Sequence Multi-task Learning to Forecast Mental Wellbeing from Sparse Self-reported Data

Dimitris Spathis
University of Cambridge, UK
Dept. of Comp. Sci. & Tech.
ds806@cam.ac.uk

Sandra Servia-Rodriguez
University of Cambridge, UK
Dept. of Comp. Sci. & Tech.
ss2138@cam.ac.uk

Katayoun Farrahi
University of Southampton, UK
Dept. of Electr. & Comp. Sci.
k.farrahi@soton.ac.uk

Cecilia Mascolo
University of Cambridge, UK
Dept. of Comp. Sci. & Tech.
cm542@cam.ac.uk

Jason Rentfrow
University of Cambridge, UK
Dept. of Psychology
pjr39@cam.ac.uk

ABSTRACT

Smartphones have started to be used as self reporting tools for mental health state as they accompany individuals during their days and can therefore gather temporallyfi ne grained data. However, the analysis of self reported mood data offers challenges related to non-homogeneity of mood assessment among individuals due to the complexity of the feeling and the reporting scales, as well as the noise and sparseness of the reports when collected in the wild. In this paper, we propose a new end-to-end ML model inspired by video frame prediction and machine translation, that forecasts future sequences of mood from previous self-reported moods collected in the real world using mobile devices. Contrary to traditional time series forecasting algorithms, our multi-task encoder-decoder recurrent neural network learns patterns from different users, allowing and improving the prediction for users with limited number of self-reports. Unlike traditional feature-based machine learning algorithms, the encoder-decoder architecture enables to forecast a sequence of future moods rather than one single step. Meanwhile, multi-task learning exploits some unique characteristics of the data (mood is bi-dimensional), achieving better results than when training single-task networks or other classifiers.

Our experiments using a real-world dataset of 33, 000 user-weeks revealed that (i) 3 weeks of sparsely reported mood is the optimal number to accurately forecast mood, (ii) multi-task learning models both dimensions of mood –valence and arousal– with higher accuracy than separate or traditional ML models, and (iii) mood variability, personality traits and day of the week play a key role in the performance of our model. We believe this work provides psychologists and developers of future mobile mental health applications with a ready-to-use and effective tool for early diagnosis of mental health issues at scale.

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330730

KEYWORDS

multi-task learning, sequence learning, recurrent neural networks, mood forecasting

ACM Reference Format:
Dimitris Spathis, Sandra Servia-Rodriguez, Katayoun Farrahi, Cecilia Mascolo, and Jason Rentfrow. 2019. Sequence Multi-task Learning to Forecast Mental Wellbeing from Sparse Self-reported Data. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330730

1 INTRODUCTION

Mood and general wellbeing have been associated with clinical outcomes. Self-reported sadness was found to be an indicator of depression [5], while self-reported happiness is linked to longevity [21] and reduced mortality risk [1]. The pervasiveness of smartphones and wearable devices has enabled timely monitoring of mental health and wellbeing, allowing a near real-time detection of clinical outcomes and relapses. The penetration of mobile devices has also introduced scale: many more individuals could be reached and assessed. However, most of the studies on investigating how mood reports collected using smartphones can help to improve mental health and wellbeing have been conducted through controlled experiments with limited number of participants and observations [11, 12, 20, 23]. It is not clear whether previousfindings and methodologies can be transferred to large, noisy and sparse datasets collected in the wild [16, 18]. Robust methodologies for anticipating mood issues from sparse data are key to the widespread adoption of smartphones for mental health support.

Hidden Markov Models (HMMs), autoregressive models and regression algorithms have been applied to sequence prediction. HMMs and autoregressive models operate by default on single sequences, being unable to learn patterns from several users. Traditional feature-based ML algorithms such as linear regression, random forests or support vector regressors, can solely predict one scalar value. They do not support an extended forecast horizon without feeding through the previous prediction as its new input [22], which unavoidably introduces compounding errors that skew the input distribution for future prediction steps. As a special kind of Neural Network, Recurrent Neural Networks (RNN) have become increasingly useful in modeling sequential, high-dimensional,

non-linear data [8]. Simple RNNs work by mapping an input sequence to an output sequence of the same length. By incorporating encoder-decoder architectures, recent RNN models (called sequence-to-sequence or *seq2seq*) can map an input sequence to an output sequence of any arbitrary length, converting RNNs the state-of-the-art in Natural Language Processing for machine translation and speech processing [19] since they can map, for example, a phrase in French to a phrase in English of different length.

Psychologists have proposed tools or scales that facilitate users to assess their mood such as the Affect Grid [15] scale, a 2-dimensional grid, where the x-axis indicates the feeling in terms of its positiveness or negativeness while the y-axis indicates its intensity. Independently of the scale used, forecasting mood requires to forecast more than one dimension. The modularity of deep neural networks enables learning similar tasks in parallel (e.g. predict both happiness and calmness in the same model). They do so by either shaping the data as multidimensional tensors so that thefi nal layers output multiple sequences, or by building a network with individual *branches* or *forks* that optimize different losses and backpropagate the error to the shared layers [14].

One caveat of deep neural networks is their inability to explain why the input data leads to a specific output. Although the scope of model interpretability is very wide, including causality, informativeness, and transparency, at least post-hoc interpretations and visualizations are needed to qualitatively evaluate what a model has learned. This is especially relevant in clinical setups where clinicians can only rely on interpretable models to make informed decisions.

In this paper, we propose a deep encoder-decoder, multi-task sequence model to forecast users' future sequence of moods from noisy and sparse self-reported moods collected in the wild using an affect grid scale. By exploiting the similarity between both components of the mood –valence and arousal, our multi-task model is able to more accurately learn both variables. We evaluate our model on a large scale dataset that includes 177, 111 total mood self-reports from 566 users collected with a smartphone application during more than 3 years [16]. By using a sequence of self-reported moods collected during the past 3 weeks, our model outperforms state-of-art feature-based ML methods on next day's forecasting, while also being able to forecast further into the future.

