Towards Estimating Missing Emotion Self-reports Leveraging User Similarity: A Multi-task Learning Approach

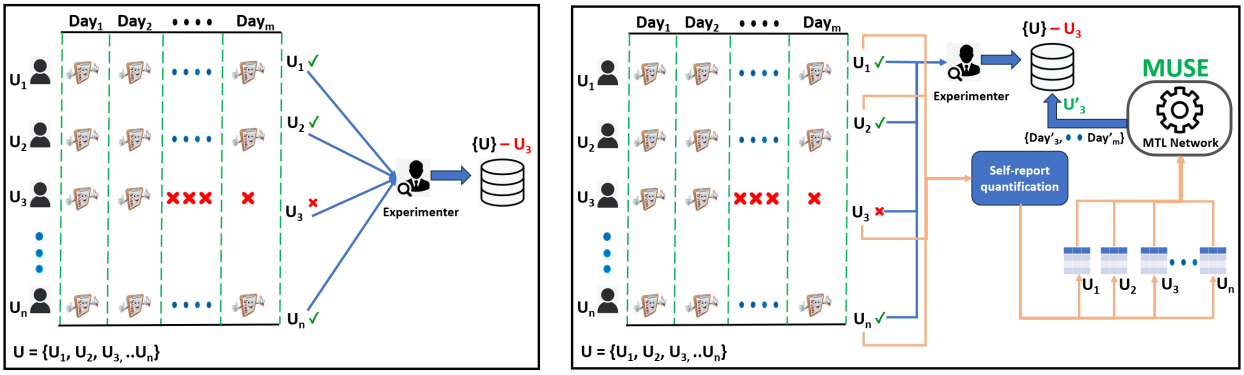
Surjya Ghosh
Computer Science and Information Systems, APPCAIR, BITS Pilani Goa
Goa, India
surjya.ghosh@gmail.com

Salma Mandi
Computer Science and Engineering
IIT Kharagpur
West Bengal, India
salmamandi@kgpian.iitkgp.ac.in

Sougata Sen
Computer Science and Information Systems, APPCAIR, BITS Pilani Goa
Goa, India
sougatas@goa.bits-pilani.ac.in

Bivas Mitra
Computer Science and Engineering
IIT Kharagpur
West Bengal, India
bivas@cse.iitkgp.ac.in

Pradipta De
Microsoft Corporation
Atlanta, USA
prade@microsoft.com



(a) Traditional ESM study (with no self-report estimation)

(b) ESM study with MUSE (for missing self-report estimation)

Figure 1: Schematic diagram comparing the ESM study in traditional approach and with MUSE (for missing self-report estimation). (a) In traditional approach of ESM study, the participants provide the emotion self-report over the study duration. However if a participant (e.g., U3) drops out in between, the researcher (or experimenter) has no option but to discard the self-reported data from the dropout participants. (b) On the contrary, in MUSE, if a participant (e.g., U3) drops in between after recording data for a reasonable time, the framework can estimate the missing self-reports (using the MTL network of MUSE) for the remaining days, thus assisting the researcher (or experimenter) to deal with the possible data loss and save from rerunning the user study.

ABSTRACT

The Experience Sampling Method (ESM) is widely used to collect emotion self-reports to train machine learning models for emotion inference. However, as ESM studies are time-consuming and burdensome, participants often withdraw in between. This unplanned withdrawal compels the researchers to discard the dropout participants' data, significantly impacting the quality and quantity of the self-reports. To address this problem, we leverage *only* the self-reporting similarity across participants (unlike prior works that apply different machine learning approaches on additional modalities) for missing self-report estimation. In specific, we propose a Multi-task Learning (MTL) framework, MUSE, that constructs the missing self-reports of the dropout participants. We evaluate MUSE in two in-the-wild studies (N1=24, N2=30) of 6-week and 8-week duration, during which the participants reported four emotions (happy, sad, stressed, relaxed) using a smartphone application. The evaluation reveals that MUSE estimates the missing emotion self-reports with an average AUCROC of 84% (Study I) and 82% (Study II). A follow-up evaluation of MUSE for an emotion inference (downstream) task reveals no significant difference in emotion inference performance when estimated self-reports are used. These findings underscore the utility of MUSE in estimating missing self-reports in ESM studies and the applicability of MUSE for downstream tasks (e.g., emotion inference).

