Imputing Missing Social Media Data Stream in Multisensor Studies of Human Behavior

Koustuv Saha[1]*, Manikanta D. Reddy[1], Vedant Das Swain[1], Julie M. Gregg[2], Ted Grover[3], Suwen Lin[4], Gonzalo J. Martinez[4], Stephen M. Mattingly[4], Raghu Mulukutla[5], Kari Nies[3], Pablo Robles-Granda[4], Anusha Sirigiri[6], Dong Whi Yoo[1], Pino Audia[6], Andrew T. Campbell[6], Nitesh V. Chawla[4], Sidney K. D'Mello[2], Anind K. Dey[7], Kaifeng Jiang[8], Qiang Liu[9], Gloria Mark[3], Edward Moskal[4], Aaron Striegel[4], and Munmun De Choudhury[1]

[1]*Georgia Institute of Technology,* [2]*University of Colorado, Boulder,* [3]*University of California, Irvine,* [4]*University of Notre Dame,* [5]*Carnegie Mellon University,* [6]*Dartmouth College,* [7]*University of Washington,* [8]*Ohio State University,* [9]*University of Texas at Austin*

Contact Author: *koustuv.saha@gatech.edu

***Abstract*—The ubiquitous use of social media enables researchers to obtain self-recorded longitudinal data of individuals in real-time. Because this data can be collected in an inexpensive and unobtrusive way at scale, social media has been adopted as a "passive sensor" to study human behavior. However, such research is impacted by the lack of homogeneity in the use of social media, and the engineering challenges in obtaining such data. This paper proposes a statistical framework to leverage the potential of social media in sensing studies of human behavior, while navigating the challenges associated with its sparsity. Our framework is situated in a large-scale in-situ study concerning the passive assessment of psychological constructs of 757 information workers wherein of four sensing streams was deployed — bluetooth beacons, wearable, smartphone, and social media. Our framework includes principled feature transformation and machine learning models that predict latent social media features from the other passive sensors. We demonstrate the efficacy of this imputation framework via a high correlation of 0.78 between actual and imputed social media features. With the imputed features we test and validate predictions on psychological constructs like personality traits and affect. We find that adding the social media data streams, in their imputed form, improves the prediction of these measures. We discuss how our framework can be valuable in multimodal sensing studies that aim to gather comprehensive signals about an individual's state or situation.**

***Index Terms*—social media, imputation, multisensor, wellbeing**

I. Introduction and Background

Understanding *why* and *how* individuals feel, think, and act is a key topic of interest among researchers from a variety of academic disciplines, such as psychiatry, psychology, sociology, economics, and anthropology [22]. Typically, studies of human behavior have largely relied on self-reported survey data. In recent years, several limitations have been noted with these approaches, for example, survey data suffers from subjective assessments, recall and hindsight biases. These surveys are often retrospective in nature — information is gathered after an event or experience [50].

A variety of active and passive sensing technologies overcome such biases by recording psychological states and behavior in-the-moment [4]. However, such approaches require diverse, extensive, and rich data via a variety of complementary sensors to provide comprehensive information about an individual's state and context [4]. However, it is not all the sensing modalities are always present for an individual, for instance, active sensing techniques such as ecological momentary assessments (EMAs), suffer from compliance issues, and are difficult to implement longitudinally at scale. Many of these limitations are overcome by passive sensing, such as logging device use [13], [39], [54]. However, despite the dense, high fidelity data that they capture, passive sensing paradigms alone are still challenged by resource and logistical constraints; thus being limited to capturing behavioral data only during the study period [43]. Such drawbacks could be overcome by leveraging the social media data of an individual. Social media provides an inexpensive and unobtrusive means of gathering both present and historical data [35], overcoming some of the challenges posed by active and passive sensing, and providing complementary information about an individual in their natural settings [7], [36]. Further, being self-recorded, social media data also serves as a *verbal* sensor to understand the psychological dynamics of an individual [37].

However, the availability and quality of retrospective data via social media widely vary depending on social media use behavior. Passive consumption is often more prevalent than active engagement, leading to sparsity in data over extended periods of time. Consequently, studies either focus on a very active participant cohort — hurting *generalizability* and *recruitment*, and introducing *compliance bias*, or disregard those with no or only limited social media data — hurting *scalability*. Additionally, everybody is not on social media, and its use is typically skewed towards young adults [27]. Yet many sensing studies focus on other demographics where social media is less prevalent. Further, gathering social media data also presents engineering challenges due to platform-specific restrictions, thereby, posing significant challenges in long-term longitudinal studies of human behavior.

This paper, therefore, makes a case to overcome the challenges of missing sensing streams (here, social media) in multimodal studies of human behaviors. The premise of this work is theoretically grounded in the Social Ecological Model [5], that posits human behaviors have social underpinnings. It suggests that behaviors can be deeply embedded in the complex interplay between an individual, their relationships, the communities they belong to, and the societal factors.

In particular, we examine: *How to leverage the potential of social media data in multimodal sensing studies of human behavior, while mitigating the limitations of acquiring this unique data stream?* We address this question within [Anonymized] project, a multisensor study that aims to predict

psychological constructs using longitudinal passive sensing data of 757 information workers.

