Semantic Gap in Predicting Mental Wellbeing through Passive Sensing

Vedant Das Swain
vedantswain@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Victor Chen
vchen36@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Shrija Misra
shmi@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Stephen M. Mattingly
smattin1@nd.edu
University of Notre Dame
Notre Dame, Indiana, USA

Gregory D. Abowd
g.abowd@northeastern.edu
Georgia Institute of Technology
Atlanta, Georgia, USA
Northeastern University
Boston, Massachusetts, USA

Munmun De Choudhury
munmund@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

When modeling passive data to infer individual mental wellbeing, a common source of ground truth is self-reports. But these tend to represent the psychological facet of mental states, which might not align with the physiological facet of that state. Our paper demonstrates that when what people "feel" differs from what people "say they feel", we witness a *semantic gap* that limits predictions. We show that predicting mental wellbeing with passive data (offline sensors or online social media) is related to how the ground-truth is measured (objective arousal or self-report). Features with psychosocial signals (e.g., language) were better at predicting self-reported anxiety and stress. Conversely, features with behavioral signals (e.g., sleep), were better at predicting stressful arousal. Regardless of the source of ground truth, integrating both signals boosted prediction. To reduce the semantic gap, we provide recommendations to evaluate ground truth measures and adopt parsimonious sensing.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; • **Applied computing** → *Law, social and behavioral sciences*; *Psychology*.

KEYWORDS

Passive Sensing, Mental Wellbeing, Social Media, Activity Patterns

ACM Reference Format:
Vedant Das Swain, Victor Chen, Shrija Misra, Stephen M. Mattingly, Gregory D. Abowd, and Munmun De Choudhury. 2022. Semantic Gap in Predicting Mental Wellbeing through Passive Sensing. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3491102.3502037

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '22, April 29-May 5, 2022, New Orleans, LA, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3502037

1 INTRODUCTION

As technology use permeates through society, researchers have abundant opportunities to infer an individual's mental wellbeing in everyday settings by modeling data from different passive sensing sources like smartphones, wearables, and social media [27]. A standard assumption across such studies is that ground truth labels are unquestionable [20, 55]. However, an individual's mental state is complex and self-reported methods to define ground truth only represent the respondent's interpretation [112]. Even for the same mental state, individuals can respond to surveys differently because of self-presentation bias [58] and non-response bias [48]. These nuances are loaded into the ground truth labels but are often ignored by passively sensed data. This lack of information, or abstraction of it, leads to a mismatch between model estimates and the actual mental state of the individual [99]. Some areas of computing refer to the abstraction between a computational model and the variable of interest as the "Semantic Gap" [16]. An individual's mental state can "appear" different despite referring to the same semantic concept [99]. This gap is stark when the signals gleaned from computational data do not coincide with the factors affecting the ground truth. This paper's position is that, for real-world longitudinal studies of mental-wellbeing, the semantic gap limits certain passive sensing models due to the nature of ground truth measures.

Consider a dominant form of mental wellbeing ground truth, self-reports. This method has been widely used in social and ubiquitous computing for predicting anxiety or stress [4, 23, 24, 75, 93]. While self-reports can approximate the psychometric component of stress (e.g., nervousness or apprehension), they do not reflect the physiological one (e.g., increase heart rate) [111]. During the self-report, describing the exact momentary effects of a stressor is more likely to align the psychometric and physiological components. However, a survey response may be influenced by the retrospective psychological effect at the time of reporting [51, 110]. Importantly, self-reports are sensitive to psycho-social factors such as recall bias, impression bias, and self-censorship [48, 58, 97]. An individual's self-report can be disconnected from their behavior because they are uncomfortable disclosing the severity of their

state [48, 112]. These psychosocial influences on self-reports are invisible to typical approaches of passive sensing, which focus on individual physical behaviors, such as activity duration, mobility, and device usage. Despite the same bodily response to watching an intense horror movie compared to being reprimanded by a supervisor; a survey response could report different stress severity for each experience. On the other hand, physiological measures of ground truth might indicate the same severity, but ignore the negativity associated with the experience. Therefore, predictive models trained on an individual's activity data can be limited simply because of a mismatch between the choice of feature representations and the type of ground truth measurement. We believe this represents a semantic gap. We aim to empirically demonstrate that this gap exists in our domain and prescribe approaches to mitigate it.

Other computing areas plagued by the semantic gap teach us that this gap is narrower when the low-level representations (e.g., passively sensed features) and high-level representations (e.g., ground truth values) for the same concept are semantically analogous [99]. This fundamental informs our inquiry. For example, posts on online social media can encapsulate the same psycho-social influences that interfere with self-reports, in terms of self-disclosure [37] and censorship [25]. By contrast, physical behaviors from offline sensing are semantically closer to physiological aspects, such as arousal [76, 83, 86]. Yet, it is challenging to observe the semantic gap in practical deployments because most efforts to predict mental wellbeing, focus on limited sensor streams and a limited set of corresponding ground truth measures. To mitigate this, we investigate the gap by leveraging a unique dataset that includes a variety of ground truth measures for mental wellbeing states and a variety of passively sensed data.

We employ the triangulation method [36] to investigate if this gap actually exists by demonstrating the predictive efficacy of different passive sensing approaches on different measures of ground truth for mental wellbeing. One relies on the offline physical activities sensed from smartphones, wearables, and Bluetooth. The other relies on the online language extracted from posts on social media. With these we build models to predict two different interpretations of an individual's mental state — the first is self-reports of state anxiety and stress, and the second is a measure of physiological arousal through a wrist-worn sensor. Our paper addresses two questions:

- **RQ1.** Compared to *behavioral signals*, do *social signals* have a smaller semantic gap with psychological interpretations of wellbeing?
- **RQ2.** Compared to *social signals*, do *behavioral signals* have a smaller semantic gap with physiological interpretations of wellbeing?

Primarily, this paper presents a case that characterizes the semantic gap in passive sensing for predictive wellbeing and demonstrates an approach to reduce it. By highlighting this semantic gap, our aim is neither to identify the most credible instrument of ground truth nor is it to deplore particular sensor streams. Instead, we intend to clarify why passive sensing models of mental wellbeing appear to work or fail. Acknowledging the semantic gap in our domain leads to several key implications. Overall, we provide empirical evidence to support how different passive sensing modalities are naturally coupled to different interpretations of ground truth. Next, we encourage researchers to consciously understand the nature of ground truth labels and what factors influence that measure. And finally, in cases of limited sensing affordances for field study deployments, our findings motivate a more theoretical approach to sensor and modality selection for efficacious predictive studies.

2 BACKGROUND & RELATED WORK

As a background for the research presented in this paper, we note that the existing literature demonstrates multiple examples of inferring mental states with personal devices that capture behavior, such as smartphones and wearables [23, 75, 93], as well as with language used by people online [4, 24]. However, many such studies rely on self-reports that do not adequately represent the multidimensional nature of mental-wellbeing constructs [111]. This abstraction between the ground truth and the actual mental state can render a semantic gap, which limits the performance of passive sensing. In the subsections below, we discuss relevant literature on semantic gap, representations of ground truth, and how social and behavior signals have been used for measurement of mental wellbeing. We use this background to propose our hypotheses.

2.1 Semantic Gap in Computing Problems

A "semantic gap" refers to the loss of information when machines try to formalize a concept that humans could interpret naturally [16, 50, 99]. This idea stems from classical frameworks of natural language processing, human cognition, and the challenge of translating real-world expressions, personal experiences, and cultural context into specific computational models [16]. The most notable explorations of semantic gap are in the image-retrieval and computer vision communities [50, 99]. Consider a model that infers the mood of an individual by analyzing the image of their face. When the faces are present in defined angles and backgrounds (e.g., front-facing, white background, fixed lighting), the model can interpret the scene in terms of certain formal parameters [99]. In contrast, when the same expression is captured in an image of a natural context (e.g., weather, crowds) the model takes a broader, more subjective, and error-prone interpretation [99]. In cases like the latter, the semantic gap is particularly stark. It requires a computer to interpret an object in the real world based on human labeled ground truth. Yet, human labels can provide rich high-level explanations of a situation that computers struggle to glean from low-level parametrized data [50]. Human interpretation of scenes often considers ecological factors, novel variations, and simply common sense. However, a computer refers to the same semantic construct with feature vectors (e.g., regions, frequency, and segments) and tries to relate these operators [50]. In this paper we broaden the definition of semantic gap provided by Smeulders et al. and adapt it:

> **Semantic Gap.** *The lack of coincidence between the information that one can extract from the data and the interpretation that the same data have for a user in a given situation.*

Research in passive sensing also attempts to infer mental state based on data-driven features [10, 17, 19, 30, 36, 42, 65, 71, 74, 88, 91]. Oftentimes, these investigations take place in natural settings [18], where researchers use self-reports for ground truth. However, *in*

situ self-reports are high-level human evaluations and are, therefore, sensitive to multiple social effects that are invisible to passive sensors [20]. This drives us to study if the semantic gap exists for passively sensing mental states. We believe describing individual mental wellbeing with passive sensing is analogous to using computer vision to describe a visual scene. While computational models are expected to explain the target construct, they are insufficient in explaining the high-level semantics [50]. **This paper is motivated to investigate the loss of information between high-level interpretations of mental wellbeing — such as self-reports — and the low-level computational interpretations of behavior — such as inferences from passive sensing.**

2.2 Representations of Ground Truth for Mental Wellbeing

Since mental wellbeing is a complex concept, human interpretations of it tend to be high-level representations. Even when collected *in situ* [20], the label that is collected is only an evaluative judgment of the participant's mental state [112], which tends to describe the psychological aspect of it. Weiss describes, "true affective states, moods, and emotions have causes and consequences distinguishable from the causes and consequences of evaluative judgments" [112, p. 176]. By contrast, mental states also have physiological artifacts, which might not be reflected in self-reports [112].

In social and ubiquitous computing, field studies to infer mental wellbeing rely on self-reports [55] as ground truth. These reports are sensitive to factors that influence reporting and self-perception. For instance, Chan et al. qualitatively studied participant experiences with ecological momentary assessment (EMA) tools for wellbeing and found external factors (e.g., commuting, social situations) to influence the quality of self-reports [20]. Even traditional survey literature states that factors like social desirability can impact participant responses [58]. This can lead to participants over-reporting certain types of experiences and under-reporting others. In some cases, social norms are stronger indicators of participant reports than personal attitudes [54]. Moreover, perceptions of wellbeing are often blurred by the subjectivity of memory. For example, participants often overestimate their sleep time in comparison to observed measurements [97]. Some self-reports describe exceptional events, some describe every episode, while some summarize multiple incidents [113]. These factors create a misalignment between what participants report, even though participant activities convey the same mental state.