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642833

CCS CONCEPTS

• **Human-centered computing** → **User studies**; • **Computing methodologies** → *Multi-task learning*.

KEYWORDS

Experience Sampling Method (ESM), Emotion self-report, Multi-task learning

ACM Reference Format:

Surjya Ghosh, Salma Mandi, Sougata Sen, Bivas Mitra, and Pradipta De. 2024. Towards Estimating Missing Emotion Self-reports Leveraging User Similarity: A Multi-task Learning Approach. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA*. ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3613904.3642833

1 INTRODUCTION

In recent times, many emotion-aware applications (e.g., online meeting, gaming, affective tutoring) aim to improve the user experience based on user emotion [23, 49]. These applications typically use a machine learning model to infer the user emotion. However, the major challenge to develop these machine learning models is to collate the emotion ground truth labels that are generally collected as emotion self-reports by performing a long-term user study, commonly termed as the Experience Sampling Method (ESM) [35]. As the self-reporting process is repetitive, burdensome, and time-consuming, the users often drop out from the study in between [3, 48, 53]. This unplanned dropout impacts the researchers (or study designers) as they often need to discard the partial data collected from the dropout users, which significantly deteriorate the quality and quantity of the emotion self-reports. Therefore, efficient strategies to counter this data loss are essential.

In the existing literature, broadly two types of approaches are practised to address this issue. First, *preventive* - in this approach, different types of rewards (or incentives) are provided to the users to keep them engaged in long-term user studies. For example, providing one-time rewards (e.g., monetary, gift cards) [56, 84] at the end of the study, or micro-incentives [50] during the study for making small progress are commonly followed. Other form of rewards (e.g., community reward [16], leaderboard entry [32]) are also adopted in various studies. However, researchers still face the challenge of participants dropping out in between [67]. Therefore, a second type of approach, (*remedial*) is adopted. In this approach, missing data from users who withdrew from the study (dropout participants) are constructed. For example, in ESM-based studies, peer-assisted self-report collection strategies have been proposed, which collect self-reports from a set of designated peers when self-reports from a participant is not available [4, 10]. More recently, with advances in machine learning methods, similarity among different but related tasks are explored using a multi-task learning (MTL) framework [9] to tackle such situations. The key idea of MTL is that if two or more tasks are similar, then data among these tasks can be shared to obtain better performance for the individual tasks [28]. This sharing addresses the data scarcity issue for individual tasks. In ESM study (for emotion self-report collection), the missing self-report estimation of dropped out users can be considered a task so that the data sharing among similar users (tasks) allows to estimate the missing self-reports accurately. In the existing literature, researchers have explored the MTL-based approaches that utilize the sensor data to reduce the annotation effort for different applications (e.g., HAR [31, 62]). Although effective, these approaches rely on sensor data. However, a sensor stream may not be available for a given ESM study [72], or it may have privacy issues [59], or it may incur significant resource cost (e.g., GPS) [8]. To reduce this dependency, we estimate the missing emotion self-reports by using *only* the self-reporting characteristics of similar users.

Developing an MTL-based approach by sharing data among similar users to construct the missing self-reports of dropout participants poses multiple challenges. First, the amount of emotion self-reports collected from a participant (before the participant drops out from the study) is usually small in number. As a result, it becomes challenging to train any model using the few self-reports. Thus, it is essential to engage the participants for a reasonable period during the study. Second, as the self-reports collected from the dropout participants are limited in number (because they are manually recorded, not automatically sampled from sensors), state-of-the-art machine learning models may not be able to automatically extract signatures that correlate well with the missing (to-be estimated) self-reports. Therefore, it requires a rigorous engineering effort to identify such properties. Finally, the emotion self-reports are subjective, person-specific, and vary among individuals of different regions [6, 11, 46, 54]. Due to this variability in self-reporting behavior, it becomes challenging to quantify the self-reporting patterns and therefore identify the similar users based on self-reporting characteristics.

