Title

Evaluation of Methods for Data-Driven Tools that Empower Mental Health Professionals

Author

Demasi, Orianna Amy

Publication Date

2019

Peer reviewed|Thesis/dissertation

Evaluation of Methods for Data-Driven Tools that Empower Mental Health Professionals

by

Orianna Amy Demasi

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Benjamin Recht, Chair
Professor James Demmel
Professor Marti Hearst

Spring 2019

Evaluation of Methods for Data-Driven Tools that Empower Mental Health Professionals

Abstract

Evaluation of Methods for Data-Driven Tools that Empower Mental Health Professionals

by

Orianna Amy Demasi

Doctor of Philosophy in Computer Science

University of California, Berkeley

Associate Professor Benjamin Recht, Chair

It is estimated that nearly one in five adults in the United States live with mental illness, and for individuals who struggle with mental health, the experience can be excruciating. The rise of mobile devices presents a unique opportunity to improve mental health outcomes, in part through empowering mental health professionals. Because many individuals always have their smartphones with them, smartphones may be able to enable health professionals to identify when an individual is in need and provide immediate care, rather than the current model of delaying care until a scheduled appointment. To this end, I investigate the feasibility of two new data-driven tools: one to identify when care is needed and the second to help train counselors to intervene when care is needed. The first tool I consider seeks to use a smartphone to sense an individual's well-being. Such a tool could be used to inform health professionals of patients' states, evaluate the efficacy of therapies, and deliver just-in-time interventions. To evaluate the potential accuracy of such a tool, I collect students' passive smartphone data and self-reported well-being measures, and then consider predicting well-being on a daily basis and detecting significant changes over a period of time. As this approach seems unreliable for most individuals, I further explore for which individuals such an approach may be reliable and develop a framework for evaluating longitudinal sensing quality. I find that while correlations between smartphone-sensed measures and reported wellbeing scores exist, these relationships are often too weak to reliably predict wellbeing. The second tool I explore seeks to help suicide prevention counselors practice intervening over chat for individuals in crisis. For this, I collect and leverage synthetic conversation transcripts and show how to evaluate a baseline system for counselors to practice crisis de-escalation strategies in a no-risk environment. While text retrieval and generation methods can return responses that make sense in limited context most of the time, i.e., in greater that 50% of examples, generated responses are shorter than retrieving full messages, implying that generation may potentially be a less engaging approach. Overall, I find that significant consideration of context is needed to provide meaningful evaluation of methods for the tools envisioned. While popular algorithms and methods may hold potential to develop the tools

discussed, rigorous evaluation and further work is needed to ensure reliability within the application context.

For my mother and father

who have taught me to love learning

Contents

List of Figures

List of Tables

Acknowledgments

Many people have helped me on this journey and I am grateful for all the help, support, funding, and inspiration that I have received. First, I would like to thank my collaborators Adrian Aguilera, Alexandra Paxton, Aluma Dembo, Emma Bruehlman-Senecal, Katie Stasaski, Kevin Koy, Konrad Kording, Patricia Avila, Sidney Feygin and Crisis Support Services of Alameda County, who I could not have done my research without and who taught me so much. I have also received generous fellowships from ARCS and the Berkeley Institute for Data Science (BIDS) and project funding from the CITRIS Tech for Social Good program. The EECS and BIDS administrative staff have really made my life easier, especially Damon Hinson and Stacey Dorton. I would like to thank my colleagues from the BeBop group who welcomed me and helped me get started at Berkeley, in particular Andrew Gearhart and Professor Kathrine Yelick. During my final years at Berkeley, I was lucky to work with members of the Hearst Group and really appreciate all that they have taught me about NLP and HCI, especially Andrew Head, Nate Weinman, and Philippe Laban. Throughout my time at Berkeley, members of Modest Yachts (Recht Lab) have been so supportive as colleagues and friends. Thank you, in particular, Ashia Wilson, Becca Roelofs, Evan Sparks, Eric Jonas, and Esther Rolf for reading all those drafts I shared with you, giving feedback on talks, coming over for dinner, getting coffee, telling me about your research, and listening to my reasonable and my ridiculous ideas. I have really enjoyed learning from many more colleagues scattered throughout the EECS department, including Shiry Ginosar, Vidya Muthukumar, Nathan Pemberton, and Patrick Li. Sheila Humphreys, thank you for all the support you provide to everyone in the department. In general, people at Berkeley have been so wonderful and supportive, but Eric Love, Jordan Freitas, and Michael Driscoll, you really made it rock. Albert Magyar, thank you for being so wonderful and for listening, even when you have already heard it. While my extended Berkeley family has been great, I really could not have gotten here without my family. My sister and father have been wonderful, and I'll never forget how many late night calls my mother answered to hear about how my classes or research were going, even though she was on east coast time and had likely heard the same thing the day before. Thanks for always being there for me. Finally, I would like to thank Marti Hearst, James Demmel, and Benjamin Recht. Thank you Marti for all you have taught me, the help you provided, and in particular for helping me follow such a wonderful project. Thank you Jim for taking a chance on me and being so supportive, despite the winding path. Ben, you have been so supportive and patient with me and my research. It is incredible. I cannot thank you enough for believing in my path before I did. Although it has been challenging, it has also been an honor to be part of such an incredible community.

Chapter 1

Introduction

Mental illness is of growing concern, as it is expected to become the leading cause of global disease burden by 2030 [107]. As such, there are increased calls for urgent action to improve treatments and expand access to care [35]. While many efforts are needed, tools that help support, inform, and train care providers and tools that help individuals manage their own mental health may be a component of meeting the growing need. Data-driven tools in particular may be of use, as they can tailor to individuals and environments.

1.1 Motivation

Data and algorithms have shown profound impacts on our daily lives and many researchers have begun exploring the role that they can play in mental health. For example, gaining insights on mental health through novel data sources and machine learning methods is of growing interest and potential. Studies have considered the ability to track mental health signals on a population scale from Twitter [42], and on individual scales for many digital traces [31, 38, 46, 64, 129]. Online behaviors have also yielded insights into language use with mental health conditions [43, 60]. Many of these insights were previously not possible to glean from sparse data that was only available from infrequent visits to mental health professionals.

Insights on possible mental health conditions are valuable, but even more exciting are the tools and interventions that they can inspire and inform. For example, researchers have explored building tools that screen online forums for potential addiction recovery lapse risk [78], medical record screening tools for suicide risk [47], therapeutic agents [100], or making recommendations for positive behavior change, e.g., increasing activity [123, 124] or decreasing insomnia [122]. Such tools have the potential to positively influence individual's lives and are starting to emerge.

Because these tools are designed explicitly to impact human lives, it is important that they are developed responsibly so that they deliver their intended positive impact. There is growing discussion of how to use algorithms responsibly and equitably [105], and fear about algorithms replacing human jobs [56]. These discussions are particularly important for developing health technologies that are designed with the intent of interacting with humans. One way of developing beneficial tools may be to focus on enhancing the work humans do and critically evaluating how and when tools can augment human efforts.

In this dissertation, I look at developing two algorithm-based tools that can empower mental health professionals. The first tool seeks to inform mental health care providers by providing clinicians with information on patients' wellbeing. This information could enable detecting times of need and thus the delivery of treatments that were previously not possible, e.g., just-in-time, personalized behavioral interventions [122, 123, 124, 137], and improve therapy evaluation. The second tool I consider seeks to ease the training of humans through engaging simulations to help humans intervene when an individual is in need. Simulations could both improve the quality of training through increased practice and free up experienced counselors to provide more care to those in need by decreasing the time they need to spend training others.

An underlying theme within this dissertation is a quest for rigorous evaluation of complex methods. In addition to designing new tools, it is imperative that they work, i.e., they reliably achieve their intended goals. Towards this end, I consider tools that are inspired by the potential of algorithms, but I take a cautious view of employing algorithms before deploying them into individuals' daily use.

Another theme within this dissertation is working with human participants to collect realistic datasets. Such datasets enable studying complex human behavior that is difficult to approximate with surrogate or simulated data. Here, in collaboration, I collected two datasets to study and develop monitoring and training technologies. The first dataset included subjective human input on wellbeing in addition to passively collected mobile phone data. The second dataset modeled realistic counseling conversations by employing experienced suicide prevention counselors.

1.2 Automatic Journaling of Wellbeing for Monitoring Patients and Just-in-time Interventions

The first problem I consider is developing a mobile phone application for automatically monitoring mental wellbeing.

Mood journaling is an important tool for managing mood disorders, as it can often be a first step towards understanding one's wellbeing [74, 30] and then influencing it. Especially for varying mood disorders, such as bipolar disorder, journaling can help individuals' be more aware of their emotions, better understand their conditions, and eventually identify patterns or triggers, that them help them build actionable plans for navigating their experience [16]. Unfortunately, continual journaling of wellbeing for life-time conditions is not only difficult, but unrealistic.

Automatic tracking of some behavioral patterns has become possible with smart devices, such as smartphones and smartwatches. These devices stream continual data on many personal aspects of our lives and a growing corpus of research is showing relationships of these data with highly personal information. Early work showed that complex social dynamics could be reconstructed from mobile phone bluetooth logs [52]. Later work has expanded this result by showing that smartphone data may be telling of even more nuanced behaviors, such as sleep [32, 97] and physical activity patterns [11, 85, 91].

Some of the behaviors that smartphones can infer have also been shown to be related to mental wellbeing. For example, sleep has been shown in multiple studies to have a relationship with depression [143, 153] and other studies have found different relationships between sleep quality and sleep quantity on mental wellbeing [117]. Research has also indicated that physical activity is related to mental wellbeing [113, 144] and even suggested physical activity may influence mood to the point of helping control depression [57]. Additional work has suggested automatically tracking behaviors with a smartphone application as a way to influence wellbeing [82].

As a result of these relationships between mobile data and behavioral signals, there is growing interest into whether smart devices can track wellbeing [28, 71, 87, 99, 132, 133]. If a smart device could track someone's mental wellbeing, clinicians could better understand their patients' conditions and design novel treatment interventions. For example, just in time behavioral interventions could be designed to intervene in specific ways, given an individual's state [109]. Another use for passive monitoring would be to better evaluate interventions, such as pharmaceutical therapies, over long periods of time.

In this dissertation, I consider how the accuracy of such a technology can be evaluated and whether or when prediction might be possible, i.e., for which individuals such an app might be relevant. Better evaluation and understanding of when a technology will be reliable is necessary to make it more reliable and thus safer. Prediction may be more difficult for some individuals, than for others, as behavior can differ greatly between individuals.

1.3 A Crisis Simulation for Training Suicide Prevention Text Line Counselors

The second problem I consider is developing a crisis simulation tool for training suicide prevention text hotline counselors so that they can practice intervening before actually counseling a human in need.

In acute times of need, individuals must be able to reach out to someone, but with insufficient access to quality healthcare – especially for mental health – there are often feelings of nowhere to turn and no one to listen [36, 158]. To provide critical support to individuals, helplines, such as the National Suicide Prevention Lifeline[1], provide opportunities for individuals in crisis to reach out, be heard, and get help. Several studies highlight the op-

[1] https://suicidepreventionlifeline.org

portunity for helplines to support the high volume of individuals in need and to expand services over text [61, 127]. While such a larger number of people using helplines indicates that helplines are providing necessary care, it also presents a challenge to fulfill the need with trained human counselors.

Helplines rely on counselors who are trained in empathy, active listening, assessing risk of suicide, de-escalation, and connecting individuals to longer term solutions. Properly training counselors is critical yet difficult, as, time and human resource costs aside, crisis counselors need to practice and develop expertise in realistic environments. However, counselors in training need environments that are low risk, i.e., they do not put distressed individuals in danger. Novice counselors are unable to assume full responsibility for a crisis situation until they have some experience.

Some organizations are considering the use of automated interactive agents (chatbots) to replace human counselors. However, unable to mirror the empathy of a human and still making conversation fumbles, bots like Woebot[2], Wysa[3], and X2AI[4] have been utilized on a limited basis to administer therapy for mild conditions or narrow tasks like walking a human through Cognitive Behavioral Therapy exercises [55]. While therapy bots hold potential for some situations, individuals in crisis are perhaps too vulnerable for helplines to rely on artificial agents now.

Individual role-playing, such as between inexperienced and experienced counselors, could help mitigate the need for practice in a safe or fail proof yet realistic environment and have been shown to improve crisis intervention skills [40]. However, according to a helpline counselor we consulted, a role-play can take an hour of both counselors' time and a novice needs multiple role-plays. Thus, for crisis lines that offer national service, such intense supervision for volunteers who might not stay long is difficult and an automated system could help lower this burden on experienced counselors.

In an effort to provide an environment for novice counselors to learn how to generate and practice generating appropriate responses, I propose building an interactive automated tutoring system that novice counselors can use to learn and, more importantly, practice de-escalation strategies. Similar systems are being considered for training therapists, but not for helpline counselors, who have a different protocol to follow [148]. I study the first component of the system, which seeks to provide a chat interface for counselors in training to practice conversing with simulated individuals in crisis. The system can be expanded by adding a second component of automated feedback that helps trainees improve.

By developing such a system, crisis helplines could expand their services to help more people in need, and new counselors could feel more confident starting out.

[2] https://woebot.io/

[3] https://www.wysa.io/

[4] https://www.x2.ai/

1.4 Summary of Contributions

In this thesis I make contributions towards assessing methods for developing data-driven tools that could empower mental health professionals to provide better care during times of need. I do this by studying both the use of mobile sensing for wellbeing tracking and exploring a chat system for training crisis counselors. I contribute a framework for assessing longitudinal predictions, use this framework in a case study, question when monitoring might be possible, and consider detecting changes over time. I also look towards the potential of building a crisisbot for training crisis counselors and assess a baseline approach that can be used to develop and evaluate comparing future systems.

Framework for evaluating predictions on longitudinal data

To evaluate algorithms for longitudinal predictions, i.e., tracking health over time, I show in Chapter 2 how traditional methods of evaluation can lead to false positive conclusions on freely available and widely used datasets. I also show how such methods are widely used in related literature. Then I propose a framework that can be used to better evaluate algorithms for health tracking applications. This new framework improves the evaluation of mental health prediction algorithms, but can be broadly used to better assess the accuracy of algorithms on longitudinal data.

Exploring the role of activity and sleep measures to monitor mood from mobile phone sensors

In Chapter 3, I use the framework I propose to explore the extent to which smartphone-sensed measures of physical activity and sleep can be used to monitor mental wellbeing. On an undergraduate student cohort, I find that physical activity and surrogates for sleep can be measured to a fine enough granularity with a smartphone to reveal correlations found in previous studies, i.e., increased physical activity and sleep duration were positively correlated with reported mood. However, modest gains in predictive capability above strong baselines indicate that smartphone-measured activity and sleep are insufficient for accurate monitoring and must either be considered in conjunction with other measures or collected with other sensors which may better measure physical activity, sleep quality, and sleep duration.

Considering when monitoring with smartphone data may be possible

There can be considerable behavioral variation between individuals. It is possible that this variation may result in smartphone mood tracking being possible with some individuals, e.g., those who use their phone frequently, and not for others, e.g., those who often leave their phone turned off. In Chapter 5, I look at whether there are patterns that are indicative of whether an individual's phone will be more or less successful at predicting their wellbeing. I find that there is considerable variation in the ability to predict individuals' wellbeing, but that location variability is related to whether a phone can predict an individual's wellbeing above a personal baseline. By evaluating when smartphone monitoring may be accurate for an individual, it may be possible to provide increased reliability.

Detecting changes in wellbeing over periods of time

As I find that challenges remain with using smartphones to monitor wellbeing on a daily basis, I also question whether smartphone-sensed measures could instead be used to predict significant changes in wellbeing over time. In Chapter 4, I consider whether smartphone-sensed features, alone or in conjunction with other features, are related to detecting changes in depression measures over the course of multiple weeks. I consider a cohort of undergraduate students and find that this approach also has limitations. For example, I find that Big 5 personality features [95] are more related to detecting changes than other smartphone-sensed features.

Exploring retrieval methods as a baseline approach for evaluating a counselor training system

In addition to identifying when an individual may be experiencing a time of distress, technology may be useful for training humans, such as suicide prevention counselors, to intervene and aid individuals during their times of need. Towards developing a training system for counselors to practice counseling before interacting with distressed individuals, I explore in Chapter 6 the use of text retrieval as a baseline method for responding to counselor input. Statistical language generation methods have seen remarkable improvements recently. However, so have embedding methods for representing natural language. To provide better comparisons for assessing whether responses generated for a crisis simulation are better than methods retrieved from a limited corpus, I explore and re-evaluate retrieving responses based on recent embedding methods. I find that novel embedding methods have improved the relevancy or coherency of responses retrieved. While the coherency is competitive with the coherency of generative models in this setting, I find that the retrieved responses are much longer and thus presumably more engaging. By re-evaluating retrieval methods, I establish stronger baseline methods to compare with generative methods, especially when considering engagingness of conversations.

Chapter 2

Meaningless Comparisons Lead to False Optimism in Medical Machine Learning

Originally published as Demasi, et al. "Meaningless comparisons lead to false optimism in medical machine learning." PloS one (2017).

A new trend in medicine is the use of algorithms to analyze big datasets, e.g. using everything your phone measures about you for diagnostics or monitoring. However, these algorithms are commonly compared against weak baselines, which may contribute to excessive optimism. To assess how well an algorithm works, scientists typically ask how well its output correlates with medically assigned scores. Here we perform a meta-analysis to quantify how the literature evaluates their algorithms for monitoring mental wellbeing. We find that the bulk of the literature (~77%) uses meaningless comparisons that ignore patient baseline state. For example, having an algorithm that uses phone data to diagnose mood disorders would be useful. However, it is possible to explain over 80% of the variance of some mood measures in the population by simply guessing that each patient has their own average mood - the patient-specific baseline. Thus, an algorithm that just predicts that our mood is like it usually is can explain the majority of variance, but is, obviously, entirely useless. Comparing to the wrong (population) baseline has a massive effect on the perceived quality of algorithms and produces baseless optimism in the field. To solve this problem we propose "user lift" that reduces these systematic errors in the evaluation of personalized medical monitoring.

2.1 Introduction

Health care should be tailored to individuals to maximize their wellbeing and health [34]. There is considerable hope that data collected from emerging data sources, such as smartphones and smartwatches, can be used to extract medical information and thus improve the tailoring of monitoring, diagnostics, and treatments for personalizing health care [41]. In particular, mental health care could particularly benefit from automated monitoring, as many mental health conditions need long-term monitoring and clinical monitoring is expensive, but automatically tracking a user with ubiquitous sensors is cheap [27, 94, 104].

Machine learning algorithms are commonly being used in an attempt to extract medical information from easy to collect data sources [17, 28, 108, 132]. These algorithms are attractive as, by automating information extraction, they promise to provide rich analyses cheaply and objectively based on collected data. Machine learning works by taking data that are easy to collect, building a model, and then using the model to make predictions for data that are harder to collect [58]. As an example, social media posts may be used to predict individuals' depressive symptoms or future suicidal ideation [44, 45]. However, without sufficient evaluation, the outputs of algorithms may be meaningless and mislead clinicians.

Whenever algorithms are used to make predictions, they must carefully be evaluated to ensure that their predictions meaningfully represent medically relevant information. Evaluation must be specified for each problem [165]. For example, if an algorithm is being used to predict one of two things, such as whether a patient is depressed, then it could be evaluated by the percent of predictions that are correct [132]. Alternatively, it could be evaluated by the percent of times that it correctly identified depression, i.e. sensitivity or true positive rate, and ascribe less importance to false positives [22]. In contrast, if an algorithm is trying to predict a value, such as someone's level of depressive symptoms, one could consider the degree to which predictions differ, i.e., the mean squared error. There are myriad additional methods for evaluating algorithms because, without sufficiently evaluating algorithms, it is easy to generate misplaced optimism about the utility of algorithms [54, 65, 120, 162].

Regardless of how the correctness of an algorithm is quantified, algorithms must be compared to a baseline approach that simply makes guesses to prove that the algorithm makes better predictions than guessing. For example, if an algorithm is trying to predict a rare event, such as a mental breakdown or suicide, an approach that simply guesses that the event never happens will usually be correct and thus will have high accuracy [76, 103, 152]. However, such an approach is entirely useless for medicine. If algorithms are not compared with reasonable guesses, the accuracy of the algorithm's predictions can appear to be good, when in reality the algorithm is doing no better than guessing and is thus medically useless.

Here we review, for modeling of longitudinal individual state, what baselines algorithms are commonly compared against and how much of the apparent success of algorithms can be ascribed to poor comparisons. We focus on the example of mental wellbeing and demonstrate in two popular datasets that individuals exhibit little variance over time. Typical wellbeing prediction algorithms seem to work well, but we find that this is simply because they are basically always guessing individuals' personal average states. This example highlights how falsely optimistic results can easily be obtained by comparing machine learning with population as opposed to personal baselines. We perform a systematic literature review and find that most studies ($\sim$77%) compare with the population baseline. By not comparing with personal baselines, studies are prone to making falsely optimistic conclusions that can unintentionally mislead researchers' perspectives and delay progress on important medical applications. We argue for a new measure, "user lift," that measures the benefit of an algorithm relative to the single-person model.

2.2 Methods

Algorithm Evaluation

There are many ways to evaluate how good an algorithm's predictions are [65, 58, 162, 165]. The general approach is a two step process of measuring an algorithm's error, or how inaccurate its predictions are, and then comparing the algorithm's error with the error of simply guessing answers. These guesses form a baseline approach and could be specific to each patient, or they could use other trivial factors, e.g. the time of the day. Regardless, because it is totally useless for medicine to simply guess answers based on subject and other trivial factors, algorithms must have lower error than such baselines to be of any use.

Algorithm Error

To evaluate how well an algorithm predicts a binary outcome, e.g., whether an individual is having a happy vs. sad or stressed vs. relaxed day, we consider the classical measure of prediction error. Prediction error is the percent of observations that were incorrectly predicted (percent incorrect). To evaluate how well an algorithm predicts an individual's level of happiness or stress, we consider the root mean squared error (RMSE), which considers how different the predicted levels are from the true reported levels [58]. With both prediction error and RMSE, lower values indicate that an algorithm is doing better at predicting an individual's state. Higher values indicate more significant prediction error.

Baselines

We consider two baseline methods that simply guess how an individual is doing: personal baselines and population baselines. Both baselines are simple approaches that always guess individuals are at the same state. The personal baseline always guesses that each individual is at a constant state, but that state can differ between individuals. The population baseline predicts that all individuals are always at the same state.

When an algorithm is attempting to predict whether an individual is having a stressed (or happy) day or not, we consider personal baseline error, which is the prediction error of always guessing that each individual is always at their most frequently reported state (mode). We also consider the population baseline error, which is the error of estimating that all individuals are always at the most frequently reported state of the population (mode).

When an algorithm is attempting to predict an individual's level of happiness or stress, we consider the personal baseline RMSE, which is how far predicted levels were from always guessing each individual to be at their average level of stress or happiness. RMSE indicates a model with higher error and thus worse predictions. We also consider the population baseline RMSE, which is the RMSE of estimating that all individuals are always at the average state of the population.

User lift

We propose the measure of user lift as a way to evaluate whether an algorithm is making better predictions than simply guessing an individual's state. The user lift is the improvement of an algorithm's predictions over the personal baseline, or the amount that error is decreased by adding better features and a model. User lift is the difference between personal baseline error and model error in RMSE or in prediction error (personal baseline error - model error). The user lift can be thought of as the increase in accuracy of an algorithm over the null accuracy of guessing an individual to be at their average state. The average user lift is the mean user lift across the individuals in the dataset.

User lift framework

As a stricter measure of whether algorithms have any utility, we suggest the user lift framework instead of comparison with a single weak baseline, such as the population baseline. With this framework, researchers calculate user lift for each study participant. The user lift quantifies whether an algorithm is better than the simple personal baseline on each user. We propose then reporting descriptive statistics on the distribution of user lift and utilizing statistical tests to determine whether the average user lift is greater than zero. Nonparametric permutation tests are appropriate and powerful tests for considering whether a single sample, such as of user lifts on study participants, has a mean greater than zero. A permutation test is appropriate here so that no assumptions on distributions are needed. While other nonparametric tests, such as the paired Wilcoxon signed-rank test, may be appropriate for comparing two samples, permutation tests have been reported to be more reliable than paired non-parametric tests [75, 141].

Machine Learning Example: Predicting Subjective State from Location and Mobility

We present an example of how falsely optimistic conclusions can be reached about algorithms' performance. For this example, we follow previous works that have used and suggested that smartphone GPS location data can predict individuals' mental wellbeing [28, 71, 132]. We follow these studies' methodologies only for constructing features, or describing individuals' daily behavior from their GPS location logs. The dataset, testing methodology, and prediction tasks that we explore differ from these prior works.

Datasets

We consider two well established datasets that are freely available. Both datasets collected individuals' smartphone data, specifically GPS location, and their stress and happiness levels. The StudentLife dataset [160] followed a cohort of students at an American university during the course of a semester. Data that was collected included daily measures of stress on a five point Likert scale. Of the initial 48 students with data accessible, we consider data for the 15 students who had sufficient data available: stress level and at least 35 GPS location observations for at least 30 days of the study period.

The second dataset we consider, the MIT Friends and Family dataset [2], resulted from a project that collected various types of data on a cohort of university affiliates and their families at another American university. The data collected included daily wellbeing measures. Here we consider the nine point Likert scale of happiness and seven point Likert scale of stress that were collected. Of the 116 participants included in the available dataset we consider data for the 31 individuals who had measurements of stress or happiness, respectively, for at least 30 study days and at least 35 GPS location measurements on those days.

Data Processing: Location and Mobility Features

To derive meaningful features of location and mobility, we follow three previous studies [28, 71, 132]. All features from these studies that were reproducible (due to the data available) were included. Before constructing features, we used two preprocessing methods.