Now, let us consider the fact that wellbeing has physiological aspects [112], which are not explicitly captured by self-reports. Compare a case where a participant is running on a treadmill and another where they are chased by a bear. Despite both cases having many similar physiological effects, such as elevated heart rate, a self-reported evaluation of stress is likely to be lower in the treadmill scenario. According to Russell's *Circumplex Model of Affect*, mental states such as stress can be described on the basis of arousal (or alertness) and valence (or pleasure) [85]. However, self-reports are known to be imprecise in describing the arousal aspect [51] as it is momentary and very sensitive to the stress event [110]. In an experimental study of public speaking, Hellhammer and Schubert found that self-reports of stress were only correlated with the physiological state during the stressful event but not before or after it. A related study found that similar stressors affect the heart rate of participants similarly, but their self-report of stress is highly correlated to higher trait anxiety [110]. In fact, prior work has encouraged incorporating physiological changes in an individual as a different type of gold standard for ground truth [52].

This does not imply that only one representation of wellbeing is "true", nor does it imply that these representations are mutually exclusive. What these works indicate is that the state of a participant's wellbeing can be interpreted differently based on how it is measured. And these different abstractions can lead to a loss of information in computational models because each representation is affected by different kinds of signals. Given these distinctions, our study investigates the related mental wellbeing constructs of *anxiety* and *stress*. These constructs are selected because they have different abstractions [111], one that is *psychometric* and the other that is *physiological*. Moreover, different instruments can measure these different abstractions, i.e., self-reports are skewed towards psychological interpretations while arousal measurements are skewed towards the physiological interpretations. As a result, the first question of the paper (RQ1) is focused on self-report assessments while the second relates to the arousal measurements (RQ2).

2.3 Social Signals and Self-Reported Wellbeing

Participant self-reports are sensitive to many factors [20, 54, 58, 97], which might not actually affect the participant's mental state, but only their interpretation or their report [112]. A majority of physical sensing work models behaviors on self-reported measures [10, 17, 19, 42, 71, 104]. However, prior work recognizes that models need to consider features that capture the variability in the representation of the target construct [118]. In mental wellbeing self-reports, this variability (e.g., self-presentation, social norms, and memory specificity) is not captured by physical behaviors, and, therefore, strips such modeling approaches of their true efficacy. By contrast, studies harnessing posts on social media to predict individual wellbeing [30, 36, 65, 88, 91] have used features that are semantically similar to self-reports.

Despite the popularity of self-reports as ground-truth, a common limitation that is invariably unchecked is the *self-presentation bias* [1, 58]. This often leads to "deliberate impression management" that is geared towards projecting an appearance based on personal motives. What participants are willing to disclose often gets confounded with how they would like to be perceived [53]. Similarly, users of social media are constantly juggling with self-presentation issues depending on their audience [98]. For instance, Ernala et al. found that users vary the depth of their mental wellbeing disclosure based on audience engagement. Relatedly, a common issue with ground truth is *response bias* (or *non-response* bias) [48]. Based on individual differences, certain participants can have a reluctance to respond to certain survey items. For instance, in surveys around alcohol, the non-response of heavy alcohol consumers' is influenced by fear of embarassment [59]. A corresponding phenomenon on social media is self-censorship, which describes online expressiveness as a function of the social norms of the perceived audience [25]. Das and Kramer have discussed how gender, age, and the diversity

of the audience can affect the content posted online. This literature leads us to believe that using passively sensed data from social media can capture some of the ecological factors that influence self-reports. To address our first research question, *RQ1: "Compared to behavioral signals, do social signals have a smaller semantic gap with psychological interpretations of wellbeing?"*, our work investigates two specific hypotheses:

H1a. Features extracted from social media posts are more predictive of self-reported *anxiety* than features extracted from sensors of offline physical activity

H1b. Features extracted from social media posts are more predictive of self-reported *stress* than features extracted from sensors of offline physical activity

Physical activity data harnessed from offline devices observe a continuous stream of data that is neither segmented nor self-vetted by social effects. Therefore, we speculate that the semantic gap in using the latter to predict self-reported wellbeing will yield a larger semantic gap. Nevertheless, physical activity data can witness some of the social effects discussed earlier as self-presentation can manifest in the offline too [46]. However, since social media posts are explicitly laced with such factors, we expect it to approximate self-reports more effectively.

2.4 Behavioral Signals and Physiological Measurement of Wellbeing

RQ1 challenges the efficacy of passively sensing physical activity to predict psychological wellbeing. Meanwhile, it is also important to comprehend which aspects of wellbeing do physical activity features actually explain. We know that mental states like stress lead to many physiological changes in an individual, which are not captured by self-reports [51]. Therefore, to complement the previous question, we question if the semantic gap between physical activities and mental wellbeing reduces when the ground truth is measured via physiological assessments.

When facing a fight or flight situation, the body prepares for dynamic actions to cope with it. This manifests as a stress response in a variety of discernible physiological reactions, e.g. the heart beats faster, with more regularity, and therefore vigilance increases [41]. Today, commodity wearable devices can track the physiological correlates of stress and related constructs, such as heart rate (HR), and these constructs in the long term are important for measuring health [8, 92]. Thus, heart rate monitors and other mobile devices can record objective measures of stress such as increased heart rate and decreased heart rate variability [51, 73]. Moreover, many ubiquitous devices can capture behavioral signatures of the user required to infer such markers. Smartphones and other mobile technologies are capable of rendering features to describe sleep, step count (movement), and device usage. All of which have previously been harnessed to indicate anxiety and stress [11, 34, 86].

Unlike self-reported measures of ground truth, which are sensitive to social effects, physiological measures (e.g., variation in arousal or cortisol) are relatively robust to such factors. Conversely, while individuals are aware of increased physiological changes they may not always report it as stress. Despite similar physiological effects, self-reports of stress reports are often related to the negative emotion [101]. In this regard, the physical activity of the individual remains consistent with the physiological experience of stress. For instance, exercise is a positive way to use available energy released by physiological stress responses and is therefore associated with decreased stress and better wellness [6, 83]. Similarly, reduced sleep is known to reflect increased stress because heightened arousal due to stressors augments alertness and thus disrupt sleep [86]. Even seemingly unrelated behaviors of an individual like phone usage are related to stress [76] because they can represent the sleep health of individuals [62]. These works lead us to our next research question, *RQ2: "Compared to social signals, do behavioral signals have a smaller semantic gap with physiological interpretations of wellbeing?"*, which we investigate with a specific hypothesis:

H2. Features extracted from physical activity sensors are more predictive of *high arousal duration* than features extracted than social media posts

Since physical activity sensors tend to be an unobtrusive, unfiltered, and longitudinal characterization of an individual's behavior we expect them to approximate physiological changes in the individual [78, 82]. Even though features extracted from social media content can be used to infer physiological markers of wellbeing such as heart rate and arousal [90], such information is limited to what people say, not what people do, and therefore poorly approximate physiological changes.

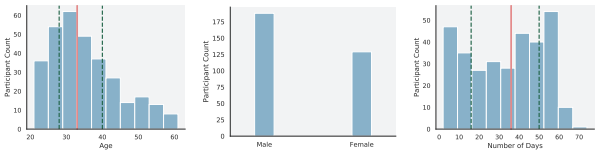
H2 can be viewed as a corollary of H1. While H1a and H1b attempt to expose evidence that high-level factors influencing mental wellbeing self-reports can exacerbate the semantic gap, H2 aims to find evidence that other interpretations of mental wellbeing can exhibit a narrower gap. We would also like to reemphasize that these hypotheses are not intended to assert that one approach is incapable of measuring wellbeing. On the contrary, these are motivated to disentangle how different interpretations of, or methods of measuring, the ground truth are biased towards different approaches based on the underlying factors certain modalities capture [118].

3 STUDY AND DATA

3.1 The Tesserae Project

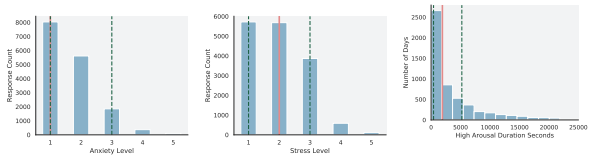
This work relied on data collected from a large multimodal sensing effort known as the Tesserae Project [69, 87]. This project investigated worker performance and wellbeing in-the-wild by using passively sensed data acquired from off-the-shelf technologies [26, 72, 90], furthering prior work in the community that leverage multimodal sensing to determine individual mental health, wellbeing and related outcomes [67, 74, 81, 88, 108]. Such a dataset is particularly appropriate for our research questions because it contains different interpretations of the ground truth (self-reported and physiological) as well as multiple sources of passive data (physical activity and social media posts).

The dataset contains a set of 757 information workers (involved in fields like engineering, consultancy, and management) recruited from various field sites in the United States in a rolling enrollment from January 2018 through July 2019. This study was approved by the Institutional Review Board (IRB) at the researchers' institutions and the data was de-identified and stored in secured databases with regulated access privileges. On enrollment, participants completed an initial assessment to record individual differences, such as demographics. Subsequently, for purposes of passive sensing, a phone



(a) Age (b) Gender (c) Study Period

Figure 1: Summary of participants. The solid red line indicates the median and the dotted green lines indicate the inter-quartile range.



(a) Anxiety (b) Stress (c) High-Arousal Duration

Figure 2: Distribution of Ground Truth. The solid red line shows the median and the dotted green lines show the inter-quartile range.

application was installed in participants' smartphones [108] and they were provided a wearable device (*Garmin Vivosmart*) along with Bluetooth beacons (*Gimbal*) [69]. These devices captured offline behaviors such as phone usage, locations, steps, sleep, and presence at home. Moreover, a subset of participants explicitly consented to the study of their historical and prospective social media data [87]. These different data sources represented the different comparative models analyzed in this paper. At the same time, the phone agent facilitated Ecological Momentary Assessments (EMAs) to capture daily variations in mental wellbeing states. The self-reports for anxiety and stress form the basis of the perceptual representation of ground truth (RQ1). Similarly, the wearable provided daily estimates of the duration an individual's physiology was in a state of high arousal (RQ2) [8].

This paper only considered those 317 participants that consented to data collection of offline behaviors as well as the social media collection. 129 participants reported they were female and 188 reported they were male (Figure 1b). Figure 1a shows the age distribution and the average age was about 35 years (stdev. = 9.27). Figure 1c depicts the study period for the participants. On average, participants provided self-reports for 33 days (stdev. = 18.44)

3.2 Ground Truth

Our central argument is that different ground truth for wellbeing are associated with unique ecological factors, which are differentially represented in passive sensing modalities. In particular, this work was scoped to predict anxiety and stress because these wellbeing constructs can be represented both psychologically and physiologically. While the former was captured with self-reports, the latter was measured through precise changes in bodily responses.

3.2.1 Self-Reports. Self-reports are considered a gold-standard method to retrieve ground truth labels in many studies that use passive sensing to infer individual wellbeing [10, 17, 19, 30, 42, 65, 71, 74, 88, 91]. However, as described in Section 2.2 and 2.3, a self-report is merely an evaluation of the individual's state [112] and these evaluations face interference from multiple factors that do not necessarily impact the actual state of the individual's wellbeing [20, 54, 58, 97]. Notably, values recorded in self-reports are subject to self-presentation bias [1, 58] and non-response bias [48]. For RQ1, we studied if certain passive sensing modalities were more predictive of self-reports because they inherently reflect an individual's attitude towards disclosure and censorship.