We envision that while the challenges are significant, there are a number of emotion self-reporting characteristics (among the users participating in an ESM study) that can be leveraged to address the challenges. First, as highlighted in earlier studies, a few emotions are more frequently reported than others in an in-the-wild study [19, 38]. Therefore, even within a reasonable amount of emotion self-report dataset, the transition pattern among these frequently occurring emotions become more prevalent. Second, studies in psychology observed that different emotions persist for different amounts of time, commonly termed as the *persistence effect* of emotion [17, 75]. As a result, if the persistence period of an emotion is long, that emotion could recur repeatedly in the self-report sequence. All these properties (emotion state transitions, persistence period, and emotion recurrence) provide cues about the self-reporting pattern of a user. For example, it may be possible to observe sufficient amount self-report sequence of a user and make an estimate about the future emotion self-reports. However, as obtaining sufficient amount of data from an individual can be challenging, data sharing among similar users can be explored. Notably, all the aforementioned characteristics can be expressed quantitatively (e.g., emotion transition patterns can be expressed as transition probabilities, persistence period and emotion recurrence length can computed) to construct a user profile and used for identifying the similar users in terms of the self-reporting behavior.

We, in this paper, propose a multi-task learning based emotion self-report estimation framework, MUSE (**M**ulti-task Learning Framework for **U**ser **S**imilarity based **E**motion Self-report Estimation), that estimates the missing self-reports from dropout participants leveraging the aforementioned intuitions. It exploits the

similarity in self-reporting behavior across participants and applies a Neural Network (NN) based multi-task learning model for estimating the emotion self-reports. In specific, MUSE (a) quantifies the interaction characteristics of emotion self-reporting behavior in terms of emotion state-transition probabilities, emotion persistence times, and emotion recurrence length, and (b) allows the framework to share knowledge among similar users based on the self-reporting behavior to construct a self-report estimation model. The quantification of self-reporting characteristics allows identifying similar users and thus enabling the data sharing among the similar users using the MTL framework. We train the MTL network considering the self-report estimation of every user as a separate *task*. This allows sharing data among similar users for a more efficient learning, and reduces the individual user data requirement for estimating the missing self-reports. We present the working of MUSE in Fig. 1 using a schematic diagram. Unlike traditional approach of an ESM study, where the researcher has to discard data from a dropout participant, MUSE provides the flexibility that if a participant has recorded the emotion self-reports for a reasonable amount of time before dropping out, the missing self-reports can still be constructed using the framework. This approach counters the data loss and saves the researcher's effort from rerunning another ESM study.

We performed two in-the-wild studies to evaluate the performance of MUSE. We developed an Android application that allowed capturing and storing the emotion self-reports of study participants. We used the application as the experiment apparatus in both the studies. The application sends self-report probes (multiple times a day) to the participant to record one of the four emotions (*happy, sad, stressed, relaxed*). As we sample instantaneous feeling, we are capturing emotions. We selected these four discrete emotions (also used in earlier works [20, 58]) as they represent each quadrant of the Circumplex model [61] (i.e., having unique valence-arousal representation and any discrete emotion and its unambiguous representation on the valence-arousal plane are equivalent [43]). In the first study (Section 3), we recruited 24 university students, who participated in a 6-week data collection. The collected dataset from this study is termed as the *Homogeneous* dataset (Section 3.3). The analysis of the dataset reveals the similarity among the users in terms of the emotion self-reporting behavior (i.e., emotion transition probabilities, persistence period, and the sequence length). However, as the the profile homogeneity (all the study participants were student) of the participants can act as a confounding, we performed another study involving a diverse population. The second study (Section 7) was performed involving 30 participants with diverse profile (in terms of age, gender, geographic location, and professional background) for 8 weeks. The dataset collected from this study is termed as the *Heterogeneous* dataset (Section 7.2). The evaluation of MUSE on these datasets reveals that MUSE can estimate the missing self-reports with an average AUCROC of 84% (in the *Homogeneous* dataset) and 82% (in the *Heterogeneous* dataset). Finally, we evaluated MUSE on a downstream task (smartphone keyboard based emotion inference), which reveals no significant difference in emotion inference performance by using original or estimated self-reports. In summary, the key contributions of this paper are:

- We demonstrated that in an ESM study, the missing emotion self-reports from the dropout participants can be constructed to ease the data collection process (from the researcher's perspective). To achieve this, we propose a MTL framework, MUSE, that leverages the similarity in self-reporting behavior across participants so that the self-reports can be estimated reliably.
- We showed that self-reporting characteristics can be expressed quantitatively in terms of emotion state transition probabilities, persistence period, and emotion recurrence length. All of these help identify similar users in terms of the self-reporting behavior.
- We presented the empirical findings from two in-the-wild studies to demonstrate that MUSE can leverage the quantitatively expressed self-report behavior to estimate missing emotion self-reports (*happy, sad, stressed, relaxed*) with an average AUCROC of 84% and 82% on two different datasets (Homogeneous and Heterogeneous respectively).
- We performed an in-the-wild user study to evaluate the utility of MUSE for missing self-report estimation in context smartphone keyboard interaction based emotion detection. The evaluation based on the data collected from the user study reveals that there is no significant difference in emotion detection performance if missing self-reports are estimated (using MUSE) and used for training the emotion inference model.

2 RELATED WORKS

In this section, we discuss the related works with respect to engaging participants in long-term ESM studies to obtain reliable data from the study. We discuss the preventive approaches (to engage the participants with various incentives) and the remedial approaches (to construct the ground truth labels for dropout participants). Within remedial approaches, we highlight the utility of peer-assisted ESM and AI-based approaches for constructing the annotations reliably.

2.1 Engaging Participants in ESM Studies with Rewards

The emotion ground truths are usually collected as manual self-reports in an Experience Sampling Method study [12, 24, 35]. However, as responding to the survey questionnaires over a long time induces fatigue, causes interruption, and increases dropout rate, one of the most common approaches adopted to engage the participants over the study duration is to incentivize them. The most common from of incentive provided to the participants is monetary rewards, which has been adopted in many works [41, 42, 56, 84]. In a few studies, micro-incentives were also adopted (instead of one-time rewards), where the participants are rewarded throughout the study for completing a small task [50]. To motivate the participants in an ESM study and improve the data quality, other schemes like community reward [16], leaderboard entry [32], day reconstruction method [50], visualization of data [21, 27] are also practised. More recently, gamification was also considered a possible avenue to keep the participants engaged in long-term studies. The participants play the mobile-based game and during this process the self-reports are being automatically collected [73]. Despite the effectiveness of these approaches, researchers noted that the participants may dropout from the studies if they are not intrinsically motivated,

which affects the quality and quantity of the collected data from ESM study [67]. Therefore, to counter the data loss in ESM studies, researchers considered the remedial approaches that we discuss next.

2.2 Countering Missing Self-reports with Peer-Assisted ESM

One of the recently developed approaches to deal with missing self-reports in ESM studies is the peer-assisted ESM [4]. In this approach, if the user does not respond to an ESM probe at a given moment, the response is collected from the designated set of peers with different levels of confidence; which is later used to complement the lack of data from a participant [4, 10]. Researchers identified that by collecting highly confident responses from peers, it is possible to increase the quantity of the responses. Moreover, the presence of the peers is also found to increase the compliance rate [10]. However, the *key* requirement in this approach is the presence of a set of designated peers, from whom the responses are collected in the absence of self-reports. On the contrary, our proposed framework automatically identifies a similar group of users and uses their self-reports to train a model, which is used to estimate the future self-reports of a user (who may have been dropped in between).

2.3 Annotations with Similarity-based AI-driven Approaches

The problem of obtaining ground truths for unlabelled dataset is prevalent across domains (e.g., Human Activity Recognition (HAR), Natural Language Processing (NLP), Medical Imaging (MI)), where securing annotations is time-consuming and expensive [15, 37, 83, 85].