Focusing on those participants whose social media data is not available, this paper proposes a statistical framework to model the latent dimensions which could have otherwise been derived, had their social media data stream been available. Specifically, we impute *missing social media features* by learning their observed behaviors from other passive sensor streams (bluetooth beacons, wearable, and smartphone use). We employ a range of state-of-the-art machine learning models, such as linear regressions, ensemble tree-based regression, and deep neural network based regression. After having demonstrated that the imputed social media features closely follow actual social media features of participants (average correlation of 0.78), we evaluate the efficacy of this social media imputation framework. We compare pairs of statistical models that predict a range of common (or benchmark) individual difference variables (psychological constructs like personality, affect, and anxiety) — one set of models being those that use imputed social media features alongside other passive sensor features, and the other set that does not use these imputed signals. Our findings suggest that the imputed social media features significantly improve the predictions by 17%.

Summarily, this paper shows that our proposed framework can augment the range of social-ecological signals available in large-scale multimodal sensing studies, by imputing latent behavioral dimensions, when one sensor stream (that is, social media data stream) is *entirely unavailable* for certain participants. We discuss the implications of our work as a methodological contribution in multimodal sensing studies of human behavior, within the sensing research community.

II. Related Work

Social Media as a Passive Sensing Modality. With the ubiquity of smartphones and wearables, passive sensing modalities enable convenient means to obtain dense and longitudinal behavioral data at scale [54], [55]. However, such a data collection is prospective — after enrollment, during the study period. To obtain historical or before-study data, researchers have recently started to use social media as a "passive sensor", which enables unobtrusive data collection of longitudinal and historical data of individuals that were self-recorded [35], [36].

Social media provides low-cost, large-scale, non-intrusive means of data collection. It has the potential to comprehensively reveal naturalistic patterns of mood, behavior, cognition, and psychological states, both in real-time and across longitudinal time [12]. Relatedly, social media has facilitated analyzing personality traits and their relationship to psychological and psychosocial well-being, through machine learning and linguistic analysis [19], [28], [42].

Together, passive sensing modalities in conjunction propagate the vision of "people-centric sensing" [4], although each one of them may have its own limitation. Social media suffers from data sparsity issues, and it can function as a "sensor" only on those who use it. This leads to a common problem that many multimodal sensing studies of human behavior face [21], [35], [54]— they either examine a larger pool of participants with fewer sensors, or a smaller pool of participants who comply with all sensing streams. This compromises the combined potential of multiple sensors or the wide spectrum of individual behaviors. Our work is motivated by computational approaches to infer latent behavioral attributes [8], [26], [30]. We model latent behavioral states as captured by multimodal sensing to impute the missing sensing stream.

Data Imputation Approaches in Sensing Studies. Data imputation is the process of replacing *missing* data with substituted values [47]. Imputation techniques commonly include *dropping missing data*, *substituting with mean or median values*, *substituting with random values*, etc [29]. These approaches are typically employed during data cleaning and pre-processing, and their downstream influence in the results largely remain understudied being overshadowed by the objectives of the studies. A number of studies have used statistical and machine learning based modeling techniques to impute missing values [6], [40], [41], [51]. In an early work, [33] proposed probabilistic approaches to handle missing data, and recently Jaques *et al.* used deep learning to impute missing sensor data and found better mood prediction results [17].

Although addressing missing data challenges has been studied in the literature, problems surrounding *missing sensing streams* remain understudied. Besides proposing a framework to impute a missing stream (social media), this paper shows the effectiveness of this imputation through the lens of predicting psychological constructs (a problem that has been widely studied in the multimodal sensing literature) through a variety of algorithms. We demonstrate the robustness in the imputation efficacy by comparing our findings with permutation tests and random- and mean- based imputation techniques.

III. Study and Data

Our dataset comes from the [Anonymized] study, that recruited 757 participants[1] who are information workers in cognitively demanding fields (e.g. engineers, consultants, managers) in the U.S. The participants were enrolled from January 2018 through July 2018. The study is approved by Institutional Review Board at the researchers' institutions.

The participants responded to self-reported survey questionnaires, and provided us their passively sensed behavioral data through four major sensing streams, bluetooth, wearable, smartphone agent, and social media [34]. They were provided with an informed-consent document with descriptions of each sensing streams and the data being collected via them. They were required to consent to each sensing streams individually, and they could opt out of any stream. The data was de-identified and stored on secured databases and servers physically located in one of the researcher institutions, and had limited access privileges.

The enrollment process consisted of responding to a set of initial survey questionnaires related to demographics (age, gender, education, and income). The participant pool consists of 350 males and 253 females, where the average age is 34 years (stdev. = 9.34). In education, the majority of the participants belong to have college (52%) and master's degree (35%) education level. Participants were additionally required to answer an initial set of survey questionnaires that measure their self-reported assessments of personality, cognitive ability, affect, anxiety, stress, sleep, physical activity, and alcohol and

[1]Note that this is an ongoing study and this paper uses sensed data collected until August 21^{str}, 2018 [23]. Randomly selected 154 participants has been "blinded at source", and their data is put aside only for external validation at the end of the study. The rest of the paper only concerns the data of the remaining 603 "non-blinded" participants in the study.

tobacco use. Relevant to the focus of the present paper, we outline the psychological constructs below:

Personality. The BFI-2 scale [45] measures personality traits across the five dimensions of personality traits on a continuous scale of 1 to 5. In our dataset, the average value of neuroticism is 2.46 (std. = 0.78), conscientiousness is 3.89 (std. = 0.66), extraversion is 3.44 (std. = 0.68), agreeableness is 3.87 (std. = 0.56), and openness is 3.82 (std. = 0.61).