Anxiety. The emotional state that represents an exaggerated unpleasantness, negativity, or fear of future events is known as anxiety [63]. While anxiety can manifest as a trait, reflecting an individual's predispositions to react to certain stimuli, it also occurs as a transitory feeling, or a state [63, 102]. This paper is concerned with the state aspect of anxiety. In fact, state anxiety can be characterized by both conscious perceptions (of apprehension or nervousness) and physiological arousal [102]. These aspects position anxiety as a viable construct for this study because it can vary in interpretation based on the measurement method. In this dataset, participants responded to a daily single item instrument developed by Davey et al. to record the changes in anxiety [28]. Figure 2a shows the distribution of self-reported state anxiety. On a scale of 1-5, the mean response was 1.68 (stdev. = 0.81).

Stress. The way an entity responds to adverse ecological demands is known as stress [95]. The environmental factors that elicit stress are known as "stressors", and these contain both psychological or physiological components [56]. While humans tend to witness stressors that are some amalgam of both components, the effects of each component have differential outcomes in terms of response behaviors and duration [56]. Therefore, stress is another wellbeing construct that can be interpreted differently based on how it is measured. Particularly, self-reported measures of stress are more indicative of perceived stress or psychological stress [61]. Participants responded to a daily single-item omnibus question to explore this phenomenon, *"Overall, how would you rate your current level of stress?"*. This instrument was internally validated within the program metrics of the overall project by robustly correlating it with other measures, to establish concurrent validity [69]. Figure 2b shows the distribution of self-reported stress. On a scale of 1-5, the mean response was 1.97 (stdev. = 0.90).

3.2.2 Physiological Measurement. Although self-reported methods to acquire ground truth have dominated studies of passive sensing, wellbeing can be measured in-situ through physiological metrics [52]. For instance, both anxiety and stress are tied to physiological responses. These physiological responses may intersect with psychological ones, but can be dislocated from them as well (Section 2.4. Accordingly, in RQ2, this paper studies if particular passive sensing modalities are more effective predictors of physiological measures, which are robust to social variances but tightly coupled with behavioral changes in the individual.

High-Arousal Duration Both state anxiety [102] and stress [56] are linked to arousal. When an individual anticipates or is subjected to a stressor they experience physiological changes such as

Table 1: Activity features derived from offline sensors; * : includes features aggregated by epochs, i.e, 24 hours, 12am - 6am, 6am - 12pm, 12pm - 6pm and 6pm - 12am

Category	Features	Stream
Activity Label	Still duration*, walking duration*, running duration*, unique activity count	Smartphone
Movement	Steps count, steps goal, floors climbed, floors goal, distance covered	Wearable
Mobility	Unique location count, total location count, inter-location distance	Smartphone
Sleep	Sleep duration, sleep debt, time of wakeup, time of bedtime	Smartphone, Wearable
Screen	Unlock Duration*, Unlock Count*	Smartphone
Presence	Work session duration, desk session duration, desk session count, percentage time at work, percentage time at desk, 30 minute break count	BT Beacons

Table 2: Language features derived from social media

Category	Features
LIWC	*Affective attributes*: anger, anxiety, negative and positive affect, sadness, swear; *Cognitive attributes*: causation, inhibition, cognitive mechanics, discrepancies, negation, tentativeness; *Perception*: feel, hear, insight, see; *Interpersonal focus*: first person singular, second person plural, third person plural, indefinite pronoun; *Temporal references*: future tense, past tense, present tense; *Lexical density and awareness*: adverbs, verbs, article, exclusive, inclusive, preposition, quantifier; *Biological concerns*: bio, body, death, health, sexual; *Personal concerns*: achievement, home, money, religion; *Social concerns*: family, friends, humans, social
Sentiment	Positive score, negative score, neutral score
N-Grams	Top 500

heightened heart rate (HR) and changes to their heart rate variability (HRV) [51, 73]. Such reactions are linked to a "fight-or-flight" response through the *Sympathetic Nervous System* (SNS). The participants in our dataset were equipped with wearables that could measure the HR and HRV of the individual. Based on this, we obtained Garmin's *HealthAPI*'s [8] estimates of the daily duration for which the user was in a high arousal state [29]. As per *Firstbeat Analytics* (the framework powering the API) a simultaneous increase in an individual's HR along with reduced HRV triggers the SNS to activate their body into a stress state [103]. Prior work has shown that Garmin's HR-based inference of maximal oxygen update — a key physiological indicator of stressful arousal, known as VO_2 max — was highly correlated (r = 0:84) with measurements from clinical instruments [57]. The Garmin *HealthAPI* leverages *Firstbeat Analytics*, which reported that even in free-living conditions, using HR to infer stress demonstrated low error (approximately 5% Mean Absolute Percentage Error) in estimating VO_2 max [114]. Garmin wrist-worn devices have been used by researchers in the domain to provide physiological ground truth for modeling passively sensed behavioral signals [26, 39, 90]. In this study, we only considered the high arousal duration for days the participants' wearable provided more than 18 hours of data, in order to get a representative sample [69]. Figure 2c shows the distribution of the daily high-arousal duration. On average, for a given day the participant experiences a high-arousal state for 4193 seconds or 1.16 hours.

3.3 Passively Sensed Data

Researchers have used several unobtrusive methods to retrieve low-level digital traces or "markers" that describe individuals and use these features to make inferences for mental wellbeing. To illustrate the semantic gap, we focused on two different sources of passive sensing, offline sensors that capture behavioral signals and language on social media that reflects social signals 2.4). By using these sources we built comparative models to predict different interpretations of the ground truth (self-report and arousal duration) and tested our hypotheses to address our research questions.

3.3.1 Physical Activity. To collect continuous physical activity data, participants had of three different sensor streams, (i) smartphone, (ii) wearable, and (iii) Bluetooth beacons, as introduced above. The application installed in the smartphone [108] measured screen activity (or device use), tracked GPS location, and provided activity labels [7]. The wrist-worn wearable estimated activity duration, step counts and was combined with the screen usage to yield sleep features. Lastly, the Bluetooth beacons were placed on the front door of the participant's residence and on their work desk. These beacons were observed by the individual's phone agent [9] to infer the time they spent on their desk, when they came into work, and how frequently they were away from the desk [26]. Table 1 summarizes the features derived from this set of sensors. These features are grounded in prior works of passive sensing [12, 72, 108]. For this paper, we only analyzed the data collected on days that the user provided self-reports. Every feature was aggregated at a day-level (e.g., daily mean) for each day in the sample.

3.3.2 Social Media Language. For this paper, we specifically focused on the data on participant posts on Facebook, given Facebook was the most widely used platform in our participant pool [87]. Social media data provided psycholinguistic attributes of the participant posts using LIWC (Linguistic Inquiry and Word Count) [79] This lexicon has been used in prior work to study mental health and wellbeing through social media [30]. Based on this [30], we also used 50 categories of LIWC that De Choudhury et al. segregated into the 9 different groups, *affective attributes*, *cognitive attributes*, *perception*, *interpersonal focus*, *temporal references*, *lexical density and awareness*, *biological concerns*, *personal concerns*, and *social concerns*. Additionally, posts were characterized with sentiment analysis (score for positive, negative, and neutral label) [66]. Lastly, this data provided a large set of open vocabulary features, i.e., the usage of the top 500 n-grams [24] within the corpus of all posts in the study. These features were sparse because n-grams do not appear consistently on all posts but are still a mainstay in language-based predictions of mental wellbeing [3, 91, 109]. Table 2 summarizes these features.

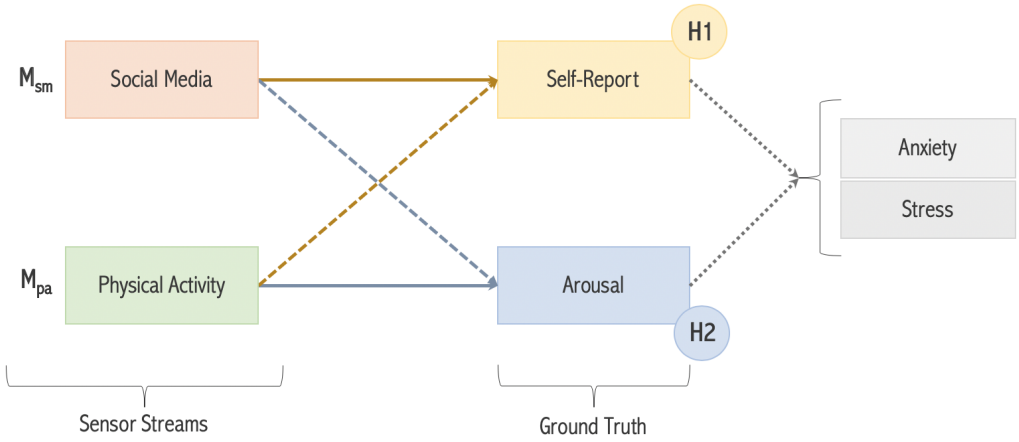


Figure 3: The triangulation framework helps compare different prediction models. For H1a and H1b, M_{pa} and M_{sm} predict self-reports of anxiety and stress respectively. For H2, M_{pa} and M_{sm} predict high arousal duration.

4 METHOD

To test our hypotheses, we compared the performance of two approaches for predicting different ground truth for mental states. The first used modalities with psycho-social signals, M_{sm}(social media language features), while the second used those with behavioral signals, M_{pa}(physical activity features). Neither self-report assessments nor arousal measurements alone can comprehensively capture the nuances of anxiety or stress. Therefore, we compared different approaches to predict them as a means to disentangle the semantic relationship between low-level computer representations and high-level mental wellbeing constructs. This approach was motivated by previous works that used quantitative data triangulation [36]. The general framework of triangulation [31] is suitable for deconstructing predictive analysis of wellbeing using passive sensing because it "adds rigor, breadth complexity, richness, and depth to any inquiry". Figure 3 illustrates an overview of our investigation framework.

4.1 Feature Engineering

This paper refers to both anxiety and stress as 'states', because these change in short periods of time. In the scope of this work, the ground truth measures (both self-report and arousal) were collected at the day-level granularity [69]. Prior works in pervasive sensing for mental wellbeing [38, 96, 107] motivated us to analyze behavior not just during the day of ground-truth measure, but also in periods preceding it. Even theoretically, mental states are indicated by general trait behavior that changes but is less sensitive [110]. Therefore, our approach accounted for the historic sensor data to approximate the target concept. To this end, we first collated features that spanned a period prior to the prediction day.

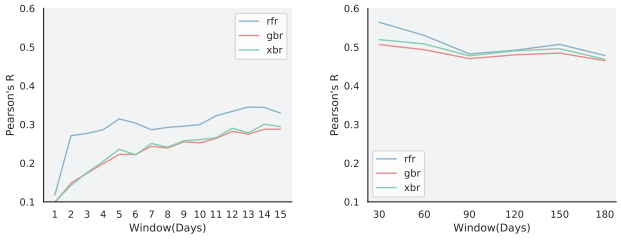
4.1.1 Feature Windows. The predictive models we built for both M_{pa} and M_{sm} to consider a window of time for the features. For instance, to predict the state-anxiety (H1a) for day n the model analyzed features in a span of d days before n. Here, d dictates the fixed window size.