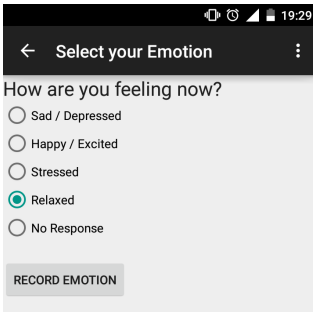
Researchers leveraged various similarity (e.g., user similarity, task similarity) based AI models to deal with this problem. For example, different types of user similarity (physical, behavioral, and sensor-data specific) have been explored in a community similarity network (CSN) so that smartphone based activity detection performance can be boosted by sharing data among similar users [33]. On other hand, similarity among various tasks have been explored by specific machine learning algorithms such as transfer learning. In transfer learning, the domain knowledge gained from one domain is applied in another domain for a similar task [51]. In case of HAR problems, Fallahzadeh et al. explored the cross-user similarity in multiple activities and construct a network-level, feature-based representation of the data in the source and target user [15]. Similarly, in UnTran, Khan et al. used the source domain's pre-trained activity model and transferred the initial two layers of the network in the target domain to generate the common feature space for both source and target domain activities [31]. More recently, in SelfHAR [69], the authors adopted a semi-supervised model to label large amount of mobile sensing data using a small amount of labelled dataset for activity recognition. The approach combines teacher-student self-training and self-supervision so that the knowledge extracted from unlabelled and labelled data can be used to reduce the labeling effort. Similarly, Saeed et al. proposed a self-supervised approach for feature learning from large amount of unalabelled sensor data so that semnatic labels are not required and the learnt feature representation can be used for down-stream HAR task [62].

In various NLP tasks (such as machine translation, affective text analysis, word sense disambiguation), crowdsourcing has been used to obtain labels from non-experts and then applying some aggregation strategy (e.g., voting policy, minimal variance method) to obtain the final ground truth [64, 65, 85]. In the case of voting policy, the score is determined as the one as agreed by half of the practitioners [36], whereas for minimal variance method, the ground truth is modeled in such a way the variance between estimation and ground truth is minimized [37]. More recently, inspired by the performance of large language models (LLM) in text-based tasks, Liu et al. applied the LLMs as a few-shot learner (so it requires fewer annotations) in context of various health tasks (cardiac signal analysis, metabolic calculation) to provide meaningful inferences using the wearable and physiological data [39].

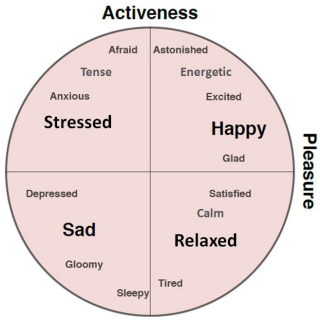
2.3.1 Leveraging User Similarity for Emotion Inference. In the domain of affective computing, a few attempts have been made to leverage user behavior similarity for emotion inference. For example, Sun et al. developed iSelf, which demonstrates that using transfer learning methods, emotion labels can be estimated from a few self-reports and other auxiliary data, like usage statistics and sensor details [68]. In [1], Alam et al. automatically label the sequence of emotions in a dyadic conversation using different feature sets such as acoustic, lexical, and psycholinguistic features. Piana et al. presented a framework that uses generic layer followed by sparse coding layer to infer emotion from different gestures [55]. Xu et al. proposed a collaborative-filtering based approach to predict the depressive symptoms among students [79]. In this approach, the authors created mobile-sensed behavior features and calculates personalized relevance weights, which are used to impute the missing label at the different time periods. While the approach aims to estimate missing labels, they rely on other sensor modalities unlike the proposed approach in this paper. Bangamuarachchi et al. explored a community level data aggregation approach exploring the mobile sensing data similarity so that the mood inference during eating can be addressed with limited personal data [2].

