Evaluating Machine Learning Algorithms for Prediction of the Adverse Valence Index Based on the Photographic Affect Meter

Gatis Mikelsons*
University of Oxford
Oxford, UK
gatis.mikelsons@jesus.ox.ac.uk

Abhinav Mehrotra
University College London, The Alan Turing Institute
London, UK
a.mehrotra@ucl.ac.uk

Mirco Musolesi
University College London, The Alan Turing Institute
London, UK
m.musolesi@ucl.ac.uk

Nigel Shadbolt
University of Oxford
Oxford, UK
nigel.shadbolt@cs.ox.ac.uk

ABSTRACT

In recent years, numerous studies have explored the use of machine learning algorithms for supporting applications in social and clinical psychology. In particular, there is an increasing prevalence of smartphone-based techniques for collecting data through embedded sensors and efficient in-situ questionnaires. Models are then built to explore the patterns between these data types.

In this paper, we study the application of machine learning for the task of predicting mental states of adverse valence, based on the Photographic Affect Meter data. We present a technique for daily aggregation, which is designed to detect significant negative events. A variety of features is used as input, including GPS-based metrics and features assessing social interactions, sleep and phone usage. Experimental evidence is presented, which suggests that machine learning algorithms could successfully be employed for such a prediction task.

CCS CONCEPTS

• **Human-centered computing** → *HCI design and evaluation methods*; *Empirical studies in ubiquitous and mobile computing*.

KEYWORDS

mobile sensing, behaviour modelling, digital mental health, Adverse Valence Index

ACM Reference Format:
Gatis Mikelsons, Abhinav Mehrotra, Mirco Musolesi, and Nigel Shadbolt. 2019. Evaluating Machine Learning Algorithms for Prediction of the Adverse Valence Index Based on the Photographic Affect Meter. In *The 5th ACM Workshop on Mobile Systems for Computational Social Science (MCSS'19), June 21, 2019, Seoul, Republic of Korea.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3325426.3329948

*The research was initiated during an internship at the Alan Turing Institute.

1 INTRODUCTION

Today's mobile phones and wearables have become highly personal devices able to assist us in a variety of day-to-day situations. They feature sophisticated sensors capable of capturing different types of contextual information such as location, movement, audio environment, proximity to other objects, collocation with other devices and many others [3, 5, 10, 15]. Recent studies have demonstrated the potential of exploiting mobile sensing data to learn and, potentially, predict users' mood and well-being [1, 4, 11, 12, 21, 23, 25]. Indeed, smartphones and wearables are increasingly seen as very powerful tools for research in social and clinical psychology [17]. A variety of modalities has been used as features for building machine learning (ML) models for this class of applications. In particular, mobility data have been shown to be very promising for effectively training such prediction models [4, 16, 23]. More recent approaches in the field include the usage of photos-based mood questionnaires and exploring potential mechanisms for sending feedback to the user [2, 6, 7, 20].

In this study, we consider the use of the Photographic Affect Meter (PAM) [19] as a quantitative indicator of users' mood. PAM provides an alternative to Likert-scale-based verbal questionnaires by presenting users with a series of photographs and asking them to pick the one that best captures their current mood. Being a simple and quick one-item questionnaire, the PAM test is a good example of Ecological Momentary Assessment (EMA) [18], whereby the participants are assessed at opportune moments during their normal routines, so as to capture more genuine mood phenomena.

Following research on human affect [22, 24], instant mood in the PAM inventory is conceptualized as a two-dimensional phenomenon characterized by "valence", or quality of feeling, and "arousal", or degree of activation. A range of values for these two independent dimensions is considered possible, forming a two-dimensional grid. In particular, valence can be positive or negative, indicating the quality of feeling, and arousal can be high or low, indicating the level of energy. A four-by-four grid is used in the PAM inventory to cover this space. Figure 1 (a) provides an illustration of the PAM questionnaire, with the 16 photos arranged on the grid. Figure 1

(b) is a response histogram for our dataset, with the valence and arousal dimensions explicitly labelled.