This paper makes the following contributions:

- We propose and adapt an end-to-end, stand-alone model inspired by video frame prediction [17] and machine translation [19], to forecast sequences of future moods –valence and arousal– from previous self-reported moods.
- Our evaluation on real world data reveals that (i) our model forecasts tomorrow's mood with ±0.14 minimum error, and 7 days later with ±0.16 error on the affect grid, and (ii) that the multi-task model trained to learn valence and arousal simultaneously is more accurate than independent models trained on each dimension separately, especially for arousal.
- We show the internal learned *black-box* representations of the deep neural networks and observe that different neurons learn different non-linear sequential patterns, which helps understand the complex trajectories of future mood.
- An exploratory post-hoc analysis reveals that the accuracy of the learned model is related to the day of the week, personality traits and mood variability. Specifically, our model performs better for *open* users and on weekends.

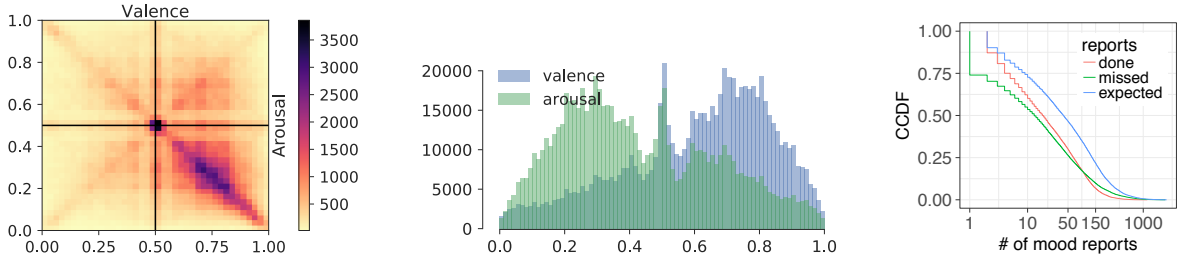
We believe this work provides psychologists and developers of future mobile mental health applications with a ready-to-use and effective tool for early diagnosis of mood issues at scale.

2 THE PROBLEM AND THE DATA

We start by analyzing self-reporting behaviour in smartphone applications for mood monitoring. To do so, we consider a large scale dataset of users' sensed and self-reported data gathered with a mobile phone application for Android designed to study subjective wellbeing and behavior in the wild [16]. From February 2013 until October 2016, this application collected 735, 778 self-reported data points from 17, 251 users through surveys presented on the phone via experience sampling, and passive behavioral data from physical and software sensors in the phone (accelerometer, microphone, location, text messages, phone calls, etc.). For this analysis, we solely consider self-reported mood collected graphically using the Affect Grid scale [15]. This is a square grid that measures the valence (pleasant-unpleasant feelings) and arousal (sleepiness-activeness) in the horizontal and vertical axes, respectively. Twice per day, between 8AM and 10PM and with a difference of at least 120 minutes, participants were asked to report their mood. Figure 1 shows the distribution of self-reports in the affect grid (a) and the distribution of each dimension –valence and arousal (b). At different stages, participants were requested to complete profile-related questionnaires covering a broad range of topics such as demographics, personality and sociability using Likert scales. We will only use such metadata during post-hoc analysis in order to gain insights about model performance at user and group levels.

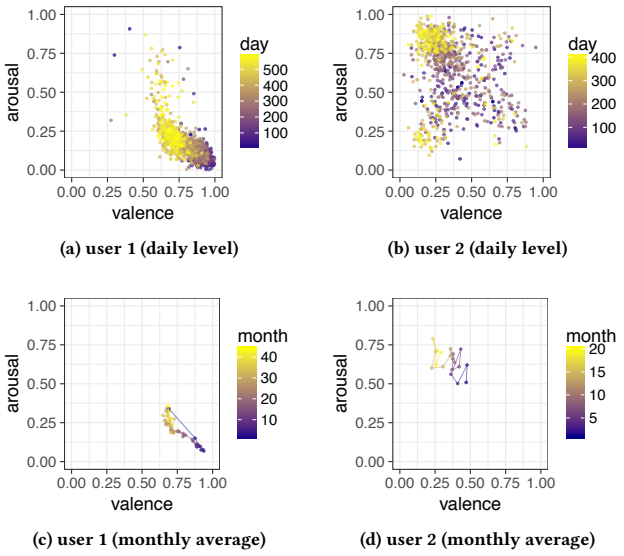
Sparsity of mood reports. A quick inspection of the dataset revealed that users did not always report even if they were prompted to do so. Figure 1c shows the complementary cumulative distribution function (CCDF) of moods reported per participant, including those they were prompted tofill *(expected)*, the ones they were prompted tofi ll but did not *(missed)* and the ones theyfi lled (*complete*). This is also true for users that used the app for large periods. Indeed, those that used the app for 45 or more consecutive days (16, 8% of the users) reported, on average, less than half of the expected times. The absence of mood reports might be a symptom of boredom or dissatisfaction with the app, but could also be indicative of mental disorders, especially in cases where users have been reporting anger and depression related feelings.

Variability of mood reports. A longitudinal exploration of the mood reported shows large differences between users in the way they report, in terms of both specific positions on the grid and area covered. Figure 2 shows moods reported by two different individuals who have self-reported for, at least, 300 days, and who are representative of two different behavioral patterns we identified. Thefi rst user (user 1 in Fig. 2) reports consistently over time, both in the short and long term, and her reports are concentrated on the positive and calm area of the grid. As time goes, her reports progressively become more negative (but still in the positive area)



(a) Heatmap of valence and arousal. (b) Distribution of valence and arousal. (c) Sparsity of mood self-reports.

Figure 1: Aggregate 735,778 data points of self-reported mood scores from 17,251 users. (a) Most users report neutral and calm-happy mood on the affect grid. (b) The two multi-modal distributions have different skews. (c) CCDF of mood reports.



(a) user 1 (daily level) (b) user 2 (daily level)

(c) user 1 (monthly average) (d) user 2 (monthly average)

Figure 2: Longitudinal mood monitoring for 2 users.

and active. The second user (user 2 in Figure 2) has quite the opposite behavior. That is, at the beginning (purple dots), she reported mixed affect states during consecutive days (purple dots are almost all over the grid), but, as time goes, her reports concentrate mainly in the negative and active area.

Given the longitudinal variability of user moods, forecasting current and future mood entails different level of difficulty for different users. For example, it is expected to be much easier to predict for user 1 than for user 2. We will go back to this in §6.