Multi-task learning (MTL) is a variation of transfer learning, where models are built simultaneously to perform a set of related tasks [9]. Learning multiple tasks together helps to share knowledge among similar tasks, thereby often yielding superior performance [9, 28]. A few researchers also applied Multi-task learning (MTL) for affect determination by identifying different related tasks. For example, Xia, and Liu proposed an MTL framework for recognizing continuous and discrete emotions from speech as two separate tasks [78]. In order to predict mood, stress, and health from the data collected from surveys, wearable sensors, smartphone logs, and the weather logs Taylor et al. used a personalized MTL framework [29, 70]. Similarly, to counter the issue of labeled data, an MTL based pain recognition model has been developed [40].

Key Takeaways: Summarizing the discussion of the related works, we note that while the preventive measures are effective to engage the participants in long-term user studies, users still drop out [67]. As a result, the researchers face the challenge of obtaining high quality data. To address this problem, missing labels can be constructed using similarity-based machine learning models. While these kind of approaches are performed in other domains (e.g., HAR [62], depressive symptom prediction [79]), they rely on other



(a) Self-reporting UI



(b) Circumplex model [61]

Emotion	Valence (Pleasure)	Arousal (Activeness)
Happy	High	High
Sad	Low	Low
Stressed	Low	High
Relaxed	High	Low

(c) Emotion and VA relationship

Figure 2: Experiment Apparatus - (a) the self-report UI was used to collect the emotion self-report, (b) the Circumplex model of emotion, which guides the self-report UI design, (c) relationship among the selected emotions and the valence-arousal (VA)

sensor data (e.g., mobile sensing dataset [2, 33]) for quantifying and identifying the similar users. Although inspired by these works, the proposed approach, leverage only the self-reported emotions to extract similarity characteristics implicitly for constructing the missing emotion self-reports of dropout participants.

3 USER STUDY I: HOMOGENEOUS POPULATION

In this section, we discuss the field study including data collection apparatus, participants, study procedure and the collected dataset. This work has been approved by our institute's ethics committee, and we have obtained the IRB approval prior to the user study.

3.1 Experiment Apparatus

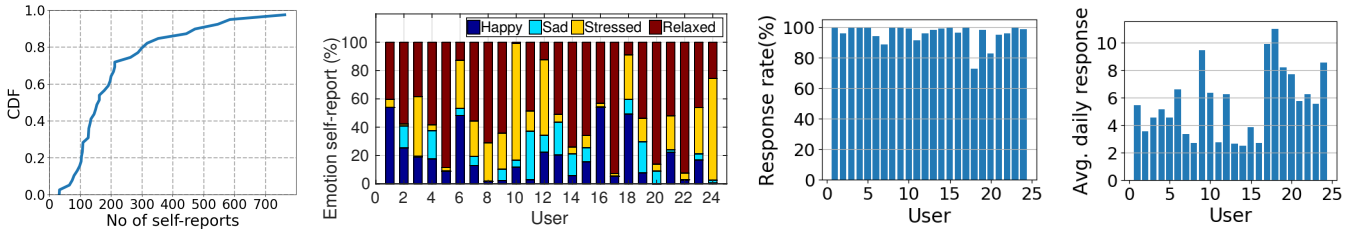
We implemented an Android-based smartphone application (Android version $\geq$ 6.0) for collecting the emotion self-reports. The self-reporting UI (Fig. 2a) consists of four emotions (*happy, sad, stressed, relaxed*); the users had to select one emotion at a time based on what they are experiencing at the moment, and press the 'Record Emotion' button to log the data. We select these emotions based on the Circumplex model (Fig. 2b) of emotion [61]. According to this model, human emotion comprises two dimensions - valence (indicating the pleasure) and arousal (indicating the activeness). As a result, the Circumplex model represents emotions in a 2D plane in four quadrants. Selecting a representative emotion from each quadrant allows to cover the different spectrum of valence and arousal. Therefore, we select these four emotions (*happy, sad, stressed, relaxed*), which belong to different quadrant of the Circumplex plane. We show the mapping between these emotions and their valence and arousal (based on the position on the Circumplex plane) in Fig. 2c. Additionally, we kept the interface simple by explicitly recording the emotion. We did not consider the intensity of perceived emotion, which can make self-reporting difficult. Notably, the same emotions have been used in earlier works for emotion modeling tasks [20, 58]. We also keep the provision of *No Response*, so that the user can skip self-reporting by selecting this option.