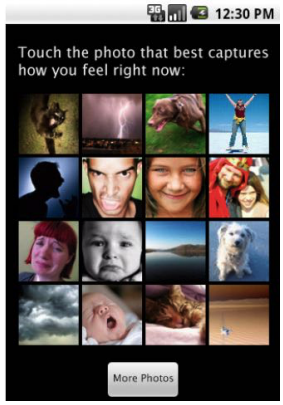
We decompose the PAM score, focusing on the valence dimension only, and develop a measure for assessing daily well-being, termed the *Adverse Valence Index* (AV index). The aim of the analysis presented in this work is to predict the value of the AV index for a particular user on a specific day, using a variety of daily input features. To evaluate our approach, we use the public version of the StudentLife dataset [26] that contains rich multimodal data, collected about 49 student participants over >10 weeks, covering a single academic term at Dartmouth College. The dataset offers information related to a variety of dimensions such as physical activity, mobility, social interactions, phone usage and others. In this study, both self-reported and objectively measured data types are harnessed to form the set of features.

Three popular ML models are then trained on the dataset to show modest evidence of learning despite the limited size of data, the complexity of the learning task, and numerous potential sources of noise. An additional contribution of this work consists in the proposal for a time-frame that should be chosen for the calculation of the features, given the fact that PAM is a testing tool concerned with instant mood.

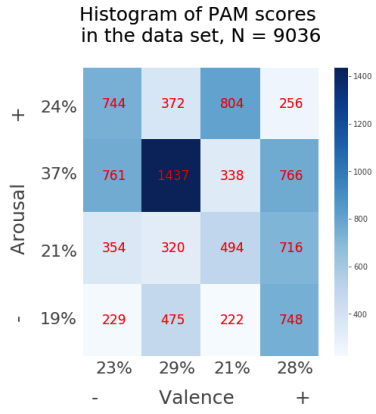
2 THE PAM DATASET

The StudentLife dataset contains ~9000 individual PAM responses, with the PAM questionnaire being sent to students typically 1-5 times per day. The photos the recipients receive (chosen from a curated set) are always arranged within the two-dimensional grid of valence and arousal, each measured on a scale of 4 units. As the first step in our procedure for daily aggregation, the variation in per-user reporting baseline is considered. Figure 1 (c) plots the mean reported arousal and valence per participant, together with error bars of 2 standard deviation (2σ) total length, with the scores translated to a 0-4 scale for clarity of presentation. It is observed that participants vary noticeably in terms of the average valence and arousal they report.

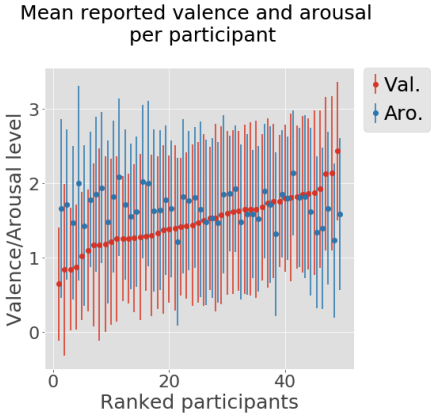
We consider the valence dimension to be the more relevant one for health applications. Previous work [16] has demonstrated the potential for treating the stress level prediction problem as a three-class classification task, based on the per-user median level of stress. In our case, such a transformation yields a very peaked distribution (>60% weight in the middle category) that is difficult to model. Instead, we convert the problem into a two-class classification task by introducing the AV index. Designed specifically to indicate adversely negative mental events, which we think is the most clinically important aspect of the PAM score, the AV index is defined to equal 1 for days with at least one below-1σ valence report and 0 otherwise. A fractional AV index value can be interpreted as the probability of a user encountering adverse valence (AV) on a particular day. It is worth noting that even a single strongly negative AV experience can potentially have a lingering quality, affecting the user over the time-frame of the day. With this transformation, the dataset is balanced, with the AV index equalling 1 in ~60% of the cases. For 3 users, even the lowest valence score is within the 1σ range; they are excluded from the analysis.



