Detecting Change in Depressive Symptoms from Daily Wellbeing Questions, Personality, and Activity

Orianna DeMasi[1], Adrian Aguilera[2,3], and Benjamin Recht[1,4]

[1]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley
[2]School of Social Welfare, University of California, Berkeley
[3]Department of Psychiatry, University of California, San Francisco
[4]Department of Statistics, University of California, Berkeley

Abstract— **Depression is the most common mental disorder and is negatively impactful to individuals and their social networks. Passive sensing of behavior via smartphones may help detect changes in depressive symptoms, which could be useful for tracking and understanding disorders. Here we look at a passive way to detect changes in depressive symptoms from data collected by users' smartphones. In particular, we take two modeling approaches to understand what features of physical activity, sleep, and user emotional wellbeing best predict changes in depressive symptoms. We find overlap in the features selected by our two modeling approaches, which implies the importance of certain features. Characteristics around sleep, such as change and irregularity of sleep duration, appear as meaningful predictors, as does personality. Our work corroborates prior results that sleep is strongly related to changes in depressive symptoms, but we show that even a very coarse measure has some predictive capability.**

I. Introduction

With the advancement in the sophistication and ubiquity of computing, the notion of real-time monitoring of behavior and emotional states has become plausible [1], [2]. Monitoring behavioral and emotional states via user input has already become relatively convenient with a proliferation of smartphone applications that can automatically remind users to log information about their state throughout the day. Logging and sharing data, particularly with health providers, can be beneficial because it can detect mood states that can benefit from intervention, either via mobile interventions or interventions from health providers.

To mitigate dropout, researchers have considered the possibility of smart apps that sense a user's behavior and automatically log their inferred state from data that is collected by a smart device without any user input [1]. The goal of automatic journaling has been attempted, in particular, for monitoring mood disorders, such as bipolar and depression [3], [4], [5], [6], [7], [8], [9], [10]. Such prediction capability would enable automatic long-term monitoring of emotional states, which is particularly applicable to mood disorders.

Research in automatic mood or emotion prediction has used simple single or double scales of wellbeing, such as "happiness" or the Circumplex model of affect and valence (wellbeing and energy) [11] as ground truth. These scales are implemented in basic user interfaces that automatically and randomly query the user throughout the day as ecological momentary assessments (EMA) of their wellbeing. Because the scales are simple, users comply more frequently, e.g., multiple times a day, for longer studies. While these scales are easy to measure, a disconnect arises with their relation to longer-term more thorough scales of mood and depressive symptoms.

In this study, we explore the ability to predict long-term changes in depressive symptoms, as measured by Beck's Depression Inventory (BDI) [12], from simple daily user input scales of affect and valence (the Circumplex model) and passively sensed data on user activity. We also compare the utility of daily Circumplex surveys with the utility of passively sensed user activity behavior. In particular we consider overall increase of Beck's Depression Inventory (BDI) [12] in an undergraduate cohort over the course of an academic semester. We ask two questions: whether daily self-reports of affect and valence during the semester can be indicative of overall changes in self-reported BDI scores from baseline to followup and whether passively sensed behavioral patterns are correlated with long-term mood changes, as quantified by changes in BDI scores. In addition to daily self-reports and activity behavior, we consider Big 5 [13] personality features: openness, extraversion, neuroticism, agreeableness, concienciousness.

This approach of predicting long-term changes in wellbeing is useful for developing targeted interventions. Detecting long-term changes would also be beneficial for monitoring wellbeing, especially of a population, such as in a randomized control trial of a treatment. Predicting absolute levels of depression from smartphones has proven difficult [3], [10], so we narrow to an equally useful goal of predicting changes.

We find that the relationship between daily reported affect and valence measures with changes in long-term measures of mood is complex. Other features, such as passively sensed user activity level and sleep duration are far more predictive of increases in depressive symptoms than features on daily surveys. We also find that the openness of a user's personality is very strongly correlated with whether they experience an increase in depressive symptoms. The strength of correlations between features and changes in BDI is established by considering small p-values on coefficients in linear regression models and being selected with a large coefficient in a Lasso

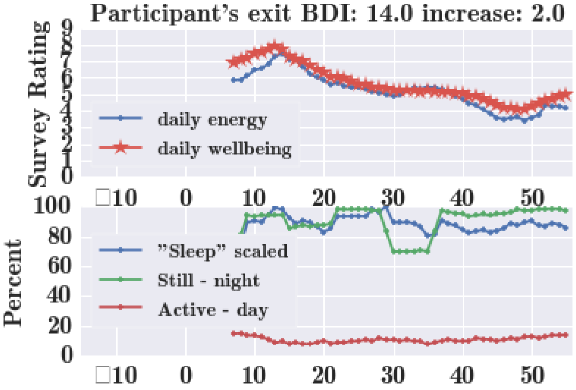

Fig. 1. Example behavior of an individual. Note apparent decrease in average daily wellbeing and energy. This decrease corresponds with a reported increase (in two points) to depressive symptoms (BDI score). The sensed activity and sleep behavior is relatively consistent during the study. Sleep is scaled by the maximum duration sample to make units comparable to daily percents.