We issued the probe when the phone was in use (screen was unlocked). To keep the interruption low, the probe was issued once the user completed tasks in one application and launched the next one. Additionally, there was a gap of at least two hours between two consecutive probes. We selected the inter-probing interval of at least two hours as this is often considered adequate in ESM studies [86]. Collected emotion self-reports are temporarily stored on the phone and later uploaded to the server. We performed off-line analysis on this dataset.

3.2 Study Participants and Procedure

We recruited 30 university students (24 M, 6 F, aged between 20 – 33 years) for the field study. Each of these participants were awarded a gift voucher worth 10 USD. We installed the application on their smartphones and instructed them to use it for 6 weeks to record their emotions. We also informed them that they would receive a self-report pop-up at different time periods of the day. Once the self-report pop-up was delivered to the participant, it remained in the foreground. There was no timeout for the pop-up. A participant could dismiss the pop-up either by recording the emotion self-report or by swiping it away. The participant was required to record their perceived emotion from one of the four available options. The participants were further instructed that if they wished to skip answering a probe, they should select the *No Response* button instead of dismissing the pop-up. This approach allowed to deal with any inadvertent swipe on the screen and helped to calculate response rate accurately by considering the *No Response* options.

We observed that 2 participants left the study in between. Among the remaining 28 participants, 4 participants recorded less than 100 self-reports during the 6-weeks of the study. To identify the minimum number of self-reports required from a participant, we looked at the CDF of the self-reports of the 28 participants (Fig. 3a). We observed that $\approx$ 15% of the participants recorded less than 100 self-reports. In a normal distribution, as $\approx$ 15% entries are there below $\mu - \sigma$, (μ implies mean, σ implies SD) [26], we decide to use 100 self-reports as the minimum one. Notably, our goal is to discard participants having fewer self-reports, therefore, discarded participants falling below one σ (not above one σ). As a result, we did not use data from the four participants. We ran all analysis on the remaining 24 participants (20 M, 4 F).



(a) CDF of self-reports **(b) User-wise emotion self-report dist.** **(c) User-wise response rate** **(d) Avg. daily responses**

Figure 3: Homogeneous dataset details - (a) CDF of 28 users' self-reports reveal that ≈ 15% users are having less than 100 self-reports. This helps to drop four users with less than 100 self-reports and carry out the analysis on remaining 24 users. (b) User-wise distribution of emotion self-reports. All but 7 users (1, 5, 8, 16, 17, 20, 22) have recorded all four emotion states. Overall, we have 17% *happy*, 8% *sad*, 25% *stressed*, and 50% *relaxed* self-reports. (c) User-wise response rate reveals a very high response rate across users (mean: 95.75%, SD: 6.5) and (d) User-wise daily average response (mean: 5.6, SD: 2.4)

3.3 Homogeneous Dataset Description

We have collected a total of 5,677 emotion self-reports, out of which only 5% were marked as *No Response*. For all but 6 users, the amount of *No Response* labels was less than 5%. We removed all the *No Response* labels before further processing. Therefore, we have 5,393 valid emotion self-reports. We refer to this dataset as the *Homogeneous* dataset due to the profile homogeneity of the participants (all were same university student).

We obtained on average 224 (std. dev 94) valid self-reports per user. In Fig. 3b, we present the distribution of the four emotion self-reports of each user. All but 7 users (1, 5, 8, 16, 17, 20, 22) recorded four emotion states. For most of the users, *relaxed* was the most dominant emotion. We also observed that all the emotion self-reports were not uniformly distributed. The non-uniform distribution pattern has been encountered in earlier studies also due to in-the-wild nature of the studies [38]. Overall we recorded 17%, 8%, 25%, 50% self-reports labeled with *happy, sad, stressed* and *relaxed* emotion states respectively from the participant responses.