Physical Activity. In our dataset, since offline sensors could continuously monitor individuals we varied the window size between $1-15$ days. This results in $f \times d$ dimensions if f is the original set of features computed for each day and d is the window size. Importantly, these sensors were not active before the first self-report was collected for any participant [69], therefore this modality was limited in how far the window could stretch retrospectively. Since the average participants had 31 labels the upper limit of 15 days was chosen to balance the remaining days for evaluation. If a window of, say, 31 days was chosen then in most cases, only the most recent label would have physical activity features for every day while all days before that would have empty data.

Online Language. Unlike physical activity data, which provided a near-continuous and contiguous signal, the online language data obtained from social media is extremely sporadic. Social media can be considered a form of "virtual sensor" that capture rich momentary events, which occur irregularly [106]. This is inherent to the approach as people do not post regularly, thus making social media platforms approximate event-based sensors. Thus, the window size for this modality varied between $30-180$ days, with a shift of 30 days between each window. In contrast to offline sensors that were only instrumented after enrollment, social media allowed us to access data prior to enrollment and could, therefore, support a much broader window [87].

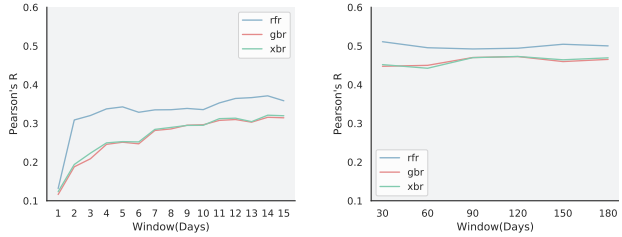
4.1.2 Prepossessing. This section elaborates on our methodology for imputing missing values and standardizing features in windows.

Physical Activity. On certain days particular features could be missing due to participant compliance (e.g., the participant did not charge a device or data failed to log). Consequently, we imputed the missing values of a feature by substituting it with the mean of that feature for an individual for a given window. To demonstrate, if a feature value was missing for an original feature f^a on day d_j, then the average will be $\sum_{i=1}^{d} f_i^a/d$, where d is all days the



(a) M_{pa} **(b)** M_{sm}

Figure 4: Comparing models with different window length to predict anxiety



(a) M_{pa} **(b)** M_{sm}

Figure 5: Comparing models with different window length to predict stress

feature was not null. After this the features were standardized by subtracting the mean of the feature values and dividing it by the standard deviation. Similar to the imputation, the standardization procedure was also applied within windows, i.e., the average and standard deviation for any feature f_i^a, was calculated on $[f_1^a, f_d^a]$ where d is the window size.

Online Language. Empty values occur more frequently because most participants did not post everyday. Because of this limitation, filling in missing values with averages could lead to washing out any true variations. Therefore, we heuristically rejected windows that have fewer than 1 post per week on account of low density. This was followed up by the approach described earlier where both imputation and standardization are applied within windows.

4.2 Feature Processing and Model Training

We developed different non-linear regression models for M_{pa} and M_{sm} to estimate self-reported state anxiety (H1a), perceived stress (H1b), and objectively measured high-arousal duration (H2). In particular, we trained models with both modalities using estimators that rely on ensemble learning because these approaches "reduce the variance — thereby improving the accuracy" of estimates [117, p. 1]. The *Random Forest* regressor aggregates independent decision trees, each of which learns on a random sample of input features [64]. *Gradient Boost* learns incrementally over time by increasing the importance of poorly estimated observations in every subsequent iteration [35]. An additional variation to this is Extreme Gradient Boosting (*XGBoost*), which is both robust to noise and designed to deal with sparse input features [21], such as those extracted from social media data. Moreover, a different model was

Table 3: Summary of between models comparison for self-reported anxiety
('-':$p < 1$, '·':$p < 0{:}1$, '*':$p < 0{:}05$, '':$p < 0{:}01$, '***':$p < 0{:}001$)**

Regressor	Pearson's R		SMAPE	
	M_{pa}	M_{sm}	M_{pa}	M_{sm}
Random Forest	$0{:}34^{***}$	$0{:}56^{***}$	0.18	0.16
Gradient Boost	$0{:}27^{***}$	$0{:}51^{***}$	0.19	0.17
XGBoost	$0{:}27^{***}$	$0{:}51^{***}$	0.19	0.17
Window Size (days)	13	30	13	30

built for every window size and each model was trained using a 5-fold cross-validation method. Additionally, the grid search approach tuned the parameters for each model [100]. Since the information used to predict the target value for each day was spread across a window of d days, it leads to $f \times d$ dimensions, which can sabotage the training because of the curse of dimensionality. To tackle this we employed certain feature transformation and reduction techniques to improve the model training. These processing approaches are applied to each model separately, i.e., it is unique to the window size. Given our cross-validation approach, these feature processing steps were "fit" only to the training data without incorporating any of the observations in the testing folds.

4.2.1 Coefficient of Variance. First, we estimated the variance explained by each dimension measuring the *coefficient of variance* (CV) [91]. With a conservative bound, we remove dimensions that are beyond 1 standard deviation of the average CV. For the linguistic features included in the M_{sm} models, this typically led to a dimension reduction by $20-26\%$ with windows varying between $30-180$ days. For the physical activity features in M_{pa}, this led to a reduction of $32-14\%$ for windows of size $1-14$ days. *Note: The M_{pa} model used in H2 does not use this aspect of the pipeline because it produces a better model without this selection.*

4.2.2 Principal Component Analysis. Next, we further reduced the dimensions by performing PCA on the remaining dimensions [115]. This approach identifies latent components in the data (linear combinations of existing dimensions) that explain maximum variance. The first set of principal components that can cumulatively explain more than 90% of the variance in the data were selected as dimensions going forward. For M_{pa}, between window sizes of $1-14$ days, this process reduced dimensions by $62-84\%$ respectively. Similarly, for M_{sm}, between window sizes of $30-180$ days we observed a reduction of $51-86\%$.

4.2.3 Mutual Information. Lastly, for M_{sm} we included a final shortlist of dimensions based on mutual information between the input dimensions and the target variable [105]. Based on the mutual information scores, this process selected the top 10 percentile dimensions. It is important to note that this procedure is both unnecessary and detrimental to apply on the features of M_{pa} as these models had lower dimensionality to begin with and reduction beyond the PCA described earlier generated weaker models.

5 RESULTS

This paper studies the semantic gap by comparing different prediction approaches with an analytic process grounded in the data triangulation framework [36]. This framework enabled us to methodologically evaluate heterogeneous approaches to understand the same phenomenon [31]. The approaches we compared in this study differ in terms of both data source and methodology. Therefore, for each , this paper addresses the research questions on the basis of the best models for M_{pa} and M_{sm}.

Within Modality Comparisons. The best model was chosen based on the highest pooled *Pearson's* correlation between the true values and the predicted values. Specifically, we pooled together the predictions from each cross-validation fold and then computed the correlation with the ground truth. This approach is robust to heterogeneity in target variables' distribution between folds and provides a more generic measure of performance [2]. We used the *Pearson's* correlation coefficient because it spans all samples to describe a complete relationship, is not sensitive to the distribution of samples, and does not assume normality [77]. This correlation contrasts a model's input features and the target variable. For internal validity of the regression models, we compared the *Symmetric Mean Absolute Percentage Error* (SMAPE) against an arbitrary regression model that always predicted the mean of the training data.

Between Modality Comparisons. Once the best models of M_{pa} and M_{sm} were identified we validated comparisons between M_{pa} and M_{sm} by performing a permutation test [5, 105]. Essentially, we attempted to reject the null hypothesis that a random set of features in a similar feature space (range and dimensionality) will still perform better than the worse model [91]. We permuted these random features and computed the probability ($p-$value) of such an arbitrary model improving over the benchmark.

5.1 RQ1: Semantic Gap in Predicting Psychological Aspects of Wellbeing

5.1.1 H1a: M_{sm} is a better predictor of self-reported anxiety. We find language on social media to be more indicative of self-reported state anxiety when compared with physical activity from offline sensors. With physical activity features, we find the best model for M_{pa} to be with a window length of $d = 14$ and using the Random Forest regressor (Figure 4a). This model recorded a *Pearson's* $r = 0{:}34$. In comparison to an arbitrary regressor, which demonstrated a *SMAPE*= 0:20, this model shows a *SMAPE*= 0:18, a 10% improvement over the baseline. By contrast, for the same target variable, the best M_{sm} model was at $d = 30$ with a Random Forest regressor (Figure 4b), which yields a *Pearson's* $r = 0{:}56$. In comparison to the baseline (*SMAPE*= 0:21), this model improves by 30% (*SMAPE*= 0:14). Between models, we see the *Pearson's* r in the anxiety values predicted by M_{sm} to be 64% better than values predicted by M_{pa}. To test the robustness of this comparison we ran the pipeline for M_{sm} 1000 times with randomly generated permutations of the feature values and find the probability of improvement over M_{pa} to be less than 0:001. As a result, this asserts M_{sm} is more predictive of self-reported state anxiety than M_{pa}, and this supports hypothesis H1a (Table 3).

Table 4: Summary of between models comparison for self-reported stress
('-':$p < 1$, '.':$p < 0{:}1$, '*':$p < 0{:}05$, '':$p < 0{:}01$, '***':$p < 0{:}001$)**

	Pearson's R		SMAPE	
Regressor	M_{pa}	M_{sm}	M_{pa}	M_{sm}
Random Forest	$0{:}37^{***}$	$0{:}51^{***}$	0.18	0.17
Gradient Boost	$0{:}31^{***}$	$0{:}44^{***}$	0.18	0.18
XGBoost	$0{:}32^{***}$	$0{:}45^{***}$	0.18	0.18
Window (days)	14	30	14	30

5.1.2 H1b: M_{sm} is a better predictor of self-reported stress. Similar to the previous result, language on social media is more predictive of self-reported stress than physical activity from offline sensors. In the case of M_{pa}, the window length of $d = 14$ with the Random Forest regressor (Figure 5a) emerged at the best model with a *Pearson's* $r = 0{:}36$. This improved on the baseline (*SMAPE*= 0:20) by 10% (*SMAPE*= 0:18). On the other hand, the best M_{sm} model was at a $d = 30$, also with Random Forest shows a *Pearson's* $r = 0{:}51$ (Figure 5b). Compared to the baseline (*SMAPE*= 0:20), this model had a *SMAPE*= 0:17, i.e., a 15% improvement. When comparing the two models, we find the *Pearson's r* of M_{sm} to be 37% better than that of M_{pa}. The permutation test was run 1000 times for random versions of M_{sm} and improved over M_{pa} less than 0:001 of the time. Based on the results, M_{sm} was a better predictor of self-reported stress than M_{pa}, and therefore the hypothesis H1b holds (Table 4).