3 METHOD

We now describe our methodology to build a mood prediction framework able to cater for the level of noise and sparsity of this kind of data. It consists of a sequence to sequence neural network that learns from previous mood sequences to predict future ones (Fig. 3). The main advantage of this approach is that, unlike traditional regression, it allows to regress to multiple steps into the future by mapping the input sequence to an arbitrary output sequence. The model is composed of an encoder and decoder, each of which are RNNs. The individual units that build up the recurrent networks are *Long Short-Term Memory* units. We use a simplified adaptation of the sequence to sequence model proposed for machine translation [19] as we know exactly how many steps in the future we want to predict, while in translation this length varies (e.g. a sentence in English might have different length in French).

Long Short-Term Memory (LSTM). RNNs are well known to be hard to train especially when employed on sequences with long-term dependencies and patterns [8]. LSTMs overcome this problem by introducing *memory cells*.

Each LSTM unit has a cell composed of state c_t at time t, also called memory unit. Sigmoid gates allow the reading and modification of this unit via the input gate i_t, the forget gate f_t, and the output gate o_t. Each unit has four paths, the three gates and the input. At every time-step the unit receives at its four paths inputs coming from two sources: the current mood $\mathbf{x}_t$ and the previous hidden states of all the units in the same layer $\mathbf{h}_{t-1}$. Internally, each gate has another source, the previous cell state c_{t-1}. The inputs are summed along with a bias term b and the total input goes through a sigmoid logistic function. The total input of the input path goes through a non-linearity (*tanh*). The result is multiplied with the activation of the input gate, and then added to the current cell state after multiplying the previous cell state c_{t-1} with the forget gate activation f_t. Thefi nal output h_t is calculated by multiplying the output gate o_t with the updated cell state c_t passed through a non-linearity. This happens in a single layer of LSTM units during training (Fig. 3). Our encoder and decoder layers are LSTM layers like the ones described here. The above updates are summarized as follows:

$$
\begin{aligned}
\mathbf{i}_t &= \sigma(W_{xi}\mathbf{x}_t + W_{hi}\mathbf{h}_{t-1} + W_{ci}\mathbf{c}_{t-1} + \mathbf{b}_i) \\
\mathbf{f}_t &= \sigma(W_{xf}\mathbf{x}_t + W_{hf}\mathbf{h}_{t-1} + W_{cf}\mathbf{c}_{t-1} + \mathbf{b}_f) \\
\mathbf{c}_t &= \mathbf{f}_t\mathbf{c}_{t-1} + \mathbf{i}_t tanh(W_{xc}\mathbf{x}_t + W_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c) \\
\mathbf{o}_t &= \sigma(W_{xo}\mathbf{x}_t + W_{ho}\mathbf{h}_{t-1} + W_{co}\mathbf{c}_t + \mathbf{b}_o) \\
\mathbf{h}_t &= \mathbf{o}_t tanh(\mathbf{c}_t)
\end{aligned}
\tag{1}
$$

where $\sigma(\cdot)$ is the sigmoid function, $\mathbf{i}_t$, $\mathbf{f}_t$, and $\mathbf{o}_t$ are the input, forget and output gates, respectively. Since we predict precise mood scores and not binary outcomes, we use the Mean Squared Error (MSE) as the evaluation metric and the loss function to train the model:

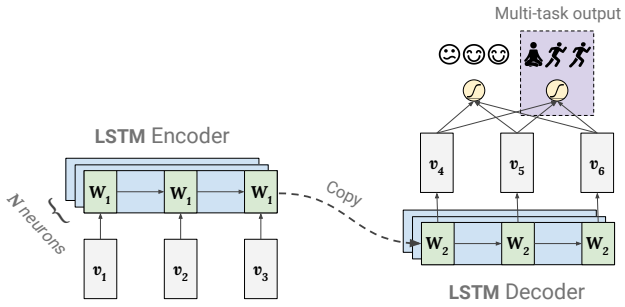


Figure 3: LSTM Encoder-Decoder model. The mood sequence (v_1,v_2,v_3) passes through an LSTM (states W_1), gets transformed to a single vector (dotted) and decoded through another LSTM (W_2) that predicts future mood sequences (v_4,v_5,v_6). Two fully-connected layers are applied to every time-step of the output (yellow circle), one for valence and one for arousal (purple box).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (2)$$

where Y_i is the vector of n predictions and $\hat{Y}_i$ is the ground truth.

Encoder-Decoder LSTM. The above structure of the LSTM unit outputs the same number of time-steps as the input sequence. Hence, $\mathbf{h}_t$ has to connect to additional fully-connected layers to reach the desired dimension of thefi nal output. However, by using simple fully-connected layers we dismiss the sequential nature of the data. Instead, we use a standard LSTM layer as an *Encoder* in order to map the past mood into afi xed length representation with the size of the prediction, and then another LSTM layer as a *Decoder* to reconstruct the original sequence in future steps. Thefi xed length representation is feasible through a layer (dotted arrow in Fig. 3) called copy (or *repeat*), which repeats the Encoder 2D output as many times as the output length, in order to create a 3D input for the Decoder. For example, given a week of past mood, we may want to forecast the next 2 days: the encoder learns to map the past week sequence into a decoded vector of the next 2 days. A similar model has been applied successfully to video frame prediction, which the authors called *LSTM Future Predictor Encoder-Decoder* [17].

Multi-task Encoder-Decoder LSTM. Multi-task learning is a transfer learning method in which a model learns to predict simultaneously two or more similar tasks. It has been used to reduce overfitting (with *auxiliary targets*), produce better data representations, and in general to improve accuracy in neural networks [14]. Specifically in deep neural networks, this multi-target setup forces the shared weights of the network to optimize both tasks and consequently learn internal representations that reflect on both.