Affective Measures. The PANAS-X scale [56] measures the positive and negative affect values on a continuous scale between 10 and 50 each. The STAI-Trait scale [46] measures anxiety on a continuous scale between 20 and 80. In our dataset, positive and negative affect averages at 34.61 (std. = 5.95) and 17.47 (std. = 5.34) respectively, and anxiety averages at 38.11 (std. = 9.29).

To passively collect data about participants' behavior, our study deployed four major sensor streams:

Bluetooth Beacons. Participants were provided with two static and two portable bluetooth beacons (Gimbal [3]). The static beacons were to be placed at their work and home, and the portable beacons were to be carried continuously (e.g., keychains). The beacons track their presence at home and work, and also help us assess their commute and desk time.

Wearable. Participants were provided with a fitness band (Garmin Vivosmart [2]), which they would wear throughout the day. The wearable continually tracks health measures, such as heart rate, stress, and physical activity in the form of sleep, footsteps, and calories lost.

Smartphone Application. The participants' smartphones (android and iPhones) were installed with a smartphone application (also used in [54]). This application tracks their phone use such as lock behavior, call durations, and uses mobile sensors to track their mobility and physical activity.

Social Media. Participants authorized access to their social media data through an Open Authentication (OAuth) based data collection infrastructure that we developed in-house. Specifically, we asked permission from participants to provide their Facebook and LinkedIn data, *unless they opted out, or did not have either of these accounts*. We asked consent from only those participants who had existing Facebook or LinkedIn accounts from before the study.

Passively Sensed Data. The participants were enrolled over 6 months (February to July 2018) in a staggered fashion, averaging at 111 days of study per participant. Table I reports the descriptive statistics of the number of days of passively sensed data that we collected per participant through each of the sensor streams. Per participant, we have an average of 42 days data through bluetooth beacons, 108 days data through wearable, and 101 days of data through a phone application.

Out of the 603 non-blinded participants, 475 authorized their Facebook data. This data can be broadly categorized in two types—ones that were self-composed (e.g., writing a status update or checking into a certain location), and ones that they received on shared updates on their timeline. Comprehensively, Facebook data consists of the updates on participants' timelines, including textual posts, Facebook apps usage, check-ins at locations, media updates, and the share of others' posts. The likes and comments received on these updates on the participants' timelines were also collected. Note that as per our IRB approval, we did not collect any multimedia data or private messages. Table II summarizes the descriptive statistics of our Facebook dataset. Temporally, our data dates back to October 2005, and the number of days of data per participant averages at 1,898 days — giving us a sense of the historical data that Facebook allows us to capture.

Table I: Descriptive statistics of # days data collected.

Type	Range	Mdn.	Std.
Study Period	16:205	99	46.7
Bluetooth	1:159	37	32.6
Wearable	5:206	94	46.9
Smartphone	1:206	93	52.4
Social Media	110:4756	2923	1474

Table II: Descriptive statistics of the Facebook dataset.

Type	Mdn.	Std.
Likes Rcvd.	1,139	5,277.85
Comms. Rcvd.	316	1,383.69
Self-posts	137	511.80
Self-comments	55	334.16
Self-Words	2,374	13,718.56

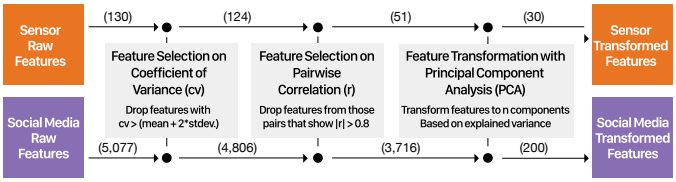


Figure 1: A schematic overview of the feature engineering pipeline to obtain transformed features from sensor and social media derived features (The numbers in brackets represent the number of features in each step in our dataset).

IV. Feature Engineering

We derive 130 features from sensor data and 5,077 features from social media data. The choice of our features is motivated by prior work on predicting psychological constructs [54]:

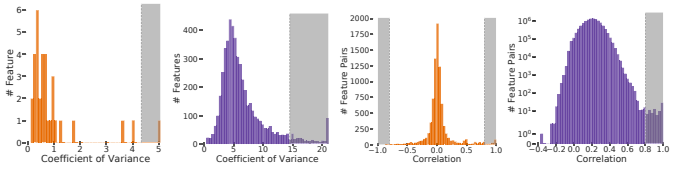
Sensor Raw Features. From the sensor datastreams, we obtain a variety of features that broadly correspond to heart-rate and heart-rate variability, stress, fitness, physical activity, mobility, phone use activity, call use, desk activity, and sleep.

Social Media Features. From the social media dataset, we obtain a variety of features corresponding to psycholinguistic attributes [49], open vocabulary n-grams (top 5,000), sentiment, and social capital (such as number of check-ins, engagement and activity with friends, etc.).