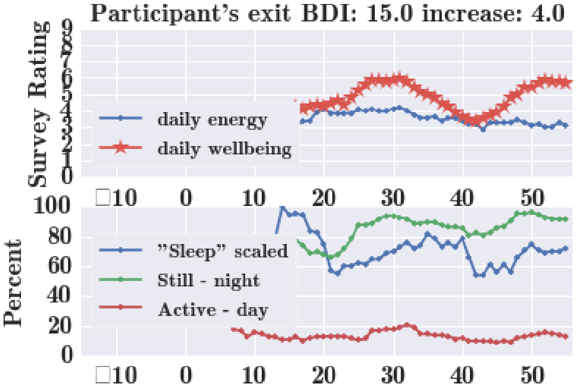

Fig. 2. A user's behavior. Note apparent increase in daily wellbeing and energy measurements, but a reported increase in long-term depressive symptoms (BDI score), which contrasts with the previous user in Figure 1. This user also has considerable fluctuation in their daily activity level and sensed sleep duration.

penalized linear model.

Our work supports prior studies that used more precise predictions of sleep duration (via collecting data on more sensors than we consider). We show that the correlation of sleep duration is so powerful that perhaps more coarse measures, i.e., loose predictions from a single sensor rather than an ensemble, are sufficient. We also find a significant impact from one outlying user, which highlights the need for larger populations with more variance to protect from overfitting artisanal datasets.

We will begin by placing the contribution of this study in the context of previous related work. We continue by briefly describing the dataset that we collected during our user study and then discuss the data processing, features extracted, and how the features could be related to the objective. We then explicitly state the two modeling approaches that we take and their merits. Our observed results on these two modeling approaches are described then followed by a discussion of the results and final conclusions.

II. Contributions

We make two significant contributions. First, we build on prior work that looked at utility of simple daily measures of wellbeing [14]. Rather than attempting to reproduce daily measures of wellbeing as ground truth, we look for relations of the daily measures with long-term changes in depressive symptoms. We would like to understand if features derived from daily measures of wellbeing are correlated with long-term changes in more thorough scales.

Our second contribution is an exploration of whether passively sensed behavioral features, particularly physical activity and sleep, are more predictive of long-term changes than the simple daily surveys of affect and valence (wellbeing and energy). We identify which behavioral features are most strongly correlated with long-term changes and could be used eventually as potential indicators of increase in depressive symptom expression. These data could improve the identification of depressive symptoms that could lead to targeted mobile or live intervention.

III. Related Work

There is a growing body of research that looks at using smartphones as sensors, particularly for mood. Various authors have shown correlations of daily emotion with call and SMS logs [4], [8], [9], [15], phone processes [9], Bluetooth [4], GPS location traces [3], [6], [7], [8], [9], [10], [15], sound data [4], physiology sensors (from wristbands) [8], [16], and macro-activity data [15], [16]. The majority of these authors have looked at predicting simple daily measures of mood over long periods. However, some authors have looked at more clinical measures of mood such as the PHQ-9 [3], [6], [7], [10]. Few authors have tried to predict values of long-term mood measures [10] or changes in outcome measures [3] from passively sensed data.

Here we focus on the long-term outcome measure (change in depressive symptoms, as measured by the BDI,) as the most important signal to predict. These longer term measures are more widely accepted as impactful from a medical community and the utility of simplistic daily emotional measures has yet to be confirmed.

In this work, we utilize physical activity as the behavioral input due to the large body of research that supports that there is a strong relation of mental wellbeing with activity levels and sleep [17], [18], [19], [20], [21], [22], [23], [24], [25]. Further there has been a large body of work that has shown that smartphone accelerometer data can be used to sense both physical activity through activity recognition [26], [27], [28], as well as sleep [29], [30]. Other pilot projects implied that mental states can be recovered from accelerometer data on small populations in artificial settings [2], [31].

IV. Motivational Example

In a variety of studies, simple scales of user emotional wellbeing have been used as ground truth and, more importantly, as a surrogate for more meaningful measures of

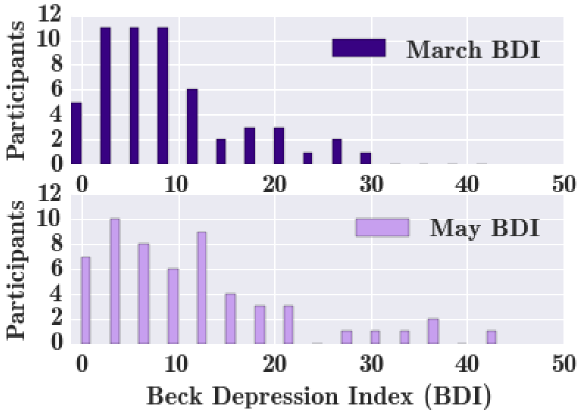

Fig. 3. Distribution of BDI scores that participants reported in the entry survey in March and in the exit survey in May. Note the slight drift of the distribution to higher BDI scores (more expressed depressive symptoms) over the course of the semester. The study ended the week before finals.