(a)



(b)



(c)

Figure 1: (a) A sample instance of the PAM questionnaire on a participant's phone screen. Reproduced with permission from the creators of the PAM test [19]. (b) A histogram of the PAM responses found in the StudentLife dataset [26]. Both the valence and the arousal dimensions are shown. (c) Plotting the per-user mean valence and arousal, along with error bars of 2 standard deviation total length.

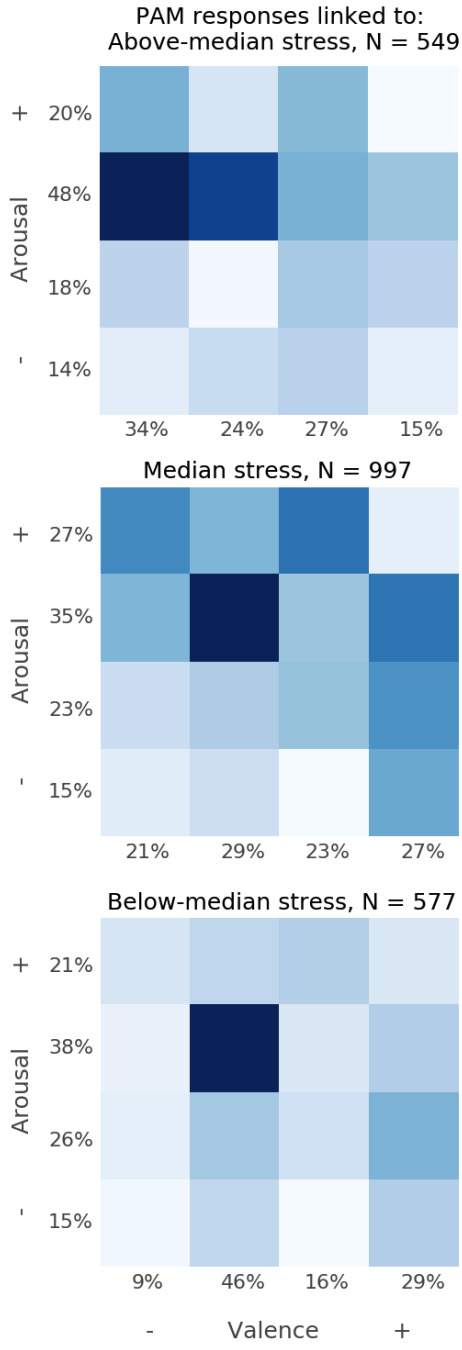


Figure 2: Linking the students' PAM responses to their stress EMAs within the same hour. The Likert scale stress responses have been re-classified after per-user median subtraction.