4 EVALUATION

We now evaluate our deep encoder-decoder, multi-task sequence model to forecast users' future sequence of moods. Wefi rst consider a simplified, single-task version of this model and study the optimal length of the input sequence, i.e., number of days in the past, that minimizes the prediction error (§4.1). We then explore

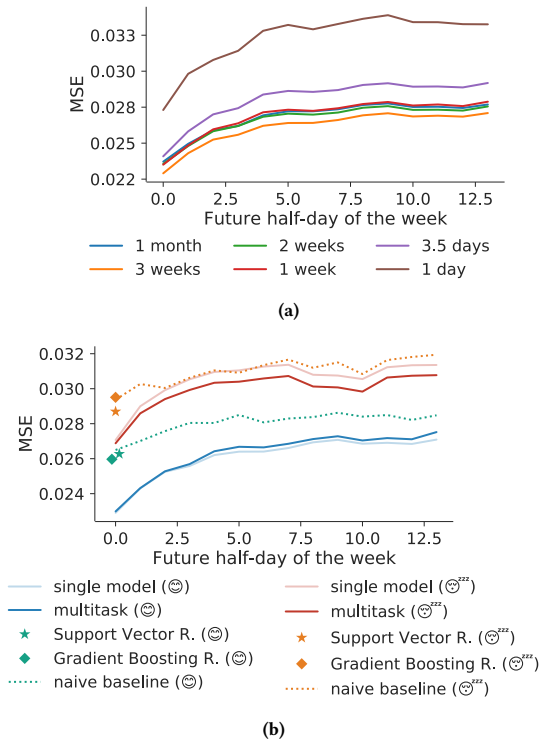
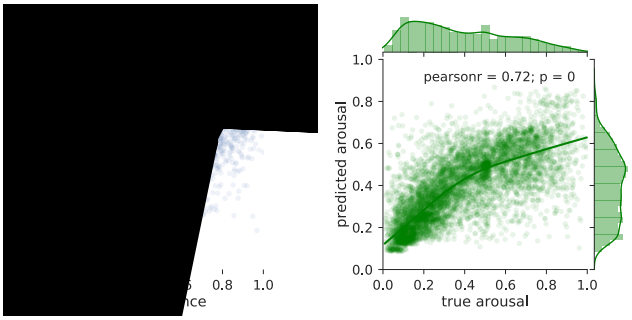


Figure 4: (a) How many days should we look into the past for accurate valence prediction? (b) Which is the best model to forecast mood using 3 weeks of past data (smiley=valence, sleepy face=arousal)?

the performance of multi-task learning for predicting valence and arousal simultaneously (§4.2).

Data preprocessing. We selected users who reported more than 100 days (> 200 half-days reports) between May 2013 and October 2016 –the period when the application was most active, ending up with 177, 111 unique self-reports from 566 participants. This is the sample we used in our experiments. This subset has similar statistics to the initial sample: $\mu_{val} = 0.57$ (±0.17) and $\mu_{aro} = 0.46$ (±0.19) for the initial dataset, and $\mu_{val} = 0.60$ (±0.17) and $\mu_{aro} = 0.43$ (±0.18) for the subset (± denotes one standard deviation).

We used a sliding window with step 1 over the mood sequences for each user, obtaining consecutive sequences of 4 weeks of past and 1 week of future moods. We then remove those samples whose *future* moods contained missing values, since that would make training more difficult, resulting in 33, 461 final sequences of past and future moods. For the past weeks, we found that only 6k out of 33k (20%) user-weeks had no missing values. Every sequence has on average 15% missing values (i.e. $\mu_{spar} = 6.36$ (±8.69) missing time-steps out of 42 steps for a past of 3 weeks). For these *past* sequences, we replaced the missing values with zeros. We tested other data imputation methods likefi lling with the median of the sequence, or min-max scaling to [0.05, 1] but we did not observe any considerable gain on the validation set. To be able to distinguish between real and missing values, we used a Masking layer to skip



(a) True and predicted valence. (b) True and predicted arousal.

Figure 5: Predictive performance of the multi-task model for thefi rst future mood forecast ($p = 0$ denotes a $p < 0.001$).

the missing values during training. In order to prevent over-fitting, we split the data to 20% testing and 80% training, ensuring that the users in the test set are completely disjoint, and do not overlap with those in the training set. During training, we use 10% of the training set as validation set to tune our models' hyper-parameters.

Implementation. Our implementation is based on Keras (with Tensorflow backend). We trained two separate models, one for valence and one arousal, for the Encoder-Decoder LSTM model. The input and output data is a matrix $\mathbf{M} \in \mathbb{R}^{s \times t}$, where s are the samples and t the time-steps. After grid-search we found the best-performing number of LSTM units for the Decoder and the Encoder (80 units each). The input layer is a standard Masking layer that skips the time-steps of missing values. In every LSTM layer, a rectified linear unit (ReLU) as well as recurrent dropout of 0.5 probability is applied, to prevent overfitting. Thefi nal layer is a standard feed forward neural layer (*Dense*) with a linear activation, that is being applied to every time-step. The objective function minimizes the MSE since this is a regression problem, while the backpropagation optimizer is *Rmsprop*. We train for 300 epochs or until the validation loss stops improving for 15 consecutive epochs.

Baselines. We compared our proposed model against a naive baseline based on simply using the average of the past days for predicting future moods (excluding the missing values), a Support Vector Regressor (SVR) and a Gradient Boosting Regressor (GBR). We used the Python's library *sklearn* implementations of an SVR with a radial basis function (RBF) kernel, and a tree-based ensemble model for the GBR, which is reportedly the state-of-art in feature-based machine learning [13]. SVRs and GBRs do not operate on sequences and assume feature independence, so we extracted 8 representative features from the time-series (non-missing counts, mean, std, min, max, and 25%, 50% and 75% quantiles) and normalized them column-wise to $[0, 1]$. We again exclude the missing values when we calculate those features. We only report the prediction for the *first* future mood since these models cannot regress to sequences.

4.1 How many days should we look back?

Building on previous research[18] we now investigate *how many days we should look back* (i.e., how much history to consider) to predict the sequence of next week's moods. We conduct different experiments tofi nd out which period, −4 weeks, 3 weeks, 2 weeks, 1 week, 3.5 days, or 1 day, predicts the moods self-reported on the next week with the lowest error. Our assumption is that by using fewer days, the prediction error will increase. By using only the valence axis on the affect grid for training and prediction, we trained a single-task, Encoder-Decoder LSTM model and tested its performance on a test set of disjoint users. Fig. 4a shows the MSE for each half-day of the next week for different training sequences. We make the following observations. First, the error increases as we forecast more days into the future. Second, the lowest reported error is $0.022 MSE$ when using 3 weeks of data for training, which corresponds to ±0.14 error on the affect grid (see Fig. 1). Although 1 month includes more time-steps, the length of the optimal sequence of past moods for training is 42 half-days, which corresponds to 3 weeks. This could be attributed to either the inability of the model to learn such long sequences, or that the fourth week of the past does not contain informative and predictive patterns in this dataset. We observed similar behaviors by testing that assumption with our baseline models. Third, our model achieves the highest error when it is trained with just one day of data −2 mood self-reports− (±0.18 error on the affect grid at its worst), followed by half-week.

