Predicting students' happiness from physiology, phone, mobility, and behavioral data

Natasha Jaques, Sara Taylor, Asaph Azaria, Asma Ghandeharioun, Akane Sano, Rosalind Picard
MIT Media Lab, Cambridge, MA 02139, USA
Email: {jaquesn, sataylor, azaria, asma_gh, akanes, picard}@media.mit.edu

***Abstract*—In order to model students' happiness, we apply machine learning methods to data collected from undergrad students monitored over the course of one month each. The data collected include physiological signals, location, smartphone logs, and survey responses to behavioral questions. Each day, participants reported their wellbeing on measures including stress, health, and happiness. Because of the relationship between happiness and depression, modeling happiness may help us to detect individuals who are at risk of depression and guide interventions to help them. We are also interested in how behavioral factors (such as sleep and social activity) affect happiness positively and negatively. A variety of machine learning and feature selection techniques are compared, including Gaussian Mixture Models and ensemble classification. We achieve 70% classification accuracy of self-reported happiness on held-out test data.**

I. Introduction

Not only have rates of depression in the United States notably increased in the last century, but a greater number of young adults are becoming depressed [1]. Depression is prevalent on college campuses, and is also the most frequent precursor to suicide [2]. Addressing depression among college students has become a major concern for some universities, especially given the fact that 18-24-year-olds have the highest incidence of suicidal ideation, and suicide has become the third leading cause of death among college-aged individuals [3].

For these reasons, it is important to understand the factors that contribute to resistance to depression. A body of research has shown that overall wellbeing, including factors like self-reported happiness, social support, and engagement with work, contribute to an individual's resiliency and ability to handle negative life events without becoming depressed [4]. Physiological factors also affect vulnerability to depression. Numerous studies have shown a significant link between sleep disturbances and subsequent depression [5], and physical health is strongly correlated with depression and happiness [6].

This study advances understanding of the role of affect in resiliency and wellbeing by investigating the relationship between factors like sleep, social and physical activity, stress, and happiness. Ideally we would like to investigate the factors that affect an individual's overall wellbeing both positively and negatively. Since wellbeing cannot be measured directly, we rely on self-reported measures that are known to affect wellbeing, including stress, health, and happiness. Because self-reported happiness is strongly correlated to measures of depression [6], we will focus heavily on happiness so that we can not only discover factors that contribute to wellbeing, but also use machine learning methods to build a system that can automatically detect when a college student is becoming vulnerable to depression. This system could be used to guide timely interventions so that serious consequences of depression — such as suicide — can be prevented. To aid our investigation we examine a wide range of data sources:

- Physiological data: electrodermal activity (EDA) (a measure of physiological stress), and 3-axis accelerometer (a measure of steps and physical activity)
- Survey data: questions related to academic activity, sleep, drug and alcohol use, and exercise
- Phone data: phone call, SMS, and usage patterns
- Location data: coordinates logged throughout the day

In this paper we will develop a machine learning algorithm to distinguish between happy and unhappy college students, assess which measures provide the most information about happiness, and evaluate the relationship between different components of wellbeing including happiness, health, energy, alertness and stress.

II. Background and related work

There is a growing literature showing the connection between social support and wellbeing. Social support has been shown to mediate stress [7], protect against depression [8], and even improve overall health and recovery from illness [9]. In fact, positive social relationships have been found to be the single most important factor in wellbeing in studies across ages and cultures [10]. Conversely, people who lack social support are at risk for a range of mental health issues, including depression, anxiety, and suicide [11]. Because smartphone logs provide a record of the number, duration, and type of communications with social contacts, they may provide insight into an individual's social support and therefore stress, health, and happiness. Further, simply using the phone itself may affect wellbeing through sleep quality; the phone screen emits a large amount of artificial light, which has been shown to adversely affect the circadian rhythm and sleep [12].

In fact, smartphone data (e.g. location, proximity, and communication) have been explored in a variety of studies which are surveyed in [13]. Dong and colleagues used smartphone data to model the underlying structure of social interactions in a student dormitory [14], while [15] uses this data to explore the relationship between sleep and mood. Predicting stress from smartphone logs has been attempted by multiple

researchers (e.g. [16] [17] [18]). There have been preliminary studies demonstrating that mood can be classified using smartphone data [17] [19], and Bogomolov and colleagues have successfully predicted happiness from a combination of smartphone data, personality, and weather patterns [20].