We conduct feature selection and transformation to overcome problems related to multi-collinearity, covariance, etc. among the features — issues that can potentially affect downstream prediction tasks [10]. Because our features are obtained from multimodal data streams, there is a high likelihood that some features are related, or are redundant, or show high variance, or lack predictive power [14], [30]. For example, the activity and stress-related features as captured by our wearable, are both intuitively and theoretically correlated [2]. We adopt three techniques of reducing the number of features and consequently transforming them:

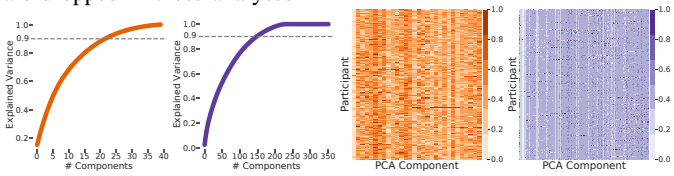
1) Selecting Features on Coefficient of Variation: First, we reduce the feature space on the basis of explained variance using the measure of coefficient of variation (cv), that essentially quantifies the ratio of standard deviation to the mean for each feature. We drop those features that are outliers in the cv (beyond threshold cv of two standard deviations away from mean). Six sensor features occur above the threshold cv of 8.6, and 271 social media features show a cv greater than the threshold cv of 14.5. Dropping these features, our feature space reduces to 124 sensor derived features and 4,806 social media derived features (Fig. 2 (a&b)).

2) Selecting Features on Pairwise Correlations: Correlated features typically affect or distort machine learning prediction models by potentially yielding unstable solutions or masking



(a) Sensor: cv (b) SM: cv (c) Sensor: r (d) SM: r

Figure 2: Feature selection stage using Coefficient of Variance (cv) and Correlation (r). The greyed-out region include those features that are dropped in these analyses



(a) (b) (c) (d)

Figure 3: (a&b) Scree plots of explained variance and number of PCA components to transform features in sensor and social media feature space. These plots help us to determine the number of components. (c&d) Transformed features' distribution across participants.

the interactions between significant features [14]. To select uncorrelated features, on the above 124 sensor and 4,806 social media features, we obtain Pearson's correlation between every feature pairs. With a threshold absolute value of 0.8, we drop those features that are highly correlated with another feature. Fig. 2 (c&d) plot the correlations of all the 15,376 (124^2) sensor feature pairs, and 23,097,636 ($4{,}806^2$) social media feature pairs. 73 sensor feature pairs and 1,090 social media feature pairs occur outside the absolute correlation of 0.8 — leading to exclusion of 73 sensor features and 1,090 social media features. At the end of this step, we are left with 51 sensor features and 3,716 social media features.

3) Transforming Features using Principal Component Analysis: On the above features, we employ Principal Component Analysis (PCA) using a singular value decomposition solver [53], where we select the number of components on the basis of explained variance [15]. This method reduces the dimensions in the feature space by transforming features into orthogonal or principal components [18], [58]. Fig. 3 (a&b) plot the *scree plots* of the explained variance of the principal components in the feature space. We find that 95% of the feature space is roughly explained at 30 principal components in the sensor features space, and 200 principal components in the social media feature space. *Note that we build the PCA models only on the training samples, and transform the features in the held-out samples with the PCA models. This way there is no data leakage in our statistical framework.*

Finally, our final feature set consist of 30 sensor-derived features and 200 social media-derived features.

V. Feature Learning Framework

Our feature learning framework broadly addresses the challenge of missing social media data stream for 128 participants in the study. Fig. 4 shows a schematic overview of the prediction models of psychological constructs that are used to evaluate the effectiveness of the imputing missing social media transformed features. We briefly mention the three algorithms that we consistently use throughout the paper.

Linear Regression (LR) Linear regression adopts a linear approach to model the relationship between the independent

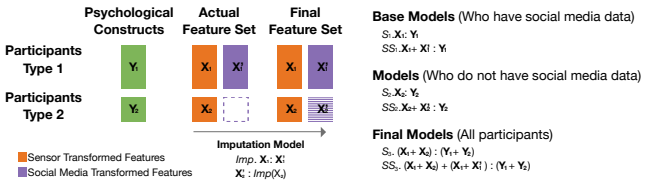


Figure 4: A schematic overview of the statistical models built to evaluate the effectiveness of imputation.

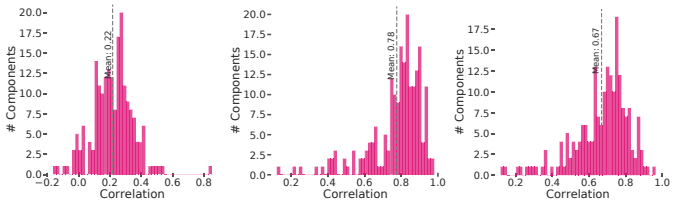
and dependent variables [44]. Specifically, wherever applicable, we employ linear regression with L1/L2 regularization to prevent overfitting and to avoid bias introduced due to the inter-dependence of independent variables [59].

Gradient Boosted Regression (GBR) Gradient boost technique conducts regression in the form of an ensemble of weak prediction models, which are typically decision trees [11], [24]. It optimizes the cost function by iteratively choosing a function that points in the negative gradient direction. In our case, we used gradient boost on an ensemble of decision tree regressors, by varying the number of decision trees between 100 and 1000, with each tree of maximum depth as 3.