4.2 How effective are multitask LSTMs?

Motivated by the moderate correlation between valence and arousal (*pearson r*=-0.23) but on a significant level (p<0.00001), we experiment with learning the two sequences simultaneously in a joint model. Our assumption is that, given this similarity, a multi-task model trained to simultaneously predict valence and arousal would perform better than a single model trained on each one separately. To this aim, we train a single multi-task model with the input and output containing the aligned sequences of valence and arousal in a tensor $\mathbf{T} \in \mathbb{R}^{s \times t \times f}$, where s are the samples, t the time-steps, and f the two features or sequences of mood. The only modification to the single-task model is on thefi nal feed-forward layer, which now has two units, one for each task.

We use the sequence of moods of the previous 3 weeks to predict the sequence of moods in the next week. We do so because in the previous experiment 3 weeks was found to be the period that produces the lowest error in the prediction. From now, we will refer to those 3 past weeks as *user-weeks*. We use the same data split as in the previous setup and compare different algorithms and approaches. To allow for comparison, we also trained a single-task model for the arousal axis using the setup followed earlier with the valence 4.1, a SVR, a GBR, and a naive baseline that predicts just the average of the past self-reports.

Figure 4b shows the MSE for each half-day of the next week for different training sequences and algorithms. Similar to the previous experiment, we observe that the error increases over time. The most interesting result comes from the multi-task learning, which improves the performance of the arousal when trained jointly with the valence, but not the opposite. In general, the arousal axis throughout all of our experiments is more difficult to predict, which reflects on higher errors in all the models. We posit that users might not be as confident evaluating their calmness as they are with their happiness, hence the relationship between the two axes might not be linear. We showed in §2 that the heatmap of the two axes forms

a *X-shape* (Fig. 1). There is evidence that there is a *V-shaped* relation of arousal as a function of valence [10]. This is in line with previous studies that found that happy/unhappy feelings usually co-occur with higher arousal for some people (reflecting joy/stress), but with lower arousal for others (relaxation/sadness) [9].

Regarding the baselines, we observe that the error on the next day's prediction using single-task and multi-task models is lower than the ones achieved with feature-based algorithms, which even fail to improve the performance of the naive heuristic. In fact, the maximum MSE of 0.032 (±0.17 on the affect grid), makes them equivalent to using only one day of data for training in the previous experiment (brown line in Figure 4a). This motivates the need of using non-linear models like LSTMs. However, and regarding their utility, we believe simple baselines like these should be encouraged more in time-series forecasting since they provide a fast lower bound. In a more systematic comparison, we compare the error distributions (squared error of predicted and ground truth) of each classifier with a Welch's t-test. Our hypothesis states that multi-task learning will outperform the rest of the classifiers. Indeed, for the valence axis at thefi rst future forecast, the multi-task model presents statistically significant results over the naive baseline ($p < 0.001$), the SVR ($p < 0.001$), and the GBR ($p < 0.001$). Similarly, for the valence axis, the multi-task model outperforms the naive baseline ($p < 0.05$) and the GBR ($p < 0.05$), and shows a weaker significance against the SVR ($p < 0.10$). For both valence and arousal there is no statistical association between the single-task and multi-task models for thefi rst forecast. However, even if the multi-task models are not better than the single task for thefirst day, they show lower error during the week for the arousal axis (red lines in Fig. 4a). Because of that, we test the forecast of the whole future sequence by taking the median of the week (since the error is not normally distributed) and compare the models. Indeed, the arousal of the multi-task model is significant over the single-task one ($p < 0.05$).

Finally, we inspect the relationship between the predicted and the ground truth scores of valence and arousal for thefi rst future day using our multi-task model (Fig. 5). We observe a significant approximation of the two distributions. A non-parametric lowess model (locally weighted linear regression) isfi tted in order to illustrate the trend. Almost linear trends appear for high valence (happy users) and low arousal (relaxed users), which is also the area with the highest density in the dataset (see Figure 1). This is corroborated by large and significant ($p < 0.001$) Pearson correlations of 0.76 and 0.72 for valence and arousal, respectively.

5 UNDERSTANDING THE ROLE OF LSTM ENCODER AND DECODER

We now analyze the role of the LSTM encoder and decoder in predicting sequences of future mood. To do so, we pass the test-set through the multi-task LSTM model and use Principal Component Analysis (PCA) to visualize the response of the network after the encoder and decoder.

Learned representations vs next day's mood. Our test-set is a tensor $\mathbf{T} \in \mathbb{R}^{s \times t \times f}$ where s are the samples, t the time-steps, and f the two features or sequences (for valence and arousal). Fig. 6 shows the visualization results for the valence feature and thefi rst time-step in this tensor. Results for arousal and other time-steps follow a similar pattern, and are omitted here due to space limitations. Data points in thefi gure are coloured according to thefi rst mood to predict (ground truth). We observe that as we move into deeper layers, the network lays out the continuum of positive-negative mood, even though it has been trained to solely predict the next week's mood. Although after the encoder we can already see this continuum, this is more evident after the decoder layer. Apart from qualitative measures, the explained variance of the projections, i.e., the sum of variances of all individual principal components, or more intuitively how much information is lost by going from N to 2 dimensions, increases up to 40% after training (from 0.52 to 0.88 after the encoder, and 0.92 after the decoder).

Learned patterns of individual neurons. We now inspect how the individual neurons of the *decoder* layer *fire* as we pass the test-set through them. Fig. 7 shows the mean and standard deviation (denoted with dark and light green respectively) of the activations of the test samples (vertical axis in each subplot) for the 14 time-steps (horizontal axis). We make the following observations. First, the decoder learns various non-linear sequence patterns of future moods. Second, some neurons, such as the 4^{th} and 5^{th} in the 6^{th} column,fi re almost always with the same exponential decay slope (low deviation), while others, such as the 1^{st}, 3^{rd} and 4^{th} in the second column, are more conservative with almostfl at lines (high deviation). Since the decoder is the penultimate layer before thefi nal feed-forward layer that performs the regression, we may interpret it as a proxy for the predictions. For example, one neuron that alwaysfi res like the 3^{rd} in the 7^{th} column might be specialized in future mood that rapidly drops and then slowly improves.