Physiological measures such as electrodermal activity (EDA) are also frequently used in studies related to affect and wellbeing (e.g. [21], [22], [23], [24]). EDA measures sudomotor innervation and sweat gland activity, which is increased through activity of the sympathetic nervous system (SNS) [25]. Because the SNS is influenced by the hypothalamus and the limbic system (structures in the brain that deal with emotion) EDA can be an effective technique for measuring emotion and stress. The link between EDA and stress was directly explored in a pilot version of the study used to collect the data analyzed in this paper [16]. Other research has investigated the link between EDA and sleep quality [26] [27].

III. User Study

This research is based on data from an ongoing longitudinal study investigating the impact of behavioral and physiological measures on wellbeing; more details and descriptive statistics about the data can be found in [28]. The data were collected over two 1-month (30-day) experiments in which 20 and 48 MIT undergraduates were recruited to participate, respectively. Each participant wore an Affectiva Q-sensor nearly continuously for the entire 30-day period to gather Electrodermal Activity (EDA), skin temperature, and 3-axis accelerometer data. Additionally, each participant downloaded an app on his or her Android Phone that logged calls, SMS messages, times the screen was turned on and off, and location information. Surveys were completed by participants two times each day (in the morning and evening), and asked questions about the participants' behaviours, activities, and wellbeing.

IV. Analysis of Wellbeing Measures

Participants in the study self-reported on five scales twice a day related to wellbeing: stress, health, energy, alertness, and happiness. Although we would ideally like to be able to predict overall wellbeing, how to create a ground-truth wellbeing measure from these scales is an open question. A first impulse might be to compute a composite measure from some of the relevant scales, for instance by computing the ratio between happiness and stress. However this schema would treat a highly happy and highly stressed state as equivalent to a low happiness and low stress state. Not only is this assumption not empirically validated, but a low happiness and low stress state could be indicative of depression (sadness and apathy), whereas a high happiness high stress state could actually represent greater underlying wellbeing than a non-stressed state, given the contribution of personal achievement and engagement with work to overall wellbeing [4]. For these reasons we first attempt to understand the relationship between these measures in order to frame our classification problem.

Table I shows the Pearson's correlation coefficients between all pairs of wellbeing measures. Note that stress was actually reported on a scale where a low score indicated a highly stressed state and a high score indicated calmness, so for consistency we report this as Calmness in the following explanations. We can see that all of the measures are highly related, with all correlations reaching significance at the $p = .01$ level, even after applying a Bonferroni correction to account for alpha inflation. Happiness has the highest correlation coefficients, suggesting that if we were to limit our predictions to Happiness alone, it would give the most insight into the remaining scales. We are also most interested in Happiness, as it has been shown to relate directly to depression [6].

TABLE I
Correlation matrix for measures of wellbeing

	Happiness	**Health**	**Calmness**	**Energy**	**Alertness**
Happiness	-				
Health	0.537	-			
Calmness	0.664	0.480	-		
Energy	0.480	0.410	0.389	-	
Alertness	0.374	0.318	0.324	0.721	-

V. Machine Learning Methods

Based on the analysis of the previous section as well as our theoretical interests, we chose to use the Happiness scale as our ground-truth measure; it was reported using a slider from "Sad" (a value of 0) to "Happy" (a value of 100). We frame the problem as binary classification; days on which a participant reported a Happiness score in the top 30% of all Happiness scores are labeled as a positive day, and days in which participants reported a Happiness score in the bottom 30% are labeled as a negative day. We do not include the middle 40% of scores. This reduces the size of our dataset to a possible 1110 points, and thus reduces our classification power. However, the behaviors on these days do not appear to have a strong effect on Happiness or wellbeing, and are thus not informative for this problem.

Using the remaining data points, we randomly partitioned a training, validation and testing dataset. The training and validation sets were used to perform feature and model selection for each data source; we refer to these sources as modalities, and discuss each in greater detail in Section VI. Given the complexity of our data, we used an iterative feature design process. After designing an initial feature set based on a review of the literature, we assessed the relevance of each feature by measuring information gain and through Wrapper Feature Selection (WFS) [29]. Irrelevant features were removed in order to prevent overfitting, and more features were repeatedly added and assessed, until we arrived at a final feature set for each modality. The number of features eventually selected optimized accuracy on the validation set.