Multilayer Perceptron Regression (MLP) Neural network regression suits in problems where a more conventional regression model cannot fit a solution. We use the multi-layered perceptron (MLP) technique that works in a feed-forward fashion (no cycles) with multiple internal layers [32]. The model learns through a method called backpropagation [20], and follows a fully connected (dense) deep neural network architecture. Wherever applicable, we use two internal layers and tune the number of nodes in them between 36 and 216 for our neural network regression models.

Our choice of the above three algorithms is motivated by the fact that they essentially cover a broad spectrum of algorithm families spread across linear regression, non-linear regression, decision trees, ensemble learning, neural networks, and deep learning. We quantify the prediction accuracy of psychological constructs in terms of the Symmetric Mean Absolute Percentage Error (SMAPE), which is computed as a mean percentage relative difference between predicted and actual values, over an average of the two values [16]. SMAPE values range between 0% (minimum error) and 100% (maximum error), and lower values of error indicate better predictive ability. To obtain these, we first divide their datasets into five equal segments, and then iteratively train models on four of the segments to predict on the held-out fifth segment. We average the testing accuracy metrics to obtain the pooled accuracy metrics for the above algorithms. This paper refers to this technique as *pooled accuracy technique* and the corresponding outcomes as *pooled accuracy or error measures*. Within the training segments, we tune the hyper-parameters using a k-fold cross-validation ($k = 5$) technique.

Baseline Prediction with Passively Sensed Data. We first seek to establish if the presence of social media features improves prediction accuracy. On the same set of 475 participants who have social media data, we compare two models of predicting psychological constructs — 1) $\mathtt{S}_1$ uses 30 sensor features, and 2) $\mathtt{SS}_1$ combines 30 sensor features and 200 social media features. Table III reports the relative decrease in error for $\mathtt{SS}_1$ compared to $\mathtt{S}_1$. The relative decrease in error averages at 21% for LR, 26% for GBR, and 21% for



(a) LR (b) GBR (c) MLP

Figure 5: Correlation distribution between PCA Components and Predicted PCA Components of Facebook

MLP. In sum, adding social media features improves the predictions by an average of 22.4% across all the models and the psychological constructs.

A. Imputing Missing Social Media Features

The baseline prediction suggests that adding social media features *indeed* improves the prediction task of the psychological constructs. However, about one-quarter of the participants do not have social media data (see section III). This restricts us from leveraging a rich feature stream to predict such attributes for these individuals. To overcome this constraint, we aim at *learning* certain latent behaviors that we could have otherwise inferred if we had access to their social media data.

We impute the social media features using the sensor features. For this, we build learning models on the sensor stream of the social media participants to predict their latent social media dimensions. That is, for every 200 social media feature, we build a separate model that uses the sensor features as the independent variables to predict the social media feature. We adopt k-fold cross-validation based hyper-parameter tuning. We use LR, GBR, and MLP to find the best algorithmic model, and quantify the pooled accuracy of the prediction models in terms of Pearson's correlation (r) between actual and predicted social media features. Levene's test between all the actual and predicted features reveals homogeneity of variance in the feature set [25]. This statistically indicates that the imputed social media transformed features are not arbitrarily generated.

Fig. 5 plots the distribution of the pooled Pearson's correlation (r) between the actual and predicted values of social media transformed features. We find that the mean correlation across the components is 0.22 in LR, 0.78 in GBR, and 0.67 in MLP. All of these correlation measures are statistically significant at $p < 0.05$. Comparing across the algorithms, GBR performs the best in predicting the latent social media dimensions. For the rest of the analyses, we used the GBR algorithm to impute the social media transformed features.

B. Evaluating the Effectiveness of Imputation

On those 128 participants whose social media data we did not have, we compare two prediction models of psychological constructs— 1) $\mathtt{S}_2$uses only sensor features of these participants, and 2) $\mathtt{SS}_2$ combines sensor features and imputed social media features (as obtained above).

We compare the accuracy metrics of $\mathtt{S}_2$ and $\mathtt{SS}_2$ to deduce if imputing the social media features improves our task of predicting psychological constructs. Table III compares the prediction errors (SMAPE) for the three algorithms that we run in each of the models $\mathtt{S}_2$ and $\mathtt{SS}_2$. We find that for LR, the relative decrease in the error ranges between 6% (for openness) and 17% (for positive affect), averaging at 11%; for GBR, the relative decrease in the error ranges between 16% (anxiety) and 20% (extraversion), averaging 17%; and for MLP, the relative decrease in the error ranges between 6% (extraversion) and 21% (anxiety). Therefore, the imputed social media features improved the prediction by an average of 14% across all models and measures.

Table III: Relative % decrease in SMAPE in prediction models using both sensor & social media features from ones using only sensor features. Positive values mean better prediction in $\mathtt{SS}_n$ than $\mathtt{S}_n$.