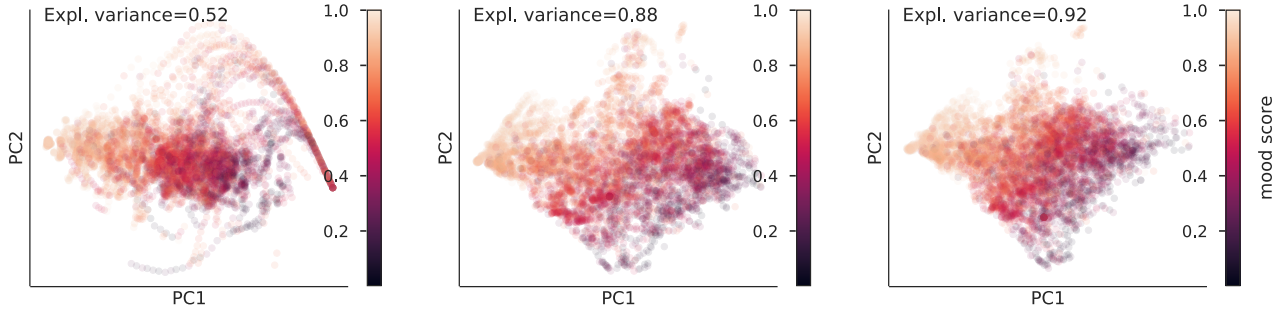
6 ERROR ANALYSIS

We have shown that mood reports might vary within a single user, and especially across a population (§2). Previous research has also found a link between mood variability and personality traits such as emotional stability [6], and that people tend to exhibit more positive affect on Saturdays than on Mondays [2]. To better understand the performance of our model and assist clinicians in taking informative decisions based on its output, we now investigate how it performs for different mood variability (§6.1), psychological traits (§6.2), and days of the week (§6.3).

To do so, wefi rst average the errors of the predicted sequence on the test-set, obtaining two long tail distributions for valence and arousal (Fig. 8). These appear because some user-weeks have errors even higher than 0.20 MSE (±0.44 on the affect grid), but the majority of the distribution resides below 0.025 MSE. Indeed, for valence, more than 10% of the user-weeks have MSE close to zero. We then divide the MSE-distributions in 3 equally-sized samples, and consider the 1^{st} quantile as the top-performing user-weeks, and the 3^{rd} quantile as the worst-performing ones.

6.1 Mood variability

Wefi rst investigate the influence of the mood variability in the best and worst performing user-weeks. We assess the mood variability for each user-week in these two groups by computing the standard deviation (std) of both the mood of the 3 past weeks (ignoring the



(a) PCA before training

(b) PCA on the LSTM *Encoder* layer activations after training

(c) PCA on the LSTM *Decoder* layer activations after training

Figure 6: Visualization of the *Encoder* and *Decoder* responses on thefi rst time-step for the valence axis.

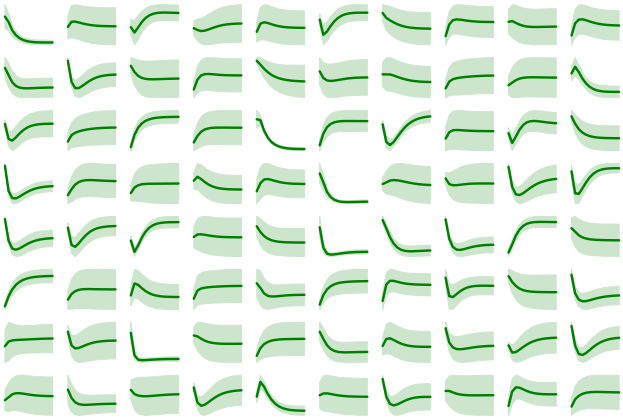


Figure 7: Visualization of the responses of the 80 neurons of the *Decoder* (one per subplot) for each of the 14 time steps for the valence axis.

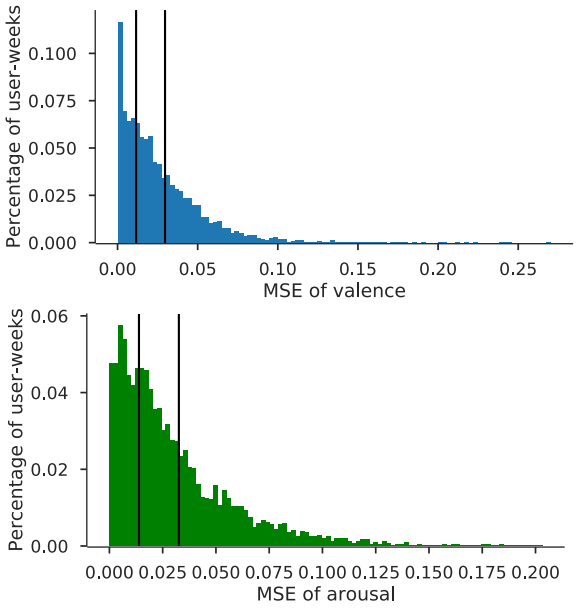


Figure 8: Distribution of *avg* MSE for valence and arousal. The MSE corresponds to the average predictions of all future days of the week for the multi-task model. Black bars denote the 1^{st} and 3^{rd} quantiles (at 33% and 66%, respectively).

missing values) and the mood of the future week. The boxplot in Fig. 9 shows the difference between these two groups. We observe that (i) the MSE increases with the variability of the (past or future) mood, and (ii) valence and arousal have similar median deviation, although the median of the arousal is slightly higher. The lowest deviation is on the future top-weeks, where there are no outliers in the boxplot, which means that the model is very reliable for those user-weeks with more stable future mood. Finally, the absolute mood differs between the bottom (≈ 0.2 std) and top quantiles (≈ 0.1 std) of the error. Specifically, the model is more reliable for user-weeks with high valence and low arousal as we saw in Fig. 5.