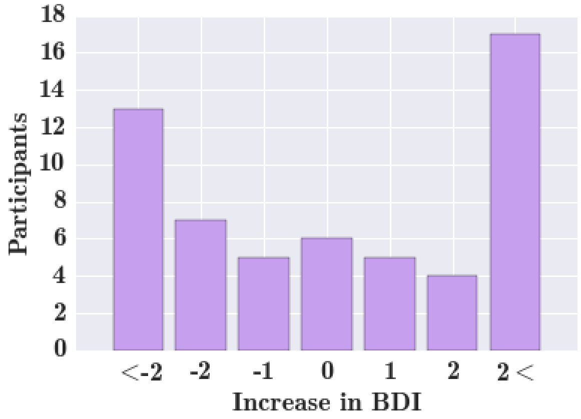

Fig. 4. The distribution of individuals' change in BDI score from the entry to the exit survey. A positive increase indicates an increase in BDI score (expression of depressive symptoms). During the course of the semester more students experienced an increase in depressive feelings than a decrease.

mood [4], [8], [9]. However, it's not clear whether, and if so how, these daily emotion measures are related to long-term mood. For example, two users' behavior and input is displayed in Figures 1 and 2. In Figure 1 the user's average daily wellbeing inputs appear to generally decrease during the course of the eight week study. (Mean daily reports are smoothed across the preceding week based on previous results which found this weekly average to be correlated with weekly PHQ-9 scores [14].) The user in Figure 1 reported a two point increase in BDI score (depressive symptoms) between the entry and exit surveys. The behavior of another user is displayed in Figure 2. This user's average emotional wellbeing displays significant fluctuation during the study period, but does not clearly decrease. However, the entry and exit surveys indicated that the user's depressives symptoms (BDI score) increased four points during the study, which was a greater increase than the user in Figure 1 reported.

These two figures give an example of how relations of daily emotion input has a complex relation to overall changes in mood. These two users' behavior imply that mappings from daily input to long-term change may be difficult to construct.

V. Field Study

To answer our research question of how daily self-reports of emotion and daily measurements of activity and sleep are related to overall changes in mood, we conducted a field study. We recruited 107 students at the University of California, Berkeley. These students were required to be native english speakers, have their own Android smartphone, and install our custom built app. The application would prompt the users to enter their wellbeing and energy level (Circumplex affect and valence) four times a day during the eight week study period from mid March through the beginning of May. We elected to use the Circumplex model of emotion [11] to align with previous work that has adopted this model [1], [9], [16], [32]. The application also collected a variety of data from sensors on the participants' smartphones, including from the accelerometer motion sensor. Data was collected from the accelerometers for three seconds every 5 minutes. The study period was chosen to conclude shortly before finals so that students would be more likely to participate through the full study. Students received compensation and the study was approved by the Internal Review Board.

VI. Dataset

Due to high attrition and missing data from the entry and exit surveys, we consider a dataset comprised of 44 participants, 27 of whom were female. The distributions of BDI scores reported by the participants for the entry (March) and exit (May) surveys are shown in Figure 3. While most participants reported a score less than 10, a few participants reported higher scores and the average score increased from 11.5 reported in March to 12.4 reported in May. A larger BDI score corresponds to increased depressive symptoms reported, so the majority of our study population reported minimal depressive symptoms. The distribution of changes in BDI scores between the entry and exit survey is shown in Figure 4. This figure shows that there was a broad experience among participants during the semester and some experienced a significant change in their response during the academic semester.

Three entry surveys and one exit survey were missing the response to one BDI question due to user error or a user electing not to answer. In these four cases, the difference in BDI score was calculated between answered questions. The entry and exit BDI scores were computed by scaling the weight of all other questions to be slightly more significant, so that the total possible sum of the 20 answered BDI questions was the same as the full 21 question survey.

Our study composed of three phases of user input: an entry survey, daily prompts, and an exit survey. During the entry and exit survey users were asked to self-report their responses to 20 questions from the Beck Depression Inventory. The question regarding suicidal thoughts was omitted due to concerns from the Internal Review Board. The entry and exit survey also collected Big 5 personality scores [13] and demographic information. All questions were optional.

In addition to the user input data, we collected data from sensors on the participants' smartphones, including from the accelerometer sensor. We collected accelerometer data using funf [33] at intervals of three seconds every five minutes. The accelerometer data was collected continuously from install time. Quality and volume of data varied greatly between participants and phone models. Some of the difficulties encountered included entirely missing observations, nonuniform readings during an observation interval, and insufficient duration of sampling, i.e., too few readings during an observation interval.