A variety of machine learning algorithms were tested in order to find the most appropriate model for each type of data. These include Support Vector Machines (SVM), Random Forests (RF), Neural Networks (NN), Logistic Regression (LR), k-Nearest Neighbour (kNN), Naïve Bayes, and Adaboost. After finding the best classifier, the parameter space

of the classifier was searched, and the parameters which optimized performance on the validation set were selected.

VI. Feature Design by Modality

A. Physiology

Physiological measurements collected include EDA measured as skin conductance (SC) in microSiemens (μS), 3-axis accelerometer, and temperature, recorded at a sampling rate of 8 Hz. Following [16], we compute each set of physiological features over different time periods during the day (12am-6am, 6am-12pm, 12pm-6pm, 6pm-12am), as well as periods when the participant was asleep vs. awake (determined through the survey responses). To compute the features, we first apply a 1Hz low-pass filter to the SC signal, then compute the normalized signal according to the mean, max, and min for each participant, i.e. $SC_i = \frac{SC_i - \mu}{max - min}$ [22]. We include statistical features related to the raw signal, normalized signal, and signal derivative computed over each time period.

When a person experiences a physiological stress response, they may simultaneously experience a skin-conductance response (SCR), in which their SC signal peaks rapidly and then decays at an exponential rate (see Figure 1 for an example of typical SCRs). Because of the relationship between stress and wellbeing, EDA will be most useful if we can determine when SCRs occur and compute features related to them. Our initial feature design treated SCRs as points at which the derivative of the SC signal exceeded a threshold, as per [26] and [30]. In later iterations, we detected SCRs based on several criteria involving the amplitude, duration, and shape of the SCR; this proved to be more effective. We include features related to the number of SCRs occurring over each period, as well as statistics related to the amplitude, rise time, and area under the curve (AUC) of the detected SCRs.

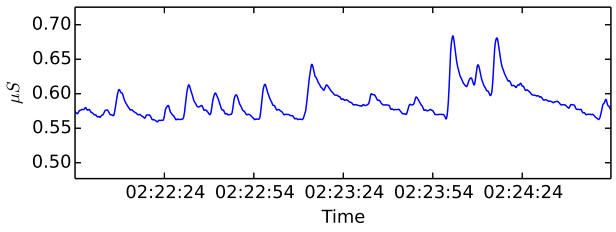

Fig. 1. An example of typical SCRs

Given the 24-hour-a-day, ambulatory nature of the EDA recordings in this study, the signal is vulnerable to artifacts and noise. Further, increases in SC are not always due to stress; they could be the result of physical activity or increasing temperature. Therefore we computed the magnitude (Mag) of the accelerometer signal, $Mag = \sqrt{acc_x^2 + acc_y^2 + acc_z^2}$, and used this to weight the strength of the SC signal. This was accomplished by normalizing the Mag values to be between 0 and 1, and computing the inverse signal by subtracting this value from 1. We then took the product of this inverse accelerometer with the SC signal. Thus, the new signal represents one in which the effects of SC increase due to physical activity are diminished. A similar procedure was used to compute a temperature-weighted SC signal. Features related to these signals were then computed, and both types proved useful in the feature assessment process. In order to detect recording artifacts we applied an algorithm described in [31], and removed peaks that were identified as possible artifacts from the analysis.

Physical activity is also strongly related to wellbeing, given its profound impacts on happiness and stress [32]. Therefore we also included a feature that approximates of the number of steps taken by counting the number of times the Mag signal crossed a small threshold [27].

B. Survey

The survey features relate to the number and duration of academic, exercise, and extracurricular activities, the amount of time spent studying, sleeping, napping, and trying to fall asleep, whether participants woke up during the night or overslept, whether they interacted with someone in person or digitally before falling asleep (referred to as *pre-sleep interaction*), and whether they had a positive or negative social interaction that day. Additionally, students indicate whether they consumed caffeine, alcohol, or drugs that could make them alert, sleepy, or tired. We are interested in how these behavioral choices and habits affect wellbeing. The other self-reported measures of wellbeing (stress, health, energy, and alertness) are not included as features; these are what we would like to eventually detect automatically.