The first preprocessing method fit a Gaussian Mixture Model (GMM) to all of the location samples for each participant collected to identify locations frequented by participants [71]. The number of clusters was chosen to be the number, up to twenty maximum, that minimized the Bayesian Information Criterion [138]. It was assumed that participants would frequent at most twenty locations during the course of the study. The home location of a participant was determined to be the location where the participant spent the majority of their time during the evening hours (11pm - 6am) and the work location was similarly determined to be where the participant spent the majority of their time during working hours (11am - 4pm). In contrast to prior work, we did not interpolate the location observations to a regular time sampling, as we did not find this beneficial to prediction accuracy [71]. We consider this first set of clusters to be the full clustering.

The second preprocessing method used K-means clustering on stationary points only [6, 132]. The StudentLife dataset included a prediction of whether the participant was moving or stationary at each observation, but the Friend and Family dataset did not. To determine whether participants in the Friends and Family dataset were stationary at each observation, we approximated movement speed with the time derivative at each observation and used a threshold. We attempted to set the threshold to be about 1km/h [132]. We consider this second set of clusters as the stationary clustering and the night cluster to be the cluster where each individual spent the most time between midnight and 6am.

To protect participants' anonymity, the GPS location data in the Friends and Family dataset was subjected to an affine transform before being released. Because this transform purposefully changes the space, but collinearity should be preserved, we approximated features in one dimension on the Friends and Family dataset.

Utilizing the two set of location clusters that resulted from the GMMs and Kmeans, the features of mobility and location that we derived for each participant each day of the study are as follows:

1. The fraction of a day that a participant spent not stationary.
2. The average displacement of a participant between two observations during the day, i.e., average speed.
3. The standard deviation of displacements between points.
4. The location variance (on log scale), i.e., the sum of the variance of location coordinates in each dimension.
5. The "circadian movement" of a participant [132], which we adapted to our daily monitoring setting as the Euclidean distance of the vector of fraction of time a participant spent in each of their stationary location clusters with the participant's mean location distribution. The mean location distribution of a participant was calculated as the average fraction of a day that a participant would spend in each stationary location cluster during the study.
6. The location entropy, which was calculated as the entropy of the vector where each entry represented the fraction of the day that a participant spent in each stationary location cluster.
7. The radius of minimum circle enclosing the participant's location samples.
8. The fraction of time a participant spent at their GMM home cluster.
9. The fraction of time a participant spent at their GMM work cluster.
10. The fraction of time a participant spent at their stationary night cluster.
11. The log likelihood of a day from the GMM to estimate how routine the day was.
12. The AIC and BIC of the GMM evaluated with the day's coordinates, to also determine how typical the day was.
13. The number of GMM clusters visited in a day.
14. The number of stationary clusters visited in a day.

Experimental Framework

We present two prediction tasks:

1. Predicting whether a participant was happy or stressed or not on a given day.
2. Predicting the average level of happiness or stress that a participant reported on a given day.

To construct levels of stress or happiness on a given day, we average all the Likert scale responses that a participant reported on that day. Whether the participant was happy (or stressed) or not is defined by a threshold on the daily average on a value to distinguish when students reported any stress versus no stress. For the StudentLife user inputs, we use “A little stressed” as the threshold. For the Friends and Family dataset, we use the middle value of the Likert scale as the threshold, as the Friends and Family scales were defined from negative to positive values, where the middle value was supposed to indicate a neutral state.

For both problems, we attempt to predict the stress or happiness from the location and mobility data with a variety of standard machine learning methods. For regression we consider: linear regression with an Elastic Net penalty, and Lasso regression [151, 168]. For the binary classification task, we consider: logistic regression with L2 penalty, support vector machines with radial basis function kernels, and random forests [23, 58]. Hyperparameters were chosen with 10-fold cross-validation on the training data. The methods that return the lowest error are presented.

We consider both population models, which could also be referred to as global, general, or all-user models and utilize all the individuals' data to make predictions, and personal models, which use only a single individual's data to make predictions for that individual. Prediction error is measured with leave-one-out cross-validation, which is commonly used for estimating an algorithm's prediction error [58]. To perform leave-one-out cross-validation on population models, we combine data from all of the participants into a single set. Then one observation is withheld, a model is trained on all of the other observations, and then that model is used to make a prediction for the held out observation. The process is repeated until every observation has been withheld exactly once. The model error reported from this process is the average error across the predictions for each data point. Population models assume that some of each participant's data is seen during training, in addition to data from other participants [133]. For personal models, we similarly hold out one observation, but we only train on the remaining observations of that individual's data and then repeat only for the number of observations that we have on that individual. Personal models only attempt to extrapolate predictions for an individual from their own data. Alternative cross validation schemes, such as N-fold, offered no benefit to the results, so are omitted for brevity.

Literature Review

In addition to an example on two real datasets of how false machine learning results can be arrived at by comparing to weak baseline models, we perform a systematic literature review to investigate how algorithms are commonly evaluated and whether baselines are sufficiently reported. Our literature review took three steps:

1. Find relevant literature.
2. Establish whether a baseline (personal or population) was compared with.
3. Identify the error of the baselines and the best reported machine learning algorithms.

Finding Relevant Literature

While baselines are needed to evaluate all machine learning algorithms on personal data, we make our literature review tractable by focusing on studies similar to the machine learning example we present. We utilize GoogleScholar to find publications that attempted to automatically infer an individual's subjective states, similar to the example we presented. The studies we include meet the following criteria:

- Relate to subjective personal data, as denoted by having one of the following words in the title: depression, depressive, stress, mood, mental, happiness, or wellness.
- Attempt a machine learning prediction task and report prediction accuracy by having the following word "accuracy" somewhere in the text of the publication.
- Attempt prediction on participants' longitudinal data, where personal baselines are defined, by containing the words "participant" or "user".
- Collect data from sensors by requiring one of the following words to be included somewhere in the text: smartphones, sensor, sensors, or sensing.
- Were published since 2010.

Because of a particularly strong focus on stress in previous work, we break the query into two queries: one that requires the word "stress" to be in the title and another search that requires any of the other wellbeing words to be in the title. We perform this joint search with the following GoogleScholar queries:

(participant OR user) accuracy (sensor OR sensors OR smartphones OR sensing)intitle:stress (from 2010)

(participant OR user) accuracy (sensor OR sensors OR smartphones OR sensing) (intitle:mental OR intitle:depression OR intitle:depressive OR intitle:mood OR intitle:happiness) (from 2010)

To be considered relevant, studies need to attempt to predict user input data of users' subjective state from other collected data, i.e. sensors. Examples of studies that are returned by our query, but are excluded from our analysis are:

- Correlational analyses that reveal certain data or behaviors are correlated with subjective state.
- Studies of one-time user surveys (in contrast to repeated prompts) or where the goal is to separate subjects, i.e., each subject was a data point.
- Literature summaries or reviews.
- Randomized control studies of intervention efficacy.
- Other evaluations of treatments on subjective state.
- Collection and presentation of a dataset collected without a prediction task.
- Measurements of behaviors without attempting prediction.
- Descriptions of tools and systems implemented with user reviews of the systems.
- Non-peer reviewed publications, such as reports and book chapters.
- Prediction of non-subjective states, e.g., prediction of labels coded by researchers who intuit what state the user was in from observational data, or labels of stimulus exposure when studies attempted to induce a given emotional state such as stress.

We only consider studies where labels are for multiple observations of a participant's subjective input state.

Establishing Comparison with a Baseline

Some studies do not report any baseline model for comparison, so we begin by noting which studies reported a baseline model. For studies that provided sufficient detail, we did the following:

- When baseline models are reported we recorded the baseline performance metrics directly from the text and the type of baseline used, e.g., a population baseline or a user baseline.
- When baseline models are not provided, but confusion matrices are provided we manually calculate the baseline performance.
- When individuals baselines are reported, we take the average user baseline performance.
- When only mean squared error are reported, we note whether the mean squared error is also provided for a constant baseline.

Comparing the accuracy of different models

There are a wide variety of performance metrics authors report when evaluating their models. We extract model prediction error for multi-class classification problems according the to following criteria:

- When results are broken down for personal models by individual, the average is used.
- When accuracy results are given for multiple objectives, e.g., different dimensions of mood, the best results for each objective is recorded.
- When multiple feature sets and models were tried, only the best performing model is considered. Models that utilize user input as features were excluded when possible.
- The number of folds in the cross validation scheme used is not incorporated into our analysis. We considered 10-fold, leave-one-out, and leave-user-out cross validation schemes to all be "population" models. Both 10-fold and leave-one-out cross validations on personal data only are considered to be "personal" models.
- The uniform baseline was calculated by noting the number of classes that the study reported using in their measurement scale.

2.3 Results

We want to consider to what extent the choice of baselines matter in medical machine learning and how baselines are used in practice. To quantify the importance of baselines, we use two publicly available datasets and compare the performance of machine learning algorithms to two different baselines: a population baseline and a personal baseline. More specifically, we use the StudentLife [160] and Friends and Family [2] datasets and analyze machine learning predictions of stress and happiness, which we compare to both personal and population baselines. To understand how the field generally uses baselines, we perform a systematic literature review. These two complementary analyses will allow us to meaningfully inform the debate about machine learning in medicine.

Initially, we find, that individual subjects have little variance over time, relative to the variance across the population, i.e. low personal baseline error relative to higher population baseline error (Figures 2.1 and 2.2). Thus, comparing learned models with population baselines can obscure whether a model is better (lower error) on individuals than constant personal baseline models. We find the same pattern when we ask about RMSE and binary predictions. This gives us an intuition that guessing each subject's mean value should produce relatively low errors.

Motivated by the intuition that there is little within-subject and more across-subject variance, we now ask how machine learning algorithms compare to the two baselines. In line with prior literature [28, 71], the algorithms predict whether an individual was having a particularly stressful or happy day from their GPS location and mobility data (Figures 2.1 and 2.2). Our binary results are comparable (max difference=6%) to past studies predicting binary stress or emotion from similar datasets [19, 20, 29, 71, 70, 90, 102], as are the errors of the personal models [28, 49, 67, 90, 135, 154, 155, 166]. Similarly, the difference between RMSE of personal models and personal baselines is comparable to the differences reported in prior publications [7, 87]. Our algorithms are much better than the population baseline and population models. They are not, however, lower than the personal baselines. This shows how good performance relative to the population baseline can be entirely meaningless.

To prevent comparing with the wrong baseline and to control against obscuring the range of how well algorithms do on individuals with aggregate statistics, we propose using statistical tests with the metric of user lift to prove that an algorithm is doing significantly better than the personal baseline. User lift is the difference of the personal model with the personal baseline, as described above. Positive user lift indicates that a model is better than the personal baseline, that the algorithm's predictions are more accurate than always assuming an individual is at their average state. Indeed, user lift shows that our naive model is useless while our moderately careful model at least adds something (Table 2.1). Using the wrong baseline, may make bad machine learning with a performance that is by any meaningful definition useless seem impressive underscoring the importance of meaningful baselines.

To understand how algorithms are typically evaluated, we perform a systematic literature review of related studies that attempted to predict emotion and stress from sensed data, such as from smartphones or smartwatches (Figure 2.3). Just like in our example datasets, participants report surprisingly little variation (Figure 2.4). As a result, guessing that an individual was at the same state incurred low personal baseline error and machine learning

Table 2.1: Statistical significance for user lift of personal models in Figures 2.1 and 2.2. The user lifts are the differences of personal baselines with personal models, in terms of prediction error or RMSE. The p-values are for permutation tests considering whether the user lifts were larger than zero. In every case the user lifts are not significantly greater than zero - the models are not doing better than constant personal baselines.

Dataset	Problem	Model	Avg. Personal Baseline Error	Avg. Personal Model Error	Avg. User Lift (Error)	p-value
SL - Stress	binary	Log.Reg.	29.19%	29.09%	**0.10**	.481
FaF - Happiness	binary	SVM(rbf)	16.51%	18.67%	**-2.17**	.967
FaF - Stress	binary	SVM(rbf)	25.17%	23.35%	**1.82**	.240
SL - Stress	regression	Elastic Net	0.75	0.78	**-0.03**	.988
FaF - Happiness	regression	Elastic Net	0.81	0.83	**-0.02**	.999
FaF - Stress	regression	Elastic Net	1.10	1.13	**-0.03**	1.000

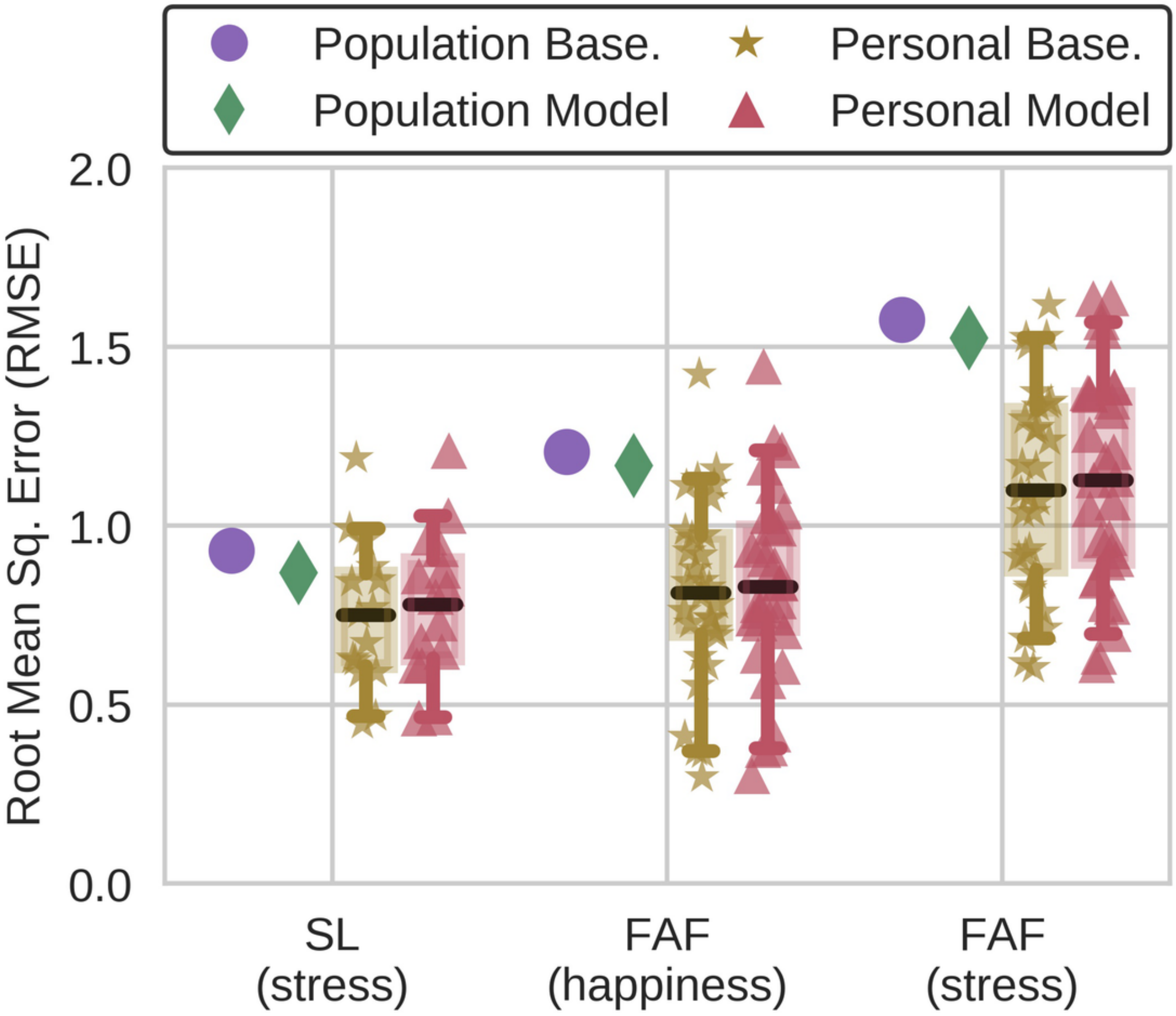

Figure 2.1: Results of machine learning regression models on StudentLife (SL) and Friends and Family (FAF) datasets. Bars represent the 5th and 95th percentiles, black lines indicate means, and boxes indicate the 1st and 3rd quartiles of error incurred on individuals. Personal models yield lower error than population models and population baselines, which often leads researchers to the conclusion that personal models are successful. Comparing personal models with personal baselines reveals that their error is no lower, so algorithms are doing no better than predicting individuals to be their most frequently reported state. The models presented are those with lowest error.

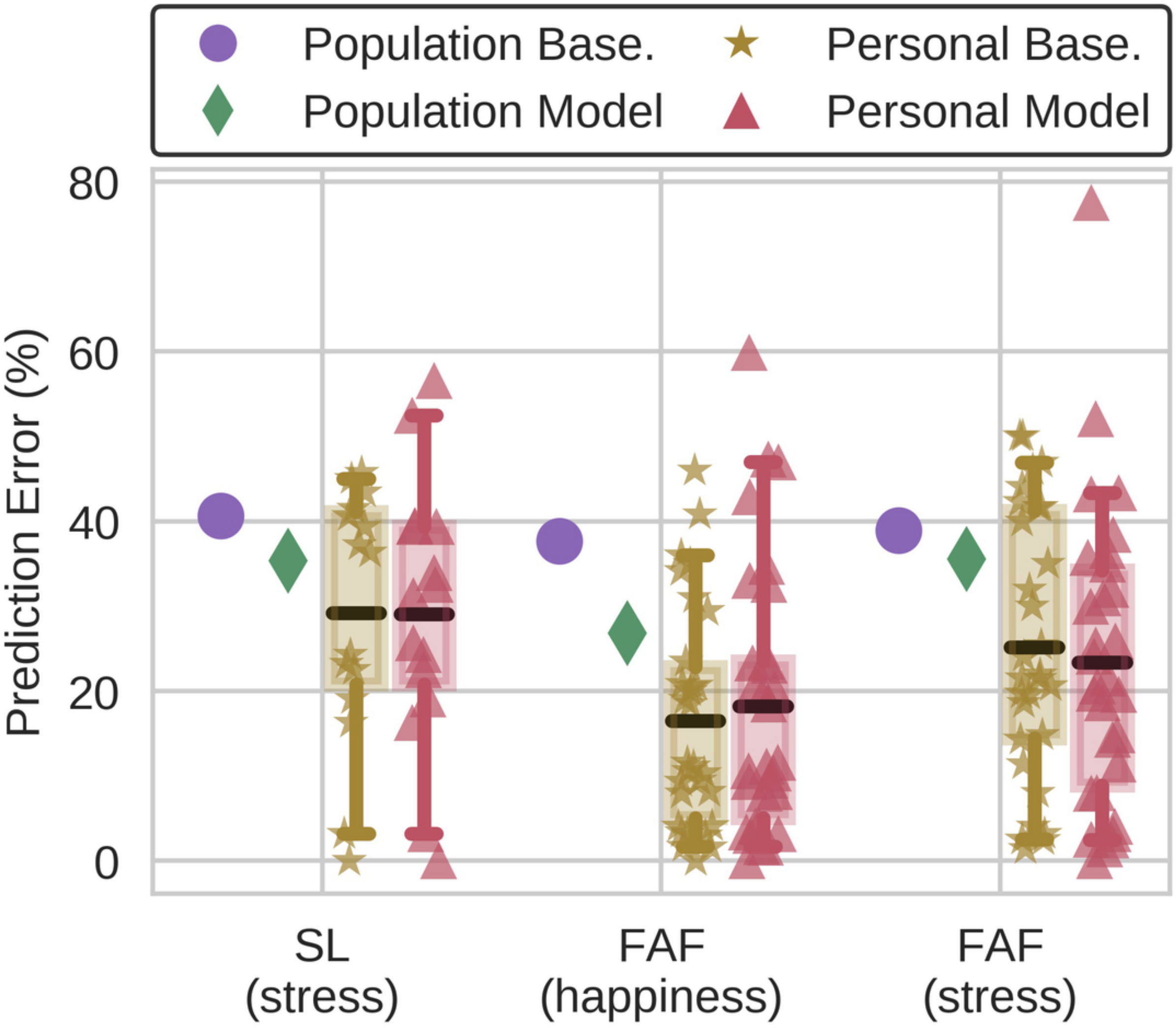

Figure 2.2: Results of machine learning classification models on StudentLife (SL) and Friends and Family (FAF) datasets. Bars represent the 5th and 95th percentiles, black lines indicate means, and boxes indicate the 1st and 3rd quartiles of error incurred on individuals. As with the regression example in Figure 2.1, personal models yield lower error than population models and population baselines. The models presented are those with lowest error.

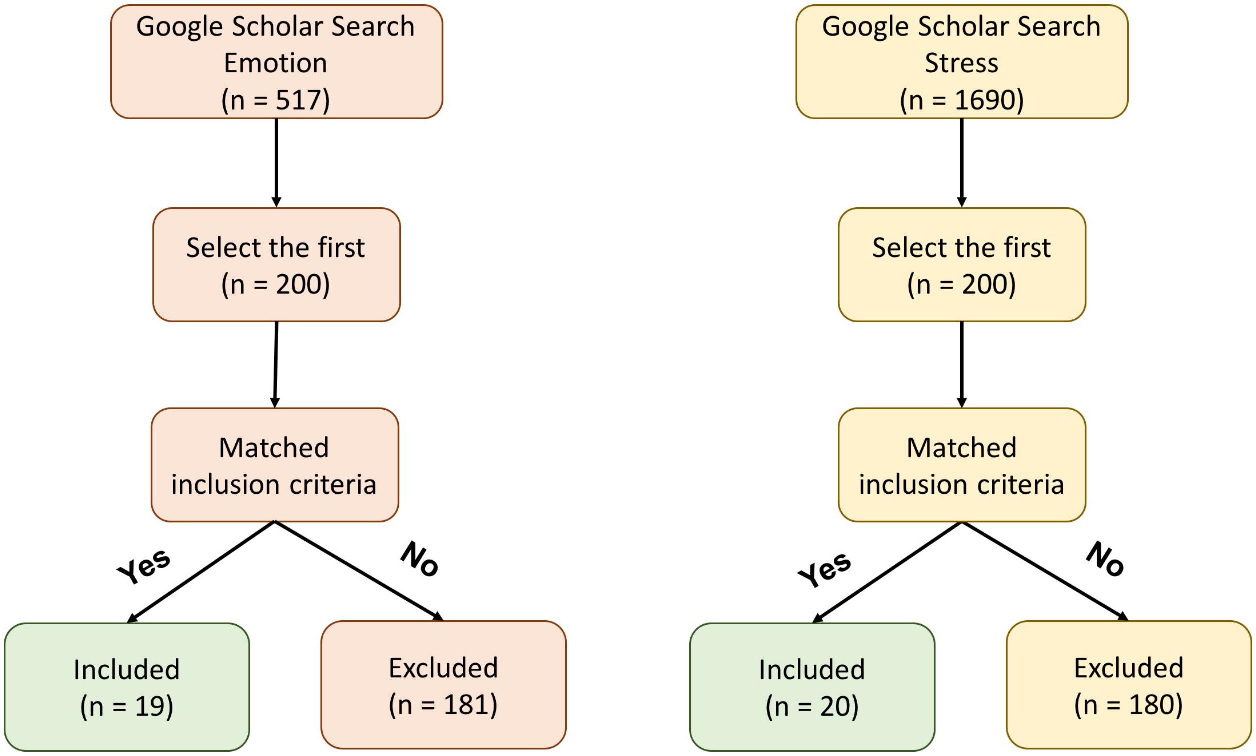

Figure 2.3: Diagram of literature review process.

algorithms typically had only slightly lower error than the personal baseline (Figure 2.5).

Studies that do report personal averages sometimes have negative user lift (Figure 2.6).When personal baselines are reported, they are usually reported in aggregate, which can be misleading by obscuring negative user lift on some individuals. Aggregation also precludes statistical tests on user lift and the only study that did report a statistical rank test on improvement across individuals found that there algorithms were no better than a naive model (using an historical averages of individuals' states) [7]. However, the bulk of studies only report population baselines making it impossible to know if they have any user lift (Figure 2.7). As such, it seems that the bulk of papers have questionable results, at best.