	$\mathtt{SS}_1$-$\mathtt{S}_1$			$\mathtt{SS}_2$-$\mathtt{S}_2$			$\mathtt{SS}_3$-$\mathtt{S}_3$		
Psy. Construct	**LR**	**GBR**	**MLP**	**LR**	**GBR**	**MLP**	**LR**	**GBR**	**MLP**
Personality Traits (BFI-2)									
Extraversion	10.6	28.4	16.6	8.4	20.1	6.4	12.8	19.5	3.6
Agreeableness	8.3	27.5	30.4	5.9	17.9	17.2	3.2	14.4	20.2
Conscientious.	11.8	26.0	28.2	9.4	17.4	13.5	15.0	21.2	12.1
Neuroticisim	11.2	24.9	17.6	7.6	16.9	13.4	6.0	17.5	-13
Openness	10.0	25.1	33.8	6.1	15.6	16.9	5.4	15.3	3.1
Affective Measures									
Pos. Affect	33.8	26.2	8.06	16.6	18.1	18.4	8.6	14.5	21.5
Neg. Affect	38.8	24.7	24.04	16.1	15.7	9.7	8.4	11.8	16.4
Anxiety (STAI)	39.4	24.3	7.5	14.1	15.7	20.8	6.4	16.8	34.4
Mean	20.5	25.9	20.8	10.5	17.2	14.5	8.2	16.4	12.3

Finally, on our entire dataset, we build two (*Final Models*) to evaluate the overarching effectiveness of imputation— 1) $\mathtt{S}_3$ incorporates sensor features of all participants, 2) $\mathtt{SS}_3$ incorporates Facebook features of all participants. In this model, for those who have Facebook data, we use their Facebook features, and for the rest, we use their imputed Facebook features.

We compare the prediction accuracy of the $\mathtt{SS}_3$ and $\mathtt{S}_3$— this gives us an estimate of how this sort of imputation framework influences the overarching task of predicting psychological constructs in multimodal studies (see Table III). We find an average improvement in prediction by 8.2% in LR, 16.4% in GBR, and 12.3% in MLP.

C. Hypothesis Tests for Robustness

After evaluating our imputation models, we measure its robustness. We compare the effectiveness of our imputed sensing stream against two other imputation approaches applied to those 128 participants without social media data.

Mean Imputation. This approach imputes social media features as the mean value of the corresponding feature sets. We build prediction models of psychological constructs as described in the previous subsections. This method draws on prior studies which adopted similar approaches of imputing missing features using static measures of central tendencies, such as mean or median of the feature sets [9].

Randomized Imputation. This approach imputes the social media transformed features as random values from the corresponding feature sets. For robustness, we repeat such a randomization for a 1000 times, and in each case compare the prediction accuracy with the *Final Model* $\mathtt{S}_3$. This method emulates a permutation test [1], and checks for robustness of the imputation effectiveness testing for the null hypothesis that the randomly imputed sensor streams are better than that imputed by our statistical framework.

Fig. 6 shows the SMAPE of these models as compared to that by $\mathtt{S}_3$. While our imputation shows an average improvement in SMAPE by 16% on the *Final Model* ($\mathtt{S}_3$) (see Table III), the same improvement for *Mean Imputation*-based model is -3.10% and *Randomized Imputation*-based model is 5.34%, clearly suggesting minimal (or no) improvement in these two models. Permuting on the randomized imputations

(a) Mean Imp. (b) Random Imp. (c)

Figure 6: (a&b) SMAPE comparing prediction models using sensor features (S_3) vs. those using sensor and (a) mean- and (b) random-imputed features, (c) Reduction in SMAPE in several permutations of randomly imputed social media features, as compared to S_3.

a thousand times, we observe that in terms of prediction error, our imputations are *never* outperformed by the randomized imputations in those thousand permutations. Essentially, this *rejects* the null hypothesis that *our imputation is only more effective than randomly generated imputations by chance*.

In conclusion, our findings suggest that by using passively sensed multimodal data streams, we can not only impute the latent social media dimensions, but also augment these latent features in inferring psychological constructs to build better prediction models. We consistently observe similar trends in the improvement of prediction accuracies by integrating the social media features (both actual and imputed) with the sensor-transformed features.

VI. Discussion and Conclusion

Theoretical and Practical Implications. This paper proposes an analytical framework of imputing a missing sensing stream (here social media) in multimodal sensing studies. We evaluate the effectiveness of this imputation by predicting psychological constructs through a variety of state-of-the-art algorithms. At a higher level, the imputation framework is grounded on the Social Ecological Model that construes interdependence among individuals, their behavior and their surroundings and environment [38], [52]. This implies its applicability not only in theory but also in practice (context and activity as captured and observed through passive sensing modalities). Our findings reveal the robustness of imputation by comparing with permutation tests and random- and mean- imputation. We believe such a framework can potentially be used in studies where there is similar theoretical grounding (around a focus on comprehensive social ecological signals), and an opportunity to infer psychological attributes.

We find that integrating social media features improves the prediction of psychological constructs. This aligns with prior work on the potential of social media (both individually as well as in tandem with other passive sensors) in predicting these measures [7], [35], [42]. However, social media data may not be available for the participants. Our proposed imputation method addresses this gap by computing latent social media dimensions, which can be used to improve such machine learning-based prediction tasks of human behavior.