C. Phone

The phone log data consist of information about the timing, type, and duration of phone calls and SMS messages, and times the screen was turned on and off. An example of this type of data is shown in Figure 2. We see two mechanisms through which screen and communication information can affect wellbeing; light from the screen can disrupt circadian rhythms and therefore sleep [12], and the amount of social support in a person's life is strongly linked to resilience to depression (see Figure 2 for a possible example of a subject's social network potentially helping the subject move from a sad mood back to a happy mood.) [4] [8] [10]. Therefore we sought to create features that would capture these factors.

We discarded days with fewer than 5 screen on events, reasoning that the app must have been malfunctioning. Previous research has shown that people interact with their phone between 10-200 times a day [33]. The total number and basic statistics (mean, median, std. dev.) related to timestamp were computed for calls, SMS messages, and screen-on events. Total duration and duration statistics were computed for calls and screen activity. As with physiology, the features were computed over time intervals spanning the course of the day. These may be important because we wish to capture the time when blue light from one's phone is experienced relative to the natural rhythm of sunlight.

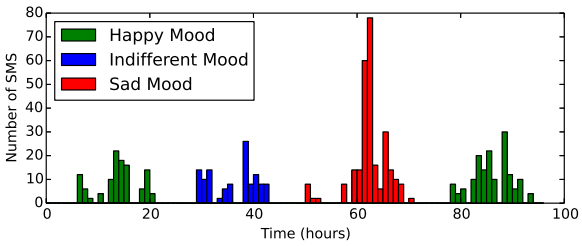

Fig. 2. Example of a participant's SMS data for 4 consecutive days. Note that there is a large increase in SMS during the negative day, which is followed by a return to a positive day

Because we can determine whether an SMS is incoming or outgoing, we also compute the above features for these specific type of events. This could be informative, for example because incoming messages may relate more strongly to social support than outgoing messages. Finally, we compute the number of unique callers and unique messengers for the entire day and also for each call/SMS type. Some researchers have hypothesized that diversity of social interactions with a range of individuals is linked to wellbeing [17].

In summary we computed 289 phone features based on statistics related to the timing, duration, and frequency of screen, call, and (incoming and outgoing) SMS events, computed over various time intervals throughout the day.

D. Mobility

In addition to communication and screen events, the phone app logs the participants' GPS coordinates throughout the day, as well as whether they are using Wifi or cellular data. Location is sampled whenever available in different frequencies on different devices, so we began by downsampling the signal into one set of coordinates for every 5 minute segment, computed using the median of the longitude and latitude samples within it. Segments that contained no samples were interpolated according to neighboring samples. We allowed interpolation of no more than three consequent segments (15 minutes), marking segments as missing data when necessary.

Building on previous studies [16] [17] [18], we extracted statistical descriptors of the subjects' distances traveled throughout the day. For each day, we computed the radius of the minimal circle enclosing the subject's location samples, as suggested by [17]. The source of the location data (WiFi or cellular) was used to compute an approximation of the time spent indoors and outdoors. Finally, we used the latitude and longitude coordinates of the university's campus to compute the time spent on campus each day.

Noticing that many students spent most of their time either at home or campus, we set out to model their location in a way that would better capture irregularities in this routine. We postulate that these irregularities would have a significant effect on measurements of their wellbeing. Therefore we computed a Gaussian Mixture Model (GMM) for each participants' typical location behavior. A GMM learns the number and location of Gaussian distributions required in order to collectively represent a probability distribution; in this case, the distribution over each participants' possible locations in 2-dimensional space. More formally, each participant's location distribution was modeled with K Gaussian components, as in:

$$p(x_i|\theta) = \sum_{k=1}^{K} \pi_k \mathbb{N}(x_i|\mu_k, \Sigma_k)$$

The GMM was trained on the latitude and longitude coordinates of the participant that were collected throughout the entire study (rather than just those seen on positive and negative days), since we are using the GMM to model routine behavior. The model selection process varied both the number of components, K, as well as the type of covariance matrix (spherical, diagonal, tied and full, each with different degrees of freedom). The trained model learns K Gaussian components that represent Regions of Interest (ROIs) that the participant commonly visits. We restricted the number of components K to 20, as we believed it is unlikely for an individual to have repetitive interest in more than 20 locations within a month. The best model fit was chosen using the Bayesian Information Criterion (BIC):

$$BIC = -2 \log p(D|\theta) + df(\theta) \log N$$

where θ is the MLE for the model and $df(\theta)$ is the number of degrees of freedom in the model [34]. Figure 3 illustrates a GMM fitted to the location data from one subject. We were able to verify by inspecting a map of the area that the identified components correspond to locations on the university campus and the participant's residence (specific coordinates have been redacted for privacy).