2.4 Discussion

We have shown, with examples of stress and happiness on two popular datasets how easily machine learning algorithms can appear promising when compared with meaningless baselines. Individuals report surprisingly little variation in state, so always guessing that an individual is at their most frequently reported state is correct most of the time. As a result, when an algorithm is compared with a population baseline that always predicts all users are

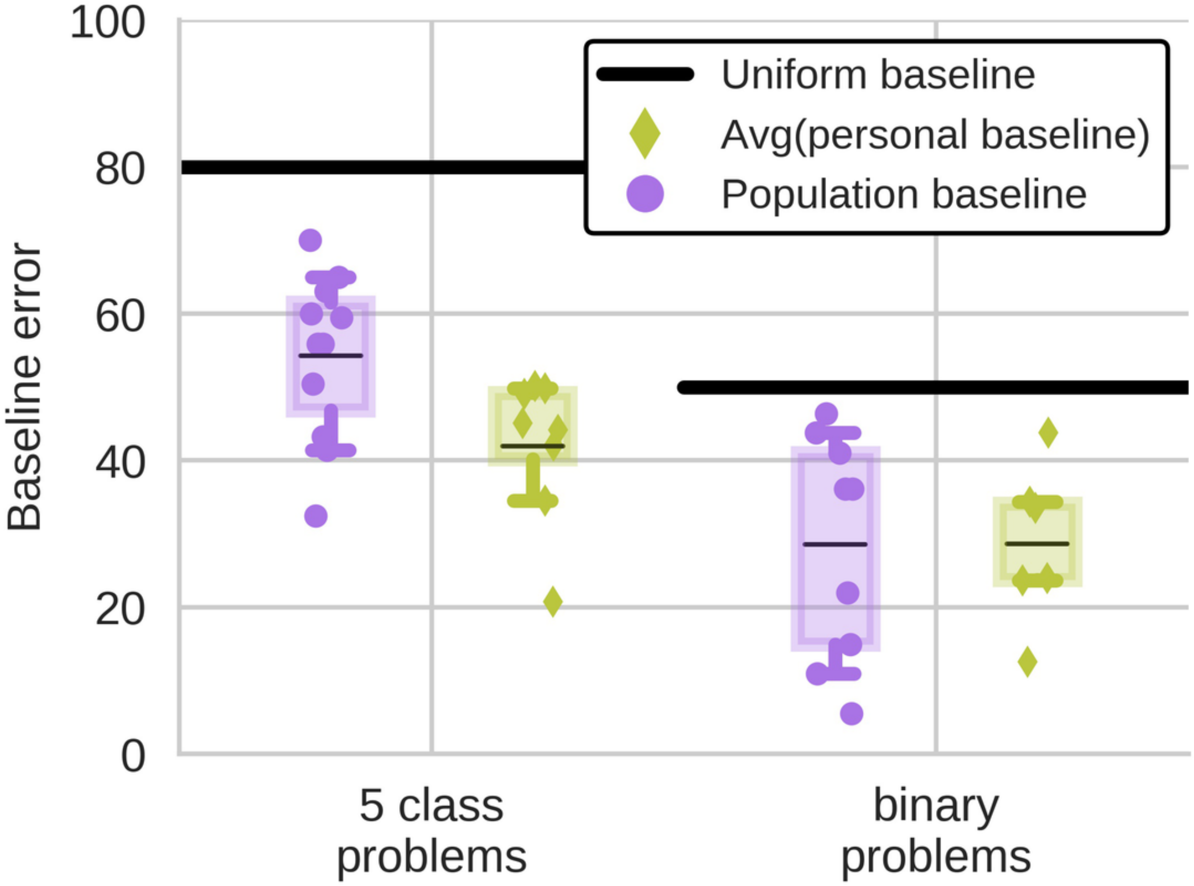

Figure 2.4: Participant variability reported in related studies. Reported results reveal little participant variability. Population and personal baselines reported by studies that had participants report their state on two point and five point scales. The black bars indicate the what the baseline would have been if participants were to report every state equally often, e.g., happy and sad each half the time. Boxes denote 1st and 3rd quantiles, bars indicate 5th and 95th percentiles, and lines the average of the markers.

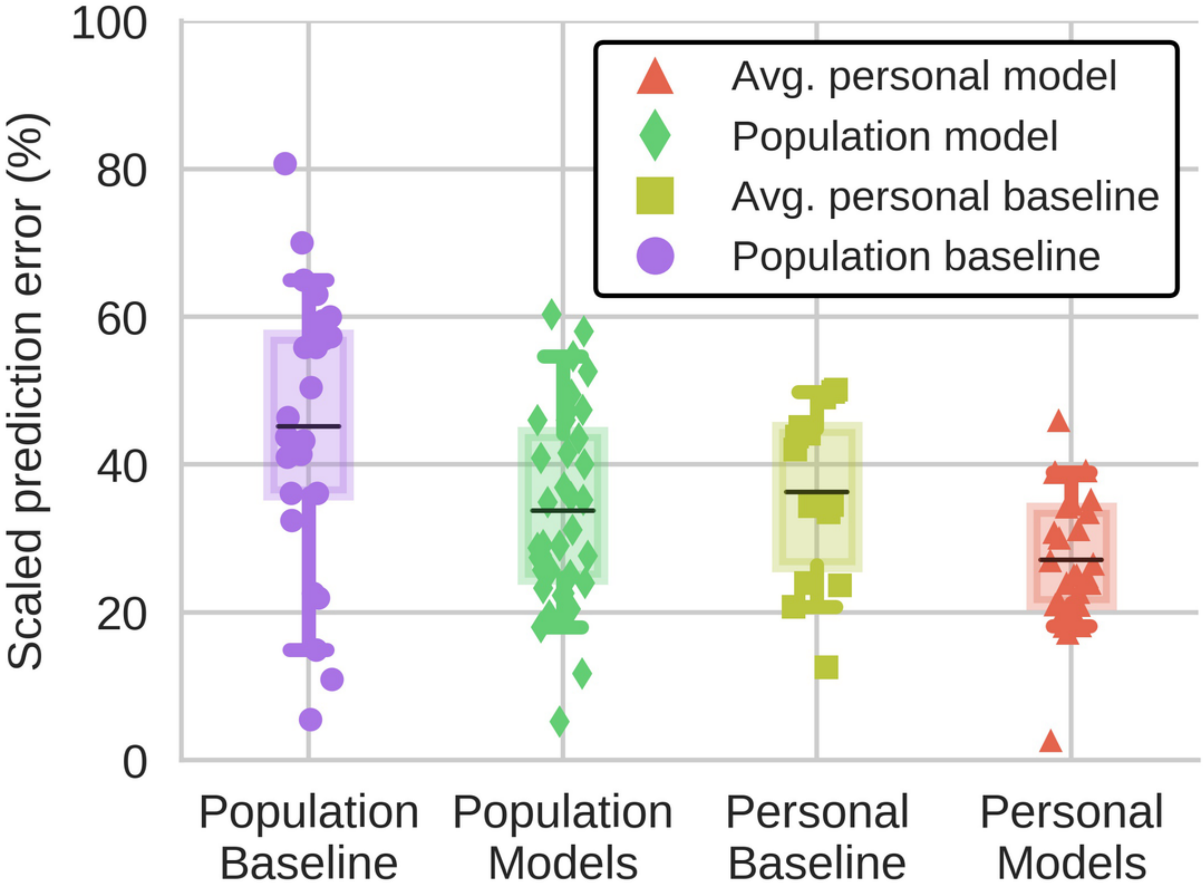

Figure 2.5: Model performance reported in related studies. Reported results reveal that models do not dramatically improve upon personal baselines. Population and personal baselines and model error reported in literature reviewed. Performance (prediction error) is scaled by the minimum class imbalance to compare studies that asked participants to report their states on scales with different numbers of points. Boxes denote 1st and 3rd quantiles, bars indicate 5th and 95th percentiles, and lines the average of the markers.

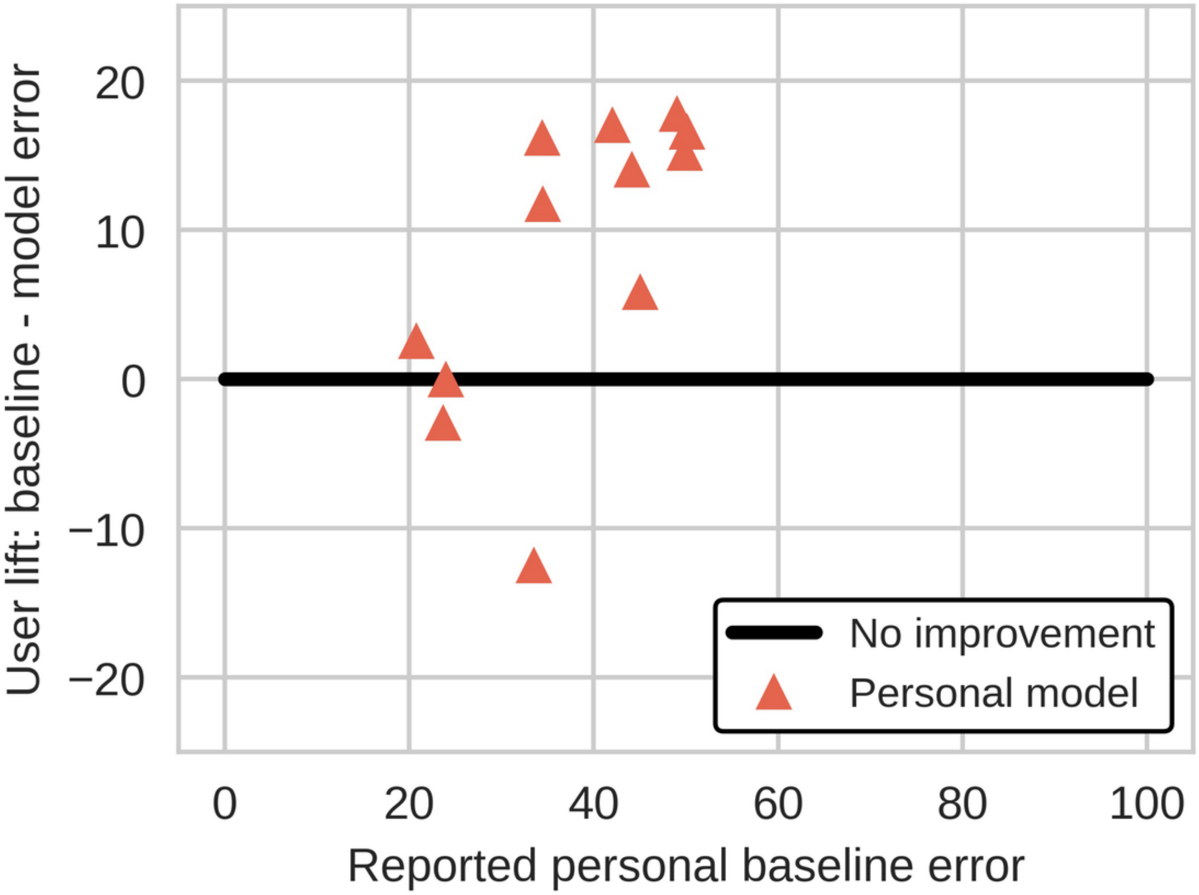

Figure 2.6: Calculated user lift and prevalence in studies reviewed. User lift calculated for studies where error of baselines and algorithms were both reported. Algorithms sometimes have no improvements over baseline guessing, and these figures are biased to studies that reported sufficient information.

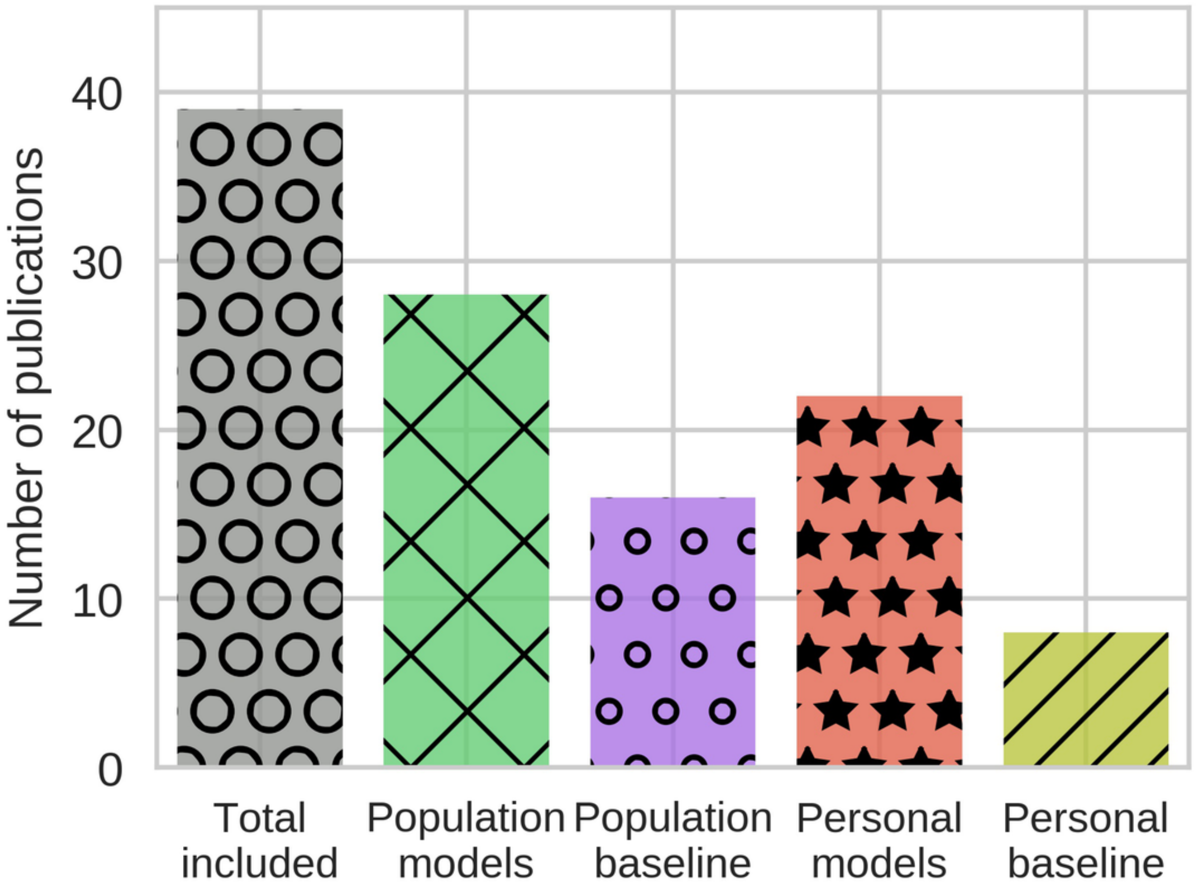

Figure 2.7: Prevalence of baselines reported in studies reviewed. Population baselines are reported in roughly half the publications reviewed while personal baselines are infrequently reported (approximately 23% of publications).

always at the same state, the algorithm's predictions can seem accurate even if they are no better than predicting each individual to be at their most frequently reported state. Despite the possibility for falsely optimistic results, we found in a systematic literature review that population baselines are commonly compared with in roughly 77% of publications reviewed. We also find that when personal baselines are reported that the algorithms often add little or nothing over these baselines (and in fact they sometimes do worse).

A limitation of the datasets that we explored, and most of the literature we reviewed, was that the study cohorts were not clinical populations, the sample size was small, and the study duration was limited. However, the study characteristics of the datasets presented are characteristic of many studies. While target populations for the monitoring we have discussed are typically individuals with mood disorders, study cohorts are frequently small in size and from the general population. It is possible that individuals with mood disorders would report more variability in state than the general public. More variability would reduce the likelihood of falsely optimistic results, but our proposed evaluation method would still be appropriate for showing that algorithms are an improvement over always predicting that individuals are at their average state. Finally, the user lift evaluation framework that we suggest would complement a larger dataset, despite being demonstrated on fewer subjects here.

While we reviewed a representative portion of relevant literature, we had to focus the scope and present a reproducible search that aligned with public datasets. We constructed general search queries to include pertinent studies, but inconsistencies in terminology between communities made it impossible to included all relevant studies and some known related works were not covered.

In addition to coverage, there were a variety of features that we could not control in the literature review. Studies recruited from disparate populations and had different study protocols. In addition to collecting different data and conducting different analyses, studies reported results in an variety of ways. We did our best to standardize across studies and present results favorably and comparably.

The proposed user lift evaluation framework is more generally applicable to predicting longitudinal patient state than we have shown here. We have focused our review on a narrow, important application of mental wellbeing, as this is a nascent and exciting application for machine learning algorithms. However, user lift would apply to any application predicting longitudinal data, such as monitoring blood sugar level, body weight, or daily sleep duration. The importance of statistical tests on user lift becomes greater for applications where individuals are expected to exhibit less variation and descriptive statistics must also be reported to quantify the size of any statistically significant user lift.

While we have calculated personal baselines here over the entire dataset, in principle this is not necessary. Because personal baselines are calculated with respect to individuals' most common state, personal baselines are easy to quickly approximate, with minimal sampling. Personal baselines could potentially vary over an extended period of study, but such scales are outside the scope of most studies and require further investigation.

An ability to predict meaningful personal signals for medical monitoring, such as mental wellbeing, could greatly improve personalized medicine by enabling novel approaches to just in time and personalized interventions. However, we have highlighted some pitfalls of evaluating algorithms for this application that can easily result in falsely optimistic results and unintentionally provide baseless optimism. To reduce falsely optimistic results, we have suggested an alternative evaluation framework using statistical tests on our proposed metric of user lift, which takes an individual-centric approach. As was shown, there is a range of model predictive capability across individuals, so we suggest statistically testing for significant improvement on the population. This framework of evaluation can help researchers to focus efforts and thus help advance progress on this application.

Chapter 3

Well-Being Tracking via Smartphone-Measured Activity and Sleep: Cohort Study

Originally published as DeMasi, et al. "Well-being tracking via smartphone-measured activity and sleep: cohort study." JMIR mHealth and uHealth (2017).

Automatically tracking mental wellbeing could facilitate personalization of treatments for mood disorders, such as depression and bipolar disorder. Smartphones present a novel and ubiquitous opportunity to track individuals' behavior and may be useful for inferring and automatically monitoring mental wellbeing. We assess the extent to which activity and sleep tracking with a smartphone can be used for monitoring individuals' mental wellbeing. Methods: A cohort of 106 individuals was recruited to install an app on their smartphone that would track their wellbeing with daily surveys and track their behavior with activity inferences from their phone's accelerometer data. Of the participants recruited, 53 had sufficient data to infer activity and sleep measures. For this subset of individuals, we related measures of activity and sleep to the individuals' wellbeing and used these measures to predict their wellbeing. We found that smartphone-measured approximations for daily physical activity were positively correlated with both mood ($P < .01$) and perceived energy level ($P < .001$). Sleep duration was positively correlated with mood ($P < .05$), but not energy. Our measure for sleep disturbance was not found to be significantly related to either mood or energy, which could imply too much noise in the measurement. Models predicting the wellbeing measures from the activity and sleep measures were found to be significantly better than naive baselines ($P < .01$), despite modest overall improvements. Measures of activity and sleep inferred from smartphone activity were strongly related to and somewhat predictive of participants' wellbeing. While the improvement over naive models was modest, it reaffirms the importance of considering physical activity and sleep for predicting mood and for making automatic mood monitoring a reality.

3.1 Introduction

A goal of personalized medicine is to tailor treatments to individuals. To aid the tailoring of treatments, it is necessary to monitor how an individual is doing and to evaluate whether they are responding to a treatment [30, 74]. However, monitoring can be a tedious, expensive process and, as a result, yields low adherence [79]. To overcome low patient adherence, automatic monitoring can help mental health disorders, such as depression and bipolar disorder, that benefit from monitoring symptoms over time to identify symptom relapse and possibly prevent symptoms due to higher self-awareness [73].

The proliferation of personal electronics has enabled continuous personal monitoring [119]. For example, activity recognition has enabled tracking to monitor physical exertion and sleep patterns [53]. Recent studies have started examining whether these smartphone-measured behavioral patterns can be used to infer and then automatically track signals that aren't explicitly measured by the smartphone, such as mental wellbeing.

Many studies have looked at inferring measures of mental wellbeing from smartphone-measured behavioral patterns [99]. In particular, researchers have considered using measures of location and mobility from GPS logs to infer depression [28, 71, 132], bipolar state [63], stress [17], and wellbeing measures related to schizophrenia [159]. These studies have shown that daily self-reported levels of stress are related to geospatial activity and sleep [17] and that mobility data can improve predictions of whether a participant is happier, or less depressed than usual [28, 71] and their bipolar state or transition between states [63]. Researchers have also found that regularity of an individual's daily mobility is significant when predicting depression symptom severity [132, 134].

Additional studies have explored the relationships of social signals, such as phone usage, call logs, and SMS logs, with wellbeing. Two recent studies found that phone usage measures were correlated with depressive symptom severity [132, 134]. Another study found that using social signals, such as emails, SMS and call logs, internet usage, app usage, and location frequency was predictive of mood and energy when previous observations of mood and energy were included [87]. However, a similar follow-up study was unable to reproduce these results. This follow-up study did not find sophisticated models considering high accelerometer activity, call and SMS logs, screen events, app usage, and number of images taken to be better than guessing each individual's wellbeing [7].

While this body of literature has established that relationships between measures of mental wellbeing and smartphone-measured behaviors may exist, the above literature has not focused extensively on physical activity in uncontrolled environments (i.e., outside of a lab without constraints on participants, such as where the phone must be located). For example, studies have explored predicting bipolar states and state transitions via accelerometers on small populations [63] or mood in constrained environments where the phone had to be in a fixed position [125, 66] or activities had to be performed in a lab [26]. One study has looked at a measure of total daily physical activity and sleep (as measured with multiple sensors), but within the context of stress and not wellbeing more broadly and did not attempt to predict wellbeing [17].

Despite these few studies' limited focus on activity and sleep, there is a body of literature external to mobile health (mHealth) that has established a strong relationship of better mood with increased activity [50, 57, 96, 113, 144] and sleep quality [117, 118]. There is also mounting evidence that a smartphone accelerometer measures physical activity to a sufficient extent to be useful for monitoring wellbeing. Several studies have demonstrated that individuals' sleep and physical activity can be somewhat accurately tracked with smartphones [32] and activity recognition [69, 83, 91], respectively. As a result, it seems probable that an individual's activity and sleep, as tracked by their smartphone's accelerometer, could be related to and potentially predictive of their mood and wellbeing more broadly.

If possible, tracking mental wellbeing with an accelerometer could have benefits over using other sensors. For example, an accelerometer could provide more privacy than previously considered sensors, such as GPS location [17, 28, 63, 71, 132] and call logs [7, 87]. Another advantage to using an accelerometer is that the sensor is always available when the phone is turned on, including when the individual is out of service or, e.g., in a tunnel. While accelerometers embedded in a wearable device might have more potential to accurately track activity, smartphones are more ubiquitous and thus more realistic for long-term tracking.

Here, we are interested in focusing on and better understanding the relationships of physical activity and sleep, as measured by a smartphone accelerometer, with emotion for improving automatic mood tracking. We are particularly interested in understanding whether the relationships are predictive, especially from data collected with ordinary participant-owned smartphones in unconstrained environments (i.e., not imposing constraints on participants about where they need to keep the phone or if they need to have a special device with an accelerometer attached to their body). To explore these research questions, we conducted a field study, extracted measures of physical activity and sleep from smartphone accelerometer logs, related these measures to participants' self-reported wellbeing, and attempted to infer participants' wellbeing with classification and regression models. We expect that increased physical activity and better sleep quality will be related to improved self-reported mood and wellbeing.

3.2 Methods

Field Study

We recruited 106 participants from the university community through the Experimental Social Science Laboratory (XLab) for an eight-week field study to pilot methods. Participants were eligible if they owned an Android smartphone, were native English speakers, were undergraduate students, and agreed to the consent form. The study was approved by the University of California, Berkeley Internal Review Board. The participants were asked to take an entry survey, respond to daily wellbeing prompts on their smartphone, allow passive collection of sensor data from their smartphone, and take an exit survey.

Data Collection

Data was collected from participants through a custom Android app that used the Funf Open Sensing Framework [2]. This app was installed by participants before the study period and collected both passive sensor data as well as daily participant input. The participants were instructed and reminded to uninstall the app at study completion.

To quantify wellbeing, we followed prior studies and asked participants to repeatedly fill out a two question survey on their phone. Participants could enter information about their state on two nine point Likert scales, one for energy and one for mood. Scales were labeled with opposite poles, such as unhappy to happy and unenergetic to energetic. Participants could select the specific words from short lists of relative synonyms for each pole, such as unhappy, negative, sad, bad versus happy, positive, good. Participants were queried for their state four times a day. Each of the four daily surveys occurred at a random time within a predefined period between 8am and 10pm. The purpose for randomizing within periods was to ensure distribution of surveys throughout the day without having participants anticipate them. All responses given in a day were averaged into a daily level of perceived mood and energy.

To measure activity, we sampled the smartphone's accelerometer for intervals of three seconds every five minutes. This data was collected continuously from the time the app was installed. There were compatibility issues with phone models and network connections, so the amount of data collected on each subject varied. Quality of accelerometers also varied between phone models, which contributed to variance in the amount and quality of data collected on each individual. Some of the difficulties we encountered with sensor data collection included entirely missing observations, non-uniform readings during an observation interval and insufficient duration of sampling, i.e., less than three seconds. Participants were excluded from the analyses if they did not have complete data (wellbeing responses and activity readings) for at least 14 days of the study.

Data Processing

Preprocessing

The smartphones' 3-axis accelerometers measured the acceleration of the device in three directions. Following prior work, we considered the magnitude of the acceleration minus gravity [98]. Gravity for each segment was estimated as the average of coordinates in each of the directions. To account for irregular sampling and reduce noise in the sensor readings during a sampling interval, we interpolated the available data points and took regular sampling from the interpolation. Quadratic and cubic splines gave irregularities with missing readings, so a linear spline was identified as performing the best. This regular sampling allowed us to compute discrete Fourier transforms on the approximated signal and approximate the spectral density using Welch's method, i.e., averaging between Fourier transforms on multiple overlapping segments of the full observation window.

Activity Inference

We inferred activity from features summarizing the orientation-invariant magnitude of acceleration deviation and the spectral density of the magnitude of deviation of acceleration. The acceleration deviation was computed by subtracting the estimated gravity from all readings in the interval. This approach was taken, to allow for finer grain analysis of movement than is presented here. Much prior work with accelerometers, predicting both mental wellbeing [26, 63, 125] and activity [69, 83, 91], utilized features on coordinate-wise acceleration. However, such approaches were not applicable here, as our participants' phones were not in a fixed position during the study. We followed prior work that considered features on the magnitude and power spectrum of the magnitude of acceleration during the sample period [91]. The features we used were the average and standard deviation of the magnitude of acceleration and the dominant frequency, entropy of the normalized power spectrum, power in the high frequencies, medium frequencies, and low frequencies of the power spectrum of the magnitude of the acceleration. These eight features were used to fit two logistic regression classifiers. One classifier was trained to identify when the phone is "still" or set down, the second classifier identified "activity" such as walking, running, or pedaling a bicycle. We did not use a classifier to explicitly identify the phone being in a vehicle, such as a car, bus, or train. We did not find a classifier to be reliable enough given the many states a vehicle can assume, e.g., idling, accelerating, and traversing a smooth or bumpy road. Such a task was also of uncertain necessity because participants do not necessarily exert extra energy while riding in transportation, and thus was less likely to correspond to elevated mood from physical exertion. As a result, we focus the current study on measures of physical activity and sleep. The goal of these two classifiers was to quantify how long the phone was set down at night, and the subject presumably sleeping, and how long the participant was physically active during the day. These classifiers were trained on an auxiliary, activity labeled dataset that was collected with the same smartphone app and data processing pipeline. The classifiers achieved 80-95% accuracy on held out subjects from the training dataset.