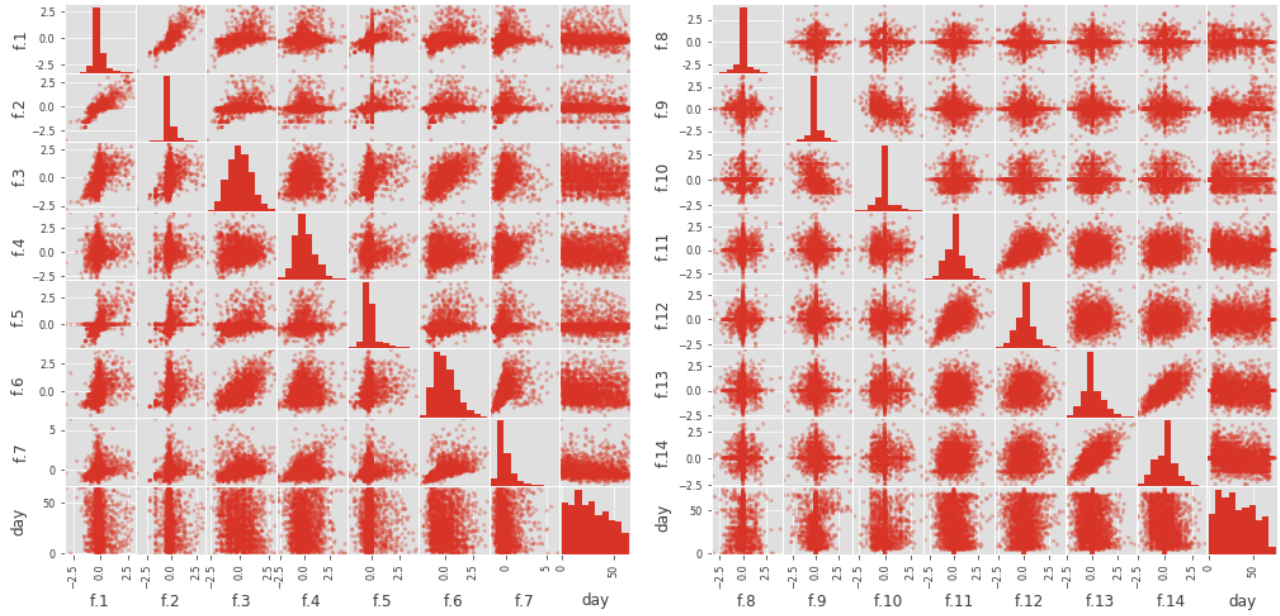
Another important measure of psychological well-being to be found in the StudentLife dataset is students' reports on their level of stress on a 5-item Likert scale (~2000 in total). Given the relative novelty of PAM, it is interesting to link these two types of mood reporting to offer further insight into the PAM inventory. Following previous research [16], the stress reports are considered after per-user median subtraction, grouping them into three classes. For each user's stress responses, we search for any PAM reports occurring within the same clock hour, averaging in the case that several responses are retrieved. Figure 2 shows the results of this analysis. Stress is generally interpreted to be a mental state of negative valence and high arousal. The prevalence of PAM responses in the top-right quadrant for the case of above-median stress gives support to this.

Considering the results presented in Figure 1(b), it appears that adverse valence in this particular dataset would typically be a high-arousal, stress-like state, rather than a low-arousal state more suggestive of depression. Moreover, the PAM response that occurs most often is a tile associated with anger or annoyance. It is interesting to note that even in low-stress situations this tile has the highest rate of response (see Figure 2). We were not able to collect any evidence for explaining this result. A possible, yet unproven, hypothesis is that students would simply report dislike towards being interrupted by the PAM questionnaire. If true, this would represent a source of noise in the PAM inventory that would need addressing in future research.

3 APPROACH

Taking the day as the unit of time, features of different kinds are aggregated for the task of AV index prediction. Based on the recent literature [4, 13, 14], we add 7 GPS-based metrics: 1) total distance covered, 2) maximum 2-point separation, 3) number of different places visited by per-user tiled area grid approximation, 4) difference in sequence of tiles covered, compared to previous day, 5) distance entropy, 6) number of non-routine clusters visited, 7) time spent on non-routine tiles. A selection has been made from the features proposed in previous studies in order to retain the metrics that least correlate with one another. The calculation of these metrics follows the steps outlined in the publications cited. We use a 50-metre radius for the DBSCAN clustering algorithm to obtain routine clusters for individuals, and tiles of area 700m^2 are used for the tiles approximation.

The self-reported corpus of the StudentLife dataset is made use of for features about sociability and sleep. These features include: 8) number of people the user has been in contact with on the day (face-to-face, phone, internet), 9) hours slept, 10) quality of sleep. Moreover, the objective sensing data are used to obtain: 11) total duration of conversations recorded on the day, 12) number of conversations registered by the sensors on the day 13) total duration for which the mobile phone has been locked for significant periods (>1 hour), 14) number of significant (>1 hour) phone lock periods. Features 8-14 are chosen as easily interpretable factors under the participant's own control. Other potentially interesting features in the dataset, such as self-reported exercise time and walking time, have been omitted due to lower response rates.