VII. Data Processing - Activity Extraction

A smartphone's accelerometer collects acceleration of the phone along three axes at every reading. These readings constructed time series that we featurized similar to the approaches found in previous work [28]. These time series features were passed to classifiers which made predictions of whether the phone was "still" or set down during an observation or whether the user was was physically active, such as walking, running, or cycling. These momentary observations of activity and stillness were collected for each day of the study and the percent of the day and previous night (1am - 7am) during which a user was physically active or the phone was still were calculated. Additionally we approximate sleep time as the longest duration that the phone was set down during the evening hours. We will refer to this duration of stillness as "sleep". This measure of sleep seems noisy, but a similar approach was found to approximate sleep to within roughly 45 minutes of true sleep time [29]. Through this process we end up with measures of the percent of time during a day and night a user spends active or still and the duration during the evening that the phone is set down and the user presumably sleeping. These measures were then averaged over seven day periods to give smoothed average activity and sleep measures. This averaging adds some robustness.

VIII. Features

The features that we use to describe participants and their behavior during the study are summarized in Table I. The behavior and self-report features were calculated on user input daily wellbeing and energy. They were also calculated on the signals we gleaned from the sensor data: percent of time the participant was active during the day and night, percent of time the phone was still during the day and night, and the "sleep" duration. The observation entropy feature was calculated on the distribution of each signal. Similar to standard deviation, it quantifies the irregularity of the signal. The difference features try to quantify changes from baseline to end of study, irrespective of intermediate fluctuations. We consider timescales of a week to stabilize daily fluctuations and to follow prior work which showed that a weekly mood average was related to weekly PHQ-9 scores [14].

Feature type	Name	Description
Personality	Neurotic	Big 5 personality test
	Extraversion	Big 5 personality test
	Openness	Big 5 personality test
	Agreeable	Big 5 personality test
	Conciencious	Big 5 personality test
Behavior and self-reports	Avg. Obs.	Mean of observations
	Obs. Stdev.	Standard deviation of observations
	Obs. Slope	Regression coefficient of observations on time
	Obs. Entropy	Entropy of observation distribution
	Diff. last week	Difference of average measurement during last week with baseline
	Diff. last 2 weeks	Difference of average measurements during last two weeks with baseline

TABLE I

Features collected and computed on each participant. The baseline of a measurement was calculated as the average over the first four weeks of the study.

IX. Methodology

In this work our goal is to understand behavioral factors that are correlated with long-term changes in participants' depressive symptoms (BDI scores) during the course of the academic semester from March to May. Our secondary goal is to use that information to successfully predict a change in depressive symptom expression. For these tasks we are interested in which features are strongly correlated and predictive of the outcome change in BDI score. To identify correlated and predictive features, we choose to use linear models because they have clear interpretations and are thus ideal for feature selection and model insight.

A. Feature Selection

To explore the relevancy of features, we use linear regression models, as these models are highly interpretable. However, we choose two methods of feature selection with these models: forward selection with the Bayesian Information Criterion (BIC [34]) to choose which of the features should be added at each subsequent step and when forward selection should terminate, and linear regression with the L1 (Lasso) penalty [35]. Both of these methods yield models with a limited number of terms and a coefficient on each terms that indicates how much that term contributes to the model.

B. Feature Comparison

To make the weights of features comparable (despite different scales), we scale all features to unit variance. While this is artificial, it yields models where features are on comparable scales and thus comparisons between feature coefficients are more insightful.

	Forward Selection		Lasso	
	All Obs.	No Outlier	All Obs.	No Outlier
Feature Name	Coefficient Value			
Openness	3.6640 (*)	3.1895 (*)	1.99	2.361
Sleep duration [Obs. Stdev.]	7.2069 (*)	x	5.599	x
Sleep duration [Slope]	-6.9844 (0.001)	x	-2.204	x
Sleep duration [Diff. last 2 weeks]	4.7048 (0.017)	x	x	x
Daytime activity [Avg. Obs.]	x	x	0.342	x
Daytime activity [Diff. last 2 weeks]	x	x	-0.067	x
Daytime stillness [Obs. Stdev.]	-3.3079 (0.001)	x	-1.019	x
Daytime stillness [Diff. last week]	1.5866 (0.053)	x	x	x
Daily energy [Entropy]	x	x	0.150	x
Daily energy [Diff. last 2 weeks]	x	x	-0.101	-0.209
Model R^2	0.785	0.404	0.704	0.392
Model MSE	61.996 [23.859]	15.849	96.149 [29.156]	16.939

TABLE II

Comparison between various modeling approaches of features selected, model fit (R^2), and mean squared error (MSE) of prediction. Approaches attempted to model the change in participants' BDI scores from the beginning of the study to the end. P-values for linear regression coefficients are in parenthesis, where appropriate, and * denotes values less than 0.001. The baseline MSE with the outlier was 83.212 and with the outlier removed was 25.184. The numbers in square brackets are the MSE calculated on the set of not outliers.