Following our framework, existing datasets that include multimodal sensing, but do not have social media streams for some participants, can now be retrained for better predictions. While our study only focuses on predicting psychological constructs, the same method can be extrapolated to predict other measures of human behavior as well. Not being limited to a single algorithm, our framework shows the consistency in the findings across a variety of algorithm families. It is not constrained by the choice of machine learning algorithms, which typically vary depending on the characteristics of the dataset and the distribution of the individual difference variables.

We believe that if there are additional sensing streams over those we consider, their features can be plugged into our framework. However, it remains interesting to study whether the additional sensors improve the imputation models. For instance, sensing technologies that capture conversations [31] among individuals in social settings would plausibly improve predicting latent social media features, on the rationale that it captures another set of dimensions in the social ecological framework — offline social interactions.

Ethical Implications. We caution against our work being misused as a methodology to surveil or infer individual behaviors. Our work intends to model latent dimensions that can assist prediction tasks in multimodal sensing studies, by being internal to the pipeline of the prediction system. However, these latent dimensions do not necessarily translate to or are indicative of actual individual behaviors on social media, and therefore such inferences cannot be drawn from the imputed social media features about the individuals.

This paper does not unpack why certain participants did not share social media data. It could be because they do not use social media, or because they do not intend to share this data for privacy reasons. Whether social media features should be imputed for the second class of individuals can constitute a debated topic. This is because such an imputation approach, when applied to make predictions of sensitive individual difference variables and incorporated into larger systems (e.g., targeted advertising), can be perceived as a violation of the very privacy considerations that spurred them to not share social media data in the first place. We envision these topics need further discussions among researchers, ethicists, and the individuals who participate in such studies.

Limitations and Future Work. There are limitations in our work, some of which open up opportunities for future research. Our findings are limited to imputing a specific type of social media data – that gathered from Facebook. Imputing social media data of other platforms, especially ones where mixed media sharing is extensive (e.g., Instagram or Tumblr) may present unique methodological challenges. In addition, because we focus on a specific cohort of participants who are information workers, whether our approach would yield similar promising imputation results in other populations needs to be explored. We, therefore, caution against making sweeping generalizable claims.

Like any other imputation approaches, our methodology is vulnerable to introduce biases within the dataset [48], [57]. Because the imputation model only learns the latent dimensions from what *it has seen*, it is unlikely to learn unknown and deviant behavioral patterns. While such occurrences are less likely to occur in a large-scale multimodal sensing studies like ours (where the participant pool is diverse), this factor should be considered in smaller scale studies or when the study population lacks representativeness and has a greater selection bias. The present work leverages a specific set, albeit a range of commercially available passive sensors. It interests future research to investigate how adding other sensing modalities can improve the imputation of social media features.

References

[1] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. Influence and correlation in social networks. In *KDD*, 2008.

[2] Garmin Health API. http://developer.garmin.com/health-api/overview/.

[3] Manager REST API. https://docs.gimbal.com/rest.html, 2018.

[4] Andrew T Campbell, Shane B Eisenman, Nicholas D Lane, Emiliano Miluzzo, Ronald A Peterson, Hong Lu, Xiao Zheng, Mirco Musolesi, Kristóf Fodor, and Gahng-Seop Ahn. The rise of people-centric sensing. *IEEE Internet Computing*, 12(4), 2008.

[5] Ralph Catalano. *Health, behavior and the community: An ecological perspective*. Pergamon Press New York, 1979.

[6] Diane J Catellier, Peter J Hannan, David M Murray, Cheryl L Addy, Terry L Conway, Song Yang, and Janet C Rice. Imputation of missing data when measuring physical activity by accelerometry. *Medicine and science in sports and exercise*, 37(11 Suppl):S555, 2005.

[7] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *ICWSM*, 2013.

[8] Mariella Dimiccoli, Juan Marín, and Edison Thomaz. Mitigating bystander privacy concerns in egocentric activity recognition with deep learning and intentional image degradation. *IMWUT*, 2018.

[9] A Rogier T Donders, Geert JMG Van Der Heijden, Theo Stijnen, and Karel GM Moons. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091, 2006.

[10] Carsten F Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J Leitão, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46, 2013.

[11] Jane Elith, John R Leathwick, and Trevor Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology*, 2008.

[12] Scott A Golder and Michael W Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 2011.

[13] Agnes Grünerbl, Amir Muaremi, Venet Osmani, Gernot Bahle, Stefan Oehler, Gerhard Tröster, Oscar Mayora, Christian Haring, and Paul Lukowicz. Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE JBHI*, 2015.

[14] Mark Andrew Hall. Correlation-based feature selection for machine learning. 1999.

[15] Robin K Henson and J Kyle Roberts. Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educ. Psychol. Meas.*, 2006.

[16] Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 2006.

[17] Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard. Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction. In *ACII*, 2017.

[18] Ian Jolliffe. Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer, 2011.

[19] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. 2013.

[20] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 2015.

[21] James Alexander Lee, Christos Efstratiou, and Lu Bai. Osn mood tracking: exploring the use of online social network activity as an indicator of mood changes. In *Ubicomp: Adjunct*, 2016.

[22] William Little, Ron McGivern, and Nathan Kerins. *Introduction to Sociology-2nd Canadian Edition*. BC Campus, 2016.

[23] 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, et al. The tesserae project: Large-scale, longitudinal, in situ, multimodal sensing of information workers. 2019.