Measure Extraction

Sleep Duration

Sleep duration was estimated as the length of the longest period during which the participant was not physically active starting after 9pm the prior evening. This period was calculated by looking at the longest contiguous series of observations when the accelerometer data predicted that the participant was "not active" and taking at the duration of that period. While this approach likely overestimates the duration of sleep, it should be representative of a period of passivity or evening rest and is preferable to the highly noisy alternative of considering the duration for which the phone was predicted to be still during the evening.

Nighttime Stillness

Sleep disturbance, or nighttime stillness, sought to capture sleep disturbance during the time when each participant's phone was most likely to be set down and the participant presumably asleep, based on their typical behavior. This measure was considered to be the fraction of time that a participant was still during their median period of late evening, or when their phone would typically be still based on their behavior during the study. The period of late evening was defined for each participant by first considering the longest contiguous set of observations during which the phone was predicted to be set down starting after 9pm for each day of the study. The median time that this period started, or presumably the phone was set down, for each day of the study defined the beginning of period and the median time that the contiguous "still" observations ended on each day of the study, was considered the end of the period of late evening. The nighttime stillness measure for each day of the study was the fraction of observations on that day of the study that occurred during the late evening period and were predicted to be "still".

Daytime Activity

For a measure of daily physical activity, we consider the daytime activity, which was the fraction of time that a participant was predicted to be physically active during their active period or the period of the day that we would expect each participant to be active, given their typical behavior during the study. The active period of the day was determined by first looking at the longest contiguous set of observations when the phone's predicted behavior was "not-physically active" starting after 9pm. The median time across all the days of the study when this physically not-active period began was considered as the end of the active period and the median end time of the not-active period was considered the beginning of the participant's typical active period. The "daytime activity" measure for each day of the study was then the fraction of time that the participant's phone predicted (with the models discussed previously) that the participant was "physically active" during the participant's active period.

Day of Study

Following prior work, we coded the day of the study as the number of days that had elapsed since the first day of the study [17]. This measure is important to account for potential participant fatigue, but also as it represents the progression of the academic semester, which may have had an effect on participants.

Weekday

The day of the week, and thus the potential effect of weekends on participants, was accounted for by coding weekdays in an ordinal variable from zero to six, Monday thru Sunday.

Type of measure	**Measure**	**How it was measured and calculated**
Time	Day of study (semester)	Coded as the number of days since the first day of the study.
	Day of week	Ordinal variable coded Monday (0) thru Sunday (6).
Sleep	Sleep duration	Longest contiguous time that the participant was not physically active starting after 9pm.
Activity	Daytime activity	Fraction of time participant was physically active during the median active period. The median active period is the time between the median hour the participant became physically active during each day of the study and the median hour that the participant stopped being active during the study.
	Nighttime stillness	Fraction of time the phone was predicted to be still, i.e., set down, during the median still period. The median still period was calculated over the course of the study to be the median hour that the longest contiguous still period started and the median hour it stopped.

Table 3.1: Daily measures of activity and sleep and how they were calculated.

Analyses

Relating Measures to Well-being

The first set of analyses sought to study the relationship of activity, sleep, and time on daily wellbeing. To account for the repeated measures design and missing data, we used mixed-effects linear models to relate reported average daily wellbeing measures to daily behavior measures [128]. We started with a maximal random-effects structure for each wellbeing measure to allow for individual variation and increase generalizability. Due to lack of initial convergence of the model, we followed suggestions in prior work by looking at the covariance of the partially converged model and removing the variable in minimum variance from the random-effects structure [12]. Using this procedure, we removed the measure of sleep disturbance, "nighttime stillness" from the random-effects structure when modeling mood and removed the scaled ordinal variable coding the day of the week when modeling energy. After this step both models converged. Activity and sleep measures were centered and normalized within individuals and time measures were scaled between zero and one before fitting the models to compare the relative sizes of effects.

To ensure the value of the model with maximally justified random-effects structure, we fit two additional models: a model with only random intercepts and no additional random-effects or fixed-effects and a model with fixed-effects and a random intercept only. Model fit was assessed with chi-square tests on the log likelihood values of different models. Model assumptions were visually checked. The linear mixed-effects models and analyses were carried out in the R programming language and environment [149] using the lme4 [14] and lmerTest [81] software packages.

Predicting Well-Being

The second set of analyses assessed if the relationships between daily mood and the activity, behavior, and time features were strong enough to be predictive. To do this we attempted two tasks. The first task was to predict if a participant was having a bad day, i.e., whether their wellbeing was lower than their median reported wellbeing. Only participants with sufficient observations of each class (at least 5 fine days and 5 bad days) were included in the analysis. The second task was to predict a participant's level of wellbeing.

Prediction Models

For the first task, predicting whether a participant was having a worse than usual day, we used logistic regressions with an L1 and an L2 norm penalty as well as support vector machines (SVM's) and random forests [39, 23]. For predicting the daily level of wellbeing, we used a linear regression model with the elastic net penalty [168] in addition to an Epsilon-Support Vector Regression and random forests. These models were used on individuals' data to build "personal models", rather than pooling all individuals' data into a "global model". Personal models were used because they have been shown to be the most successful approach to predicting individuals' responses [159]. Mixed-effect models help to model behavior within the population as a whole, while taking into account individuals have different behavior, but personal linear models are a best case scenario for predicting individuals' behavior from their own data.

Prediction Framework

For both prediction tasks, we evaluated prediction accuracy with leave-one-out cross-validation on personalized models, i.e., we trained a model on all but one of a participant's data points, evaluated the model accuracy on the held out observation, and then averaged accuracy across observations. The penalty weights hyper-parameters were set with leave-one-out cross-validation on the training data and scanning a variety of penalty weights. The predictive analysis was performed in Python with the scikit-learn library [112].

Model Evaluation: User Lift

The accuracy of predicting whether an individual was having a good day, was quantified by prediction error, or the percent of observations that were incorrectly predicted. The accuracy of predicting the level of wellbeing on a given day was quantified by root mean squared error, which is the square root of the average squared distance of a prediction from the true value. We report the accuracy of predictions compared with the accuracy of predicting each participant to be at their most common state. This measure is called "user lift"; it is the increase in accuracy, or decrease in error, that the model has relative to always predicting an individual to be at their most common state. By comparing a model with each participant's baseline, user lift reveals how much better a model is doing than guessing a participant to always be at their usual state. We then used permutation tests to assess whether user lift was significantly positive across the participants, i.e., that the models were significantly better than always guessing a participant to be at their most common state, as permutation tests are reported to be more reliable than paired non-parametric tests [75, 141].

3.3 Results

Participation

Of the 106 participants recruited, 87 installed our app and 57 completed the study, i.e., completed the exit survey at the end of the eight-week study period. However, there was only sufficient data on 53 participants to include in the analyses. Baseline characteristics of individuals included and excluded from the analyses are in Table 3.2 and indicate similar populations were included and excluded from the analyses. While some attrition was due to participation waning over the eight-week study period, there was also attrition as a result of technical difficulties and app compatibility issues on older phones.

Relationship of Sensor Data with Well-Being

From linear mixed-effects models, we found significant positive relationships of daytime activity and sleep duration with daily mood; when participants get more sleep and more daily activity they tend to report better moods (Table 3.3). Daytime activity has a stronger relative effect than sleep duration. Of note is that nighttime stillness (sleep disturbance) is not significant. This lack of significance could imply that the measurement is too noisy and that more work is needed to reliably measure sleep disturbance with a smartphone. The model with the maximal random-effects structure better accounted for the variance across

	Included participants with exit survey (n = 47)	**Included participants with no exit survey (n = 6)**	**Excluded participants due to insufficient data (n = 53)**
Age*	19.83 (1.99)	20.33 (1.60)	20.80 (4.13)
Female (number)*	26	3	28
BDI-20 score (entry)*	11.14 (9.27)	7.33 (3.54)	12.61 (7.20)
BDI-20 score (exit)*	11.98 (12.00)	-	-
Median mood rating	5.17 (1.63)	5.83 (0.90)	5.44 (1.44)
Median energy rating	5.60 (1.27)	6.67 (0.94)	5.98 (0.80)
Number of emotion surveys completed	160.51 (44.42)	139.33 (55.01)	30.25 (50.97)
Number of days with emotion ratings	49.45 (8.27	44.00 (11.06)	10.49 (15.99)
Reported typical sleep duration in hours (from exit survey)*	6.88 (1.35)	-	-
Average duration of inactive period in hours (sensed "sleep duration")	8.79 (1.22)	8.56 (0.48)	-
Number of times per month participant exercised (from exit survey)*	4.24 (5.04)	-	-
Average minutes active per day (sensed "daytime activity")	118.78 (32.67)	151.25 (59.68)	-
Number days with sensed activity and mood input	38.60 (9.15)	40.00 (9.64)	3.36 (5.15)

Table 3.2: Participant baseline characteristics. Averages across individuals are reported with standard deviations in parenthesis, except where indicated. Where appropriate, numbers represent the average across individuals of averages within individuals. "BDI-20" indicates optional self-reports to 20 questions of the Beck's Depression Inventory (the question related to suicidal ideation was omitted). Entry and exit survey questions were optional, so * indicates measures averaged only over submitted responses.

Linear mixed-effects model of mood				
Fixed-effect	**Estimate**	**Std. error**	**t value**	**P value**
Mean mood (intercept)	5.056	0.174	28.973	< .001
Day of study (semester)	-0.059	0.261	-0.226	.822
Day of week (coded 0-6, Mon-Sun)	0.040	0.076	0.528	.598
Sleep duration	0.072	0.030	2.451	.018
Daytime activity	0.097	0.032	3.062	.004
Nighttime stillness	0.040	0.026	1.528	.127

Table 3.3: Results of fixed-effects for linear mixed-effects model of mood level from smartphone measured and time variables. The measure for nighttime stillness was excluded from the otherwise maximal random-effects structure.

individual participants than the random intercept only model (Table 3.4). The main effects also remained significant, even when accounting for individual differences.

We also found a significant positive relationship of daytime activity with daily perceived energy level (Table 3.5). The relation for sleep, though negative, is not significant, revealing a potentially different relationship between the two emotions (mood and energy) with sleep. Day of the week has a significant positive fixed-effect, but had to be removed from the random-effects structure following prior suggestions about how to handle lack of model convergence [128]. This effect for day of the week indicated that participants collectively felt more energy at the end of the week and there is not sufficient evidence to support the idea that weekday affected participants differently. When we changed the variable encoding weekday to a binary variable indicating a fixed weekend of Saturday and Sunday versus the rest of the week, as has been suggested in related work [133], this relationship did not remain significant. An interaction term between a weekend indicator and daily activity was similarly not found to be significant. This lack of significance as a binary variable could be a result of weekends being less defined in our undergraduate population, some of whom may or may not have classes on Friday and thus have had extended “weekends. The lack of significance could alternatively result from insufficient observations of weekends for each participant. Again, sleep disturbance is not significant, further indicating that there might be too much noise in the variable measuring sleep quality. The model with the maximally justified random-effects structure accounted for significantly more variation across participants than having only a random intercept (Table 3.6).

Linear mixed-effects model fit for mood					
Model name	**AIC**	**BIC**	**Log likelihood**	**Chi-squared value (df)**	**P value**
Random intercept only	6522.0	6538.8	-3258.0		
Fixed-effects with random intercept only	6508.8	6553.7	-3246.4	23.177 (5)	< .001
Maximal random-effects structure	6322.0	6445.4	-3139.0	214.757 (14)	< .001

Table 3.4: Checking model fits for linear mixed-effects model of mood.

Linear mixed-effects model of energy				
Fixed-effect	**Estimate**	**Std. error**	**t value**	**P value**
Mean energy (intercept)	5.686	0.184	30.857	< .001
Day of study (semester)	-0.304	0.233	-1.303	.199
Day of week (coded 0-6, Mon-Sun)	0.196	0.067	2.912	.004
Sleep duration	-0.027	0.031	-0.858	.394
Daytime activity	0.182	0.039	4.673	< .001
Nighttime stillness	0.024	0.030	0.810	.422

Table 3.5: Fixed-effects for a mixed-effects linear model relating daily energy level from smartphone measured and time variables. The ordinal variable for weekday was excluded from the near-maximal random-effects structure.

Linear mixed-effects model fit for energy					
Model name	**AIC**	**BIC**	**Log likelihood**	**Chi-squared value (df)**	**P value**
Random intercept only	6284.2	6301.0	-3139.1		
Fixed-effects with random intercept only	6196.1	6240.9	-3090.0	98.117 (5)	< .001
Maximal random-effects structure	5972.5	6095.9	-2964.2	251.562 (14)	< .001

Table 3.6: Checking model fits for linear mixed-effects model of energy.

Problem (model)	Wellbeing measure	Average user lift	Minimum user lift	Maximum user lift	P value
Good or bad day (penalized logistic regression)	Mood (Pred. error)	5.44%	-21.74%	35.00%	.001
	Energy (Pred. error)	4.92%	-22.73%	39.39%	.008
Daily average (linear regression with elastic net)	Mood (RMSE)	0.026	-0.232	0.48	.085
	Energy (RMSE)	0.048	-0.169	0.575	.011

Table 3.7: Statistics on linear models predicting daily wellbeing from activity measures. While the models provide an improvement overall, there is a range in the ability to model individuals. The P-values are for permutation tests checking whether user lift is greater than zero, i.e., if models are significantly more accurate than always predicting each individual to be at their most frequent state.

Predicting Well-Being From Sensor Data

3.4 Discussion

Principal Findings

We found that increased daily activity, as tracked with a smartphone's accelerometer, positively correlated with participant-reported mental wellbeing over time. While a positive correlation of activity and wellbeing has been substantiated in literature external to mHealth [50, 57, 96, 113, 144], we have shown that smartphones measure individuals' daily activity to a sufficient level of accuracy to measure this relationship in everyday life. While the potential for this result has been shown in environments where constraints were placed on the participants [63, 125, 66, 26], we have found this relationship present when no constraints are placed on participants. Previous work did not find a significant correlation of the total activity in a 24-hour day with stress [17], which could indicate the need for distinguishing daytime activity from nighttime activity, as we have done, or indicate that physical behavior has unique effects on different emotions, which we have observed by considering mood and energy separately.

We also found that a simple measure of sleep duration derived solely from accelerometer data was significantly positively correlated with mood. However, it was not significantly correlated with perceived energy, which supports the idea that there are different relationships between different emotions and physical behaviors. We did not find a significant correlation of either mood or energy with our measure of smartphone measured sleep disturbance. This may imply that the measure did not sufficiently describe sleep quality and more work is needed to monitor sleep quality in a sustainable way. It is possible that a more sophisticated method for predicting sleep, such as found in prior works, would allow for a finer measure of sleep disturbance [32].

When we used the activity, sleep, and time measures to predict individuals' wellbeing, we found modest, but significant improvement over naive baseline models. It is important to emphasize that there was a range in our ability to predict individuals' wellbeing from their activity and sleep behavior. This range highlights the need for tracking approaches that tailor to the user. However, it is unclear if this effect is the result of a range in how thoughtfully individuals responded with their state, phone usage, data quality and quantity, or the strength of wellbeing and activity relationship between individuals.

Limitations

A limitation of this study is that participants', self-reported wellbeing is subjective and the population was not clinically assessed. However, the measures of wellbeing that we used have been widely used and prior research has found simple single scale measures to be related to longer clinical assessments [1]. While a better measure of wellbeing could be a longer survey, such a measure would incur significant participant fatigue and likely decrease the duration of participation.

Whether all of the participants' relevant activity was tracked with smartphones during the study is another concern. There are limitations to activity recognition, especially when the smartphone is not in a fixed position, a participant is doing a non-standard activity, or the phone is set down, e.g., left in a gym locker. However, the study cohort retrospectively reported little vigorous exercise during the study period (Table 3.2), so the underestimation of vigorous exercise is likely minor. Such limitations could possibly partially be mitigated with location tracking, but time at a location is not necessarily representative of activity and poor GPS sensitivity would remain a challenge. Wearables may provide a better facsimile of an individual's behavior when they are worn, but they have notorious compliance limitations that smartphones don't suffer.

Another limitation was the sample size and lack of clinical population. Some of the individuals in our study cohort did report elevated levels of depressive symptoms in the entry and exit survey. However, the cohort is not necessarily representative of a population with clinically diagnosed mood disorders. Depressed individuals often are less active than the general population, but even small increases in physical activity can improve symptoms [150].

Conclusions

This study examined the extent to which smartphones' accelerometers can contribute to passively tracking individuals' mental wellbeing in everyday life. We have found that smartphones measure activity and sleep with sufficient accuracy to reproduce prior findings of significant relationships between activity and sleep with mood. While models have a modest, though significant, improvement over naive baseline models in general, the range in predictive capability implies that more work is needed to tailor mood and depression tracking apps to individuals.

Our results support the promise for smartphones to be used in sophisticated and long-term monitoring of patients' wellbeing. Because smartphone use is high and their presence ubiquitous, the ability to use a smartphone for tracking mental wellbeing could have a huge impact on mental health care. Smartphone monitoring may improve self-management via smartphone apps making care more affordable and thus accessible to individuals who currently do not have access to care. Passive monitoring could also be used as an adjunct to clinician led treatment, thus increasing the quality of care and personalizing treatments.

Chapter 4

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

Originally published as Demasi, et al. "Detecting change in depressive symptoms from daily wellbeing questions, personality, and activity." IEEE Wireless Health (2016).

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.

4.1 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 [84, 125]. 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 [84]. The goal of automatic journaling has been attempted, in particular, for monitoring mood disorders, such as bipolar and depression [17, 18, 25, 28, 63, 71, 87, 132]. 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) [131] 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) [15], 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) [15] 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 [95] personality features: openness, extraversion, neuroticism, agreeableness, conscientiousness.

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 [17, 132], 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 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

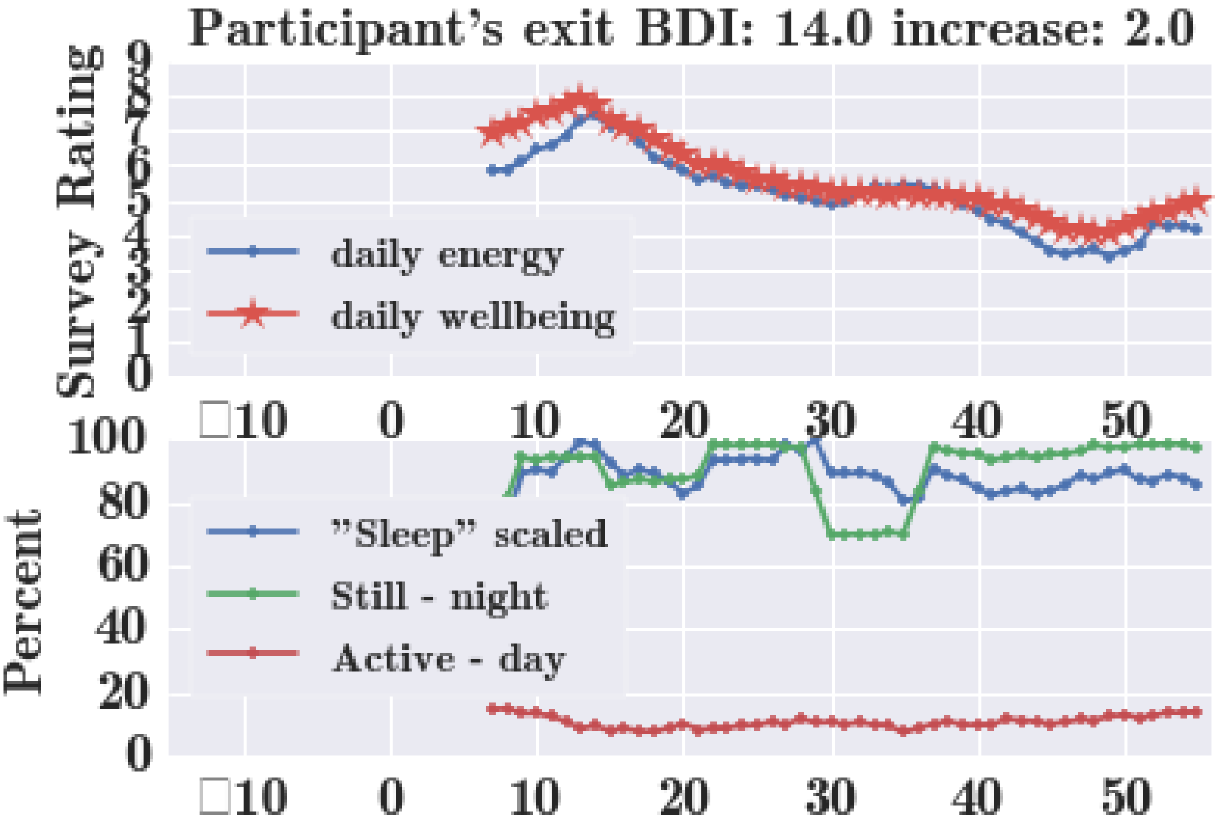

Figure 4.1: Example behavior of an individual. Note apparent decrease in average daily wellbeing and energy. This decrease corresponds with a reported increase (in two points) to depressive symptoms (BDI score). The sensed activity and sleep behavior is relatively consistent during the study. Sleep is scaled by the maximum duration sample to make units comparable to daily percents.

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.

4.2 Contributions

We make two significant contributions. First, we build on prior work that looked at utility of simple daily measures of wellbeing [1]. 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.

4.3 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 [18, 71, 87, 93], phone processes [87], Bluetooth [18], GPS location traces [17, 28, 63, 71, 87, 132, 93], sound data [18], physiology sensors (from wristbands) [71, 66], and macro-activity data [93, 66]. 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 [17, 28, 63, 132]. Few authors have tried to predict values of long-term mood measures [132] or changes in outcome measures [17] 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

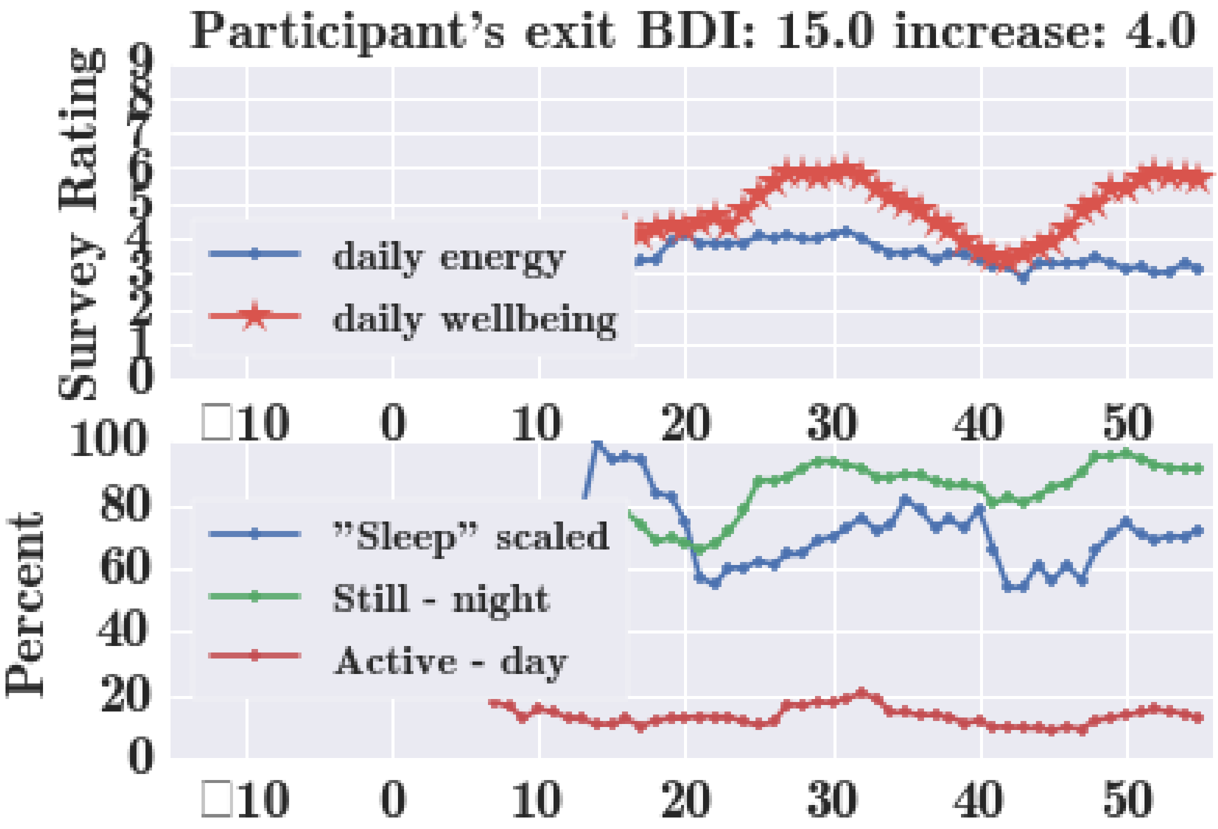

Figure 4.2: A user's behavior. Note apparent increase in daily wellbeing and energy measurements, but a reported increase in long-term depressive symptoms (BDI score), which contrasts with the previous user in Figure 4.1. This user also has considerable fluctuation in their daily activity level and sensed sleep duration.