C. Outliers

There is a single outlier in our dataset of one participant who experienced a particularly difficult semester. This outlier had a dramatic effect on the models due to our small population size. Rather than controlling for the observation, we present models with and without the outlier.

X. Results

The feature selection, model fit, and prediction accuracy using both of the regression approaches outlined above are presented in Table II. There are four models presented in Table II. Two of the models presented used forward selection with the BIC and two of the models used L1-penalized linear regression. The difference between the models using the same modeling approach is that one of the models has a single outlier removed. The models were fit with intercept terms, but those terms are omitted for brevity.

A. Linear Regression with Forward Selection and BIC

The first (left most) column of coefficients in Table II presents a linear model that was fit to the entire dataset. The features were selected by using forward selection and choosing models that minimized the BIC. This modeling procedure resulted in six features being selected, five of which were statistically significant (p-values < 0.05). The features selected were the participant's openness (as measured by the Big 5 personality survey,) and features quantifying variability and change in both the duration of "sleep" (stillness during the evening) and fraction of time still during the day. Aside from the feature quantifying one dimension of the participants' personality, all the other features result from accelerometer measurements, and particularly measurements of when the phone is not in motion, but presumably set down. The model has reasonably high R^2 of 0.785 indicating that a large fraction of the variability of the data is explained by these five features.

B. L1 (Lasso) Penalized Linear Model

The third column of coefficients (second from the right) in Table II presents the model that is selected for an L1-penalized linear regression model. With this modeling approach, features are selected by adding a penalty to the model accuracy term everytime a coefficient is included. This process drives the coefficients of unnecessary terms to zero and thus removes them from the model.

The Lasso approach selects the largest model that we observe with eight features. Again, the openness of a participant's personality is selected as highly predictive of the increase in BDI score during the semester. Features describing the change in and variability of the participant's daytime stillness and "sleep" are also selected. In contrast to the model chosen with forward selection, the Lasso penalized linear model selects two features describing the variability and change in the participants' self-reported energy levels. Two features describing the average activity level and change in average activity level during the day are also selected. It is interesting that two features on the participants' energy levels are selected, but no features on the participants' self-reported wellbeing are selected. It is also interesting that out of the eight features selected, five of them are describing the activity of the participant, as measured by the users' smartphones.

C. Removing the Outlier

In the collected dataset, there was a single outlier. The outlier resulted from a single participant experiencing a particularly difficult semester and unfortunately reporting a increase in BDI score of 50. The second largest change in score was 14, so one participant was an outlier and had significant impact on the model selection. To explore the robustness of the previous models, we used the same methodology to fit two models, one with forward selection and the second with a Lasso penalty, to the dataset with the

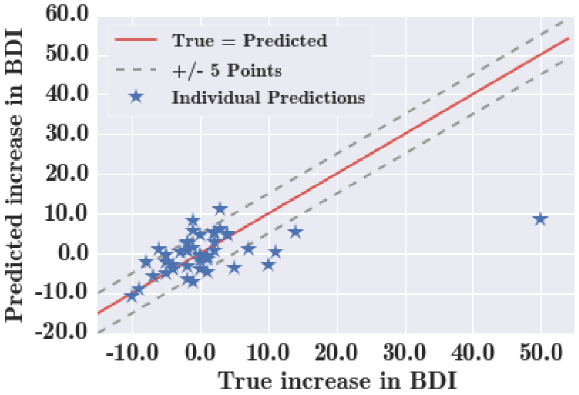

Fig. 5. The distribution of predictions from leave-one-out cross-validation. Features were preselected with forward selection and models were fit on population with the outlier included. Most predictions are within the dotted lines indicated predictions within five points of the true increase.

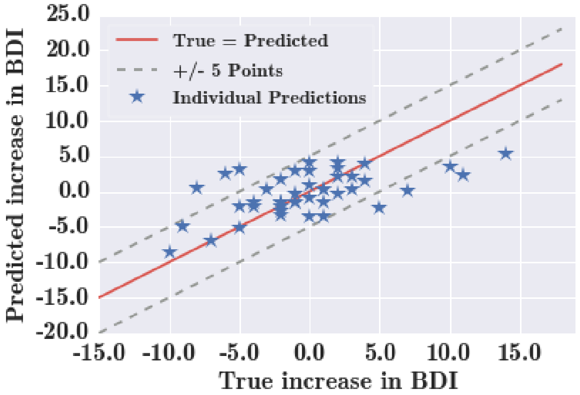

Fig. 6. The distribution of predictions with leave-one-out cross-validation. Models were fit on a population with the outlier removed and features were pre-selected by forward selection with the BIC.

single outlying participant removed. These two models are presented in the second and fourth (right most) columns of coefficients in Table II.