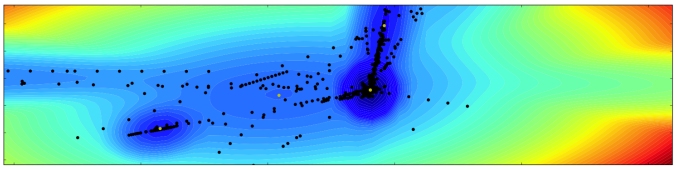

Fig. 3. GMM fitted to location data from one participant. The black points are latitude longitude coordinates, while the yellow are the means of the Gaussian components representing the ROIs. The contours mark the probability distribution induced by the model.

After fitting the GMMs for each subject, they were used to compute several features that relate to the regularity of participants' routines. First, the induced probability distribution was used to compute the log likelihood for each day; this represents whether the day varied unusually from the typical routine; we refer to it as *normality of day*. Because each model learns the number and coordinates of the locations typically visited by the participants, we can determine how many different familiar locations were visited on a given day (ROIs). This approach builds on an idea that was presented by [18], where a correlation was shown between emotional stress and a person's number of geo-location ROIs. Finally, the model BIC score and Akaike Information Criterion (AIC) [34] were computed using the data from each day; this represents how well the

TABLE II
FEATURES WITH THE HIGHEST INFORMATION GAIN FOR EACH MODALITY

Physiology	Survey	Phone	Mobility
.0560 ↑ SC median 12am-6am	.0379 ↑ Pre-sleep activity	.0602 ↓ Screen dur. med.	.0301 ↓ Time indoors
.0418 ↑ SC s.d. 12am-6am	.0240 ↑ Positive interaction	.0456 ↓ Screen dur. med. 6pm-12am	.0293 ↓ Normality of day
.0408 ↑ SCR AUC total 12am-6am	.0239 ↓ Negative interaction	.0377 ↓ Screen dur. med. 8pm-12am	
.0390 ↑ Mag Acc. s.d. wake	.0200 ↑ Exercise duration	.0367 ↓ Screen dur. med. 4pm-8pm	
.0382 ↑ Mag Acc. s.d. 6pm-12am	.0191 ↑ Exercise (true or false)	.0235 ↓ Screen dur. med. 8am-12pm	
.0381 ↑ SC med. sleep	.0190 ↑ Exercise count	.0213 ↓ Screen dur. mean 4pm-8pm	
.0378 ↑ Temp. weight. SC s.d. 12am-6am	.0140 ↓ Drugs - tired	.0210 ↓ Screen dur. med. 12pm-6pm	
.0374 ↑ SCR AUC mean 12am-6am	.0128 ↓ Studying duration	.0204 ↑ Screen total num. 8am-12pm	
.0367 ↑ SCR AUC max 12am-6am	.0106 ↑ Drugs - alcohol	.0185 ↑ Screen total num.	
.0366 ↑ SC deriv. mean 12pm-6pm	.0105 ↓ Extracurricular count	.0178 ↓ Screen timestamp s.d. 12am-4am	

model fits that particular day, and thus how much the day deviates from routine.

VII. RESULTS

The goal of this research is two-fold: 1) to understand the behavioral and physiological factors that impact wellbeing positively and negatively, and 2) to build a model that can detect when students become unhappy and thus drive interventions to mitigate the risks of depression. Therefore this section will first discuss the features found to be the most informative from each modality, and then present the classification performance in predicting positive and negative days.

A. Feature Evaluation

Despite reducing the feature set for each modality to the size that optimized validation accuracy, in total we still have 762 features. Therefore we cannot provide a thorough analysis of the usefulness of every feature selected. Instead, we will use information gain to assess the informativeness of the features. Information gain is computed according to Eq. 1, which involves the entropy function given in Eq. 2.

$$\mathbb{I}(X,Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X) \quad (1)$$

$$\mathbb{H}(X) = -\sum_i P(x_i)\log P(x_i) \quad (2)$$

Information gain can be interpreted as the reduction in uncertainty about one variable after observing the other [34]. In this case, we assess how much information each feature provides about our classification label, Happiness. We present up to 10 of the features that had the highest information gain for each modality in Table II, along with the score itself. We do not present features for which the information gain was close to zero. Since information gain is computed on each feature in isolation, it does not relate to how informative a collection of features may be when used in combination in a classification model. Therefore highly similar features (such as the multitude of screen duration features) all appear as valuable according to information gain. For this reason the information gain scores presented in Table II are not necessarily predictive of classification performance for each modality.