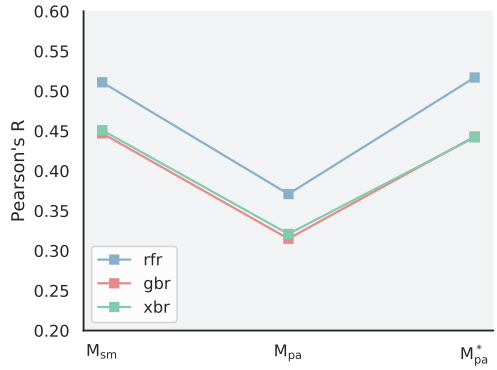
5.1.3 Post-Hoc Analysis. The results of the experiments argue that features extracted from social media posts can encapsulate analogous phenomena and therefore predict the target variable better. However, social signals can be derived from data acquired through offline signals as well. Since offline interactions are subject to similar presentation effects [46], we performed an additional experiment that augments M_{pa} with some physically sensed social features. In particular, we used the Bluetooth beacons to identify social behaviors, such as the time of first interaction, number of unique interactions, and their duration (Table 5). We included these features in the models used to test M_{pa} to predict the ground truth. The paper refers to this combined modality as, M^*_{pa}. In fact, the pipeline used for M_{pa} is the best framework for M^*_{pa} as well. For anxiety, the optimal results were produced with a random forest regressor at a window length of $d = 13$ where the Pearson's r is 0:49. Albeit still less than M_{sm} (Pearson's r is 0:56), this was markedly more than the best model for M_{pa} (Pearson's r is 0:34) by 64%. Actually, for predicting stress, the best results emerged with the same regressor and same window length (Pearson's r = 0:51). Not only was it better than M_{pa}(Pearson's r = 0:37) by 41%, it was comparable to M_{sm} as well (Pearson's r = 0:51).

5.1.4 Interpretation. The results for predicting self-reported anxiety and self-reported stress support our hypotheses towards our first research question, which investigates if features encapsulating social signals reduce the gap with self-reported measures of wellbeing (Figure 6a and Figure 6b). To reemphasize the intuition behind these hypotheses we reiterate the motivation discussed in Section 2.3. Mental wellbeing constructs like anxiety and stress

Table 5: Social features extracted from offline sensors

Category	Features	Stream
Colocation	Time of first and last interaction, number of interactions, number of unique participants, duration of interactions, percentage alone, percentage with at least one /two /three others	BT Beacons

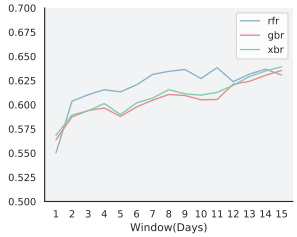
(a) Anxiety



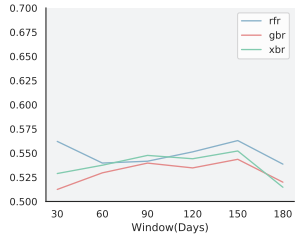
(b) Stress

Figure 6: Comparison between best models of different modalities (RQ1)

have different aspects. Self-reports are skewed to capture the psychological aspects [51] that might not be concordant with how the individual actually behaves. Moreover, self-reports are influenced by many social effects of self-presentation such as impression management [53, 58] and response bias [48]. These effects are inconspicuous to sensors that capture physical activity, even though that data can be modeled to predict such self-report (evident from the improvement on the baseline). By contrast, data sourced from social media are weaved with similar ecological effects that could influence an individual's self-report. For example, self-disclosure [37] and self-censorship [25] are both factors that affect the language posted online. This could explain why M_{sm} exhibited better results



(a) M_{pa}



(b) M_{sm}

Figure 7: Comparing models with different window length to predict high-arousal duration

Table 6: Summary of model comparison between models to predict high-arousal duration
('-':p < 1, '·':p < 0:1, '*':p < 0:05, '':p < 0:01, '***':p < 0:001)**

Regressor	Pearson's R		SMAPE	
	M_{pa}	M_{sm}	M_{pa}	M_{sm}
Random Forest	0:63***	0:56***	0.45	0.46
Gradient Boost	0:60***	0:54***	0.46	0.46
XGBoost	0:61***	0:55***	0.43	0.46
Window (days)	11	150	11	150

than M_{pa} when trying to predict self-reports. Relatedly, incorporating more explicitly social features in offline sensing also shows an improvement in the prediction (M^*_{pa}).

5.2 RQ2: Semantic Gap in Predicting Physiological Aspects of Wellbeing

5.2.1 H2: M_{pa} is a better predictor of objectively-measured high-arousal duration. While predicting high-arousal duration the model built with physical activity features was better than the corresponding model built with social media language features. We find the best model for M_{pa} to be at window length of d = 15 with the XGBoost regressor (Figure 7a), which showed a *Pearson's r*= 0:63. This surpassed the baseline (*SMAPE*= 0:54) by 16% (*SMAPE*= 0:45). In comparison, the best performing M_{sm} model occured at d = 150, also with Random Forest, which yielded a *Pearson's r*= 0:56 (Figure 7b). This model had a *SMAPE*= 0:43, i.e., a 20% improvement on the baseline (*SMAPE*= 0:54). In comparison to M_{sm} the *Pearson's r* of M_{pa} is 13% better. To reject the possibility of chance improvement, from 1000 randomly generated permutations of M_{pa} less than 0:01 feature sets improved over M_{sm}. These results indicate that M_{pa} is a better predictor of self-reported stress than M_{sm} and therefore supports hypothesis H2 (Table 6).

5.2.2 Post-Hoc. Similar to the analysis performed in Section 5.1.3, we further experiment on predicting physiological wellbeing by including offline sensed social features (Table 5). The argument to pursue such an analysis in the light of RQ1 was to estimate the effects of social signals from alternative sources to reduce the potential semantic gap. However, in RQ2 testing a prediction with M^*_{pa} is to explore how social factors interact with physical signals to predict physiological aspects of wellbeing. On experimenting

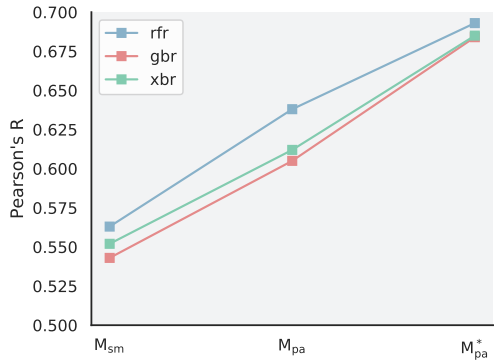


Figure 8: Comparison between best models of different modalities (RQ2)

with M^*_{pa} we find that a random forest regressor at a window length of d = 13 yielded the best result of *Pearson's r* = 0:69. Compared to the large boost we observed in predicting self-reports, adding social signals to predict objective measurements only augmented M_{pa} (Pearson's r is 0:63) by only 9%. While this is still noticeable, we believe the improvement is limited by the nature of the additional signal (psycho-social) in comparison to the representation that is being predicted (physiological). Therefore, although additional features can lead to some increment in performance, large boosts can be achieved when a model is augmented by semantically similar features (Section 5.1.3).

5.2.3 Interpretation. The findings for predicting high arousal duration support our hypotheses towards the second research question, which speculates a reduction of the semantic gap in predicting objective measures of wellbeing by modeling features with behavioral signals (Figure 8). This question was proposed to provide divergent validity to the first question and reinforce the quantitative data triangulation method of validation [36]. As discussed in Section 2.4, the physiological aspects of wellbeing can often be independent of what individuals report [51]. It can be subject to inherent beliefs, other subjective factors, and confounding mental phenomena [101]. On the other hand, the physiological experience of the individual remains consistent. Furthermore, the physical behaviors of an individual are coupled with physiological responses to wellbeing constructs like stress. For example, increased activity can reduce arousal by expending energy [83], while reduced sleep can be the result of increased arousal [86]. This kind of information is challenging for passive sensing through online traces to perceive as people only present a part of their selves on such platforms. Accordingly, M_{pa} performed better in this regard due because offline modalities that continuously capture an individual's functioning can illustrate richer representations of their behavior.

5.3 Participant-Independent Models

The models described in Section 5.1 and 5.2 were validated by using some observations in the training folds while others are used in testing folds (each participant had an average of 33 days of data). Such approaches, known as "mixed-model" or "personalized-model", account for individualized routine and trait-like propensities to predict the target variable. These have been used in prior works in longitudinal sensing with smartphone data [32, 107] and social media data [22]. An alternative approach to modeling sensor data is with participant-independent models which treat testing data as entirely unseen participants. These are expected to generalize better to new participant data. To inquire our hypotheses with this approach we first performed a participant-independent 5-fold cross-validation. We followed the same feature processing described in Section 4.2, with the only difference being that M_{sm} performed better without any additional mutual information based feature selection (Section 4.2.3). For H1, we found that the best model for M_{sm} significantly estimated both self-reported anxiety (*Pearson's r* = 0:15 with XGBoost when d = 180) and stress (*Pearson's r*= 0:08 with Random Forest when d = 90). In contrast, however, M_{pa} models did not significantly estimate the ground truth at all (*Pearson's r*= 0:02 for anxiety and *Pearson's r*= 0:02 for stress). For H2, the best model for M_{pa} significantly estimated high arousal duration (*Pearson's r*= 0:52 with XGBoost when d = 1), whereas the best performance for M_{sm} did not show significant correlation (*Pearson's r*= 0:03). Further, we also performed a leave-one-participant-out validation for both hypotheses. Again for H1, we found that only M_{sm} could significantly estimate self-reported anxiety (*Pearson's r*= 0:08 with XGBoost when d = 30) and stress (*Pearson's r*= 0:07 with XGBoost when d = 30). Similarly, for H2, only M_{pa} significantly estimated high arousal duration (*Pearson's r*= 0:52 with Gradient Boost when d = 1). Even though the performance of person-independent models was lower than the personalized ones (as shown in similar studies [32, 107]), we still found that these models demonstrated a persistent semantic gap. Table 7 summarizes the comparison between M_{pa} and M_{sm} for the different hypotheses.

6 DISCUSSION

The findings of the paper reveal the presence of the semantic gap in studies to infer mental wellbeing. This is not meant to discourage research in this space, but to highlight the untapped potential of these studies. For instance, our post-hoc analyses in Section 5.1.3 and 5.2.2 illustrated that models improve by incorporating features reflecting social signals. To this end, we propose a set of guidelines for researchers in social and ubiquitous computing, who plan to conduct passive sensing studies to infer mental wellbeing.

6.1 Ground Truth Matters

> In studies of mental wellbeing, even with validated instruments to assess ground truth, researchers need to consider how this ground truth represents or abstracts the underlying mental wellbeing construct (psychological or physiological). Consequently, researchers must account for the different factors that affect these representations (e.g., social biases or behavioral artifacts) to get optimal results.

