○	○	○	○
Your phone's app usage data (which Apps are open and how long) (5)	○	○	○	○	○

Fig. A.8. Willingness to disclose data with a Medical Professional Survey sample.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed.). Washington, DC: Autor.

Ben-Zeev, D., Scherer, E. A., Wang, R., Xie, H., & Campbell, A. T. (2015). Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric Rehabilitation Journal, 38*, 218.

Burke, J., Estrin, D., Hansen, M., Parker, A. R. N., Reddy, S., & Srivastava, M. (2006). *Participatory sensing, wsw' 06 at sensys' 06*.

Canzian, L., & Musolesi, M. (2015). Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing* (pp. 1293–1304). ACM.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *ICWSM, 13*, 1–10.

Epstein, R. M., Duberstein, P. R., Feldman, M. D., Rochlen, A. B., Bell, R. A., Kravitz, R. L., et al. (2010). i didn't know what was wrong:" how people with undiagnosed depression recognize, name and explain their distress. *Journal of General Internal Medicine, 25*, 954–961.

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on multimedia* (pp. 835–838). ACM.

Fu, C. H., Mourao-Miranda, J., Costafreda, S. G., Khanna, A., Marquand, A. F., Williams, S. C., et al. (2008). Pattern classification of sad facial processing: Toward the development of neurobiological markers in depression. *Biological Psychiatry, 63*, 656–662.

Hasan, M., Agu, E., & Rundensteiner, E. (2014a). Using hashtags as labels for supervised learning of emotions in twitter messages. In *ACM SIGKDD workshop on health informatics, New York, USA*.

Hasan, M., Rundensteiner, E., & Agu, E. (2014b). Emotex: Detecting emotions in twitter messages. In *Academy of science and engineering (ASE), social computing 2014*. May 27-31, 2014, Stanford, CA. USA.

Hasan, M., Rundensteiner, E., & Agu, E. (2018). Automatic emotion detection in text streams by analyzing twitter data. *International Journal of Data Science and Analytics*, (1–17).

Hasan, M., Rundensteiner, E., Kong, X., & Agu, E. (2017). Using social sensing to discover trends in public emotion. In *Semantic computing (ICSC), 2017 IEEE 11th international conference on (pp. 172–179)*. IEEE.
Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications, 13*, 18–28.
Ipeirotis, P. G. (2010). Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students, 17*, 16–21.
Katikalapaudi, R., Montgomery, F., Wunsch, D., Lutzen, K., & Chellappan, S. (2012). Associating internet usage with depressive behavior amongst college students. In *IEEE tech soc mag 2012*. IEEE.
Kroenke, R. L. S., Kurt, & Williams, J. B. (2001). The phq–9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16*, 606–613, 0.
Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., & Campbell, A. T. (2010). A survey of mobile phone sensing. *IEEE Communications Magazine, 48*.
M De Choudhury, E. H., & Counts, S. (2013). Predicting postpartum changes in emotion and behavior via social media. In *Proc. ACM CHI*. ACM.
M Park, C. C., & Cha, M. (2012). Depressive moods of users captured in twitter. In *Proc. ACM SIGKDD workshop on healthcare informatics*. ACM.
Manea, L., Gilbody, S., & McMillan, D. (2012). Optimal cut-off score for diagnosing depression with the patient health questionnaire (phq-9): A meta-analysis. *Canadian Medical Association Journal, 184*, E191–E196.
Mason, W., & Suri, S. (2012). Conducting behavioral research on amazon' s mechanical turk. *Behavior Research Methods, 44*, 1–23.
Ma, X., Yang, H., Chen, Q., Huang, D., & Wang, Y. (2016). Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th international workshop on audio/visual emotion challenge* (pp. 35–42). ACM.
Moreno, M. A., Egan, K. G., Cox, E., Young, H., Gannon, K. E., Jelenchick, L. A., et al. (2011). Feeling bad on facebook: Depression disclosures by college students on a social networking site. In *, Vol. 28. Depression and anxiety* (pp. 447–455).
Munmun De Choudhury, S. C., Gamon, M., & Horvitz, E. (2013). Predicting depression via social media. In *Proc in ICWSM*. ACM, 2013.
Park, M., Cha, C., & Cha, M. (2012). Depressive moods of users portrayed in twitter. In *, Vol. 2012. Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)* (pp. 1–8). New York, NY: ACM.
Park, S., Kwak, J., Cha, M., Lee, S. W., & Jeong, B. (2013). Activities on facebook reveal the depressive state of users. *Journal of Medical Internet Research, 15*(10), e217. Oct 2013.
Rapaport, M. H., Clary, C., Fayyad, R., & Endicott, J. (2005). Quality-of-life impairment in depressive and anxiety disorders. *American Journal of Psychiatry, 162*, 1171–1178.
Reece, A. G., & Danforth, C. M. (2017a). Instagram photos reveal predictive markers of depression. *European Physical Journal Data Science, 6*, 15, 2017.
Reece, A. G., & Danforth, C. M. (2017b). Instagram photos reveal predictive markers of depression. *European Physical Journal Data Science, 6*, 15.
Richards, D. A., Ekers, D., McMillan, D., Taylor, R. S., Byford, S., Warren, F. C., et al. (2016). Cost and outcome of behavioural activation versus cognitive behavioural therapy for depression (cobra): A randomised, controlled, non-inferiority trial. *The Lancet, 388*, 871–880.
Rodrigues, R. G., das Dores, R. M., Camilo-Junior, C. G., & Rosa, T. C. (2016). Sentihealth-cancer: A sentiment analysis tool to help detecting mood of patients in online social networks. *International Journal of Medical Informatics, 85*, 80–95.
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.
Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., et al. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research, 17*.
Siu, A. L., Bibbins-Domingo, K., Grossman, D. C., Baumann, L. C., Davidson, K. W., Ebell, M., et al. (2016). Screening for depression in adults: Us preventive services task force recommendation statement. *Jama, 315*, 380–387.
Ströhle, A. (2009). Physical activity, exercise, depression and anxiety disorders. *Journal of neural transmission, 116*, 777.
Taylor, A., Marcus, M., & Santorini, B. (2003). The penn treebank: An overview. In *Treebanks* (pp. 5–22). Springer.
Torous, J., Staples, P., & Onnela, J.-P. (2015). Realizing the potential of mobile mental health: New methods for new data in psychiatry. *Current Psychiatry Reports, 17*, 61.
Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., et al. (2016). Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge* (pp. 3–10). ACM.
Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., et al. (2014). Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing* (pp. 3–14). ACM.
Weist, M. D., Rubin, M., Moore, E., Adelsheim, S., & Wrobel, G. (2007). Mental health screening in schools. *Journal of School Health, 77*, 53–58.
Witters, D., Liu, D., & Agrawal, S. (2015). *Depression costs us workplaces $23 billion in absenteeism* (gallup).