(a) (b)

Figure 3: (a) Scatter plot matrix for the seven GPS-based features (features 1-7), after per-user normalization. (b) Scatter plot matrix for the seven actionable features (features 8-14), after per-user normalization. To illustrate temporary variation, the day of response is included as the final row/column in both panels.

An illustration of these features and their relationships, after per-user standardization, is provided in Figure 3 and also in the Appendix, where numeric pairwise correlation figures are given. It is interesting to note, for example, that sleep quality correlates negatively with sleep duration, and that self-reported sociability (feature 8) shows little correlation with the objective conversations-related features (features 11, 12). The day of response has been added as a "feature" here to illustrate patterns of temporary evolution. For example, there appears to be a pattern of increase in phone usage with time. As the term progresses, students also appear to be spending less time on conversations. All the GPS metrics indicate a small but consistently negative correlation with the day of the term.

A slight decrease in the average AV index with progressing weeks can also be noticed in the dataset, which is likely to be related to the dynamics of the academic term [26]. Accordingly, we introduce a temporal feature: 15) one-hot indicator for the first half of the term. No major variation is found between weekdays and weekends. Finally, to go beyond the treatment of each day's mood as an independent event, we add the user's AV index from the previous day and that of the day before as two more features. Consequently, in our analysis, 17 features in total are explored.

Assuming perfect sensor functioning and user responsiveness, collecting daily readings from 49 users over slightly more than 10 weeks would result in a dataset of ~3500 points. Because of excluding some of the users and due to the sparse nature of some of the measurement types, we are able to construct a dataset of <1600 points. In merging the different feature types, missing entries have sometimes been replaced with the population mean (0 for standardised features, 0.6 for previous days' AV index). There is a trade-off between being able to build a larger dataset, on the one hand, and introducing too much noise, on the other, by "interpolating" the data as described. However, given the uneven nature of datasets involving human participants, the deployment of some technique for interpolation, at least in parts, seems almost unavoidable.

4 EVALUATION

The results of our experimental evaluation are now presented. We use three algorithms of increasing complexity: logistic regression, the linear support vector classifier and a neural network model to explore non-linear dependencies. All three models are compared against the baseline of a features-blind classifier.

The logistic regression model and the linear support vector classifier are implemented and trained using standard packages in Scikit-learn. In the case of the support vector classifier, an optimization of validation accuracy with respect to the penalty parameter C used by Scikit-learn is performed. As far as the neural network model is concerned, we build a network made of 2 fully-connected hidden layers in addition to the input and output layers. The two hidden layers are made of 50 nodes each and use the $tanh$ activation function, batch normalization, and drop-out regularization of rate 0.3. The output layer uses the $softmax$ activation function, with batch normalization applied, and has two output nodes corresponding to the two values of the AV index. It is found with the current dataset

Table 1: Performance metrics for modelling the AV index

	Mode classifier	Neural network	Linear SVC	Logistic regression
Accuracy	0.57	0.66(2)	0.66(2)	0.66(2)
Precision	0.57	0.67(1)	0.68(1)	0.69(2)
Recall	1	0.81(4)	0.74(3)	0.76(4)

that a range of neural-network architectures (varying the number of layers and the number of nodes) yields a similar result. In addition to drop-out regularization, early stopping is employed (setting patience = 4), based on the validation loss. The network is trained using categorical cross-entropy loss and the Adam optimizer [9].

Table 1 presents the results of the prediction task using our ML models. Given the balanced nature of the two classes, accuracy is treated as the main performance metric. We employ a stratified 5-fold split for performance evaluation and the results in Table 1 provide the 1σ error computed across the folds. As the baseline, the "Mode classifier" is used, which guesses the most frequent category (AV index of 1) irrespective of the input feature values.