[24] Ananth Mohan, Zheng Chen, and Kilian Weinberger. Web-search ranking with initialized gradient boosted regression trees. In *Proceedings of the learning to rank challenge*, pages 77–89, 2011.

[25] David W Nordstokke and Bruno D Zumbo. A new nonparametric levene test for equal variances. *Psicologica*, 2010.

[26] Liangying Peng, Ling Chen, Zhenan Ye, and Yi Zhang. Aroma: A deep multi-task learning based simple and complex human activity recognition method using wearable sensors. *PACM IMWUT*, 2018.

[27] Pew. pewinternet.org/fact-sheet/social-media, 2018.

[28] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. Our twitter profiles, our selves: Predicting personality with twitter. In *PASSAT and SocialCom*.

[29] Quinten AW Raaijmakers. Effectiveness of different missing data treatments in surveys with likert-type data: Introducing the relative mean substitution approach. *Educ. Psychol. Meas.*, 1999.

[30] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. Multimodal deep learning for activity and context recognition. *IMWUT*, 2018.

[31] Tauhidur Rahman, Alexander Travis Adams, Mi Zhang, Erin Cherry, Bobby Zhou, Huaishu Peng, and Tanzeem Choudhury. Bodybeat: a mobile system for sensing non-speech body sounds. In *MobiSys*, 2014.

[32] Dennis W Ruck, Steven K Rogers, Matthew Kabrisky, Mark E Oxley, and Bruce W Suter. The multilayer perceptron as an approximation to a bayes optimal discriminant function. *IEEE Trans. Neural Netw.*

[33] Hesam Sagha, José del R Millán, and Ricardo Chavarriaga. A probabilistic approach to handle missing data for multi-sensory activity recognition. In *Workshop on Context Awareness and Information Processing in Opportunistic Ubiquitous Systems at UbiComp*, 2010.

[34] Koustuv Saha, Ayse Elvan Bayraktaraglu, Andrew Campbell, Nitesh V Chawla, et al. Social media as a passive sensor in longitudinal studies of human behavior and wellbeing. In *CHI Ext. Abstracts*, 2019.

[35] Koustuv Saha, Larry Chan, Kaya De Barbaro, Gregory D Abowd, and Munmun De Choudhury. Inferring mood instability on social media by leveraging ecological momentary assessments. *IMWUT*, 2017.

[36] Koustuv Saha and Munmun De Choudhury. Modeling stress with social media around incidents of gun violence on college campuses. 2017.

[37] Koustuv Saha, Benjamin Sugar, John Torous, Bruno Abrahao, Emre Kıcıman, and Munmun De Choudhury. A social media study on the effects of psychiatric medication use. In *Proc. ICWSM*, 2019.

[38] James F Sallis and Neville Owen. *Physical activity and behavioral medicine*, volume 3. SAGE publications, 1998.

[39] Akane Sano and Rosalind W Picard. Stress recognition using wearable sensors and mobile phones. In *ACII*, 2013.

[40] Joseph L Schafer. *Analysis of incomplete multivariate data*. Chapman and Hall/CRC, 1997.

[41] Joseph L Schafer and Maren K Olsen. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, 33(4):545–571, 1998.

[42] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[43] Christie Napa Scollon, Chu-Kim Prieto, and Ed Diener. Experience sampling: promises and pitfalls, strength and weaknesses. In *Assessing well-being*, pages 157–180. Springer, 2009.

[44] George AF Seber and Alan J Lee. *Linear regression analysis*, volume 329. John Wiley & Sons, 2012.

[45] Christopher J Soto and Oliver P John. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113(1):117, 2017.

[46] Charles D Spielberger, Fernando Gonzalez-Reigosa, Angel Martinez-Urrutia, Luiz FS Natalicio, and Diana S Natalicio. The state-trait anxiety inventory. *Revista Interamerican Journal of Psychology*, 2017.

[47] Daniel J Stekhoven and Peter Bühlmann. Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 2011.

[48] Thomas R Sullivan, Amy B Salter, Philip Ryan, and Katherine J Lee. Bias and precision of the "multiple imputation, then deletion" method for dealing with missing outcome data. *Am. J. Epidemiol*, 2015.

[49] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *J. Lang. Soc. Psychol.*, 2010.

[50] Roger Tourangeau, Lance J Rips, and Kenneth Rasinski. *The psychology of survey response*. Cambridge University Press, 2000.

[51] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 2001.

[52] Jacqueline A Walcott-McQuigg, Julie Johnson Zerwic, Alice Dan, and Michele A Kelley. An ecological approach to physical activity in african american women. *Medscape women's health*, 6(6):3–3, 2001.

[53] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer, 2003.

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

[55] Weichen Wang, Gabriella M Harari, Rui Wang, Sandrine R Müller, Shayan Mirjafari, Kizito Masaba, and Andrew T Campbell. Sensing behavioral change over time: Using within-person variability features from mobile sensing to predict personality traits. *IMWUT*, 2018.

[56] David Watson and Lee Anna Clark. The panas-x: Manual for the positive and negative affect schedule-expanded form. 1999.

[57] Ian R White and John B Carlin. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in medicine*, 29(28):2920–2931, 2010.

[58] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemom. Intell. Lab. Syst*, 1987.

[59] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J. Royal Stat. Soc.: Series B*, 2005.