For interest's sake, Table II also provides an arrow indicating the direction of the relationship between the feature and the classification label, where an up arrow indicates that the feature affects happiness positively. These directions were obtained from the direction of the correlation between the feature and Happiness. For example, we see that *time indoors* is negatively correlated with Happiness; the more time spent indoors, the less likely the participant is to report feeling happy. We seek only to provide a general trend to give the reader some idea of how the feature affects Happiness, and have not attempted to establish the statistical signifance of all of these relationships.

Many of the physiology features relate to the SC signal and SCRs that occur between midnight and 6am, presumably when the participant is asleep. Note that different sleep stages are characterized by different SC patterns; for example, SCRs are more likely to occur during slow-wave sleep or non-REM 2 sleep [26]. Therefore these features may relate to sleep quality and thus to wellbeing. The survey features confirm our hypotheses that exercise and social interaction are strongly linked to happiness, supporting current research on the topic (e.g. [8] [10] [32]). It may be surprising to see that alcohol use appears to boost happiness. Since alcohol consumption is reported in the evening at the same time as the happiness scores, this is likely a reflection of the current effect of alcohol or possibly social interaction, and does not relate to prolonged alcohol use over the long term. We see from the phone logs that using the phone for longer periods of time (screen duration) appears to be associated negatively with happiness, especially if it occurs late in the day. Conversely, checking the phone frequently (screen total num) has a positive association with happiness, especially in the morning. Finally, we can see from the mobility features that time indoors and the likelihood or normality of the day as computed by the GMM are inversely related to happiness. This implies that when a participant spends time outdoors or deviates more from their typical routine, they tend to be happier.

B. Classification

Table III presents the classification results for each modality, the relative dataset and feature set sizes for context, and the classifier and parameter settings that were found to optimize validation accuracy. We found that the SVM and RF classifiers

TABLE III
CLASSIFIER, PARAMETER SETTINGS AND ACCURACY RESULTS FOR EACH MODALITY

Modality	Dataset Size	# Features	Classifier	Parameter Settings	Classification Accuracy		
					Validation	Baseline	Test
Physiology	933	426	SVM	C=100.0, RBF kernel, $\beta = .0001$	68.37%	51.79%	64.62%
Survey	1110	32	SVM	C=100.0, RBF kernel, $\beta = .01$	71.26%	50.86%	62.50%
Phone	1072	289	RF	Num trees = 40, Max depth = infinite	66.67%	51.98%	55.95%
Mobility	905	15	SVM	C=100.0, RBF kernel, $\beta = 1$	69.95%	53.65%	65.10%
All	768	200	SVM	C=0.1, Linear kernel	72.84%	53.94%	68.48%

tended to produce the best results on this dataset. Accuracy on the held-out test set (i.e. the proportion of samples in which the classifier's prediction matches the true label) provides an estimation of the results we can expect on novel data; therefore we can conclude that our best model would be able to identify students that are unhappy with 68.48% accuracy.

Note that the size of the dataset involving all features is reduced due to missing data. Considering the volume of data collected and the length of the study, some modalities are missing data for certain participants on certain days; for example, if a participant forgets to wear their sensor. Therefore when we combine all the modalities and restrict our focus to only those days/participants for which data from each modality is available, the dataset shrinks. This could make the 'all' dataset vulnerable to overfitting; therefore we applied the same feature selection techniques and found a reduced set of 200 features to be most effective.