As seen in Table 1, all three ML models show similar performance, outperforming the Mode classifier baseline and showing modest evidence of learning. Given the limited amount of data (<1600 points), the fact that the neural network is unable to outperform the more basic models is not surprising. Rather, it offers evidence that appropriate regularisation has been deployed.

So far as judgements on the relative importance of the feature types can be made, we find tentative evidence, by experimenting with feature removal, of the temporal features (features 15-17) being the most important in our prediction task, with the GPS metrics (features 1-7) and the actionable features (features 8-14) sharing a joint second place. One possible explanation could be that the effects due to the temporal features are less affected by user heterogeneity and therefore less likely to be averaged out when many individuals are considered in a single dataset. These are early hypotheses, however, which would need further justification in the context of larger datasets and perhaps also more personalised models [8].

5 CONCLUSION

We have presented an evaluation of the use of ML algorithms for predicting the *Adverse Valence Index* (AV index), as derived from the Photographic Affect Meter, using a variety of behavioural and temporal features. We have evaluated our approach using the StudentLife dataset. Experimental results illustrating the potential of ML algorithms for such a prediction task have been discussed.

The findings presented in this study add to the evidence that information elicited from mobile phones can be exploited to predict human mental state. We hope that our work would encourage further research in social and behavioural science.

ACKNOWLEDGMENTS

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. This work was also supported in part through the award of a DPhil studentship from the Department of Computer Science at the University of Oxford as part of its contribution to the EPSRC SOCIAM Project EP/J017728/2. We would like to thank the authors of the StudentLife dataset for making it available for the research community.

REFERENCES

[1] Jorge Alvarez-Lozano, Venet Osmani, Oscar Mayora, Mads Frost, Jakob Bardram, Maria Faurholt-Jepsen, and Lars Vedel Kessing. 2014. Tell me your apps and I will tell you your mood: correlation of apps usage with bipolar disorder state. In *PETRA'14*.

[2] Min S Hane Aung, Faisal Alquaddoomi, Cheng-Kang Hsieh, Mashfiqui Rabbi, Longqi Yang, John P Pollak, Deborah Estrin, and Tanzeem Choudhury. 2016. Leveraging multi-modal sensing for mobile health: a case review in chronic pain. *IEEE journal of selected topics in signal processing* 10, 5 (2016), 962–974.

[3] Jeffrey A Burke, Deborah Estrin, Mark Hansen, Andrew Parker, Nithya Ramanathan, Sasank Reddy, and Mani B Srivastava. 2006. Participatory Sensing. In *World Sensor Web Workshop*.

[4] Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *UbiComp'15*.

[5] Tanzeem Choudhury, Gaetano Borriello, Sunny Consolvo, Dirk Haehnel, Beverly Harrison, Bruce Hemingway, Jeffrey Hightower, Karl Koscher, Anthony LaMarca, James A Landay, et al. 2008. The Mobile Sensing Platform: An Embedded Activity Recognition System. *IEEE Pervasive Computing* 7, 2 (2008), 32–41.

[6] Alex W DaSilva, Jeremy F Huckins, Rui Wang, Weichen Wang, Dylan D Wagner, and Andrew T Campbell. 2019. Correlates of Stress in the College Environment Uncovered by the Application of Penalized Generalized Estimating Equations to Mobile Sensing Data. *JMIR mHealth and uHealth* 7, 3 (2019), e12084.

[7] Sophia Haim, Rui Wang, Sarah E Lord, Lorie Loeb, Xia Zhou, and Andrew T Campbell. 2015. The mobile photographic stress meter (MPSM): A new way to measure stress using images. In *Adjunct proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2015 ACM international symposium on wearable computers*. ACM, 733–742.

[8] Natasha Jaques, Sara Taylor, Akane Sano, Rosalind Picard, et al. 2017. Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation. In *IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*. 17–33.

[9] Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR'15*.

[10] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. 2010. A survey of mobile phone sensing. *IEEE Communications Magazine* 48, 9 (2010).