Our results show that the presence of a semantic gap reflects a mismatch between what computational models represent and what different ground truth represent for the same mental wellbeing state. That said, the paper's findings are not intended to entirely disregard

Table 7: Summary of best models for participant independent models.
('-':$p < 1$, '.':$p < 0.1$, '*':$p < 0.05$, '':$p < 0.01$, '***':$p < 0.001$)**

		5-Fold CV			LOPO CV		
	Ground Truth Measure	M_{pa}	M_{sm}	M^*_{pa}	M_{pa}	M_{sm}	M^*_{pa}
H1a	Self-Reported Anxiety	0.02-	0.15***	0.02-	0.02-	0.08***	0.02-
H1b	Self-Reported Stress	0.02-	0.08**	0.02-	0.02-	0.07**	0.02-
H2	High Arousal Duration	0.52***	0.03-	0.55***	0.52***	0.03-	0.55***

certain sensors, family of features, or instruments for ground truth. On the contrary, the primary motivation of this paper is to bring to attention the nuances of ground truth measurements and what they represent. Specifically, this paper demonstrates how the ground truth labels are merely an abstraction of the actual mental wellbeing state and reflect limited aspects of it [36, 112]. Since measurements of anxiety or stress can be influenced by unseen factors [20, 54, 58, 97], features that encapsulate, or are associated with, analogous factors are more suitable to explain that form of ground truth. For instance, we find that modeling social media is better than modeling physical activities to predict self-reported measurements of both anxiety and stress (Section 5.1). However, this is not the case when using the same approaches to predict objective arousal-based measurements (Section 5.2). **It is not that certain modalities are more effective at explaining the mental state itself, but in fact, they are more capable at inferring that representation of the underlying mental state** (Sec 5.1.4 and Sec 5.2.3).

From the perspective of computer scientists, the ground truth is considered an unquestionable "gold standard". The literature has discussed several challenges to passive sensing [49], such as choice of device, application, duration, and sampling rate. Our findings extend this list with a focus on ground truth representations. This paper demonstrates a case that urges conscious consideration of the ground truth's sensitivity to ecological factors. Many studies in the community tend to acquire ground truth *in situ* [20, 69, 88, 108] but it distances the researchers from carefully observing the circumstances of ground truth measurement. In reference to the uncertainty of ground truth labels, Plötz's third postulate for machine-learning on sensor data states, "there is no ground truth" [80]. We situate this in the context of passively inferring mental wellbeing. Do participants respond to anxiety questions immediately after stressful incidents or do they summarize the experience of their day? Do they actually report how they felt or are their responses describing the state they wanted to be in? These concerns are not only challenging to quantify but also opaque to researchers and sensors [20]. **However, acknowledging the semantic gap can help researchers diagnose model performance by determining the mismatch between their sensor features and their ground truth representation.**

While self-reports remain a mainstay for measuring mental wellbeing constructs like anxiety and stress, many studies in mHealth have posited alternative measures. Hovsepian et al., proposed a new measure of stress in the wild, which involves a wearable device consisting of multiple biomedical sensors [52]. They found this measure to be a strong estimator of self-reported stress in the moment. Sometimes, physiological changes might not be captured in self-reports[51], but it is still valuable to characterize stressful episodes [94]. Prior work has provided evidence for these signals to trigger effective wellbeing interventions in field studies (e.g., heart-rate [47] and breathing [44]). Even though mental wellbeing constructs remain fairly subjective with respect to how they are experienced, perceived and eventually recorded [45], every kind of measurement is sensitive to different factors. For example, objective markers of physiological changes can vary with motion artifacts [52] and self-reports of psychological changes often obscure low-level details of the stressful episode [45]. **The presence of the semantic gap revealed in this work is meant to urge researchers to assess the imperceptible aspects of their ground truth measure while trying computational approaches to predict such constructs**.

6.2 Parsimonious Sensing

> For practical field deployments, the changing socio-technical landscape affects resource availability and privacy perceptions, which can limit researchers from conducting brute-force passive sensor studies with multiple complementary streams. Therefore, researchers should determine the smallest set of streams that are semantically the most representative of the ground truth measure. Less is more if studies select sensors that provide features that reduce the semantic gap in predictions.

As new sensing platforms become commercialized and other interfaces like social media become abundant, researchers have a plethora of means to digitally infer their mental wellbeing. One approach to mitigate the semantic gap is to capture more ecological information that can help explain the high-level processes that influence the ground truth. What is evident from this paper is that a single sensor stream is typically not robust enough to represent the different types of variability in ground truth. While offline sensors are skewed to represent behavioral changes (M_{pa}), online logs of virtual presence are better suited to represent social effects (M_{sm}). Therefore, a natural argument to reduce the gap between input features and target construct would be to deploy more sensors and track logs from multiple sources. In fact, combining multimodal features together can elicit new context-specific features [89, 116]. However, multimodal studies are challenging to deploy in the wild [69, 74, 81, 87], as they are expensive in terms of both instrumentation and recruitment. Moreover, additional sensors to capture the "reality" of a participant can introduce privacy concerns and generally overwhelm their experience [15, 84]. Instead, our findings suggest an alternative position to pursue parsimonious sensor deployments, or to make the most of limited resources to appropriately sense mental wellbeing constructs. We are inspired

by Plötz's fifth postulate, "data rule, models serve" [80]. For instance, if deployments intend to measure ground truth through self-reports and researchers do not have access explicit sources of social signals (such as online activity or conversations), researchers should try to accommodate for social effects in offline sensors (as demonstrated by the Bluetooth beacons used in M^*_{pa} in Section 5.1.3). Alternatively, if the study plans to estimate wellbeing with physiological changes then resources should be allocated to sense behavioral markers, such as movement and sleep. The existence of a semantic gap supports the idea of minimal sensing to predict wellbeing in comparison to conventional ideas of massive sensing. **Thus, our paper demonstrates realistic approaches to adhere to paradigms like "small data" in (critical) data science [15, 60] and passive sensing [40], and the Occam's razor metaphor for parsimony in machine learning [33]**

In the meanwhile, more sophisticated methods to identify markers for mental wellbeing from passively sensed computational data have emerged [70]. Arguably, better feature crafting can help reduce this gap even with the same set of sensors. In this regard, the semantic gap serves two functions. First, it provides a guiding rail to engineer features based on domain-driven aspects of mental wellbeing ground truth. Second, it provides interpretability to models by encouraging researchers to inquire if their features capture psychological or physiological aspects of wellbeing. Moreover, the presence of a semantic gap calls into question the objectivity of machine learning/data mining to generate inferences. Since unobtrusive sensing can capture vast amounts of information, engineering this data can often yield spurious connections with the target variable [15]. The findings of this paper encourage more critical investigations of computational models to arrive at theoretically meaningful interpretations. Researchers need to resist the allure of viewing more passive data as a *Maslow's golden hammer* [68] — a tool to solve any problem. Over-engineering the "hammer" can result in finding spurious associations in the data [15, 43]. For example, does sensing physical behaviors actually predict stress holistically or does it merely describe its physiological aspects? Conversely, does tracing online content explain what an individual experiences or does it only reflect how they project themselves? Similar to other works that critique, yet advocate, employing machine learning for health and wellbeing [13, 36], **this paper encourages researchers employing passive sensing to build models with deeper consideration of the domain and select sensors accordingly to avoid misrepresenting seemingly objective results**.

6.3 Limitations and Future Work

Although this paper provides evidence of a semantic gap in predicting wellbeing, it is only a case study specific to a particular dataset. Having said that, we believe this phenomenon can be observed in other datasets with diverse multimodal sensing streams and different sources of ground truth. In practice, such studies are challenging to implement and very few datasets with the required richness exist at the time of writing. The central concept of the paper is fundamental to other computing fields [50, 99] and our findings align with those notions of information loss in computational representations of human concepts — in this case, constructs of mental wellbeing. Relatedly, this paper only investigates two specific constructs of wellbeing, anxiety, and stress, which are also linked. While both of these are associated with many other states and constructs, an individual's wellbeing has many other components that are independent of anxiety and stress. Despite this, the implication, that the nature of ground truth can inform the choice of passive data collected, applies to other constructs of wellbeing that are vulnerable to differential interpretation because of measurement instruments (psychological or physiological).

To address the research questions, the paper models sensor data to explain daily wellbeing states over a period of time. Therefore, at its current stage, the findings are limited to dynamic constructs of wellbeing, such as state anxiety and perceived stress (as well as arousal). These constructs are expected to vary within short periods and tightly coupled with ecological changes. However, many studies use passive sensing to predict trait-based mental wellbeing constructs, such as social anxiety [14]. Since our work is motivated by the ground truth acquisition in the moment, the implications might be directly transferable to other predictions of wellbeing. This creates an opportunity for further investigating the possibility of a semantic gap in studies where the target construct is assessed in a lab setting or collected once during enrollment.

Lastly, these results support additional studies regarding the validity of *in situ* methods to collect ground truth. In particular, subsequent work can explore the contexts within which the self-reported ground truth is robust and the semantic gap becomes trivial, or tolerable. As a result, researchers can scrutinize the quality of their ground truth collection. In turn, the machine learning models built with passive data can be trained only on reliable or invariant measurements.

7 CONCLUSION

Mental wellbeing is a complex phenomenon that different measurements of ground truth can interpret it in varying ways. For instance, anxiety and stress are constructs that manifest both psychologically and physiologically. This paper is motivated to investigate a semantic gap in commonly used pervasive sensing methods to predict such wellbeing constructs. By applying the triangulation methodology this paper demonstrates evidence that based on the ground truth of mental wellbeing, certain types of passively sensed features are more skewed to explaining it. Particularly, features with social signals (M_{sm}), have a smaller semantic gap with self-reported wellbeing. By contrast, predictive modalities with physical signals (M_{pa}), have a smaller semantic gap with physiological measures of wellbeing, such as arousal. This study exposes how the gap in sensing streams and the ground truth affects predictions. The implications of this semantic gap inform passive sensing studies, particularly with respect to the nature of the ground truth and the choice of sensing for practical deployments.

ACKNOWLEDGMENTS

This research was supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2017-17042800007. The contents of this paper do not necessarily represent the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized

to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. We thank the Tesserae team for their support in enabling our investigation. Additionally, we are grateful to the members of the Social Dynamics & Wellbeing lab and the Ubiquitous Computing Group at Georgia Institute of Technology for their assistance, guidance, and feedback.

REFERENCES

[1] Leona S Aiken and Stephen G West. 1990. Invalidity of true experiments: Self-report pretest biases. *Evaluation review* 14, 4 (1990), 374–390.

[2] Antti Airola, Tapio Pahikkala, Willem Waegeman, Bernard De Baets, and Tapio Salakoski. 2011. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Computational Statistics & Data Analysis* 55, 4 (2011), 1828–1844.

[3] Saima Aman and Stan Szpakowicz. 2008. Using roget's thesaurus for fine-grained emotion recognition. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*. Asian Federation of Natural Language Processing, 312–318.

[4] Silvio Amir, Glen Coppersmith, Paula Carvalho, Mario J Silva, and Byron C Wallace. 2017. Quantifying mental health from social media with neural user embeddings. *arXiv preprint arXiv:1705.00335* (2017).

[5] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. 2008. Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 7–15.

[6] Mark H Anshel. 1996. Effect of chronic aerobic exercise and progressive relaxation on motor performance and affect following acute stress. *Behavioral Medicine* 21, 4 (1996), 186–196.

[7] Activity Recognition API. 2018. https://developers.google.com/location-context/activity-recognition/. Accessed: 2018-11-01.

[8] Garmin Health API. 2018. http://developer.garmin.com/health-api/overview/. Accessed: 2018-11-01.