6.2 Personality traits

We now study the influence of personality traits in the best and worst performing user-weeks. We consider those individuals with samples in the 1^{st} and 3^{rd} quantile of the prediction error distributions (Fig. 8) who completed the personality questionnaire. This includes questions regarding the *Big-5* personality traits [7]: Agreeableness, Conscientiousness, Emotional Stability, Openness, and Extraversion, answered through a discrete Likert scale with values

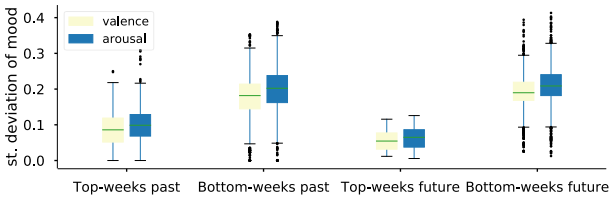


Figure 9: Deviation of past and future mood by top and bottom performing user-weeks.

normalized in [0,1]. Note that not every userfi lled the personality questionnaire. Thus, even though each original quantile contains the same number of user-weeks (2188), our sample shrinks to 701

Table 1: Differences on personality between the top and bottom quantiles, broken down by valence and arousal. Significance is represented with * = $p < 0.05$, ** 0.01, * 0.001.**

	Valence				Arousal			
	Mean				Mean			
	Top	Bottom	t-stat	Sig	Top	Bottom	t-stat	Sig
Agreeableness	0.67	0.61	6.58	***	0.68	0.60	9.53	***
Conscientiousness	0.66	0.59	5.55	***	0.70	0.64	4.72	***
Emotional Stability	0.33	0.25	6.42	***	0.38	0.25	11.89	***
Openness	0.77	0.61	16.48	***	0.78	0.66	12.55	***
Extraversion	0.24	0.29	-4.00	***	0.32	0.27	3.62	***

Table 2: Pearson's correlation (r) of the prediction error (MSE) with the personality of top and bottom quantiles, broken down by valence-arousal. Significance is represented with * = $p < 0.05$, ** 0.01, * 0.001.**

	Valence				Arousal			
	r with MSE				r with MSE			
	Top	Sig	Bottom	Sig	Top	Sig	Bottom	Sig
Agreeableness	-0.12	**	-0.06	*	0.01		-0.10	***
Conscientiousness	-0.06		0.03		-0.20	***	0.29	***
Emotional Stability	-0.19	***	-0.00		-0.11	*	0.00	
Openness	-0.03		-0.07	*	0.03		0.14	***
Extraversion	-0.06		-0.00		0.11	*	-0.06	*

(1^{st}) and 1082 (3^{rd}) user-weeks for valence when we consider only users who responded the questionnaire, and to 687 (1^{st}) and 1207 (3^{rd}) user-weeks for arousal. Some users might appear in both quantiles, but their skewed appearance in a quantile will influence that more.

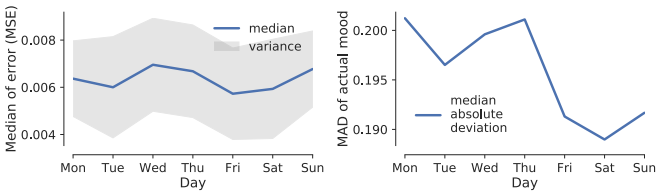
Wefi rst perform a Welch's t-test to check whether there are significant differences between the personality traits of the users in the two quantiles (Table 1). Significant differences were found for all traits, with special relevance for Openness. That is, users for which the model forecasts happiness *and* calmness more accurately tend to be more open to new ideas and showcase creativity, intellectual curiosity, and a preference for novelty.

Previous research found that Emotional Stability, Extraversion, Agreeableness, and sometimes Conscientiousness were related to decreased variability in affect [6]. In §6.1 we showed that users in the top and the bottom performing quantiles differ in terms of their mood stability, while here we see that also all their personality traits are significantly related with the performance of the model. In our case, higher Openness might be associated with the nature of our experiment and data collection since users that are more open to new technologies might use the app more honestly and therefore becoming more predictable.

Wefi nally check whether increments in personality scores increase or decrease the error. Table 2 shows the correlation of the error with the personality traits. For valence users in the top quantile we observe that our model is increasingly more accurate for Emotional Stable users (r = -0.19). We do not observe the reverse effect on the bottom quantile. For arousal, the model is more accurate for users with high Conscientiousness –self-disciplined– (r = -0.20) and a reverse effect appears on the bottom quantile (r = 0.29).

6.3 Day of the week

We now investigate the impact of the day of the week on the accuracy of our model. We consider the error of thefi rst mood in the sequence of predicted moods, grouped by the day of the week. The distribution of this error is similar to the distribution of the *average* error of the sequence of predicted moods in Fig. 8, but skewed towards lowest errors since our errors are lower for tomorrow's mood (first day in the sequence). We group and average (i) the errors by the day of the week, and (ii) the actual mood of this day across all the user-weeks in the test set. We observe that the distributions of the actual mood (Fig. 1) and the error (Fig. 8) across all user-weeks in the test set are very different. While the actual mood has a bimodal shape, the error resembles more a long tail. Thus, since we compare distributions with non-uniform shapes we use robust statistics, such as the median or median absolute deviation (the median of the absolute deviations from the data's median: $MAD = median(|X_i - median(X)|)$).



(a) (b)

Figure 10: Contribution of the day of the week to the median error for thefi rst future mood (valence) (a). Comparison with the actual mood variability (b).

We obtain the median and variance of the error for each day of the week, and the median absolute deviation (MAD) of the actual mood for each day. Fig. 10 shows the median of the MSE (a) and the MAD of the actual mood (b) across different days of the week. We observe that our model (Fig. 10 (a)) is more accurate on Tuesdays, Fridays and Saturdays, while the highest median error is on Wednesdays. This is in line with the trend observed on the variability of the actual mood (Fig. 10 (b)), where on Fridays and Saturdays there were fewer differences across the moods reported.

6.4 Discussion

Mood variability, personality and day of the week play a key role in the performance of our model. Clinicians may wish to *screen* their patients with fast personality questionnaires to assess the reliability of the sequences of moods predicted. For example, openness affects performance in both dimensions, whereas emotional stability affects valence, and conscientiousness arousal. Clinicians should also consider the day of the week when forecasting sequences of mood, trusting less outputs on Wednesdays than on weekends.

We also acknowledge the caveats of our model. The average error (MSE) of our model is low across the population (±0.14 error on the affect grid valence for valence, ±0.16 for arousal). However, the fact that it performs better for emotionally stable users, and users with low mood variability, might limit its utility in patients with mental disorders. Further analysis of the trajectories of mood reported by *unstable* individuals is required to build accurate models for this specific population. Moreover, the studied outcome should be affected by a variety of environmental and genetic factors and

additional data collected in this study could improve forecasting. We leave this for future work.