TABLE IV
CONFUSION MATRIX FOR ENSEMBLE CLASSIFIER

		Predicted	
		Happy	**Sad**
Actual	**Happy**	77	40
	Sad	31	90

Ensemble classification offers an alternative approach to training a single classifier on all of the available features. Rather, the predictions from several classifiers are integrated, often in a weighted majority vote [35]. We built an ensemble classifier which combines the predictions of the best classifier from the best modalities by weighting their predictions according to the classifier's validation accuracy. We found that using the best three modalities produced the highest validation accuracy. The ensemble allows us to deal with missing data in a more robust way; if a modality is missing data for a given sample day, then that classifier simply abstains from the vote. Each modality is able to maintain the maximum amount of training data, while the ensemble combines data from several modalities without losing information. The best accuracy achieved by the ensemble classifier on the held-out test set was 70.17%. Table IV shows the confusion matrix for the predictions made by the ensemble classifier on the held-out test set. It is slightly more likely to falsely predict that a student is sad when she is actually happy, rather than falsely predict that a student is happy when she is actually sad. This characteristic suggests the system is more sensitive to detecting sadness, which is desirable if it is to be used to detect when to intervene if a student is becoming unhappy.

VIII. DISCUSSION AND LIMITATIONS

Although the classifiers trained on each modality were able to achieve results exceeding the baseline, performance differed across modalities. Interestingly, mobility offers high performance with few features; given the features found to be the most valuable for mobility, it would appear that whether or not a person spends time outdoors and deviates from normal routine is strongly related to whether they will feel happy on that day. Physiology also offered relatively high performance, suggesting that wearable devices which can monitor a person's physiology throughout the day may be a promising way to detect changes in happiness, especially if those devices are capable of monitoring sleep quality.

A limitation of this work is that it does not consider individual differences; for example, extracurricular activities could make some students happy or be stressful for other students. In future work we would like to model these complexities using Multi-Task Learning (MTL) [36]. Another limitation is that we do not model long-term effects of behaviors on happiness; for example, drinking alcohol may affect a participant's mood over the long term.

IX. CONCLUSION

This work has demonstrated that physiological, behavioral, phone and mobility data can all be used successfully to model happiness. We have contributed to the literature on wellbeing by examining not only which features provide the most information about happiness and how they affect it, but also by investigating the relationship between happiness and other components of wellbeing, such as health, stress, and energy. The best accuracy obtained by our models on novel data, 70.2%, may be sufficient to guide interventions intended to prevent depression, especially if these interventions are only triggered after the classifier detects a consistent pattern of unhappiness over several days or weeks.

ACKNOWLEDGMENT

We would like to acknowledge the team at Brigham and Women's hospital led by Dr. Charles Czeisler, Dr. Elizabeth Klerman, and Conor O'Brien, whose help was invaluable in running the user study on which this work is based. This work was supported by the MIT Media Lab Consortium, the Robert Wood Johnson Foundation Wellbeing Initiative, NIH Grant R01GM105018, Samsung, and Canada's NSERC program.

REFERENCES

[1] G. L. Klerman and M. M. Weissman, "Increasing rates of depression," *Jama*, vol. 261, no. 15, pp. 2229–2235, 1989.

[2] J. S. Westefeld and S. R. Furr, "Suicide and depression among college students.," *Professional Psychology: Research and Practice*, vol. 18, no. 2, pp. 119, 1987.

[3] J. Kisch et al., "Aspects of suicidal behavior, depression, and treatment in college students: results from the spring 2000 national college health assessment survey," *Suicide and Life-Threatening Behavior*, vol. 35, no. 1, pp. 3–13, 2005.

[4] M. Seligman, *Flourish: A visionary new understanding of happiness and well-being*, Simon and Schuster, 2012.

[5] N. Tsuno et al., "Sleep and depression.," *J. of Clin. Psychiatry*, 2005.

[6] H. Cheng and A. Furnham, "Personality, self-esteem, and demographic predictions of happiness and depression," *Personality and individual differences*, vol. 34, no. 6, pp. 921–942, 2003.

[7] S. Cohen and T. A. Wills, "Stress, social support, and the buffering hypothesis.," *Psychological bulletin*, vol. 98, no. 2, pp. 310, 1985.

[8] R. S. Peirce et al., "A longitudinal model of social contact, social support, depression, and alcohol use.," *Health Psychology*, vol. 19, no. 1, pp. 28, 2000.

[9] S. Cohen and T. B. Herbert, "Health psychology: Psychological factors and physical disease from the perspective of human psychoneuroimmunology," *Annu. rev. of psychology*, vol. 47, no. 1, pp. 113–142, 1996.

[10] H. T. Reis and S. L. Gable, "Toward a positive psychology of relationships.," 2003.