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 [50, 57, 92, 96, 101, 113, 117, 118, 144]. 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 [69, 83, 91], as well as sleep [32, 136]. Other pilot projects implied that mental states can be recovered from accelerometer data on small populations in artificial settings [26, 125].

4.4 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 mood [18, 71, 87]. 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 4.1 and 4.2. In Figure 4.1 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 [1].) The user in Figure 4.1 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 4.2. 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 4.1 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.

4.5 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 [131] to align with previous work that has adopted this model [84, 87, 66, 126]. 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.

4.6 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 4.3. While most participants reported a score less than 10, a few participants reported

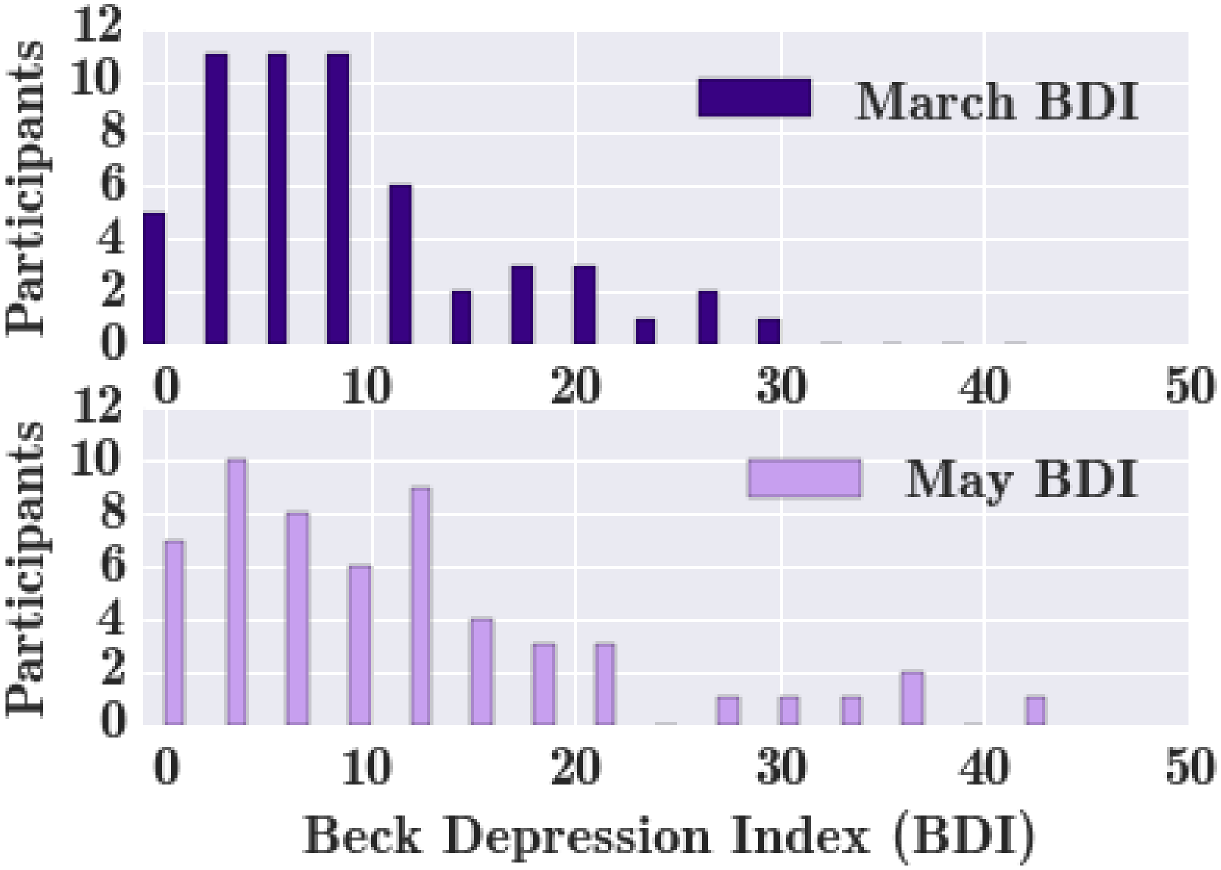

Figure 4.3: Distribution of BDI scores that participants reported in the entry survey in March and in the exit survey in May. Note the slight drift of the distribution to higher BDI scores (more expressed depressive symptoms) over the course of the semester. The study ended the week before finals.

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 4.4. 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

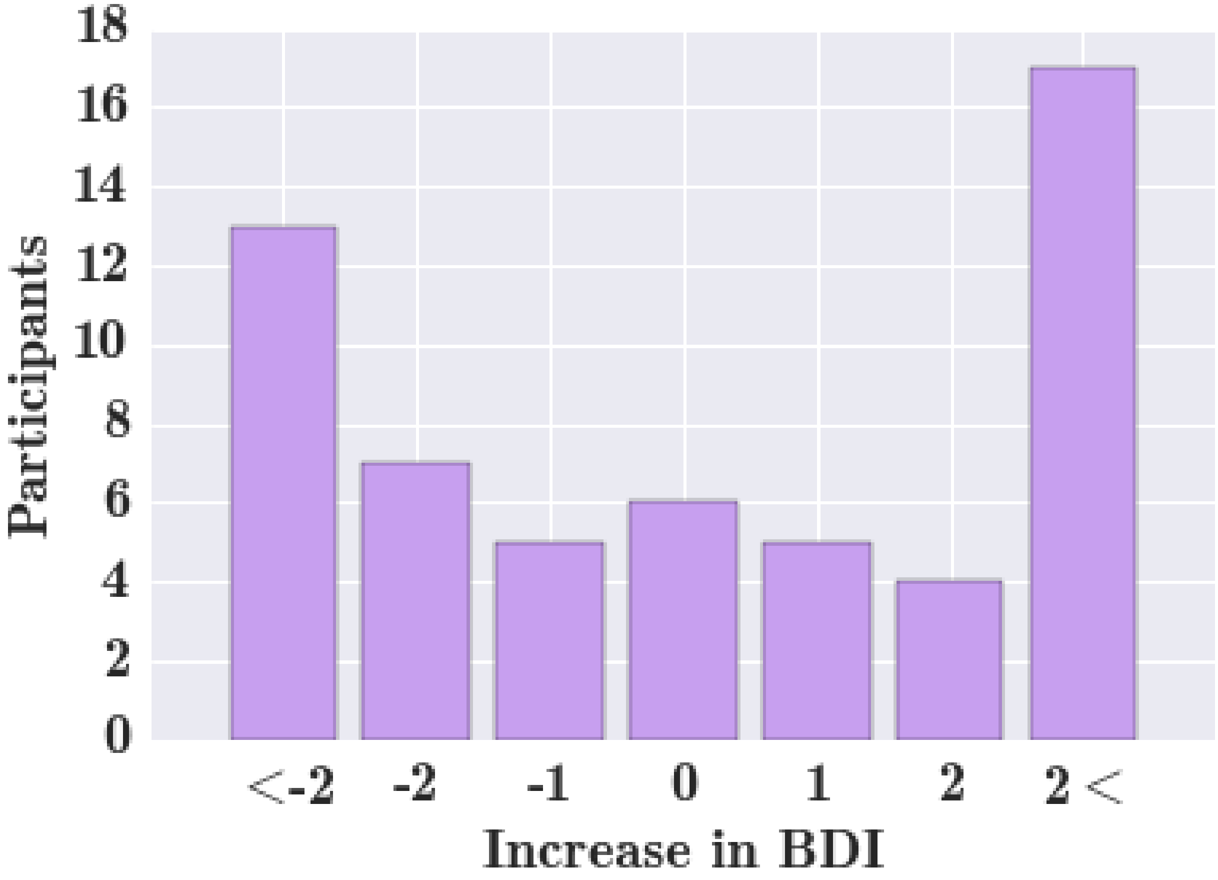

Figure 4.4: The distribution of individuals' change in BDI score from the entry to the exit survey. A positive increase indicates an increase in BDI score (expression of depressive symptoms). During the course of the semester more students experienced an increase in depressive feelings than a decrease.

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 [95] 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 [2] 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.

4.7 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 [91]. 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 [32]. 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.

4.8 Features

The features that we use to describe participants and their behavior during the study are summarized in Table 4.1. 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 [1].

Feature type	Name	Description
Personality	Neurotic	Big 5 personality test
	Extroversion	Big 5 personality test
	Openness	Big 5 personality test
	Agreeable	Big 5 personality test
	Conscientious	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 4.1: 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.

4.9 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.

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 [138]) 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 [151]. Both of these methods yield models with a limited number of terms and a coefficient on each terms that indicates

	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 – standard deviation	7.2069 (*)	x	5.599	x
Sleep duration – slope	-6.9844 (0.001)	x	-2.204	x
Sleep duration – Difference with last 2 weeks	4.7048 (0.017)	x	x	x
Daytime activity – Average of observations	x	x	0.342	x
Daytime activity – Difference with last 2 weeks	x	x	-0.067	x
Daytime stillness – standard deviation	-3.3079 (0.001)	x	-1.019	x
Daytime stillness – Difference with last week	1.5866 (0.053)	x	x	x
Daily energy – Entropy	x	x	0.150	x
Daily energy – Difference with 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 4.2: 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.

how much that term contributes to the model.

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.

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.

4.10 Results

The feature selection, model fit, and prediction accuracy using both of the regression approaches outlined above are presented in Table 4.2. There are four models presented in Table 4.2. 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.

Linear Regression with Forward Selection and BIC

The first (left most) column of coefficients in Table 4.2 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.

L1 (Lasso) Penalized Linear Model

The third column of coefficients (second from the right) in Table 4.2 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.

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 single outlying participant removed. These two models are presented in the second and fourth (right most) columns of coefficients in Table 4.2.

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.

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 4.2. 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

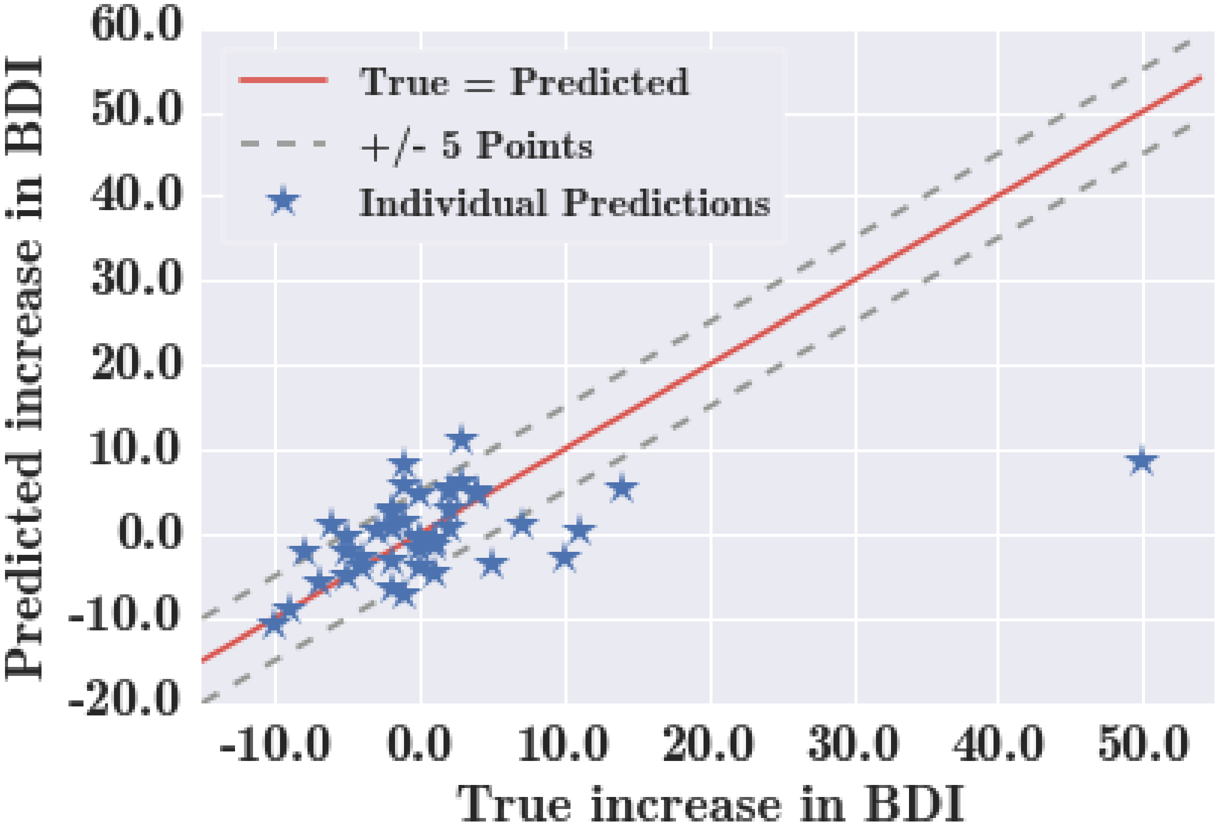

Figure 4.5: The distribution of predictions from leave-one-out cross-validation. Features were preselected with forward selection and models were fit on population with the outlier included. Most predictions are within the dotted lines indicated predictions within five points of the true increase.

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 4.5 displays the distribution of predicted BDI increases relative to the true in-

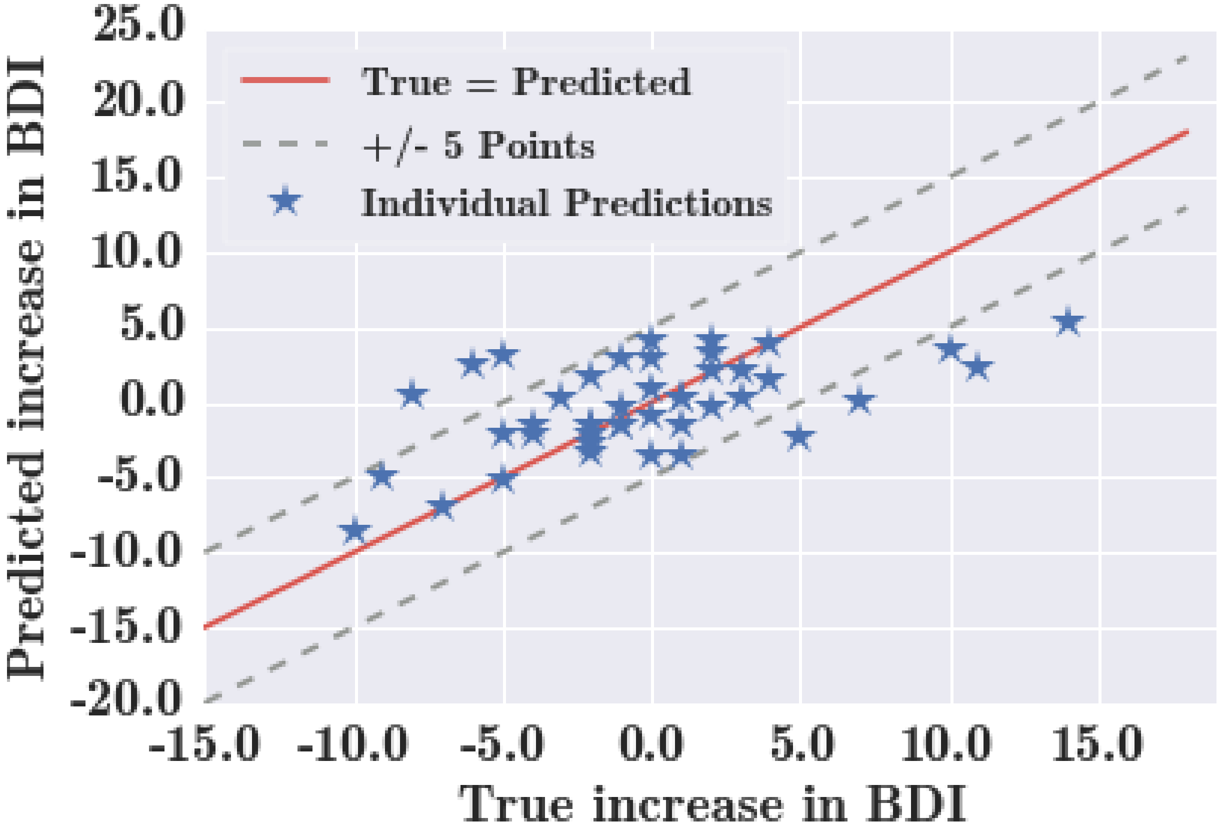

Figure 4.6: The distribution of predictions with leave-one-out cross-validation. Models were fit on a population with the outlier removed and features were pre-selected by forward selection with the BIC.

creases in BDI scores observed when the outlying participant is included in the dataset. Figure 4.6 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.

4.11 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 [17]. 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.

4.12 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.

Chapter 5

A Step Towards Quantifying When an Algorithm Can and Cannot Predict an Individual”s Wellbeing

Originally published as DeMasi, et al. “A step towards quantifying when an algorithm can and cannot predict an individual’s wellbeing.” Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers (2017).

Researchers are exploring the ability to infer complex signals, such a mental wellbeing, from easily collected smartphone behavioral data. Rather than focusing on improving overall accuracy of such an approach, we seek to understand when we are and are not capable of predicting an individual’s wellbeing. In particular, we consider the ability to predict daily wellbeing from smartphone GPS location data as a case study. We hypothesize that user characteristics, such as behavioral variability, level of depression symptoms, and amount of labeled data, are related to improvements in prediction accuracy. Our preliminary results indicate that there may be a relationship between an algorithm’s ability to successfully predict an individual’s wellbeing reports and the individual’s location behavior variability. While further work is needed to improve model accuracy and confirm this relationship in a larger study, our work is a step in the necessary direction of understanding which individuals can be monitored with smartphone data.

5.1 Introduction

Mental health disorders can be devastating to those who suffer from them and are widespread. Collectively, it is thought that mental health disorders, such as depression, are so widespread that they are a major contributor to the global disease burden [106]. Improving mental health is particularly challenging as disorders can last for a lifetime and it is difficult to collect data on and monitor individuals over such long timescales.

The recent development and adoption of personal electronics provides an exciting opportunity for mental health, as personal electronics are a frequent source of highly personal data. It has been shown that data from personal electronics, such as smartphones, can be used to infer behavioral signals, such as sleep [32] and activity [83] without any user input. In addition to physical and social behaviors, researchers have begun exploring whether personal electronics can also sense mental wellbeing from passively collected data, such as smartphone GPS location and mobility [17, 28, 71, 132]. By not needing any user input, these devices may be a sustainable way to collect data on and track individuals' behavior over longer periods than are sustainable with paper journaling.

The possibility of automatic wellbeing tracking over long periods of time offers exciting opportunities for mental health research and treatment. However, the hope of tracking every individual with a smartphone may be naïve due to the large variance in individuals' behavior. Some individuals, say those who use their phone often, have active lifestyles, or have irregular schedules may be easier to track than individuals who, for example, often leave their phone at home or turn it off.

In this work, we explore the potential for understanding, and eventually predicting, whether an individual's wellbeing can be tracked by a smartphone application through a user study. As an example, we use features of location and mobility from GPS coordinates to predict individuals' daily wellbeing. The features of location and mobility that we use are inspired by previous research that sought to diagnose depression from similar features [132]. These features were chosen due to their apparent relevance to detecting depression and their reproducibility with the collected data.

We begin by calculating location and mobility features for each participant. Using these features, we utilize machine learning algorithms to predict each individual's wellbeing and then quantify the model improvement with GPS data over a simple baseline approach. We then look at the relationship of prediction improvement with user characteristics to see if user behavior is broadly related to an algorithm's ability to model their wellbeing. The user characteristics that we consider are data quality (as measured by median GPS accuracy) and quantity, behavioral variability, depressive symptoms, and emotional variability.

We find some significant positive correlations of user characteristics with prediction improvement. In particular, we find a positive correlation of the number of data points with prediction improvement, a negative correlation of baseline accuracy (i.e., how constant a user reports their state to be) with prediction improvement, and a positive correlation of location or behavioral variance with prediction improvement. By considering location variability as a course measure for behavioral variability, this result indicates what one would expect from a statistical perspective – more varying features are better able to model signals than features that rarely vary. From a psychological perspective, this result indicates that users with unfluctutating behavior are more difficult to model, perhaps because changes are outliers. Notably, we do not find significant relationships of depressive symptoms with prediction improvement. However, these relationships are not present across models, which could be the result of the task being difficult for models. While our preliminary results indicate some promise in being able to understand which individuals' wellbeing are easier to predict, further work and a larger study are needed to confirm the relationships of prediction improvement with user characteristics.

5.2 User Study and Data Collection

To explore whether there is a relation between user characteristics and success in predicting their wellbeing, we ran a user study. For this study, we recruited undergraduates with Android phones who spoke English as a native language on the University of California, Berkeley campus. While we recruited 107 participants, only 87 installed our custom Android application and 60 took the exit survey at the end of the study period.

The study ran for eight weeks, consisted of three phases, and collected two types of data: active user input and passive smartphone sensor data. The first phase of the study was an entry survey which asked user profile information, such as personality, demographics, and the Beck's Depression Inventory (BDI) [15]. The second phase was the daily collection of user input data, ecological momentary assessments (EMA's) of user wellbeing, and passive collection of smartphone sensor data. The final phase of the study was an exit survey which collected personality, the BDI, reflection on personal behavior during the study, and study feedback.

Users were queried four times a data for their wellbeing along two axes of the Circumplex model: mood and energy level [131]. These axes were labeled with words such as "good" and "bad" or "high energy" and "low energy", respectively, and implemented as two 9-point Likert scales. The words labeling the scales were selected by users from short lists of antonyms.

In addition to data from other sensors, data were collected of users' GPS location using the Funf Open Sensing Framework [2]. A user's location was recorded every five minutes. While a multi-modal approach is ultimately desired, we focus in this study only on the location data as an exploratory study of this approach. We chose GPS location for passively collected data as prior studies have indicated positive results with such an approach [17, 28, 71, 132]

5.3 Data Processing

The first stage of data processing aggregated multiple wellbeing observations that were made each day into a daily measure of wellbeing. Then the passively collected smartphone sensor data was processed into features describing individuals' daily mobility and location. Finally, features or "user characteristics" were calculated on each user that sought to quantify behaviors that may account for variability in prediction accuracy between users.

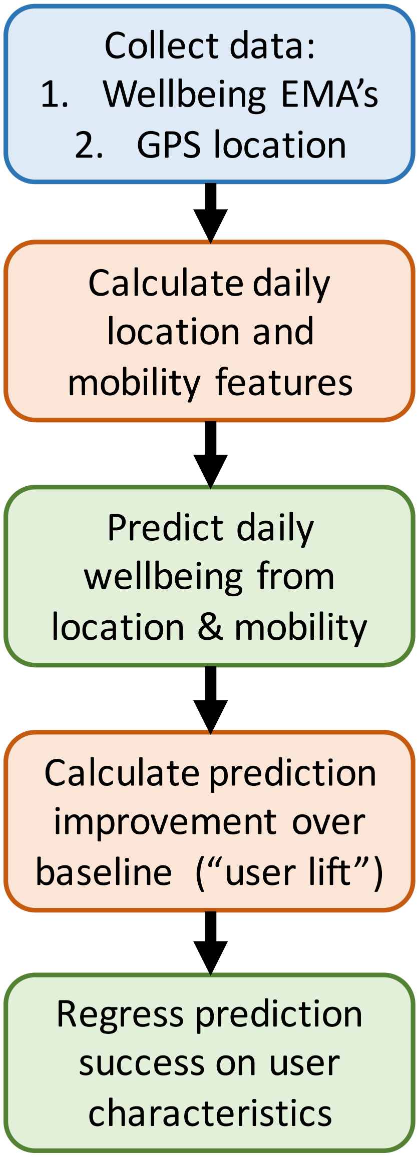
Collect data:
1. Wellbeing EMA's
2. GPS location
Calculate daily location and mobility features
Predict daily wellbeing from location & mobility
Calculate prediction improvement over baseline ("user lift")
Regress prediction success on user characteristics

Figure 5.1: Overview of data flow from collection and processing through the final analysis of whether user characteristics are related to the improvement in prediction accuracy.

User Wellbeing

Users’ wellbeing scores, which were solicited with four EMA’s per day, were averaged to give daily levels. The two wellbeing dimensions measured, energy and mood, were considered separately. The means of the daily mood an energy levels during the course of the study were taken as study-means. These study-mean levels were used to determine when a user was having a particularly good day in terms of mood or energy. A particularly good mood day was when the mood level was above the study-mean mood score and similarly for energy. Using this approach, we accumulated two wellbeing measures for each day that a user responded to any wellbeing prompts: whether the user was reporting an above-mean mood level for that day and whether the user was reporting an above-mean energy level.

Daily Location and Mobility

Here we focus on using GPS location and mobility features as predictors of daily user wellbeing. Location was intended to be collected every five minutes. However, some individuals’ locations were collected at a higher frequency. For these users, we downsampled data to roughly five minute intervals. The features we used to describe daily location and mobility are adapted from a previous study that used similar features to quantify user behavior during an entire study period [132]. We selected these features due to our ability to reproduce them, given our regular sampling approach, and their success on a related task.

Before constructing daily features, we used a preprocessing stage to determine frequented locations. The preprocessing used K-Means clustering [6] to cluster all of a user’s stationary location coordinates that were recorded during the entire study period. Points were determined to be stationary if the calculated gradient was less than 1km/hr. We chose the number of clusters for each user be such that the largest distance from any coordinate to the center of its assigned cluster was about 3km. We labeled the “home” location to be that which the user spent the most time at during the study period between the hours of midnight to 6am.