7 RELATED WORK

Several mobile applications have been proposed to monitor and study mental health. For example, *StudentLife* [23] combined sensing and self-reports to assess the impact of student workload on stress, whereas *Snapshot* [20] tracked their mood and sleep. Others focused on detecting depression by tracking medication, sleep patterns and actions [18], location [3] or keypress acceleration [4].

However, most of existing works suffer from (i) limited sample size, both in terms of number and diversity, which hampered them from drawing robust conclusions, as well as (ii) limited duration of the studies. For instance, in the *MoodScope* study [12] 32 people were monitored for 2 months; in *StudentLife* [23], 48 students were tracked for 10 weeks, whereas in *Snapshot* [20], probably the biggest general published study about mood monitoring using mobile devices, 206 students were tracked for 1 month. In contrast, we learn a mood prediction model from real-world data from 566 participants collected in the wild for more than 3 years. On a similar scale, but for (binary) depression prediction specifically, the *Deepmood* [18] study analyzed 2,382 users over 2 years. Contrary to this work, our model does not aim to distinguish between healthy and depressed patients, but to predict a sequence of real-valued moods. Binary prediction is recurrent on the mood prediction literature, where mood gets simplified to a binary state [16, 20], and extreme depression is considered in the same class as moderate unhappiness. Since neutral mood might be uninformative and make the predictions harder, authors often omit the middle 40-60% of reports. Instead, we use regression approaches to predict precise mood scores.

Even by overlooking those weaknesses, most of the proposed systems are maybe infeasible to be deployed on a real world scenario since they require training N personalized models, where N is the number of users. Also, although better performance can be achieved by averaging the individual model accuracies [3, 12], no results are reported on unseen disjoint users. Instead, we report performance from a disjoint user set that the model has not seen during training.

The majority of related literature has applied some kind of supervised learning algorithms, like Logistic Regression or Support Vector Machines, that however cannot capture non-linear combinations of features. Some recent deep learning results include the *Deepmood* study that uses RNNs for depression prediction [18], and a multi-task personalized deep architecture for the *Snapshot* dataset that looks promising [20]. We build upon this piece of literature of employing deep learning on mood prediction.

8 CONCLUSION

This paper introduces a new end-to-end, stand-alone ML model to forecast future sequences of mood from previous self-reported mood. Contrary to previous research on classifying between extremes of mood using data collected in controlled experiments with limited number of participants, we forecasts exact values of valence and arousal from noisy and sparse reports collected in the wild.

Experiments using a real-world dataset revealed that (i) 3 weeks of sparsely reported mood is the optimal number to accurately forecast mood, (ii) multi-task learning learns both dimensions of mood –valence and arousal– with higher accuracy than when training separate models, and (iii) mood variability, personality traits and day of the week play a key role in the performance of our model. We believe this work provides psychologists and developers of future mobile mental health applications with a ready-to-use and effective tool for early diagnosis of mood issues at scale.

ACKNOWLEDGMENTS

This work was supported by the Embiricos Trust Scholarship of Jesus College Cambridge, EPSRC through Grants DTP (EP/N509620/1) and UBHAVE (EP/I032673/1), and Nokia Bell Labs through the Centre of Mobile, Wearable Systems and Augmented Intelligence.

REFERENCES

[1] Stephen Aichele, Patrick Rabbitt, and Paolo Ghisletta. 2016. Think fast, feelfi ne, live long: A 29-year study of cognition, health, and survival in middle-aged and older adults. *Psychological science* 27, 4 (2016), 518–529.
[2] Charles S Areni and Mitchell Burger. 2008. Memories of "bad" days are more biased than memories of "good" days: past Saturdays vary, but past Mondays are always blue. *Journal of Applied Social Psychology* 38, 6 (2008), 1395–1415.
[3] Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Ubicomp '15*. ACM, 1293–1304.
[4] Bokai Cao et al. 2017. DeepMood: Modeling Mobile Phone Typing Dynamics for Mood Detection. In *KDD '17*. ACM, 747–755.
[5] Helen Cheng and Adrian Furnham. 2003. Personality, self-esteem, and demographic predictions of happiness and depression. *Personality and individual differences* 34, 6 (2003), 921–942.
[6] Katharina Geukes et al. 2017. Trait personality and state variability: Predicting individual differences in within-and cross-contextfl uctuations in affect, self-evaluations, and behavior in everyday life. *Journal of Research in Personality* 69 (2017), 124–138.
[7] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.
[8] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[9] Peter Kuppens. 2008. Individual differences in the relationship between pleasure and arousal. *Journal of Research in Personality* 42, 4 (2008), 1053–1059.
[10] Peter Kuppens et al. 2013. The relation between valence and arousal in subjective experience. *Psychological Bulletin* 139, 4 (2013), 917.
[11] Nicholas D. Lane et al. 2011. Bewell: A smartphone application to monitor, model and promote wellbeing. In *PervasiveHealth '11*.
[12] Robert LiKamWa et al. 2013. MoodScope: Building a Mood Sensor from Smartphone Usage Patterns. In *MobiSys '13*. ACM.
[13] RS Olson et al. 2018. Data-driven advice for applying machine learning to bioinformatics problems.. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, Vol. 23. 192–203.
[14] Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098* (2017).
[15] James A. Russell, Anna Weiss, and Gerald A. Mendelsohn. 1989. Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology* (1989).
[16] Sandra Servia-Rodríguez et al. 2017. Mobile sensing at the service of mental well-being: a large-scale longitudinal study. In *WWW '17*. 103–112.
[17] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. 2015. Unsupervised learning of video representations using lstms. In *ICML '15*. 843–852.
[18] 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*. 715–724.
[19] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS '14*. 3104–3112.
[20] Sara A. Taylor et al. 2017. Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE Transactions on Affective Computing* (2017).
[21] Ruut Veenhoven. 2008. Healthy happiness: Effects of happiness on physical health and the consequences for preventive health care. *Journal of happiness studies* 9, 3 (2008), 449–469.
[22] Arun Venkatraman, Martial Hebert, and J Andrew Bagnell. 2015. Improving Multi-Step Prediction of Learned Time Series Models.. In *AAAI*. 3024–3030.
[23] Rui Wang et al. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Ubicomp '14*. ACM, 3–14.