[9] Manager REST API. 2018. https://docs.gimbal.com/rest.html. Accessed: 2018-11-01.

[10] Dror Ben-Zeev, Emily A Scherer, Rui Wang, Haiyi Xie, and Andrew T Campbell. 2015. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric rehabilitation journal* 38, 3 (2015), 218.

[11] Stuart JH Biddle and Kenneth R Fox. 2003. The way forward for physical activity and the promotion of psychological well-being. In *Physical activity and psychological well-being*. Routledge, 166–173.

[12] Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K. D'Mello, Munmun De Choudhury, Gregory D Abowd, and Thomas Plötz. 2019. Prediction of Mood Instability with Passive Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019).

[13] Daniel Bone, Matthew S Goodwin, Matthew P Black, Chi-Chun Lee, Kartik Audhkhasi, and Shrikanth Narayanan. 2015. Applying machine learning to facilitate autism diagnostics: pitfalls and promises. *Journal of autism and developmental disorders* 45, 5 (2015), 1121–1136.

[14] Mehdi Boukhechba, Yu Huang, Philip Chow, Karl Fua, Bethany A Teachman, and Laura E Barnes. 2017. Monitoring social anxiety from mobility and communication patterns. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. ACM, 749–753.

[15] Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society* 15, 5 (2012), 662–679.

[16] Louis Bucciarelli. 2003. *Engineering philosophy*. DUP Satellite; an imprint of Delft University Press.

[17] Michelle Nicole Burns, Mark Begale, Jennifer Duffecy, Darren Gergle, Chris J Karr, Emily Giangrande, and David C Mohr. 2011. Harnessing context sensing to develop a mobile intervention for depression. *Journal of medical Internet research* 13, 3 (2011), e55.

[18] Andrew T Campbell, Shane B Eisenman, Nicholas D Lane, Emiliano Miluzzo, and Ronald A Peterson. 2006. People-centric urban sensing. In *Proceedings of the 2nd annual international workshop on Wireless internet*. ACM, 18.

[19] Luca Canzian and Mirco Musolesi. 2015. 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, 1293–1304.

[20] Larry Chan, Vedant Das Swain, Christina Kelley, Kaya de Barbaro, Gregory D Abowd, and Lauren Wilcox. 2018. Students' Experiences with Ecological Momentary Assessment Tools to Report on Emotional Well-being. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 3.

[21] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, and Yuan Tang. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2* (2015), 1–4.

[22] Farhan Asif Chowdhury, Yozen Liu, Koustuv Saha, Nicholas Vincent, Leonardo Neves, Neil Shah, and Maarten W Bos. 2021. Modeling Cyclic and Ephemeral User Behavior on Social Platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[23] Matteo Ciman and Katarzyna Wac. 2016. Individuals' stress assessment using human-smartphone interaction analysis. *IEEE Transactions on Affective Computing* 9, 1 (2016), 51–65.

[24] Glen Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in Twitter. In *Eighth international AAAI conference on weblogs and social media*.

[25] Sauvik Das and Adam Kramer. 2013. Self-censorship on Facebook. In *Seventh international AAAI conference on weblogs and social media*.

[26] Vedant Das Swain, Manikanta D. Reddy, Kari Anne Nies, Louis Tay, Munmun De Choudhury, and Gregory D. Abowd. 2019. Birds of a Feather Clock Together: A Study of Person–Organization Fit Through Latent Activity Routines. *Proc. ACM Hum.-Comput. Interact* CSCW (2019).

[27] Vedant Das Swain, Koustuv Saha, Gregory D Abowd, and Munmun De Choudhury. 2020. Social Media and Ubiquitous Technologies for Remote Worker Wellbeing and Productivity in a Post-Pandemic World. In *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 121–130.

[28] Heather M Davey, Alexandra L Barratt, Phyllis N Butow, and Jonathan J Deeks. 2007. A one-item question with a Likert or Visual Analog Scale adequately measured current anxiety. *Journal of clinical epidemiology* 60, 4 (2007), 356–360.

[29] Firstbeat Analytics All day Stress & Recovery. 2021. https://www.firstbeatanalytics.com/en/features/all-day-stress-recovery/. (2021).

[30] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

[31] Norman K Denzin. 2012. Triangulation 2.0. *Journal of mixed methods research* 6, 2 (2012), 80–88.

[32] Trinh Minh Tri Do and Daniel Gatica-Perez. 2014. Where and what: Using smartphones to predict next locations and applications in daily life. *Pervasive and Mobile Computing* 12 (2014), 79–91.

[33] Pedro Domingos. 1998. Occam's two razors: The sharp and the blunt. In *KDD*. 37–43.

[34] Jon D Elhai, Robert D Dvorak, Jason C Levine, and Brian J Hall. 2017. Problematic smartphone use: A conceptual overview and systematic review of relations with anxiety and depression psychopathology. *Journal of affective disorders* 207 (2017), 251–259.

[35] Jane Elith, John R Leathwick, and Trevor Hastie. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology* 77, 4 (2008), 802–813.

[36] Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 134.

[37] Sindhu Kiranmai Ernala, Tristan Labetoulle, Fred Bane, Michael L Birnbaum, Asra F Rizvi, John M Kane, and Munmun De Choudhury. 2018. Characterizing audience engagement and assessing its impact on social media disclosures of mental illnesses. In *Twelfth International AAAI Conference on Web and Social Media*.

[38] Sindhu Kiranmai Ernala, Asra F Rizvi, Michael L Birnbaum, John M Kane, and Munmun De Choudhury. 2017. Linguistic markers indicating therapeutic outcomes of social media disclosures of schizophrenia. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 43.

[39] Emre Ertin, Nathan Stohs, Santosh Kumar, Andrew Raij, Mustafa Al'Absi, and Siddharth Shah. 2011. AutoSense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*. 274–287.

[40] Deborah Estrin. 2014. Small data, where n= me. *Commun. ACM* 57, 4 (2014), 32–34.

[41] George S Everly and Jeffrey M Lating. 2019. The anatomy and physiology of the human stress response. In *A clinical guide to the treatment of the human stress response*. Springer, 19–56.

[42] Raihana Ferdous, Venet Osmani, and Oscar Mayora. 2018. Smartphone apps usage patterns as a predictor of perceived stress levels at workplace. *arXiv preprint arXiv:1803.03863* (2018).

[43] Kenneth R Foster, Robert Koprowski, and Joseph D Skufca. 2014. Machine learning, medical diagnosis, and biomedical engineering research-commentary. *Biomedical engineering online* 13, 1 (2014), 94.

[44] Asma Ghandeharioun and Rosalind Picard. 2017. BrightBeat: effortlessly influencing breathing for cultivating calmness and focus. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1624–1631.

[45] Martin Gjoreski, Mitja Luštrek, Matjaž Gams, and Hristijan Gjoreski. 2017. Monitoring stress with a wrist device using context. *Journal of biomedical informatics* 73 (2017), 159–170.

[46] Erving Goffman et al. 1978. *The presentation of self in everyday life*. Harmondsworth London.

[47] Mathew J Gregoski, Alexey Vertegel, Aleksey Shaporev, and Frank A Treiber. 2013. Tension Tamer: delivering meditation with objective heart rate acquisition for adherence monitoring using a smart phone platform. *The Journal of Alternative and Complementary Medicine* 19, 1 (2013), 17–19.

[48] Robert M Groves and Emilia Peytcheva. 2008. The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public opinion quarterly* 72, 2 (2008), 167–189.

[49] Gabriella M Harari, Nicholas D Lane, Rui Wang, Benjamin S Crosier, Andrew T Campbell, and Samuel D Gosling. 2016. Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science* 11, 6 (2016), 838–854.

[50] Jonathon S Hare, Paul H Lewis, Peter GB Enser, and Christine J Sandom. 2006. Mind the gap: another look at the problem of the semantic gap in image retrieval. In *Multimedia Content Analysis, Management, and Retrieval 2006*, Vol. 6073. International Society for Optics and Photonics, 607309.

[51] Juliane Hellhammer and Melanie Schubert. 2012. The physiological response to Trier Social Stress Test relates to subjective measures of stress during but not before or after the test. *Psychoneuroendocrinology* 37, 1 (2012), 119–124.

[52] Karen Hovsepian, Mustafa al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 493–504.

[53] John A Johnson. 1981. The" self-disclosure" and" self-presentation" views of item response dynamics and personality scale validity. *Journal of Personality and Social Psychology* 40, 4 (1981), 761.

[54] Timothy P Johnson and Fons JR Van de Vijver. 2003. Social desirability in cross-cultural research. *Cross-cultural survey methods* 325 (2003), 195–204.

[55] Victor Jupp. 2006. *The Sage dictionary of social research methods*. Sage.

[56] Alexandra Kavushansky, Dorit Ben-Shachar, Gal Richter-Levin, and Ehud Klein. 2009. Physical stress differs from psychosocial stress in the pattern and time-course of behavioral responses, serum corticosterone and expression of plasticity-related genes in the rat. *Stress* 12, 5 (2009), 412–425.

[57] Gina Leigh Kraft and Rachel A Roberts. 2017. Validation of the Garmin Forerunner 920 XT Fitness Watch VO 2 peak Test. *Int. J. Innov. Educ. Res* 5 (2017), 62–67.

[58] Ivar Krumpal. 2013. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity* 47, 4 (2013), 2025–2047.

[59] Vivienne MHCJ Lahaut, Harrie AM Jansen, Dike Van de Mheen, and Henk FL Garretsen. 2002. Non-response bias in a sample survey on alcohol consumption. *Alcohol and Alcoholism* 37, 3 (2002), 256–260.

[60] David Lazer and Jason Radford. 2017. Data ex machina: introduction to big data. *Annual Review of Sociology* 43 (2017), 19–39.

[61] Eun-Hyun Lee. 2012. Review of the psychometric evidence of the perceived stress scale. *Asian nursing research* 6, 4 (2012), 121–127.

[62] Uichin Lee, Joonwon Lee, Minsam Ko, Changhun Lee, Yuhwan Kim, Subin Yang, Koji Yatani, Gahgene Gweon, Kyong-Mee Chung, and Junehwa Song. 2014. Hooked on smartphones: an exploratory study on smartphone overuse among college students. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2327–2336.

[63] Aubrey Lewis. 1970. The ambiguous word" anxiety". *International journal of psychiatry* 9 (1970), 62–79.

[64] Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

[65] Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Qi Li, Jie Huang, Lianhong Cai, and Ling Feng. 2014. User-level psychological stress detection from social media using deep neural network. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 507–516.

[66] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 55–60.

[67] Gloria Mark, Shamsi Iqbal, Mary Czerwinski, and Paul Johns. 2014. Capturing the mood: facebook and face-to-face encounters in the workplace. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 1082–1094.

[68] Abraham H Maslow. 1966. The psychology of science a reconnaissance. (1966).