For each day of the study period when a user had sufficient GPS readings we used the cluster centers from the preprocessing stage and calculated the following measures:

- The sum of the variance of the latitude and the variance of the longitude coordinates, on a log scale.
- The number of locations (clusters) visited.
- The location entropy, i.e., $-\sum_i p_i \log p_i$ where p_i is the probability of the user being in location i at any point during that data.
- The fraction of time that the user spent at what we presume is their home location.
- The fraction of time the user was moving.
- The average distance that a user traversed between location readings (normalized by the time between readings).

R^2	$= 0.241$
Adj. R^2	$= 0.101$
F-stat	$= 1.717$
p	=.165

Feature	**coef**	**p-val**
Intercept	-25.49	.574
GPS radius	4.05	.324
No. days	0.34	.206
Avg. BDI	0.06	.819
Loc. var.	0.92	.208
Base acc.	-0.03	.967

Table 5.1: Linear model relating user lift (from L1-penalized logistic regressions) to user characteristics. On average, daily predictions were 3.77% less accurate than a constant baseline model.

- The “circadian rhythm”, which we calculated as the euclidian distance between the vector where entry j is the fraction of time that a user spent at location j on an average day, and the day’s vector where each entry i is the fraction of time that day that the user spent at location i.
- The radius of the minimum size circle that surrounded all of the user’s locations for the day.
- The fraction of observations during which the user was moving (as determined by the calculated location gradient).
- The fraction of observations that were “GPS” rather than “Network”, which could indicate the fraction of time that the user spent outside.

User Characteristics

We hypothesized that it is plausible that how well a user’s location and mobility behavior reflects – and thus is predictive of – their wellbeing could be related to the following five dimensions:

1. How reliable a phone is at measuring location.
2. How much data a model has to learn from.
3. How depressed a user is.

4. How much a user's daily location pattern fluctuates.

5. How much a user's emotional wellbeing fluctuates.

We quantified these potential sources of variability with the following measures:

- The median radius of confidence reported by the GPS sensor (on a log scale).
- The number of labeled data points we have for a user (i.e., days with GPS location and user wellbeing).
- The expression of depressive symptoms (as measured by the BDI and averaged between the entry and exit responses).
- The sum of the variance of longitude coordinates and variance of the latitude coordinates during the course of the study (on a log scale).
- The user's baseline accuracy: the percent of wellbeing observations that would be correctly predicted if the user were always predicted to be at their most commonly reported state.

The radius of confidence or "inaccuracy" of the GPS location data, is the radius of the circle that the sensor estimates the true location falls into with high confidence. The second to last measure, location or behavioral variance, is related to the daily location variance described previously. Instead of being calculated on the coordinates for a single day, it was calculated on all coordinates from the entire study period. This feature characterized a user's behavior during the study period instead of during a single day.

5.4 Data Analysis

We performed two stages of analysis to explore whether user characteristics relate to how successfully an algorithm can predict an individual's wellbeing. First we used standard machine learning procedures to predict daily user wellbeing from the location and mobility features. Second we related the success of these models to the user characteristics described above. An individual's emotion and location had to be observed for at least 14 days of the study for them to be included in the analyses. To quantify success, we needed to account for variability in how regularly individuals reported a single wellbeing measure.

Individual Wellbeing Baseline Models

Individuals reported different levels of emotional variance, e.g., some individuals always reported the same mood while others report different mood levels. As a result, certain individuals are "easy" to predict with high accuracy, as predicting that they are always at the

$R^2 = 0.647$
Adj. $R^2 = 0.582$
F-stat = 9.906
p ¡ .001

Feature	**coef**	**p-val**
Intercept	51.27	.032
GPS radius	2.59	.214
No. days	0.34	.015
Avg. BDI	-0.2	.130
Loc. var.	0.8	.035
Base acc.	-1.25	.002

Table 5.2: Linear model relating user lift (from L2-penalized logistic regressions) to user characteristics. On average, daily predictions were 1.97% less accurate than a constant baseline model.

same state will usually be correct. However, from an algorithm's perspective, these individuals are challenging. It is difficult for an algorithm to predict the individual's wellbeing better than a baseline model that always guesses that the individual is always at the same state. To account for individuals' base level of difficulty, we considered the "baseline accuracy", which is the percent of observations that would be correctly predicted if an individual were always predicted to be at their most frequently reported state. The "baseline error" is the percent of observations that would be incorrectly predicted by always assuming that an individual is at their most commonly reported state.

Wellbeing Prediction

In the first stage of analysis, we attempted to predict whether a user was having a particularly good day (in terms of mood or energy level) from their location and mobility measures. For these predictions, we used a variety of standard machine learning models: logistic regression (with L1 and L2 penalties), random forest classifiers, and support vector machines (SVM's) [13, 58]. Models were trained on each individual's data (personal models) with leave-one-out cross-validation. Model hyperparameters were trained with 10-fold cross-validation on the training set.

Characterizing Prediction Improvement with User Lift

To characterize prediction improvement over a naïve approach that uses no features, we considered ***user lift*** to be the difference of model accuracy with the baseline accuracy

$R^2 = 0.402$
Adj. $R^2 = 0.291$
F-stat = 3.633
p =.012

Feature	**coef**	**p-val**
Intercept	-4.67	.884
GPS radius	2.63	.366
No. days	0.22	.249
Avg. BDI	-0.02	.931
Loc. var.	1.38	.011
Base acc.	-0.25	.630

Table 5.3: Linear model relating user lift (from random forests) to user characteristics. On average, daily predictions were 5.16% less accurate than a constant baseline model.

described above [48]. User lift quantifies for each user how much better a machine learning model is than guessing.

Relating User Characteristics to Prediction Success

To better understand when users' daily wellbeing may be predicted by an algorithm, we related different algorithms' user lift for each individual to the user's above mentioned user characteristics. We related prediction improvement, i.e., user lift, to user characteristics with a multivariate linear regression. This model was chosen for interpretability.

5.5 Results

Of the individuals who participated in our field study, 33 had enough data to be included in our analyses. This limited number was in part due to compatibility issues that we encountered with the smartphone application and in part due to limited user participation. The level of depressive symptoms for each participant was quantified as their average report (between entry and exit surveys) to 20 questions of the BDI. The mean level reported across included participants was 12.68 (standard deviation: 10.66). Of the participants included in the analyses, 29.63% reported levels above 15, which could indicate mild levels of depressive symptoms.

Predicting Daily Wellbeing

In general, we found that models did not have appreciably higher prediction accuracy than the baseline approach, i.e., predicting users to be at their most common state all the time. This result is reflected in negative average user lift for predicting daily energy. It is also reflected by low correlation of model accuracy across individuals, as can be seen in Figure 5.2. The model with maximum average user lift for predicting daily energy was the support vector machine, which still had negative user lift (i.e., improvement over the baseline approach) of -1.50%. We also noted models were worse at predicting whether an individual's mood was particularly good than predicting whether an individual's energy was particularly high. However, on individual users some of the models performed considerably better than the constant baseline approach. This variance of performance between individuals motivates the second stage of analysis.

Explaining Prediction Improvement

Correlation between user characteristics was fairly low, as seen in Figure 5.3. The user characteristics that were most correlated were the variance in location coordinates and the total number of observations. The multivariate regression models relating user lift of daily energy predictions to user characteristics are summarized in Models 5.1 - 5.4. These models explore the improvement of predicting energy and not mood. While user lift for mood prediction did vary between individuals, the overall average user lift was better for energy prediction. As a result, we proceed with understanding the error of predicting energy and will investigate mood further after better models have been developed.

Despite accounting for little of the overall variability between individuals, user characteristics had significant relationships with the user lift from L2-penalized logistic regressions and random forest models (p ¡ .01) as well as from SVM's (p ¡ .05). When random forests or an L2-penalized logistic regression are used as the prediction model, we see a positive correlation of location variance with prediction improvement. This indicates that individuals who displayed more physical behavioral variance were easier to successfully predict than those with little variation. For the L2-penalized logistic regression, we also note a significant positive correlation of the number of data points with user lift and a negative correlation with the baseline accuracy. These relationships indicate that individuals who have more data are easier to learn, and those who report little fluctuation in state are harder to predict more accurately than a baseline model, which is already quite accurate.

An interesting consistency between models is a lack of significant relationship between the reported expression of depressive symptoms, as measured by the BDI, with model improvement. Higher BDI scores indicates increased depressive symptoms. No significant relationships indicate that individuals with higher depressive symptoms are either easier or more difficult to predict.

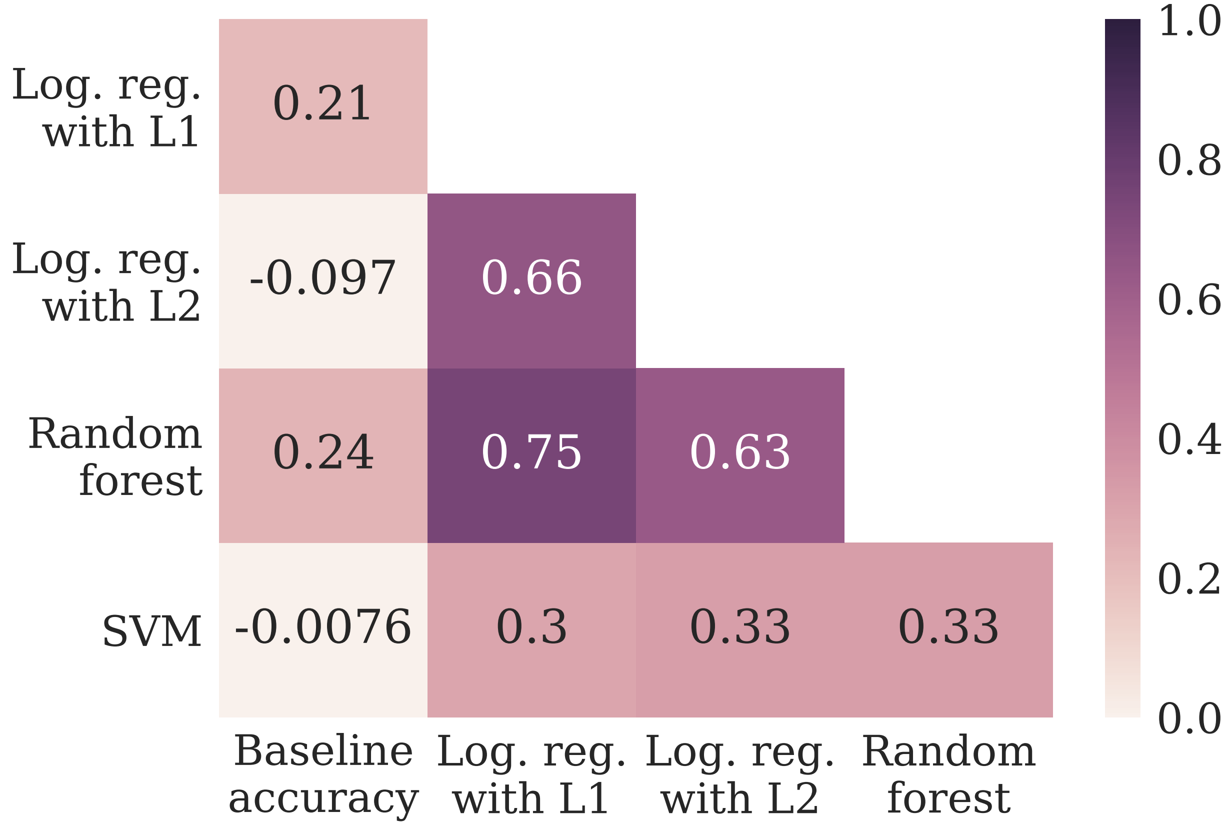

Figure 5.2: Correlation of how well different models and a constant baseline model predict individuals' daily energy. Correlation is calculated between average prediction accuracy on individuals from different models.

5.6 Discussion

In this work, we explored the potential to explain when individuals' wellbeing can and cannot be predicted by location data from their smartphone. We have focused preliminarily on the example of predicting perceived energy level from GPS location and mobility data and relating prediction improvement to user characteristics, such as emotional variability, location variance, level of depressive symptoms, and amount of data collected.

In general, it was difficult to learn models that made better predictions than predicting each individual to always be at their most common state. Daily mood was particularly difficult and insufficiently accurately predicted. There were improvements in prediction accuracy when using location data to predict energy, but this did not seem to remain consistent across models, as indicated by low correlation between average model accuracy (Figure 5.2). The variability in model improvement between individuals motivated us to compare prediction

$R^2 = 0.330$		
Adj. $R^2 = 0.206$		
F-stat = 2.656		
p =.045		
Feature	**coef**	**p-val**
Intercept	45.04	.252
GPS radius	-5.0	.160
No. days	0.43	.063
Avg. BDI	-0.04	.859
Loc. var.	0.47	.455
Base acc.	-0.91	.147

Table 5.4: Linear model relating user lift (from support vector machines) to user characteristics. On average, daily predictions were 1.50% less accurate than a constant baseline model.

improvement with user characteristics.

When we related model improvement to user characteristics, we found a significant positive correlation between location variance, which we used as a coarse measure of behavioral variance, and model improvement (user lift). From a statistical perspective, this indicates that users' wellbeing can be better learned from more varying and potentially descriptive features. From an applied perspective it indicates that more active individuals might be easier to monitor with this approach. Additional characteristics were also found to be significant, but were dependent upon which model was used. Depressive symptoms where notably not found to have a significant relationship with model improvement regardless of model.

There are limitations to this work, including a small sample size. A study with a larger cohort size is needed to validate the above mentioned relationships (and lack of relationships). We have also restricted our first step to explore GPS data, but other sensors should be included, as some sensors may be more predictive for different individuals. The limited predictive capability of location and mobility features that we found could have also constrained our ability to explain model improvement by having little model improvement in general. It is possible that with more descriptive features (or a multi-modal approach) daily wellbeing prediction would be more accurate and thus the resulting model improvement would have stronger or different relationships with user characteristics. Finally, additional user characteristics should be considered, which may improve the quantification of user variability and reveal stronger relationships between model improvement and user characteristics. In particular, different measures of depressive symptoms, other than the BDI, may better capture depressive symptoms that may influence predictive capability.

In future work we would like to explore a larger study population and incorporate more

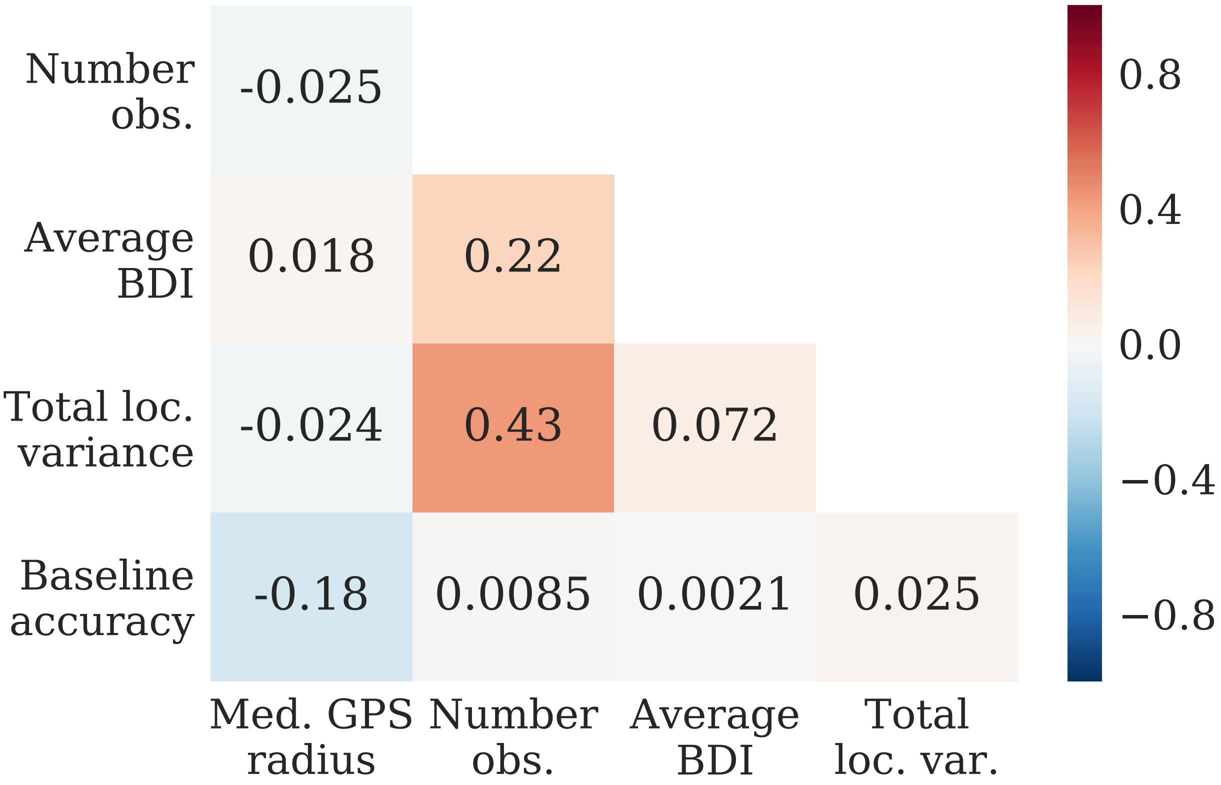

Figure 5.3: Correlation between user characteristics. "Med." denotes the median and the Beck's Depression Inventory (BDI) is a measure of depressive symptoms. In general, there isn't high correlation between user characteristics.

descriptive features for daily wellbeing prediction. Such features could include those from other sensors, such as accelerometer activity. As a result of including more daily features, we would also like to explore different user characteristics that describe the behaviors measured by other sensors. For example, when exploring the benefit of using accelerometer activity measures to predict daily wellbeing, we would like to address if the model improvement is related to the user's general activeness and variability. With sufficiently descriptive daily features that generate better models, we could also explore the relationship of user characteristics with predicting daily mood in addition to daily energy.

This preliminary work is a case study in trying to understand model discrepancies for wellbeing prediction, a problem that has been generating optimism for medical applications. Larger studies with multi-modal prediction approaches are still needed to improve monitoring accuracy. However, these studies may consider including an analysis, such as we have presented, to understand for which individuals such a monitoring approach (i.e., with a smartphone) is plausible and for whom it is unrealistic. Smartphone monitoring may be attractive for its ease of use, but it is imperative to have accurate monitoring for individuals suffering from mental health disorders. Understanding when smartphones are unable to monitor individuals, as we have attempted to do, may eventually help achieve such necessary reliability.

Chapter 6

Towards Augmenting Crisis Counselor Training by Improving Message Retrieval

Originally published as Demasi, et al. "Towards augmenting crisis counselor training by improving message retrieval." Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (2019).

A fundamental challenge when training counselors is presenting novices with the opportunity to practice counseling distressed individuals without exacerbating a situation. Rather than replacing human empathy with an automated counselor, we propose simulating an individual in crisis so that human counselors in training can practice crisis counseling in a low-risk environment. Towards this end, we collect a dataset of suicide prevention counselor role-play transcripts and make initial steps towards constructing a CRISISbot for humans to counsel while in training. In this data-constrained setting, we evaluate the potential for message retrieval to construct a coherent chat agent in light of recent advances with text embedding methods. Our results show that embeddings can considerably improve retrieval approaches to make them competitive with generative models. By coherently retrieving messages, we can help counselors practice chatting in a low-risk environment.

6.1 Introduction

Suicide prevention hotlines can provide immediate care in critical times of need [62, 61, 127]. These hotlines are expanding services to text to meet growing demands and adapt to shifts in communication trends [140]. Crisis helplines rely on counselors who are trained in a variety of skills, such as empathy, active listening, assessing risk of suicide, de-escalation, and connecting individuals to longer term solutions [61, 111].

Properly training counselors is critical yet difficult as, resource costs aside, counselors need to practice and develop expertise in realistic environments that are low-risk, i.e., they do not put distressed individuals in danger. Because novice counselors are unable to assume full responsibility for a crisis situation until they have some experience, training often includes human-to-human role-playing [4, 146]. Role-playing has been shown to improve crisis intervention training [40]. However, such training takes a lot of human time, which centers struggle to provide.

Instead of attempting to scale services by replacing human counselors and trying to automate the generation of empathetic responses, we seek to build a training tool that can augment hotline training and empower more counselors. As a first component, we develop a chat interface where novices can practice formulating responses by interacting with a simulated distressed individual.

To build such a system, we collect synthetic role-play transcripts that provide example scenarios and example messages, while protecting the identity of any individual contacting a crisis hotline. Here, we consider the one-sided case of simulating the individual in distress with the intention of eventually providing a training environment for novice counselors to practice counseling without putting anyone in danger.

In the application we consider, and in many similarly data-constrained applications, language generation methods may be challenged by the limited data that can initially be collected. To surmount this issue, we explore the extent to which retrieval methods can be improved to provide an engaging chat experience. More specifically, we consider whether improved embedding methods, which enable better representation of text, improve retrieval models through better comparisons of text similarity. Briefly stated, we ask two research questions:

RQ1 Do improved embedding methods retrieve coherent responses to a single turn of context more often than commonly-used TF-IDF or generative models?

RQ2 Can we extend retrieval baseline models to consider more than one turn of context when selecting a response?

Our results show that recent developments in embedding methods have considerably improved dialogue retrieval, which is promising for the use of these methods in data-limited applications. We also find that extending retrieval to consider additional messages of context does improve baselines. This indicates the potential for retrieval methods to benefit data-limited dialogue systems and the need to re-evaluate baselines for generative models. Within the setting that we study, our results provide promise for building a chat module that can enable crisis counselors to practice before interacting with individuals in need.

6.2 Related Work

Considerable potential for automating a counselor was shown with the initial rule-based Eliza system [163] and recent developments have sought to target systems for delivering cognitive behavioral therapy [55]. Other studies have looked at the effect of suicide prevention counselor training [61], identifying patterns of successful crisis hotline counselors [3], automating counselor evaluation [115], and building a dashboard for crisis counselors [51]. There is additional work to identify supportive and distressed behaviors and language in online forums [9, 42, 161] and support forum moderators [68]. Most similar to our study, was one study that showed the potential for an avatar system to help train medical doctors to deliver news to patients [5]. However, this study did not target counselors or train conversation strategies. To our knowledge, there has been no work on automating the individual seeking help to improve counselor training.

Text Retrieval for Dialogue Systems

Previous systems have explored the use of retrieving messages from related contexts for continuing dialogue. Some studies have looked at defining or learning scoring functions over IDF weights to construct retrieval scores [80, 130]. Most similar to our work is a system that considered similarities of full histories of dialogues in addition to a previous turn of context [10] and another study that hand-tuned weights in a scoring function on IDF weights to include additional messages of context [142]. However, these works used similarities calculated over TF-IDF [8] and bag-of-words of representations, instead of more recent embedding methods [21, 37, 114, 116, 145], which we explore.

6.3 Dataset

We collected a dataset of synthetic chat transcripts between suicide prevention counselors and hotline visitors. An example of such a conversation is shown in Figure 6.1 and additional examples are discussed in the Results section. Artificial or role-play transcripts were generated by trained counselors in order to protect the identity of any individuals who may contact crisis hotlines. We chose this approach because retrieval should not be used on datasets consisting of real conversations. Such datasets have been explored in prior work to understand effective hotline conversations [3].

Role-playing between experienced and novice counselors is a common tool for crisis counselor training, and is a task counselors are often exposed to before being approved to work on a hotline [4, 72]. In addition to expecting role-playing to be a natural task for hotline counselors, prior work on short, unstructured social dialogues between peers found that self-dialogues, i.e., where an individual would produce both sides of a two-person dialogue, generated high quality and creative example conversations [80]. We followed this work and asked experienced counselors to self-role-play scenarios of a counselor working with a hotline visitor. We collected transcripts in three phases: full role-plays, visitor-only role-plays, and counselor-paraphrase role-plays.

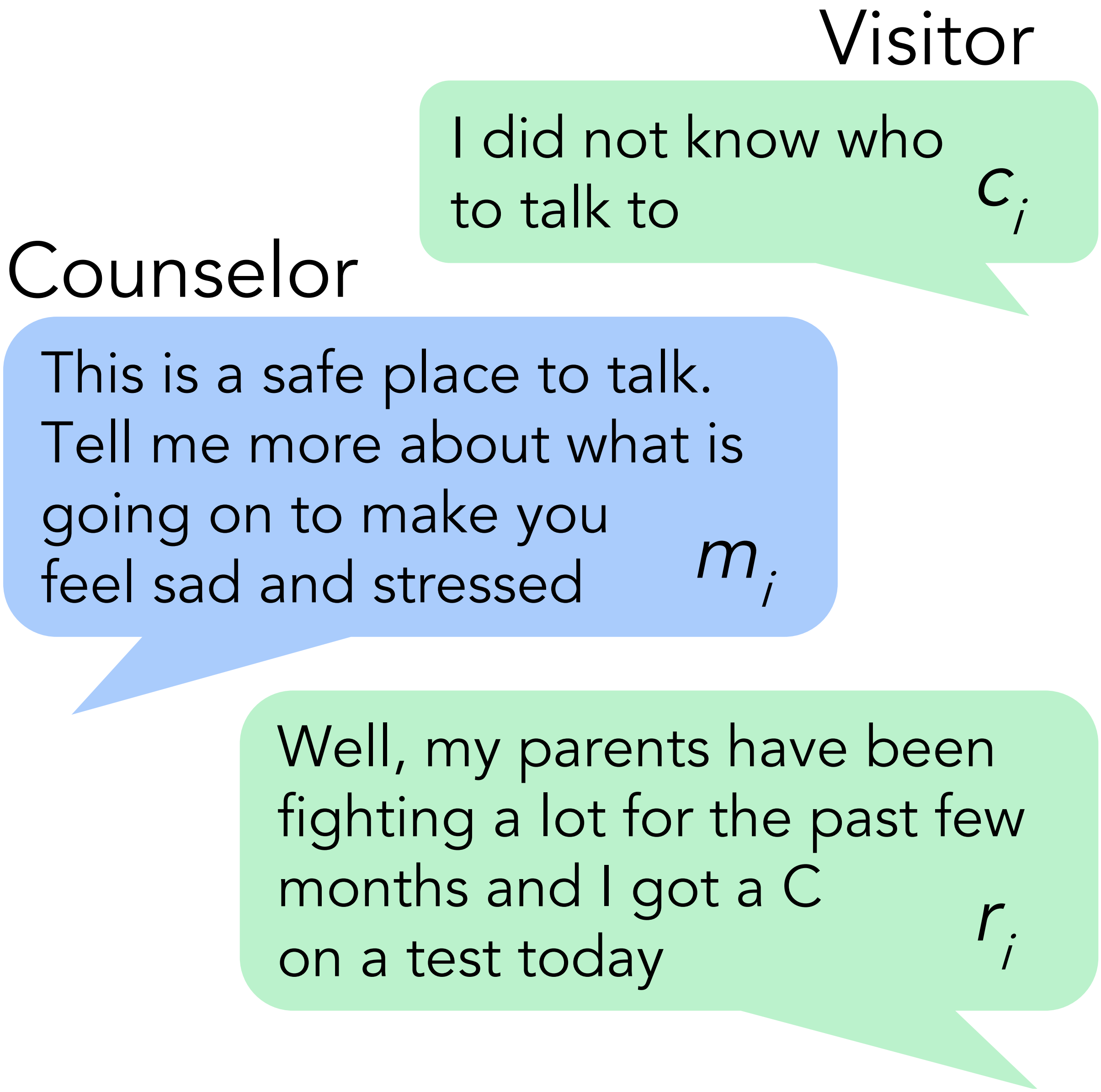
Visitor
I did not know who to talk to
c_i
Counselor
This is a safe place to talk. Tell me more about what is going on to make you feel sad and stressed
m_i
Well, my parents have been fighting a lot for the past few months and I got a C on a test today
r_i

Figure 6.1: A conversation snippet showing a visitor's response r_i to a counselor's message m_i with preceding context, i.e., a visitor's message c_i.