[11] L. C. Hawkley and J. T. Cacioppo, "Loneliness matters: a theoretical and empirical review of consequences and mechanisms," *Ann. of Behavioral Medicine*, vol. 40, no. 2, pp. 218–227, 2010.

[12] C. A. Czeisler et al., "Bright light resets the human circadian pacemaker independent of the timing of the sleep-wake cycle," *Science*, vol. 233, no. 4764, pp. 667–671, 1986.

[13] N. D. Lane et al., "A survey of mobile phone sensing," *Commun. Magazine, IEEE*, vol. 48, no. 9, pp. 140–150, 2010.

[14] W. Dong et al., "Modeling the co-evolution of behaviors and social relationships using mobile phone data," in *Int. Conf. on Mobile and Ubiquitous Multimedia*. ACM, 2011, pp. 134–143.

[15] S. T. Moturu et al., "Using social sensing to understand the links between sleep, mood, and sociability," in *Int. Conf. on Social Comput.* IEEE, 2011, pp. 208–214.

[16] A. Sano and R. W. Picard, "Stress recognition using wearable sensors and mobile phones," in *ACII*. IEEE, 2013, pp. 671–676.

[17] A. Bogomolov et al., "Daily stress recognition from mobile phone data, weather conditions and individual traits," in *Int. Conf. on Multimedia*. ACM, 2014, pp. 477–486.

[18] G. Bauer and P. Lukowicz, "Can smartphones detect stress-related changes in the behaviour of individuals?," in *Int. Conf. on Pervasive Comput. and Commun.* IEEE, 2012, pp. 423–426.

[19] R. LiKamWa et al., "Moodscope: building a mood sensor from smartphone usage patterns," in *Int. Conf. on Mobile systems, applications, and services*. ACM, 2013, pp. 389–402.

[20] Andrey Bogomolov, Bruno Lepri, and Fabio Pianesi, "Happiness recognition from mobile phone data," in *Social Computing (SocialCom), 2013 International Conference on*. IEEE, 2013, pp. 790–795.

[21] M. S. Hussain et al., "Affect detection from multichannel physiology during learning sessions with autotutor," in *AIED*. Springer, 2011, pp. 131–138.

[22] J. Healey and R. Picard, "Digital processing of affective signals," in *Int. Conf. on Acoustics, Speech and Signal Processing*. IEEE, 1998, vol. 6, pp. 3749–3752.

[23] I. Arroyo et al., "Emotion sensors go to school.," in *AIED*, 2009, vol. 200, pp. 17–24.

[24] E. Vyzas, *Recognition of emotional and cognitive states using physiological data*, Ph.D. thesis, MIT, 1999.

[25] M.Z. Poh et al., "A wearable sensor for unobtrusive, long-term assessment of electrodermal activity," *Biomedical Eng.*, vol. 57, no. 5, pp. 1243–1252, 2010.

[26] A. Sano and R. W. Picard, "Recognition of sleep dependent memory consolidation with multi-modal sensor data," in *Body Sensor Networks (BSN)*. IEEE, 2013, pp. 1–4.

[27] A. Sano and R. W. Picard, "Comparison of sleep-wake classification using electroencephalogram and wrist-worn multi-modal sensor data," in *EMBC*. IEEE, 2014, pp. 930–933.

[28] A. Sano et al., "Discriminating high vs low academic performance, self-reported sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones," in *Body Sensor Networks (to appear)*, 2015.

[29] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. of Mach. Learning Research*, vol. 3, pp. 1157–1182, 2003.

[30] S. Blain et al., "Assessing the potential of electrodermal activity as an alternative access pathway," *Medical eng. & physics*, vol. 30, no. 4, pp. 498–505, 2008.

[31] S. Taylor et al., "Automatic identification of artifacts in electrodermal activity data," in *EMBC (to appear)*. IEEE, 2015.

[32] J. J. Ratey, *Spark: The revolutionary new science of exercise and the brain*, Hachette Digital, Inc., 2008.

[33] K. Truong et al., "Slide to x: unlocking the potential of smartphone unlocking," in *Human factors in comput. systems*. ACM, 2014, pp. 3635–3644.

[34] K. P. Murphy, *Mach. learning: a probabilistic perspective*, MIT press, 2012.

[35] L. Rokach, *Pattern classification using ensemble methods*, World Scientific, 2009.

[36] Rich Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.