[69] Stephen M. Mattingly, Julie M. Gregg, Pino Audia, Ayse Elvan Bayraktaroglu, Andrew T. Campbell, Nitesh V. Chawla, Vedant Das Swain, Munmun De Choudhury, Sidney K. D'Mello, Anind K. Dey, Ge Gao, Krithika Jagannath, Kaifeng Jiang, Suwen Lin, Qiang Liu, Gloria Mark, Gonzalo J. Martinez, Kizito Masaba, Shayan Mirjafari, Edward Moskal, Raghu Mulukutla, Kari Nies, Manikanta D. Reddy, Pablo Robles-Granda, Koustuv Saha, Anusha Sirigiri, and Aaron Striegel. 2019. The Tesserae Project: Large-Scale, Longitudinal, In Situ, Multimodal Sensing of Information Workers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. 8 pages. https://doi.org/10.1145/3290607.3299041

[70] Abhinav Mehrotra and Mirco Musolesi. 2018. Using autoencoders to automatically extract mobility features for predicting depressive states. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 127.

[71] Abhinav Mehrotra, Fani Tsapeli, Robert Hendley, and Mirco Musolesi. 2017. MyTraces: Investigating correlation and causation between users' emotional states and mobile phone interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 83.

[72] Shayan Mirjafari, Kizito Masaba, Ted Grover, Weichen Wang, Pino Audia, Andrew T Campbell, Nitesh V Chawla, Vedant Das Swain, Munmun De Choudhury, Anind K Dey, et al. 2019. Differentiating Higher and Lower Job Performers in the Workplace Using Mobile Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 37.

[73] Larry W Morris and Robert M Liebert. 1970. Relationship of cognitive and emotional components of test anxiety to physiological arousal and academic performance. *Journal of consulting and clinical psychology* 35, 3 (1970), 332.

[74] Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K D'Mello, Munmun De Choudhury, Gregory D Abowd, and Thomas Plötz. 2019. Prediction of Mood Instability with Passive Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 75.

[75] Amir Muaremi, Bert Arnrich, and Gerhard Tröster. 2013. Towards measuring stress with smartphones and wearable devices during workday and sleep. *BioNanoScience* 3, 2 (2013), 172–183.

[76] Karla Klein Murdock. 2013. Texting while stressed: Implications for students' burnout, sleep, and well-being. *Psychology of Popular Media Culture* 2, 4 (2013), 207.

[77] MD Nefzger and James Drasgow. 1957. The needless assumption of normality in Pearson's r. *American Psychologist* 12, 10 (1957), 623.

[78] Jukka-Pekka Onnela and Scott L Rauch. 2016. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology* 41, 7 (2016), 1691.

[79] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54, 1 (2003), 547–577.

[80] Thomas Plötz. 2021. Applying Machine Learning for Sensor Data Analysis in Interactive Systems: Common Pitfalls of Pragmatic Use and Ways to Avoid Them. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–25.

[81] Rachael Purta, Stephen Mattingly, Lixing Song, Omar Lizardo, David Hachen, Christian Poellabauer, and Aaron Striegel. 2016. Experiences measuring sleep and physical activity patterns across a large college cohort with fitbits. In *Proceedings of the 2016 ACM international symposium on wearable computers*. ACM, 28–35.

[82] Mashfiqui Rabbi, Shahid Ali, Tanzeem Choudhury, and Ethan Berke. 2011. 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, 385–394.

[83] Ulrike Rimmele, Roland Seiler, Bernard Marti, Petra H Wirtz, Ulrike Ehlert, and Markus Heinrichs. 2009. The level of physical activity affects adrenal and cardiovascular reactivity to psychosocial stress. *Psychoneuroendocrinology* 34, 2 (2009), 190–198.

[84] John Rooksby, Alistair Morrison, and Dave Murray-Rust. 2019. Student Perspectives on Digital Phenotyping: The Acceptability of Using Smartphone Data to Assess Mental Health. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 425.

[85] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[86] Avi Sadeh, Giora Keinan, and Keren Daon. 2004. Effects of stress on sleep: the moderating role of coping style. *Health Psychology* 23, 5 (2004), 542.

[87] Koustuv Saha, Ayse E. Bayraktaroglu, Andrew T. Campbell, Nitesh V. Chawla, Munmun De Choudhury, Sidney K. D'Mello, Anind K. Dey, Ge Gao, Julie M. Gregg, Krithika Jagannath, Gloria Mark, Gonzalo J. Martinez, Stephen M. Mattingly, Edward Moskal, Anusha Sirigiri, Aaron Striegel, and Dong Whi Yoo. 2019. Social Media As a Passive Sensor in Longitudinal Studies of Human Behavior and Wellbeing. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. ACM, Article CS12, 8 pages. https://doi.org/10.1145/3290607.3299065

[88] Koustuv Saha, Larry Chan, Kaya De Barbaro, Gregory D Abowd, and Munmun De Choudhury. 2017. Inferring mood instability on social media by leveraging ecological momentary assessments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 95.

[89] Koustuv Saha, Ted Grover, Stephen M. Mattingly, Vedant Das swain, Pranshu Gupta, Gonzalo J. Martinez, Pablo Robles-Granda, Gloria Mark, Aaron Striegel, and Munmun De Choudhury. 2021. Person-Centered Predictions of Psychological Constructs with Social Media Contextualized by Multimodal Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 32 (mar 2021), 32 pages. https://doi.org/10.1145/3448117

[90] Koustuv Saha, Manikanta D Reddy, Stephen M Mattingly, Edward Moskal, Anusha Sirigiri, and Munmun De Choudhury. 2019. LibRA : On LinkedIn based Role Ambiguity and Its Relationship with Wellbeing and Job Performance. *Proc. ACM Hum.-Comput. Interact.* CSCW (2019).

[91] Koustuv Saha, Manikanta D Reddy, Vedant Das Swain, Julie M Gregg, Ted Grover, Suwen Lin, Gonzalo J Martinez, Stephen M Mattingly, Shayan Mirjafari, Raghu Mulukutla, et al. [n.d.]. Imputing Missing Social Media Data Stream in Multisensor Studies of Human Behavior. *linguistic analysis* 19, 29 ([n. d.]), 43.

[92] Virginia Sandulescu, Sally Andrews, David Ellis, Nicola Bellotto, and Oscar Martinez Mozos. 2015. Stress detection using wearable physiological sensors. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer, 526–532.

[93] Akane Sano and Rosalind W Picard. 2013. Stress recognition using wearable sensors and mobile phones. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 671–676.

[94] Hillol Sarker, Matthew Tyburski, Md Mahbubur Rahman, Karen Hovsepian, Moushumi Sharmin, David H Epstein, Kenzie L Preston, C Debra Furr-Holden, Adam Milam, Inbal Nahum-Shani, et al. 2016. Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 4489–4501.

[95] Hans Selye. 1976. Stress without distress. In *Psychopathology of human adaptation*. Springer, 137–146.

[96] Moushumi Sharmin, Andrew Raij, David Epstien, Inbal Nahum-Shani, J Gayle Beck, Sudip Vhaduri, Kenzie Preston, and Santosh Kumar. 2015. Visualization of time-series sensor data to inform the design of just-in-time adaptive stress interventions. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 505–516.

[97] Graciela E Silva, James L Goodwin, Duane L Sherrill, Jean L Arnold, Richard R Bootzin, Terry Smith, Joyce A Walsleben, Carol M Baldwin, and Stuart F Quan. 2007. Relationship between reported and measured sleep times: the sleep heart health study (SHHS). *Journal of Clinical Sleep Medicine* 3, 06 (2007), 622–630.

[98] Meredith M Skeels and Jonathan Grudin. 2009. When social networks cross boundaries: a case study of workplace use of facebook and linkedin. In *Proceedings of the ACM 2009 international conference on Supporting group work*. ACM, 95–104.

[99] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 12 (2000), 1349–1380.

[100] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*. 2951–2959.

[101] Marcantonio M Spada, Ana V Nikčević, Giovanni B Moneta, and Adrian Wells. 2008. Metacognition, perceived stress, and negative emotion. *Personality and Individual Differences* 44, 5 (2008), 1172–1181.

[102] Charles D Spielberger. 2013. The effects of anxiety on complex learning. *Anxiety and behavior* (2013), 361–397.

[103] Firstbeat Technologies Stress and Recovery White Paper. 2014. https://assets.firstbeat.com/firstbeat/uploads/2015/11/Stress-and-recovery_white-paper_20145.pdf. Accessed: 2019-04-01.

[104] John Torous, Patrick Staples, Meghan Shanahan, Charlie Lin, Pamela Peck, Matcheri Keshavan, and Jukka-Pekka Onnela. 2015. Utilizing a personal smartphone custom app to assess the patient health questionnaire-9 (PHQ-9) depressive symptoms in patients with major depressive disorder. *JMIR mental health* 2, 1 (2015), e8.

[105] Gert Van Dijck and Marc M Van Hulle. 2006. Speeding up the wrapper feature subset selection in regression by mutual information relevance and redundancy analysis. In *International Conference on Artificial Neural Networks*. Springer, 31–40.

[106] Stephen Voida, Donald J Patterson, and Shwetak N Patel. 2014. Sensor data streams. In *Ways of Knowing in HCI*. Springer, 291–321.

[107] Rui Wang, Min SH Aung, Saeed Abdullah, Rachel Brian, Andrew T Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Michael Merrill, Emily A Scherer, et al. 2016. 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*. 886–897.

[108] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Ubicomp*. ACM, 3–14.

[109] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing twitter" big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, 587–592.

[110] Travis A Wearne, Abbie Lucien, Emily M Trimmer, Jodie A Logan, JacquelineA Rushby, Emily Wilson, Michaela Filipčíková, and Skye McDonald. 2019. Anxiety sensitivity moderates the subjective experience but not the physiological response to psychosocial stress. *International Journal of Psychophysiology* 141 (2019), 76–83.

[111] Daniel A Weinberger, Gary E Schwartz, and Richard J Davidson. 1979. Low-anxious, high-anxious, and repressive coping styles: psychometric patterns and behavioral and physiological responses to stress. *Journal of abnormal psychology* 88, 4 (1979), 369.

[112] Howard M Weiss. 2002. Deconstructing job satisfaction: Separating evaluations, beliefs and affective experiences. *Human resource management review* 12, 2 (2002), 173–194.

[113] J Mark G Williams, Thorsten Barnhofer, Catherine Crane, Dirk Herman, Filip Raes, Ed Watkins, and Tim Dalgleish. 2007. Autobiographical memory specificity and emotional disorder. *Psychological bulletin* 133, 1 (2007), 122.

[114] Firstbeat Technologies Automated Fitness Level (VO2max) Estimation with Heart Rate and Speed Data. 2014. https://hubloq.net/cloud/runlog/diverse/white_paper_VO2max_11-11-20142.pdf. (2014).

[115] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2, 1-3 (1987), 37–52.

[116] Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K Villalba, Janine M Dutcher, Michael J Tumminia, Tim Althoff, Sheldon Cohen, Kasey G Creswell, J David Creswell, et al. 2019. Leveraging routine behavior and contextually-filtered features for depression detection among college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–33.

[117] Cha Zhang and Yunqian Ma. 2012. *Ensemble machine learning: methods and applications*. Springer.

[118] Rhong Zhao and William I Grosky. 2002. Bridging the semantic gap in image retrieval. In *Distributed multimedia databases: Techniques and applications*. IGI Global, 14–36.