	Phase	**Count**
Unique conversations	1	254
Visitor-only role-plays	2	182
Counselor-only role-plays	3	118
Visitor messages	1-2	9062
Counselor messages	2	5320
Counselor paraphrases	3	2999

Table 6.1: Statistics on role-play transcripts. Phase indicates the study phase during which each set of data was collected. Each counselor paraphrase reworded a single counselor message.

Collection

After consenting to participate in the study, counselors were invited to the first of three phases. In the first phase, counselors were asked to role-play both sides of a potential crisis text conversation. To be representative of common demographic of individuals who contact a helpline over text, counselors were prompted to role-play a youth experiencing trouble in school and with their parents. This persona was chosen to represent a common scenario that a counselor may encounter in a text-based conversation. The counselors were able to decide if the fictional youth was experiencing suicidal thoughts, specific issues they were having, and if they felt better by the end of the conversation. Transcripts were required to be 20 turns for each counselor and visitor (40 turns total). However, participants were able to extend the conversation to at most 60 turns total, if they chose. Messages were unconstrained in length, but it was suggested that they resemble SMS messages.

Counselors who participated in a second phase of the study were given the counselor's side of a transcript generated in the first phase of the study and asked to role-play only the youth experiencing trouble in a way that fit with the counselor's messages. Participants in the third phase of the study were given a full transcript generated in the first phase and asked to generate counselor paraphrases that reworded and possibly improved the original counselor messages. The second and third phases were designed to increase the variety of responses that might be made.

Additional data were collected for evaluating models, as will be discussed below. All study methods were approved by the university's Internal Review Board.

Dataset Statistics

In total, 32 crisis counselors participated in the study and wrote example messages. In general, the transcripts represent a broad range of scenarios. Statistics on the resulting dataset are in Table 6.1. In the following results, we do not include messages generated in the second phase of the study.

6.4 Methods

After preprocessing, we consider two tasks: how to return a visitor response to a single input counselor message and how to return a visitor response when considering a counselor input message and preceding conversation context. For responding to a single counselor input message, we consider two approaches: one based on cosine similarity of vector representations and the based other on likelihood. For responding to a counselor message when considering additional conversation context, we extend retrieval to consider additional messages of context, i.e., an additional message preceding the counselor's last message. For generating responses, we consider a popular Seq2Seq model [147, 156] and a hierarchical neural model [110].

Data Preprocessing

Names were standardized to be popular American male or female baby names from the last 5 decades. Entire messages were tokenized with appropriate tokenizers for each embedding method and converted to lowercase, as appropriate.

Response Retrieval Considering a Single Message

For the first retrieval approach we consider, let a message input to the system be m_i. Let M_N and R_N be all the N messages and responses, respectively, in the training set and m_j and r_j indicate individual messages and responses in the training set. The first method considers all the messages in the training set and returns the response $r_{j'}$ to the message $m_{j'}$ that shares the highest cosine similarity with the input message, i.e., $j' = \arg\max_j \text{sim}(m_i, m_j)$ where j indexes over the messages in the training set.

Similarity is commonly calculated as cosine similarity between TF-IDF vector representations of the input (i.e., counselor) message m_i and messages in the training set. We compare the TF-IDF representation with additional vector representations of the counselor input. Exhaustive comparison of embedding methods is not feasible, so we chose popular, successful, and diverse embeddings: GloVe [114], FastText [21], Attract-Repel [157], and ELMo [116, 59]. We also consider two sentence embeddings: InferSent [37] and GenSen [145]. Messages are embedded by summing the embeddings of their elements, e.g., across words or sentences for appropriate embeddings.

For the second retrieval approach, we select the response from the training data that is most probable, i.e, $j' = \arg\max_j P(r_j|m_i)$ where m_i is again the input message and j indexes over training examples. With this approach, which we will refer to as *S2S-retrieve*, the probability of a response is calculated by a Seq2Seq model trained on counselor-visitor message-response pairs. All Seq2Seq models were trained in the OpenNMT framework [77].

Response Retrieval Considering More than One Message of Context

When multiple messages of context are present, we propose including the additional context in the retrieval methods in three ways. For this work, we consider only one message in the conversation that precedes the counselor's input message to be additional context, as indicated in Figure 6.1.

First, we consider the response from the training data $r_{j'}$ that has the highest similarity calculated over the sum of the previous messages embeddings, i.e., considering contexts c_i and c_j that precede a test message m_i and a training message m_j respectively, we choose $r_{j'}$ such that $j' = \arg\max_j \text{sim}(m_i + c_i, m_j + c_j)$.

As a second approach, we measure context similarity as the weighted sum of context and message similarities: $j' = \arg\max_j \text{sim}(m_i, m_j) + \lambda \text{sim}(c_i, c_j)$. The weight parameter λ is found via cross-validation to optimize the similarity of embedded responses returned with true responses on a development set.

Third, for the likelihood based model, we again consider the response from the training set that returns the highest likelihood, as calculated by a Seq2Seq model. To include an additional context message, we concatenate preceding messages before encoding and decoding.

Response Generation

For generating a response to a single counselor message, we consider a Seq2Seq model [147].

When considering an additional message of context, we first use the Seq2Seq model with the preceding messages concatenated into a single input. Second, we use a Variational Hierarchical Conversation RNN (VHCR) that explicitly models prior conversation state with a hierarchical structure of latent variables [110]. This model has been shown to improve on other models that adjust for context when there is more than one preceding utterance [110]. Seq2Seq and VHCR model embeddings are initialized with GloVe vectors [114].

6.5 Experiments

For the two response selection tasks, we randomly separated transcripts into training, development, and test sets, with the training set accounting for 80% of the conversations and the rest evenly distributed between development and test sets. Counselor paraphrases were assigned to the set that their original message was assigned to. Messages were not randomly shuffled, but separated by conversation, to avoid training on data related to the test data. For both research questions, a response was either generated from a model trained on the training set or retrieved from the bank of training examples for every counselor message or paraphrased counselor message in the test set.

	Method	**Unit of embedding**	**Selection metric**	**Percent that made sense**	**Avg. tokens in response**	**Avg. tokens in MS**
	Random	–	–	25.30	15.1	12.6
retrieval	TF-IDF	word	cos-sim	60.34	13.1	12.4
	Attract-Repel	word	cos-sim	58.50	18.3	**16.2**
	ELMo	word	cos-sim	65.88	14.5	14.0
	FastText	word	cos-sim	62.71	**16.2**	15.5
	GloVe	word	cos-sim	58.63	15.9	15.1
	GenSen	sentence	cos-sim	64.16	14.5	14.2
	InferSent	sentence	cos-sim	61.79	14.9	14.0
	S2S-retrieve	–	likelihood	**67.46**	8.8	8.2
gen.	S2S-generate	–	–	64.16	11.7	10.8
	Ground truth	–	–	**89.33**	**14.6**	**14.6**

Table 6.2: Performance of methods used to return a response to a single input message. MS indicates the set of responses that crowdworkers judged as making sense in context, rather than all the responses that the method returned. Both the best performing method and ground truth results are in bold.

Evaluation

To evaluate the overall quality of responses that methods returned, we follow prior work that indicated there is currently no automatic equivalent and used human judges [88]. These judges were crowdworkers on Amazon Mechanical Turk[1] who had been granted Masters status and were located in the United States. Crowdworkers were presented with instructions, labeled examples, and batches of 10 cases where they were asked to judge responses to messages.

To evaluate methods for the first research question, crowdworkers were given a single message and a response and asked to judge the response. For the second research question, crowdworkers were given two messages of context and a highlighted response and asked to judge the response.

[1] `https://www.mturk.com/`

Decision	**Subcategory**	**Count**
Makes sense	Answers the counselor's question(s)	17
	Logical response, fits the conversation	15
	Not perfect, but conceivable someone could respond this way	7
	Agrees/disagrees with counselor's statement	2
Mismatched	Doesn't answer or respond to the question	11
	Messages are unrelated	9
	Doesn't fit, seem right, or make sense	4
	Responses answers a different question	3
	Response is a bad, incoherent message	3
	Message is from a different part of the conversation	2
Unclear	Response is vague or confusing	4
	Worker just didn't know	3
	Can't tell without more context	2
	Explanation of why worker is unsure	1
Other	Researchers were unsure what rationale meant	13
	Description of message content	4

Table 6.3: Themes in crowdworker rationales for why a response made sense or not. The count is the number of rationales out of a subset of 100 pairs that shared the theme.

In contrast to studies that rank on scales [89], we directly asked the workers to decide if a response made sense or not. In addition to indicating that a response did or did not make sense, we allowed a third class for workers to indicate if they were unsure without additional context. We found these classes to be sufficiently descriptive to consistently label messages between researchers. In preliminary trials with crowdworkers, there was insufficient agreement on labels. This instability of labels could stem from a variety of causes, including uncertainty about whether a change of topic should be considered a coherent response. To surmount this ambiguity, we asked two crowdworkers to label each response and a third crowdworker to break any ties. All cases where crowdworkers indicated that they were unsure were considered to be labeled as not coherent. With this voting approach, on a trial set of message and response pairs, crowdworker labels corresponded with researcher determined labels with a Cohen's Kappa of 0.69 [33], indicating considerable agreement.

Performance Metrics

To assess the quality of a method at returning responses, we take messages from a held-out test set and return a response to it by either selecting a message from the training set or generating a response with a model trained on the message and response pairs in the training set. The split into training, development, and test sets is held constant across methods. We ask crowdworkers to judge whether each response makes sense as a possible response to the given message and aggregate multiple crowdworker decisions into a single label for each returned response. We then use the percent of responses returned by a method that were labeled as making sense as an indicator of method performance. The higher percent of messages that made sense as responses, the better the method is at responding coherently. We also consider the number of tokens in each response returned by a method and average the number across all the responses returned as a surrogate for how interesting the responses are. Presumably, longer messages are more interesting than short responses.

Random and Ground Truth Baselines

For the first research question, we included a method that randomly selected responses from the training set to messages in the test set. This method is intended as a baseline for how easy the task was for a method to guess responses.

For both the first and second research questions, we included a method that returned ground truth visitor responses from the test set as an indicator of how hard the task was for humans to determine response quality without additional context.

Assessing Why Responses Are Coherent

To understand how crowdworkers decided if a response was coherent, we asked crowdworkers to evaluate responses on a set of 100 message-response pairs and additionally provide a rationale for their decision. For each of 50 test messages, we made two pairs: one with a response randomly selected from the training messages and the other with the ground truth response from the test set. These two methods where chosen to generate pairs that were not likely and likely to be coherent. We directly asked whether the response was coherent and “Why did you choose that option?” with an open text box for crowdworkers to enter a rationale. We read and grouped the rationales into themes of why responses did or did not make sense.

6.6 Results

We present results on two tasks corresponding to our two research questions: retrieving a response to a counselor’s message and extending retrieval to consider an additional message of context. We also consider rationales for why responses do or do not make sense.

Comparing Retrieval Methods for a Single Message of Context

Retrieval methods showed a clear benefit over randomly selecting responses, i.e., retrieval methods returned a higher percent of coherent messages, as judged by crowdworkers (Table 6.2). ELMo embeddings and three other embeddings (FastText, InferSent, and GenSen) improved on the commonly used TF-IDF retrieval baseline. This suggests that retrieval methods with recent embeddings provide stronger baselines for generative methods and may continue to improve alongside better embedding methods. However, only the ELMo embeddings surpassed the Seq2Seq generative model in percent of coherent messages returned, indicating the potential of generative models even in this data-limited setting.

Retrieving responses with a Seq2Seq likelihood score returned the highest percent of coherent responses, but the shortest messages on average. Short messages may indicate that, while coherent, the messages may be less engaging or interesting for an interactive environment.

When assessing crowdworker rationales for deciding if a response was coherent, we found 16 themes or reasons why crowdworkers would decide whether a response made sense (Table 6.3). Crowdworkers cited more ways that responses did not make sense than ways a response would be coherent. Examples of messages in the test set and the corresponding responses that each method returned for them are shown in Tables 6.4 and 6.5.

Extending Retrieval to Include Additional Messages of Context

Providing crowdworkers with an additional message of context appeared to impact their impression of whether responses made sense in context. When presented with an additional message of context, i.e, one visitor message and one counselor message, crowdworkers found a larger percent of the ground truth responses from the test set to make sense (Table 6.6). In contrast, when provided with an additional message of context to evaluate a response, crowdworkers judged a lower percent of responses returned by the ELMo-based retrieval method to be coherent (61.40%, Table 6.6) than when they were only presented with a single message of context (65.88%, Table 6.2). Incorporating a previous message of context into a similarity score increased the percent of coherent messages returned, but by less than 1%. We only considered the ELMo embeddings, as they were found to perform best in the first research question. Three out of four retrieval methods returned a higher percent of coherent messages than both generative models, indicating that including more context for generative models is challenging. Again using the Seq2Seq likelihood to retrieve responses returned the highest percent of messages that made sense. However, these responses also had the fewest tokens, implying generic, short messages that might score low on a qualitative scale of how engaging an interactive system is.

6.7 Discussion

In contrast to many popular dialogue datasets [139], the transcripts we collected have a relatively high number of turns (minimum 40 total turns per conversation), implying rich conversations. These conversations are also interesting for their unique position of having distinct roles for participants, a counselor and a distressed youth, and related themes. We find retrieval to be a competitive approach with generative models and return responses that make sense for more than 60% of input messages. We also find themes for how responses can seem to be coherent.

Giving crowdworkers an additional message of context to judge whether a response was coherent or not affected their decisions. It appeared that ground truth responses were easier to distinguish as coherent and fewer retrieved messages were judged as coherent if an additional message of context was presented. This indicates the importance of context, especially during evaluation.

The results we present are on a specific, data-limited setting, but the implications of our results may be broader both for other important applications, which commonly have data limitations, and for retrieval baselines that are used to assess generative models. As embeddings have improved, so too have retrieval baselines, which need to be updated for appropriate evaluation of generative models in any language generation setting.

Our results are not without limitations. The data-limited setting presented a challenge to training generative models, and perhaps extensive hyper-parameter tuning could influence results. However, limited data and non-exhaustive parameter tuning are common limitations. Further, as datasets increase in size, so does the potential for relevant, related contexts to be present and thus the potential for successful retrieval increases as well. Thus, even on larger datasets, competitive retrieval models, such as those we have presented, should be considered for baseline comparisons.

Another limitation of our approach is the extent to which we have considered context so far. Because the conversations we collected are long relative to some other datasets it is likely more context will be necessary to produce a coherent simulation. We have begun to methodically look at the effects of incrementally including more context and extending retrieval models beyond a single message. These initial steps indicate the impact context has and provide important baselines for comparing future, more general models.

6.8 Conclusion

Our work shows promise that data-limited applications may build initial systems with retrieval methods powered by recently developed embeddings. By collecting role-play transcripts and showing results in a data-limited context, we have demonstrated the potential to develop a successful simulation of a hotline visitor that novice counselors can practice with during training. We found that retrieval methods became more competitive with improved embedding methods and surpassed generative methods when more context was considered.

We also found that context had impact on how difficult it was for crowdworkers to evaluate responses.

As a next step, we plan to explore better leveraging rich structure in the conversations, with a focus on the protocol that the counselors are trained to follow. There has been increased interest in blending retrieval and generation approaches by modifying prototypes retrieved from training data [86, 164]. It is possible that such an approach would enable modifying and thus tailoring responses to similar contexts.

	Method	Example 1	Example 2	Example 3
Message		**Wow, not even 18 but already paying the bills. How old are you?**	**Sounds like a great plan, take care Jessica... we're here 24/7 in case you need to talk to us again**	**Hi, I'm Michael! What is your name? Location?**
Responses	**Random**	I took a small amount. (N)	Sometimes I have to. They don't really like hanging out with him. (N)	Yes, please. (N)
	Tf-idf	I'm 17. (Y)	Thanks, it's good to talk about this stuff. I have to go eat dinner now (Y)	Hi Michael, this is Christopher. I'm in Daly City and I need help. (Y)
	Attract-Repel	I feel lonely (N)	Maybe. Maybe next time. (N)	I'm Jennifer (Y)
	ELMo	14 (Y)	ok thx (Y)	i'd rather not say my name. is this space confidential (Y)
	FastText	14 (Y)	ok thx (Y)	i'd rather not say my name. is this space confidential (Y)
	GloVe	Im 15, i have a twin but hes a boy. Im a girl. They let him get away with evvvverything at home and school (Y)	Maybe. Maybe next time. (N)	i'd rather not say my name. is this space confidential (Y)

Table 6.4: Examples of three counselor messages and the corresponding visitor response output from various methods explored. Remaining methods in Table 6.5. These examples are from the first research question, where only one preceding counselor message is considered. Whether crowdworkers thought a response made sense or not is indicated parentheses as ”Y” and ”N”, respectively.

	Method	**Example 1**	**Example 2**	**Example 3**
Message		**Wow, not even 18 but already paying the bills. How old are you?**	**Sounds like a great plan, take care Jessica... we're here 24/7 in case you need to talk to us again**	**Hi, I'm Michael! What is your name? Location?**
Responses	**GenSen**	14 (Y)	Okay. I got it. Thank you. You were really helpful. I was feeling really awful but I'm feeling better now. (Y)	i'd rather not say my name. is this space confidential (Y)
	InferSent	14 (Y)	Yes please. I'll be right back. (N)	i'd rather not say my name. is this space confidential (Y)
	S2S-retrieve	I'm 15 and I'm pretty sure. Life is lame and I can't go on like this anymore. (Y)	Maybe (Y)	i'd rather not say my name. is this space confidential (Y)
	S2S-generate	I'm 15 and I'm pretty sure . Life is lame and I can't go on like this anymore . (Y)	Okay, I will message you tomorrow. Thank you (Y)	i'd rather not say my name. is this space confidential (Y)
	Ground truth	Yea it's awkward. Im 17, be 18 in 4mo (Y)	You too (Y)	My name is Christopher and I'm in Golden Gate Park. (Y)

Table 6.5: Examples of three counselor messages and the corresponding visitor response output from methods not included in Table 6.4. These examples are from the first research question, where only one preceding counselor message is considered. Whether crowdworkers thought a response made sense or not is indicated parentheses as ”Y” and ”N”, respectively.

	Method	**Incorporation of additional context**	**Percent that made sense**	**Avg. tokens in response**	**Avg. tokens in MS**
retrieval	ELMo	–	61.40	14.6	13.6
	ELMo-sum	Measure similarity of sum of embedded messages	51.78	**15.6**	**15.2**
	ELMo-weight	Weight similarities of previous messages	61.66	14.9	13.9
	S2S-retrieve	Concatenate context	**65.48**	5.5	4.6
gen.	S2S-generate	Concatenate context	58.89	8.3	7.3
	VHCR-generate	Models conversation	55.07	10.8	8.4
	Ground truth	–	**91.30**	**14.6**	**14.7**

Table 6.6: Performance of methods used to retrieve or generate responses when an additional message of context is considered, i.e., two total messages. MS denotes only responses that were considered to make sense in context. Both the best performing method and ground truth results are in bold.

Chapter 7

Conclusions

In this dissertation, I explored two approaches that could use data-driven methods to empower mental health professionals to provide better care during times of need. First, I looked at using smartphone-sensed features to monitor wellbeing or detect changes in wellbeing over time. Such approaches, if successful, would enable professionals to identify and intervene during times of need. Second, I considered text retrieval approaches for developing a baseline chat system that counselors could practice intervening before they interact with humans in need. Throughout this work, I strove for rigorous evaluation to assess whether or when data-driven tools were promising.

7.1 Contributions

Towards using mobile sensing for monitoring wellbeing, I developed a framework for evaluating longitudinal predictions. I then, in collaboration, collected a dataset on a cohort of students, used our framework to evaluate the potential for a set of features to predict wellbeing, and questioned when mobile sensing might be possible. I found that some smartphone-sensed features have correlations with wellbeing, but the relationships were not strong enough to generate reliable predictions across a cohort. I also found that personality was more strongly correlated with whether students would experience changes in depression measures over the course of a semester than the smartphone-sensed features we explored. These results are in line with results in related work, which also questions the extent to which mobile phones can be used to sense wellbeing [7, 121].

Towards a training system for crisis counselors to practice counseling, I collected a dataset of synthetic chat transcripts and explored the use of recent text embedding methods to improve text retrieval and build a baseline system. I compared these retrieval approaches with popular text generation methods. I found that, in general, messages could be retrieved or generated to make sense in context with similar reliability. However, retrieving messages provided much longer messages, which were more similar in length, on average, to the expected average length of a message found in the transcripts. This indicates that, while currently retrieval isn't sufficiently reliable at returning coherent responses, retrieval may play a role in an eventual system. In particular, retrieval and generation may be combined to tailor responses to context and thus make a more reliable and engaging system. I also found that adding another message of context when evaluating retrieved or generated responses affected the ability for crowdworkers to distinguish whether a response made sense in context. This result is important in the implications it has for how prototype systems can be effectively evaluated. While evaluating system responses with shorter contexts may be more convenient, due to lower reading requirements, evaluating responses within longer contexts or through interactively chatting with a prototype system may be necessary.

7.2 Future Work

While this dissertation explored the use of data-driven methods for developing two tools, significant work remains towards evaluating and possibly developing such tools.

First, this work explored wellbeing prediction from smartphone data on a relatively small and non-clinical cohort. It is possible that, with a larger cohort of individuals, clusters of people with similar behaviors could be found and then data on multiple individuals within a cluster could be leveraged to improve predictions of individuals' wellbeing. In the cohort I studied here, there were insufficiently many individuals to separate behavior clusters. However, a challenge for considering clusters of behavior would be developing algorithms which incorporated information on cluster affiliation into their consideration of training examples, e.g., down-weighting examples from individuals in the data collection cohort who didn't have similar behavior when predicting another individual's wellbeing.

Further, the individuals who participated in this study, were not clinically-diagnosed with mental health disorders, to our knowledge. It is possible that individuals with disorders may have more varied behavior that could be predicted from easily-collected data sources, such as smartphone sensor data. If such a population of individuals were recruited for a study, more significant changes in wellbeing could be explored, as some studies have attempted, also on limited size populations [24, 63, 121]. However, a challenge with trying to detect more significant variations in behavior is that such variations may only be present sporadically in a dataset or over considerable periods of time. For example, if trying to detect psychotic episodes or relapses in wellness, occurrences of such events may be few in a feasibly collected dataset.

To develop an engaging chat agent for training crisis counselors, more work is needed to leverage the structure of counseling protocol. For example, counselors must progress through stages and have a variety of techniques that they can draw on when counseling. More work to model these stages and techniques, and work to model dialogue acts more generally, could be used to better retrieve, modify, or generate responses.

Such work to classify or model components of the conversation could also lead towards building suggestion and feedback modules that would guide novice counselors towards improvement while they chat with a simulation. Collaboration with current counselor trainers, combined with automatic measures of conversation state, could be used to develop possible feedback mechanisms that could then be tested through user studies.

An additional remaining challenge is to construct coherent stories around a crisis that the simulation would present. A coherent story that spans multiple messages in a conversation could draw on recent work looking at maintaining personas over the course of a chat session [167]. One approach for this challenge may be to identify when messages relate to an individual crisis situation, e.g., introduce information to the conversation, and when messages are more general responses or necessary conversation acts. If identified, crisis-specific messages could be leveraged from the collected corpus of transcripts to produce vignettes that a retrieval method could select from and then a generative method could modify retrieved messages to ensure coherence within context.

In addition to the feedback modules mentioned above, more work is also needed to holistically approach a training system for counselors. While a chat simulation could be engaging for counselors with some training and some notion of desired counseling structure, it may be too challenging of an educational task for counselors in preliminary stages of training. As such, there is considerable opportunity to explore preliminary tasks that expose counselors early in training to conversation techniques and protocol without yet asking them to hold a full counseling conversation. For example, a system with tasks that ask counselors to identify a technique that is present in a prototype message could be useful early in training and later to also assess the educational value of a more sophisticated chat simulation system.