The resulting models are considerably different from the models selected with the outlier included. This result reveals that the models were very sensitive to the one participant's experience. However, the openness of participants is still selected as a feature with both forward selection and Lasso penalized regression, which implies that this personality feature is resolutely correlated with the change in the BDI score. Of note is also the selection of the change in self-reported energy from the beginning to the end of the study with Lasso penalized regression. We are trying to model the change in depressive symptoms (change in BDI), but the self-reported energy and not self-reported wellbeing is being selected as an important feature.

D. Leave-One-Out Prediction Accuracy

As a final test of our models, we tried using them for prediction and measuring the accuracy of trained models' predictions on a holdout set. Due to the constrained size of the dataset (44 participants,) we used leave-one-out cross-validation. In this approach, one user is held out, a model is trained on all the other users, and the error of the trained models is then measured by the error in prediction on the held out user. This process is repeated for all users. The error measured is mean squared error (MSE) and it is averaged across all participants to yield the MSE reported in the bottom row of Table II. A lower MSE is better with zero indicating a perfect model. When considering these numbers, one should consider the baseline. We consider the baseline MSE to be the MSE returned when the null "model" is used. We consider the null model to be when the mean of the dataset is always returned as the prediction, i.e., when no features are considered, only the population baseline. This null model results in a baseline MSE of 83.212 when the outlier is included and 25.184 when the outlier is excluded from the dataset.

We see that the models constructed with forward selection and the BIC yield MSE's lower than the baselines, which implies those models have better prediction accuracy that predicting the mean of the dataset uniformly. The model fit with the Lasso penalty does not yield a MSE (96.149) lower than the baseline (83.212). However, when the outlier is excluded, the Lasso penalized model does yield a better model (MSE = 16.939) than the baseline (MSE = 25.184) with just two features. This result highlights the strength of the correlation of the openness of a participant with the increase in BDI score they experienced during the semester.

Figure 5 displays the distribution of predicted BDI increases relative to the true increases in BDI scores observed when the outlying participant is included in the dataset. Figure 6 is similar, but displays the distribution of predictions on a dataset with the outlier excluded. Predictions in both figures were generated by fitting linear models on the features that were selected with forward selection and the BIC in leave-one-out cross-validation schemes. Both figures show little structure in the error of predictions, i.e., BDI increase is not consistently under or over predicted. Further, these figures show that the majority of predictions are within five points (the dotted lines) of the true reported increase in BDI.

XI. Discussion

In the above sections, we have explored which features, from a set of 47, were most predictive of participants' increase (or decrease) in BDI scores between the beginning and end of our eight week study. To gain insight from modeling the data, we have chosen to use linear models for their interpretability. Due to our small population, we have pursued two feature selection approaches: forward selection with the BIC and Lasso regression. By comparing these two different approaches, we hope to reduce over-extrapolation from our small population.

When modeling the full population, sleep features were not only selected, but found to be most impactful for prediction, i.e., large coefficients in both approaches and small p-values with forward selection. This result is in line with prior results which looked at more sophisticated predictions of sleep duration from multiple sensors [3]. Our result highlights how important these features are: even our coarse approximation to sleep with one sensor is significantly predictive. To a lesser extent, activity levels and irregularity of stillness during the day (7am - 1am the next day) are predictive and selected in both models.

Of notable absence is any feature derived from daily reported emotional wellbeing or affect. Only two features derived from each set of reports were loosely related (small coefficient values) to increase in BDI score when the Lasso penalized model was used. As these measurements are meant to be a brief estimate for more thorough measures, one would think they could be correlated with the increase in BDI. However, none of the features we constructed around daily wellbeing, or the change in it, were ever found to be correlated, regardless of modeling approach. The irregularity and change from baseline of daily energy was chosen to be predictive in the Lasso regression, but not daily wellbeing. This result implies that daily mood scores may be an insufficient measure, or that at least it is not straight forward to correlate such a noisy measure of emotion with longer term changes in depressive symptoms. Daily self-reports are tedious to comply with for an ongoing basis, so if their application is unclear, it is possible that alternative metrics should be considered for measurement. Another factor that could account for the lack of affect and valence features is missing data.

A major hindrance to our approach is missing data. As the study progressed, participation waned. This waning resulted in a poorly sampled or observed period before the exit survey was offered, and thus final May BDI was recorded. It is possible that with better observation immediately before recording the May BDI score, more features constructed on the daily self-reports would have been selected or found to be statistically significantly correlated.

Similar to missing data, data quality was a problem. Our population had a variety of phone models that yielded a range in the quality and regularity of data recorded on each participant. It was not possible with our limited population size to explore to what extent the quality of data recorded by individual devices affected our results.

Unfortunately very few features are left significant when the outlier was removed from the population. The only feature that is found to be significant in every model regardless of if the outlier is removed, is the Big 5 openness dimension. This result speaks to the importance of personality, or the strength of the correlation between a person's predisposition to having an increase in depressive symptoms and the observation of a change in BDI score. This strong correlation could also have impact for academic administrators who are concerned with how students fare during semesters and the stresses imposed by undergraduate life.