[11] Huitian Lei, Ambuj Tewari, and Susan Murphy. 2014. An Actor-critic Contextual Bandit Algorithm for Personalized Interventions using Mobile Devices. In *NIPS 2014 Workshop on Personalization: Methods and Applications*.

[12] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. 2016. Towards multimodal anticipatory monitoring of depressive states through the analysis of human-smartphone. In *Adjunct UbiComp'16*.

[13] A Mehrotra, S Muller, G Harari, S Gosling, Cecilia Mascolo, M Musolesi, and Peter Jason Rentfrow. 2017. Understanding the role of places and activities on mobile phone interaction and usage patterns. *IMWUT* 1, 3 (2017).

[14] Abhinav Mehrotra and Mirco Musolesi. 2017. Designing effective movement digital biomarkers for unobtrusive emotional state mobile monitoring . In *MobiSys'17 Adjunct*.

[15] Abhinav Mehrotra, Veljko Pejovic, and Mirco Musolesi. 2014. SenSocial: A Middleware for Integrating Online Social Networks and Mobile Sensing Data Streams. In *Middleware'14*.

[16] Gatis Mikelsons, Matthew Smith, Abhinav Mehrotra, and Mirco Musolesi. 2017. Towards Deep Learning Models for Psychological State Prediction using Smartphone Data: Challenges and Opportunities. In *NIPS Workshop on Machine Learning for Healthcare 2017*. Long Beach, CA, USA.

[17] Geoffrey Miller. 2012. The Smartphone Psychology Manifesto. *Perspectives on Psychological Science* 7, 3 (2012), 221–237.

[18] Debbie S Moskowitz and Simon N Young. 2006. Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology. *Journal of Psychiatry and Neuroscience* 31, 1 (2006), 13.

[19] John P Pollak, Phil Adams, and Geri Gay. 2011. PAM: a photographic affect meter for frequent, in situ measurement of affect. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 725–734.
[20] Mashfiqui Rabbi, Min Hane Aung, and Tanzeem Choudhury. 2017. Towards health recommendation systems: an approach for providing automated personalized health feedback from mobile data. In *Mobile Health*. Springer, 519–542.
[21] Kiran K. Rachuri, Mirco Musolesi, Cecilia Mascolo, Jason Rentfrow, Chris Longworth, and Andrius Aucinas. 2010. EmotionSense: A mobile phones based adaptive platform for experimental social psychology research. In *UbiComp'10*.
[22] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.
[23] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of Medical Internet research* 17, 7 (2015), e175.
[24] Klaus R Scherer. 2005. What are emotions? And how can they be measured? *Social science information* 44, 4 (2005), 695–729.
[25] Yoshihiko Suhara, Yinzhan Xu, and Alex'Sandy' Pentland. 2017. DeepMood: Forecasting Depressed Mood Based on Self-Reported Histories via Recurrent Neural Networks. In *WWW'17*.
[26] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *UbiComp'14*.

CORRELATION MATRICES

The tables below report the pairwise Spearman coefficients of correlation for a selection of features. The feature definitions are given in the text.

GPS features, N=1947:

	f.1	f.2	f.3	f.4	f.5	f.6	f.7
f.2	0.77						
f.3	0.58	0.40					
f.4	0.14	0.13	0.14				
f.5	0.59	0.53	0.31	0.06			
f.6	0.44	0.28	0.63	0.15	0.22		
f.7	0.32	0.24	0.40	0.28	0.18	0.66	
day	-0.09	-0.07	-0.14	-0.03	-0.02	-0.10	-0.10

Actionable features, N = 2401:

	f.8	f.9	f.10	f.11	f.12	f.13	f.14
f.9	-0.01						
f.10	0.01	-0.33					
f.11	0.07	-0.06	0.01				
f.12	0.03	-0.04	-0.00	0.53			
f.13	0.00	-0.03	-0.01	0.09	0.08		
f.14	0.00	-0.06	-0.03	0.15	0.09	0.71	
day	-0.02	0.06	0.01	-0.12	0.01	-0.12	-0.18