7.3 Final Remarks

In this dissertation, I have explored methods for data-driven tools to empower mental health professionals. While data-driven methods can provide enormous advantages in some applications, on the problems I have explored, they have faced some challenges. This result is not surprising given that humans exhibit remarkable variability in behavior and experience, and thus methods that look for statistically likely results may not always be successful. However, many opportunities to improve mental health care remain and data-driven methods can provide creative solutions worth exploring. Evaluation is critical for methods in these applications that seek to impact humans. As such, evaluation of methods should be critically approached and evaluation approaches may need to be tailored to the setting to ensure accuracy and reliability of novel tools.

Bibliography

[1] Adrian Aguilera, Stephen M Schueller, and Yan Leykin. “Daily mood ratings via text message as a proxy for clinic based depression assessment”. In: *Journal of affective disorders* 175 (2015), pp. 471–474.

[2] Nadav Aharony et al. “Social fMRI: Investigating and shaping social mechanisms in the real world”. In: *Pervasive and Mobile Computing* 7.6 (2011), pp. 643–659.

[3] Tim Althoff, Kevin Clark, and Jure Leskovec. “Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health”. In: *Transactions of the Association for Computational Linguistics* (2016).

[4] American Association of Suicidology. *Organization Accreditation Standards Manual.* 2012. URL: `https://www.suicidology.org/LinkClick.aspx?fileticket=Nb6eC6r2HX0%3D&portalid=14` (visited on 12/06/2018).

[5] Allen D Andrade et al. “Avatar-mediated training in the delivery of bad news in a virtual world”. In: *Journal of palliative medicine* 13.12 (2010), pp. 1415–1419.

[6] David Arthur and Sergei Vassilvitskii. “k-means++: The advantages of careful seeding”. In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms.* Society for Industrial and Applied Mathematics. 2007, pp. 1027–1035.

[7] Joost Asselbergs et al. “Mobile Phone-Based Unobtrusive Ecological Momentary Assessment of Day-to-Day Mood: An Explorative Study”. In: *Journal of Medical Internet research* 18.3 (2016).

[8] Ricardo Baeza-Yates, Berthier de Araújo Neto Ribeiro, et al. *Modern information retrieval.* New York: ACM Press; Harlow, England: Addison-Wesley, 2011.

[9] Sairam Balani and Munmun De Choudhury. “Detecting and characterizing mental health related self-disclosure in social media”. In: *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems.* ACM. 2015, pp. 1373–1378.

[10] Rafael E Banchs and Haizhou Li. “IRIS: a chat-oriented dialogue system based on the vector space model”. In: *Proceedings of the ACL 2012 System Demonstrations.* Association for Computational Linguistics. 2012, pp. 37–42.

[11] Ling Bao and Stephen S Intille. “Activity recognition from user-annotated acceleration data”. In: *International Conference on Pervasive Computing*. Springer. 2004, pp. 1–17.

[12] Dale J Barr et al. “Random effects structure for confirmatory hypothesis testing: Keep it maximal”. In: *Journal of memory and language* 68.3 (2013), pp. 255–278.

[13] Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis. “Support vector regression”. In: *Neural Information Processing-Letters and Reviews* 11.10 (2007), pp. 203–224.

[14] Douglas Bates et al. “lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-7)”. In: *Retrievable from: https://cran. r-project. org/web/packages/lme4/index. html* (2014).

[15] Aaron T Beck, C Ward, M Mendelson, et al. “Beck depression inventory (BDI)”. In: *Arch Gen Psychiatry* 4.6 (1961), pp. 561–571.

[16] Dror Ben-Zeev et al. “CrossCheck: Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse.” In: *Psychiatric rehabilitation journal* 40.3 (2017), p. 266.

[17] Dror Ben-Zeev et al. “Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health.” In: *Psychiatric rehabilitation journal* 38.3 (2015), p. 218.

[18] 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.

[19] Andrey Bogomolov et al. “Daily stress recognition from mobile phone data, weather conditions and individual traits”. In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM. 2014, pp. 477–486.

[20] Andrey Bogomolov et al. “Pervasive stress recognition for sustainable living”. In: *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*. IEEE. 2014, pp. 345–350.

[21] Piotr Bojanowski et al. “Enriching Word Vectors with Subword Information”. In: *arXiv preprint arXiv:1607.04606* (2016).

[22] Edward J Boyko. “Ruling Out or Ruling In Disease with the Most sensitiue or Specific Diagnostic Test Short Cut or Wrong Turn?” In: *Medical Decision Making* 14.2 (1994), pp. 175–179.

[23] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[24] Benjamin Buck et al. “Relationships between smartphone social behavior and relapse in schizophrenia: A preliminary report”. In: *Schizophrenia Research* (2019).

[25] Michelle Nicole Burns et al. “Harnessing context sensing to develop a mobile intervention for depression”. In: *Journal of Medical Internet Research* 13.3 (2011), e55.

[26] Richard Byrne, Parisa Eslambolchilar, and Andrew 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.

[27] Rafael A Calvo et al. "Computing in mental health". In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM. 2016, pp. 3438–3445.

[28] Luca Canzian and Mirco 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.

[29] Erin A Carroll et al. "Food and mood: Just-in-time support for emotional eating". In: *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE. 2013, pp. 252–257.

[30] Dianne L Chambless and Steven D Hollon. "Defining empirically supported therapies." In: *Journal of consulting and clinical psychology* 66.1 (1998), p. 7.

[31] Stevie Chancellor et al. "Quantifying and predicting mental illness severity in online pro-eating disorder communities". In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM. 2016, pp. 1171–1184.

[32] Zhenyu Chen et al. "Unobtrusive sleep monitoring using smartphones". In: *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*. IEEE. 2013, pp. 145–152.

[33] Jacob Cohen. "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit." In: *Psychological bulletin* 70.4 (1968), p. 213.

[34] Francis S Collins and Harold Varmus. "A new initiative on precision medicine". In: *New England Journal of Medicine* 372.9 (2015), pp. 793–795.

[35] Pamela Y Collins et al. "Grand challenges in global mental health". In: *Nature* 475.7354 (2011), p. 27.

[36] Commonwealth Fund Commission et al. *Why not the best? Results from the national scorecard on US health system performance, 2008*. 2008.

[37] Alexis Conneau et al. "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Sept. 2017.

[38] Glen Coppersmith, Mark Dredze, and Craig Harman. "Quantifying mental health signals in Twitter". In: *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*. 2014, pp. 51–60.

[39] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.

[40] Wendi F Cross et al. “Does practice make perfect? A randomized control trial of behavioral rehearsal on suicide prevention gatekeeper skills”. In: *The journal of primary prevention* 32.3-4 (2011), p. 195.

[41] Alison M Darcy, Alan K Louie, and Laura Weiss Roberts. “Machine learning and the profession of medicine”. In: *Jama* 315.6 (2016), pp. 551–552.

[42] Munmun De Choudhury and Sushovan De. “Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity.” In: *ICWSM*. 2014.

[43] Munmun De Choudhury and Sushovan De. “Mental health discourse on reddit: Self-disclosure, social support, and anonymity”. In: *Eighth International AAAI Conference on Weblogs and Social Media*. 2014.

[44] Munmun De Choudhury et al. “Characterizing and predicting postpartum depression from shared facebook data”. In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM. 2014, pp. 626–638.

[45] Munmun De Choudhury et al. “Discovering shifts to suicidal ideation from mental health content in social media”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM. 2016, pp. 2098–2110.

[46] Munmun De Choudhury et al. “Predicting depression via social media”. In: *Seventh international AAAI conference on weblogs and social media*. 2013.

[47] Marcos DelPozo-Banos et al. “Using Neural Networks with Routine Health Records to Identify Suicide Risk: Feasibility Study”. In: *JMIR Mental Health* 5.2 (2018), e10144.

[48] Orianna DeMasi, Konrad Kording, and Benjamin Recht. “Meaningless comparisons lead to false optimism in medical machine learning”. In: *PloS one* 12.9 (2017), e0184604.

[49] Yong Deng et al. “Evaluating feature selection for stress identification”. In: *Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on*. IEEE. 2012, pp. 584–591.

[50] F Dimeo et al. “Benefits from aerobic exercise in patients with major depression: a pilot study”. In: *British journal of sports medicine* 35.2 (2001), pp. 114–117.

[51] Karthik Dinakar et al. “Mixed-initiative real-time topic modeling & visualization for crisis counseling”. In: *Proceedings of the 20th international conference on intelligent user interfaces*. ACM. 2015, pp. 417–426.

[52] Nathan Eagle and Alex Pentland. “Reality mining: sensing complex social systems”. In: *Personal and ubiquitous computing* 10.4 (2006), pp. 255–268.

[53] Kelly R Evenson, Michelle M Goto, and Robert D Furberg. “Systematic review of the validity and reliability of consumer-wearable activity trackers”. In: *International Journal of Behavioral Nutrition and Physical Activity* 12.1 (2015), p. 159.

[54] Tom Fawcett. “An introduction to ROC analysis”. In: *Pattern recognition letters* 27.8 (2006), pp. 861–874.

[55] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial". In: *JMIR mental health* 4.2 (2017).

[56] Martin Ford. *Rise of the Robots: Technology and the Threat of a Jobless Future*. Basic Books, 2015.

[57] Kenneth R Fox. "The influence of physical activity on mental well-being". In: *Public health nutrition* 2.3a (1999), pp. 411–418.

[58] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. Springer series in statistics Springer, Berlin, 2001.

[59] Matt Gardner et al. "AllenNLP: A Deep Semantic Natural Language Processing Platform". In: *ACL workshop for NLP Open Source Software*. 2018.

[60] George Gkotsis et al. "The language of mental health problems in social media". In: *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*. 2016, pp. 63–73.

[61] Madelyn S Gould et al. "Impact of applied suicide intervention skills training on the national suicide prevention lifeline". In: *Suicide and Life-Threatening Behavior* 43.6 (2013), pp. 676–691.

[62] Madelyn S Gould et al. "National suicide prevention lifeline: enhancing mental health care for suicidal individuals and other people in crisis". In: *Suicide and Life-Threatening Behavior* 42.1 (2012), pp. 22–35.

[63] Agnes Gruenerbl et al. "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.

[64] Sharath Chandra Guntuku et al. "Detecting depression and mental illness on social media: an integrative review". In: *Current Opinion in Behavioral Sciences* 18 (2017), pp. 43–49.

[65] Haibo He and Edwardo A Garcia. "Learning from imbalanced data". In: *IEEE Transactions on knowledge and data engineering* 21.9 (2009), pp. 1263–1284.

[66] Jennifer Healey et al. "Out of the lab and into the fray: towards modeling emotion in everyday life". In: *Pervasive computing*. Springer, 2010, pp. 156–173.

[67] Javier Hernandez, Rob R Morris, and Rosalind W Picard. "Call center stress recognition with person-specific models". In: *International Conference on Affective Computing and Intelligent Interaction*. Springer. 2011, pp. 125–134.

[68] M Sazzad Hussain et al. "Moderator assistant: A natural language generation-based intervention to support mental health via social media". In: *Journal of Technology in Human Services* 33.4 (2015), pp. 304–329.

[69] Ozlem Durmaz Incel, Mustafa Kose, and Cem Ersoy. "A review and taxonomy of activity recognition on mobile phones". In: *BioNanoScience* 3.2 (2013), pp. 145–171.

[70] Natasha Jaques et al. "Multi-task, Multi-Kernel Learning for Estimating Individual Wellbeing". In: *Proc. NIPS Workshop on Multimodal Machine Learning, Montreal, Quebec*. 2015.

[71] Natasha Jaques et al. "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.

[72] John Kalafat et al. "An evaluation of crisis hotline outcomes. Part 1: Nonsuicidal crisis callers". In: *Suicide and Life-threatening behavior* 37.3 (2007), pp. 322–337.

[73] Sylvia Deidre Kauer et al. "Self-monitoring using mobile phones in the early stages of adolescent depression: randomized controlled trial". In: *Journal of medical Internet research* 14.3 (2012), e67.

[74] Alan E Kazdin. "Evidence-based treatment and practice: new opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care." In: *American psychologist* 63.3 (2008), p. 146.

[75] Oscar Kempthorne and TE Doerfler. "The behaviour of some significance tests under experimental randomization". In: *Biometrika* 56.2 (1969), pp. 231–248.

[76] RC Kessler et al. "Predicting suicides after outpatient mental health visits in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS)". In: *Molecular Psychiatry* (2016).

[77] Guillaume Klein et al. "OpenNMT: Open-Source Toolkit for Neural Machine Translation". In: *Proceedings of ACL*. 2017. DOI: `10.18653/v1/P17-4012`. URL: `https://doi.org/10.18653/v1/P17-4012`.

[78] Rachel Kornfield et al. "Detecting Recovery Problems Just in Time: Application of Automated Linguistic Analysis and Supervised Machine Learning to an Online Substance Abuse Forum". In: *Journal of medical Internet research* 20.6 (2018), e10136.

[79] William J Korotitsch and Rosemery O Nelson-Gray. "An overview of self-monitoring research in assessment and treatment." In: *Psychological Assessment* 11.4 (1999), p. 415.

[80] Ben Krause et al. "Edina: Building an Open Domain Socialbot with Self-dialogues". In: *arXiv preprint arXiv:1709.09816* (2017).

[81] Alexandra Kuznetsova, Per B Brockhoff, and Rune Haubo Bojesen Christensen. "lmerTest package: tests in linear mixed effects models". In: *Journal of Statistical Software* 82.13 (2017).

[82] Nicholas D Lane et al. "Bewell: Sensing sleep, physical activities and social interactions to promote wellbeing". In: *Mobile Networks and Applications* 19.3 (2014), pp. 345–359.

[83] Oscar D Lara and Miguel A Labrador. "A survey on human activity recognition using wearable sensors". In: *Communications Surveys & Tutorials, IEEE* 15.3 (2013), pp. 1192–1209.

[84] Neal Lathia et al. "Smartphones for large-scale behavior change interventions". In: *IEEE Pervasive Computing* 3 (2013), pp. 66–73.

[85] Jonathan Lester et al. "A hybrid discriminative/generative approach for modeling human activities". In: (2005).

[86] Juncen Li et al. "Delete, Retrieve, Generate: a Simple Approach to Sentiment and Style Transfer". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 2018, pp. 1865–1874.

[87] Robert LiKamWa et al. "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.

[88] Chia-Wei Liu et al. "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 2016, pp. 2122–2132.

[89] Ryan Lowe et al. "Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.* 2017, pp. 1116–1126.

[90] Hong Lu et al. "Stresssense: Detecting stress in unconstrained acoustic environments using smartphones". In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing.* ACM. 2012, pp. 351–360.

[91] Hong Lu et al. "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.

[92] Hannah G Lund et al. "Sleep patterns and predictors of disturbed sleep in a large population of college students". In: *Journal of adolescent health* 46.2 (2010), pp. 124–132.

[93] Yuanchao Ma et al. "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.

[94] Lisa Marzano et al. "The application of mHealth to mental health: opportunities and challenges". In: *The Lancet Psychiatry* 2.10 (2015), pp. 942–948.

[95] Robert R McCrae and Oliver P John. "An introduction to the five-factor model and its applications". In: *Journal of personality* 60.2 (1992), pp. 175–215.

[96] Charlotte M McKercher et al. “Physical activity and depression in young adults”. In: *American journal of preventive medicine* 36.2 (2009), pp. 161–164.

[97] Jun-Ki Min et al. “Toss’n’turn: smartphone as sleep and sleep quality detector”. In: *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM. 2014, pp. 477–486.

[98] David Mizell. “Using Gravity to Estimate Accelerometer Orientation”. In: *Proceedings of the Seventh IEEE International Symposium on Wearable Computers (ISWC’03)*. Vol. 1530. 0811/03, pp. 17–00.

[99] David C Mohr, Mi Zhang, and Stephen M Schueller. “Personal sensing: understanding mental health using ubiquitous sensors and machine learning”. In: *Annual review of clinical psychology* 13 (2017), pp. 23–47.

[100] Robert R Morris et al. “Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions”. In: *Journal of medical Internet research* 20.6 (2018).

[101] Sai T Moturu et al. “Sleep, mood and sociability in a healthy population”. In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2011, pp. 5267–5270.

[102] Sai T Moturu et al. “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.

[103] George E Murphy. “The prediction of suicide: why is it so difficult?” In: *American Journal of Psychotherapy* (1984).

[104] Miranda Olff. “Mobile mental health: A challenging research agenda”. In: *European journal of psychotraumatology* 6 (2015).

[105] Cathy O’Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2017.

[106] World Health Organization. *Depression [fact sheet no. 369]*. (Accessed April 2017), URL=http://www.who.int/mediacentre/factsheets/fs369/en/. 2017.

[107] World Health Organization et al. *The global burden of disease: 2004 update*. 2008.

[108] Alexandros Pantelopoulos and Nikolaos G Bourbakis. “A survey on wearable sensor-based systems for health monitoring and prognosis”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 40.1 (2010), pp. 1–12.

[109] Pablo Paredes et al. “PopTherapy: Coping with stress through pop-culture”. In: *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering). 2014, pp. 109–117.

[110] Yookoon Park, Jaemin Cho, and Gunhee Kim. “A Hierarchical Latent Structure for Variational Conversation Modeling”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2018, pp. 1792–1801.

[111] Amber Paukert, Brian Stagner, and Kerry Hope. “The assessment of active listening skills in helpline volunteers”. In: *Stress, Trauma, and Crisis* 7.1 (2004), pp. 61–76.

[112] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *Journal of machine learning research* 12.Oct (2011), pp. 2825–2830.

[113] Frank J Penedo and Jason R Dahn. “Exercise and well-being: a review of mental and physical health benefits associated with physical activity”. In: *Current opinion in psychiatry* 18.2 (2005), pp. 189–193.

[114] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. “GloVe: Global Vectors for Word Representation”. In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543. URL: `http://www.aclweb.org/anthology/D14-1162`.

[115] Verónica Pérez-Rosas et al. “Predicting Counselor Behaviors in Motivational Interviewing Encounters”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Vol. 1. 2017, pp. 1128–1137.

[116] Matthew E. Peters et al. “Deep contextualized word representations”. In: *Proc. of NAACL*. 2018.

[117] June J Pilcher, Douglas R Ginter, and Brigitte Sadowsky. “Sleep quality versus sleep quantity: relationships between sleep and measures of health, well-being and sleepiness in college students”. In: *Journal of psychosomatic research* 42.6 (1997), pp. 583–596.

[118] June J Pilcher and Elizabeth S Ott. “The Relationships Between Sleep and Measures of Health and Weil-Being in College Students: A Repeated Measures Approach”. In: *Behavioral Medicine* 23.4 (1998), pp. 170–178.

[119] Lukasz Piwek et al. “The rise of consumer health wearables: promises and barriers”. In: *PLoS Medicine* 13.2 (2016), e1001953.

[120] David Martin Powers. “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation”. In: (2011).

[121] Abhishek Pratap et al. “The accuracy of passive phone sensors in predicting daily mood”. In: *Depression and anxiety* 36.1 (2019), pp. 72–81.

[122] I Wayan Pulantara, Bambang Parmanto, and Anne Germain. “Clinical Feasibility of a Just-in-Time Adaptive Intervention App (iREST) as a Behavioral Sleep Treatment in a Military Population: Feasibility Comparative Effectiveness Study”. In: *Journal of medical Internet research* 20.12 (2018).

[123] Mashfiqui Rabbi, Min Hane Aung, and Tanzeem Choudhury. "Towards health recommendation systems: an approach for providing automated personalized health feedback from mobile data". In: *Mobile Health*. Springer, 2017, pp. 519–542.

[124] Mashfiqui Rabbi et al. "Feasibility and Acceptability of Mobile Phone–Based Auto-Personalized Physical Activity Recommendations for Chronic Pain Self-Management: Pilot Study on Adults". In: *Journal of medical Internet research* 20.10 (2018).

[125] Mashfiqui Rabbi et al. "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.

[126] Kiran K Rachuri et al. "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.

[127] Rajeev Ramchand et al. "Characteristics and proximal outcomes of calls made to suicide crisis hotlines in California". In: *Crisis* (2016).

[128] Stephen W Raudenbush and Anthony S Bryk. *Hierarchical linear models: Applications and data analysis methods*. Vol. 1. Sage, 2002.

[129] Andrew G Reece and Christopher M Danforth. "Instagram photos reveal predictive markers of depression". In: *EPJ Data Science* 6.1 (2017), p. 15.

[130] Alan Ritter, Colin Cherry, and William B Dolan. "Data-driven response generation in social media". In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics. 2011, pp. 583–593.

[131] J.A. Russell. "A Circumplex Model of Affect". In: *Journal of Personality and Social Psychology* 39 (1980), pp. 1161–1178.

[132] Sohrab Saeb et al. "Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study". In: *Journal of medical Internet research* 17.7 (2015).

[133] Sohrab Saeb et al. "The need to approximate the use-case in clinical machine learning". In: *Gigascience* 6.5 (2017), gix019.

[134] Sohrab Saeb et al. "The relationship between mobile phone location sensor data and depressive symptom severity". In: *PeerJ* 4 (2016), e2537.

[135] Virginia Sandulescu et al. "Stress detection using wearable physiological sensors". In: *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer. 2015, pp. 526–532.

[136] Akane Sano and Rosalind 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.

[137] Stephen M Schueller, Adrian Aguilera, and David C Mohr. "Ecological momentary interventions for depression and anxiety". In: *Depression and anxiety* 34.6 (2017), pp. 540–545.

[138] Gideon Schwarz et al. "Estimating the dimension of a model". In: *The annals of statistics* 6.2 (1978), pp. 461–464.

[139] Iulian Vlad Serban et al. "A survey of available corpora for building data-driven dialogue systems". In: *arXiv preprint arXiv:1512.05742* (2015).

[140] Aaron Smith and Dana Page. "US smartphone use in 2015". In: *Pew Research Center* 1 (2015).

[141] Mark D Smucker, James Allan, and Ben Carterette. "A comparison of statistical significance tests for information retrieval evaluation". In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management.* ACM. 2007, pp. 623–632.

[142] Alessandro Sordoni et al. "A Neural Network Approach to Context-Sensitive Generation of Conversational Responses". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 2015, pp. 196–205.

[143] Tara W Strine and Daniel P Chapman. "Associations of frequent sleep insufficiency with health-related quality of life and health behaviors". In: *Sleep medicine* 6.1 (2005), pp. 23–27.

[144] Andreas Ströhle. "Physical activity, exercise, depression and anxiety disorders". In: *Journal of neural transmission* 116.6 (2009), pp. 777–784.

[145] Sandeep Subramanian et al. "Learning general purpose distributed sentence representations via large scale multi-task learning". In: *ICLR*. 2018.

[146] Suicide Prevention Resource Center. *Applied Suicide Intervention Skills Training (ASIST).* 2007. (Visited on 12/06/2018).

[147] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems.* 2014, pp. 3104–3112.

[148] Michael J Tanana et al. "Development and evaluation of ClientBot: A patient-like conversational agent to train basic counseling skills". In: ().

[149] R Core Team et al. "R: A language and environment for statistical computing". In: (2013).

[150] Megan Teychenne, Kylie Ball, and Jo Salmon. "Physical activity and likelihood of depression in adults: a review". In: *Preventive medicine* 46.5 (2008), pp. 397–411.

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

[152] Truyen Tran et al. “An integrated framework for suicide risk prediction”. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2013, pp. 1410–1418.

[153] Norifumi Tsuno, Alain Besset, and Karen Ritchie. “Sleep and depression.” In: *The Journal of clinical psychiatry* (2005).

[154] Gaetano Valenza et al. “Mood recognition in bipolar patients through the PSYCHE platform: Preliminary evaluations and perspectives”. In: *Artificial intelligence in medicine* 57.1 (2013), pp. 49–58.

[155] Gaetano Valenza et al. “Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis”. In: *IEEE Journal of Biomedical and Health Informatics* 18.5 (2014), pp. 1625–1635.

[156] Oriol Vinyals and Quoc Le. “A neural conversational model”. In: *arXiv preprint arXiv:1506.05869* (2015).

[157] Ivan Vulić et al. “Morph-fitting: Fine-Tuning Word Vector Spaces with Simple Language-Specific Rules”. In: *Proceedings of ACL*. 2017, pp. 56–68.

[158] Philip S Wang et al. “Twelve-month use of mental health services in the United States: results from the National Comorbidity Survey Replication”. In: *Archives of general psychiatry* 62.6 (2005), pp. 629–640.

[159] Rui Wang et al. “CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia”. In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM. 2016, pp. 886–897.

[160] Rui Wang et al. “Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones”. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM. 2014, pp. 3–14.

[161] Zijian Wang and David Jurgens. “It”s going to be okay: Measuring Access to Support in Online Communities”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018, pp. 33–45.

[162] Gary M Weiss. “Mining with rarity: a unifying framework”. In: *ACM SIGKDD Explorations Newsletter* 6.1 (2004), pp. 7–19.

[163] Joseph Weizenbaum. “ELIZA - a computer program for the study of natural language communication between man and machine”. In: *Communications of the ACM* 9.1 (1966), pp. 36–45.

[164] Jason Weston, Emily Dinan, and Alexander H Miller. “Retrieve and refine: Improved sequence generation models for dialogue”. In: *arXiv preprint arXiv:1808.04776* (2018).

[165] Ian H Witten et al. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[166] Min Wu et al. "Modeling perceived stress via HRV and accelerometer sensor streams". In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2015, pp. 1625–1628.

[167] Saizheng Zhang et al. "Personalizing Dialogue Agents: I have a dog, do you have pets too?" In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 2204–2213.

[168] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320.