The result of lost significance when the outlier is removed speaks to the importance of every participant and observation in these small population, artisanal datasets. Overfitting must be carefully avoided and explored and outliers must be addressed to avoid presenting misleading results. The impact of our results is hindered by the small sample size. While our study population size is commensurate with previous studies, the population size is still small, which results in a strong tendency to overfit the dataset. We have tried to minimize overfitting by use of the BIC and forward selection and Lasso-penalized regression. Further, we have tried to limit our conclusions to insights about which features appear to have some correlation with the desired metrics (or rather which sets of features have little predictive capability). We do not focus on the overall predictiveness of the model, but which features are capable of explaining some of the variance in the observed dataset. The relatively large observed R^2 values of our two models are encouraging, but a larger sample population is needed for more definitive results. A population skewed to more clinical depression, rather than the general population that we observed, may also present different conclusions.

The loss of significance could also not speak to the lack of importance of the other features or the need for a larger population, but to the need for a population specifically with larger variation in baseline BDI scores and variability in mood, or change in BDI scores. Our student participants were selected from a general, non-targeted population. It is possible that a population more inclined to experience significant changes in mood, e.g., a clinically depressed population, would benefit from modeling with more features. However, the fact that our population did not experience a very large distribution in increase in BDI scores, means that there may have been little to predict. A single point increase or decrease in score could be little more than noise and thus very difficult to predict.

For future work, we would like to use these methods on a larger population with more depressive symptoms and where fluctuations are more demonstrative. Another approach we would like to consider is separating populations by gender, but for that a larger population is needed.

XII. Conclusions

We have explored the utility of different features for predicting increases in reported depressive symptoms (Beck's Depression Inventory). In particular, we sought to understand the utility of daily affect and valence self-reports for predicting increases in the BDI, as compared with passively collected activity and sleep features. We found relatively large R^2 values for both modeling approaches used, indicating the ability to model the data, and a variety of interesting insights into predictive features. We found that passively sensed data was actually more predictive of increases in BDI than the active user input.

While this work provides encouraging results corroborating that behavioral patterns can be measured by smartphones and used to predict meaningful metrics, more work is needed, specifically with a larger population. Comparing results on a clinically depressed population that has a different distribution of BDI scores is also an area deserving further investigation.

XIII. Acknowledgments

We would like to thank everyone who helped collect data and the study participants. This research is supported in part by DHS Award HSHQDC-16-3-00083, NSF CISE Expeditions Award CCF-1139158, DOE Award SN10040 DE-SC0012463, and DARPA XData Award FA8750-12-2-0331, and gifts from Amazon Web Services, Google, IBM, SAP, The Thomas and Stacey Siebel Foundation, Apple Inc., Arimo, Blue Goji, Bosch, Cisco, Cray, Cloudera, Ericsson, Facebook, Fujitsu, HP, Huawei, Intel, Microsoft, Mitre, Pivotal, Samsung, Schlumberger, Splunk, State Farm and VMware.

References

[1] N. Lathia, V. Pejovic, K. K. Rachuri, C. Mascolo, M. Musolesi, and P. J. Rentfrow, "Smartphones for large-scale behavior change interventions," *IEEE Pervasive Computing*, no. 3, pp. 66–73, 2013.

[2] M. Rabbi, S. Ali, T. Choudhury, and E. Berke, "Passive and in-situ assessment of mental and physical well-being using mobile sensors," in *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 2011, pp. 385–394.

[3] D. Ben-Zeev, E. A. Scherer, R. Wang, H. Xie, and A. T. Campbell, "Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health." *Psychiatric rehabilitation journal*, vol. 38, no. 3, p. 218, 2015.

[4] A. Bogomolov, B. Lepri, and F. Pianesi, "Happiness recognition from mobile phone data," in *Social Computing (SocialCom), 2013 International Conference on*. IEEE, 2013, pp. 790–795.

[5] M. N. Burns, M. Begale, J. Duffecy, D. Gergle, C. J. Karr, E. Giangrande, and D. C. Mohr, "Harnessing context sensing to develop a mobile intervention for depression," *Journal of Medical Internet Research*, vol. 13, no. 3, p. e55, 2011.

[6] L. Canzian and M. Musolesi, "Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 1293–1304.

[7] A. Gruenerbl, V. Osmani, G. Bahle, J. C. Carrasco, S. Oehler, O. Mayora, C. Haring, and P. Lukowicz, "Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients," in *Proceedings of the 5th Augmented Human International Conference*. ACM, 2014, p. 38.

[8] N. Jaques, S. Taylor, A. Azaria, A. Ghandeharioun, A. Sano, and R. Picard, "Predicting students' happiness from physiology, phone, mobility, and behavioral data," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 222–228.

[9] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong, "Moodscope: Building a mood sensor from smartphone usage patterns," in *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. ACM, 2013, pp. 389–402.

[10] S. Saeb, M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording, and D. C. Mohr, "Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study," *Journal of medical Internet research*, vol. 17, no. 7, 2015.

[11] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[12] A. T. Beck, C. Ward, M. Mendelson *et al.*, "Beck depression inventory (bdi)," *Arch Gen Psychiatry*, vol. 4, no. 6, pp. 561–571, 1961.

[13] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of personality*, vol. 60, no. 2, pp. 175–215, 1992.

[14] A. Aguilera, S. M. Schueller, and Y. Leykin, "Daily mood ratings via text message as a proxy for clinic based depression assessment," *Journal of affective disorders*, vol. 175, pp. 471–474, 2015.

[15] Y. Ma, B. Xu, Y. Bai, G. Sun, and R. Zhu, "Daily mood assessment based on mobile phone sensing," in *Wearable and implantable body sensor networks (BSN), 2012 ninth international conference on*. IEEE, 2012, pp. 142–147.

[16] J. Healey, L. Nachman, S. Subramanian, J. Shahabdeen, and M. Morris, "Out of the lab and into the fray: towards modeling emotion in everyday life," in *Pervasive computing*. Springer, 2010, pp. 156–173.

[17] F. Dimeo, M. Bauer, I. Varahram, G. Proest, and U. Halter, "Benefits from aerobic exercise in patients with major depression: a pilot study," *British journal of sports medicine*, vol. 35, no. 2, pp. 114–117, 2001.

[18] K. R. Fox, "The influence of physical activity on mental well-being," *Public health nutrition*, vol. 2, no. 3a, pp. 411–418, 1999.

[19] H. G. Lund, B. D. Reider, A. B. Whiting, and J. R. Prichard, "Sleep patterns and predictors of disturbed sleep in a large population of college students," *Journal of adolescent health*, vol. 46, no. 2, pp. 124–132, 2010.

[20] C. M. McKercher, M. D. Schmidt, K. A. Sanderson, G. C. Patton, T. Dwyer, and A. J. Venn, "Physical activity and depression in young adults," *American journal of preventive medicine*, vol. 36, no. 2, pp. 161–164, 2009.

[21] S. T. Moturu, I. Khayal, N. Aharony, W. Pan, and A. Pentland, "Using social sensing to understand the links between sleep, mood, and sociability," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, 2011, pp. 208–214.

[22] F. J. Penedo and J. R. Dahn, "Exercise and well-being: a review of mental and physical health benefits associated with physical activity," *Current opinion in psychiatry*, vol. 18, no. 2, pp. 189–193, 2005.

[23] J. J. Pilcher, D. R. Ginter, and B. Sadowsky, "Sleep quality versus sleep quantity: relationships between sleep and measures of health, well-being and sleepiness in college students," *Journal of psychosomatic research*, vol. 42, no. 6, pp. 583–596, 1997.

[24] J. J. Pilcher and E. S. Ott, "The relationships between sleep and measures of health and weil-being in college students: A repeated measures approach," *Behavioral Medicine*, vol. 23, no. 4, pp. 170–178, 1998.

[25] A. Ströhle, "Physical activity, exercise, depression and anxiety disorders," *Journal of neural transmission*, vol. 116, no. 6, pp. 777–784, 2009.

[26] O. D. Incel, M. Kose, and C. Ersoy, "A review and taxonomy of activity recognition on mobile phones," *BioNanoScience*, vol. 3, no. 2, pp. 145–171, 2013.

[27] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *Communications Surveys & Tutorials, IEEE*, vol. 15, no. 3, pp. 1192–1209, 2013.

[28] H. Lu, J. Yang, Z. Liu, N. D. Lane, T. Choudhury, and A. T. Campbell, "The jigsaw continuous sensing engine for mobile phone applications," in *Proceedings of the 8th ACM conference on embedded networked sensor systems*. ACM, 2010, pp. 71–84.

[29] Z. Chen, M. Lin, F. Chen, N. D. Lane, G. Cardone, R. Wang, T. Li, Y. Chen, T. Choudhury, and A. T. Campbell, "Unobtrusive sleep monitoring using smartphones," in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*. IEEE, 2013, pp. 145–152.

[30] A. Sano and R. W. Picard, "Comparison of sleep-wake classification using electroencephalogram and wrist-worn multi-modal sensor data," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 2014, pp. 930–933.

[31] R. Byrne, P. Eslambolchilar, and A. Crossan, "Health monitoring using gait phase effects," in *Proceedings of the 3rd International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, 2010, p. 19.

[32] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas, "Emotionsense: a mobile phones based adaptive platform for experimental social psychology research," in *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 2010, pp. 281–290.

[33] N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland, "Social fmri: Investigating and shaping social mechanisms in the real world," *Pervasive and Mobile Computing*, vol. 7, no. 6, pp. 643–659, 2011.

[34] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[35] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.