We investigate the engagement of the participants in our emotion self-report collection study in Fig. 3c and Fig. 3d. First, we define response rate as $\frac{v*100}{v+n_r}$, where v indicates the number of valid emotion labels (*happy, sad, stressed, relaxed*) and n_r indicates the number of *No Response* labels. Notably, we trigger the ESM probes judiciously (after the user completed task in an app) and instructed the participants not to dismiss the self-report pop-ups; if they are really occupied, they may select *No Response*. Hence, we expected a very few probes to be dismissed by the users, where *No Response* indicated the skipped probes. In Fig. 3c, we show that 87.5% participants have a response rate of at least 90% and obtain an average response rate of 95.75% (std dev. 6.50). We also show the average number of daily probes answered (either valid emotion label or *No Response*) by every participant in Fig. 3d. We observe that 67% of the participants have answered more than 4 probes on average on a daily basis, while all the participants have answered at least 2 probes on average on a daily basis.

4 FEASIBILITY ANALYSIS

In order to estimate the missing self-reports by sharing data among similar users, we need to find ways to identify similar users based on the self-reporting characteristics and check if these attributes can distinguish the emotions. Only by sharing data among similar users, the missing emotion self-reports can be estimated. Accordingly, we investigate the following aspects in this section - (a) quantify the self-reporting behavior of every user, (b) measure self-reporting similarity across users, and (c) distinguish emotions using the self-reporting characteristics.

4.1 Quantifying Self-reporting Behavior

We quantify self-reporting behavior of a user with the help of - (a) *emotion-state transition*, (b) *emotion persistence time* and (c) *emotion sequence length*. We decide to use these characteristics for the following reasons. First, emotion transitions indicate the probability of the next emotion from the current emotion [71]. Second, as the persistence time of an emotion indicates the continued existence of the emotion [75], therefore, if the second emotion self-report is collected within this time period, it is more likely that the same emotion is recorded. Finally, if a number of emotion self-report sequences are observed, it may be possible to find the typical recurrence length of an emotion before it changes. Next, we investigate the utility of these properties for self-report quantification.

4.1.1 Emotion State Transition. One way to represent self-reporting behavior as the *emotion-state transition*. It depicts the most likely next self-reported emotion, given the current emotion self-report. In literature, state transition is used as an indicator of future emotion [19, 71] and in related behavioral studies [47]. Precisely, we quantify the emotion state transition of every user as the probability of switching from the current emotion, to the any of the four emotions (including continuing to remain in the current emotion), in the next self-report. To compute the probabilities, we use the entire sequence of observed self-reports. We organize these probabilities in a 4×4 matrix (Fig. 4) and define it as the state-transition matrix (P). We denote the state transition probability from state x to state y using p_{xy}, where $x, y \in \{happy, sad, stressed, relaxed\}$. We compute the transition probability p_{xy} as the ratio of the total number of transitions made from emotion x to y (n_{xy}) and the total number of transitions made from emotion x to any state (n_x) (see Eq. 1). The emotion-transition matrices are calculated for every user.

$$p_{xy} = \frac{n_{xy}}{n_x} \quad (1)$$

$$P = \begin{pmatrix} p_{hh} & p_{hs} & p_{ht} & p_{hr} \\ p_{sh} & p_{ss} & p_{st} & p_{sr} \\ p_{th} & p_{ts} & p_{tt} & p_{tr} \\ p_{rh} & p_{rs} & p_{rt} & p_{rr} \end{pmatrix}$$

Figure 4: Emotion transition matrix

Sequence number	1	2	3	4	5	6	7	8
Self-report	H	H	S	R	H	H	H	T
Elapsed time (in Hr.)	2	3	2	3	6	4	2	3

Figure 5: Schematic showing the computation of emotion persistence time. For *happy* state it is the average of the average elapsed time from two blocks. The average elapsed times are 2.5, 4 respectively from the first and second block. Therefore, the persistence time is average of 2.5, and 4 i.e. 3.25 hours.

4.1.2 Emotion Persistence Effect. Self-reporting behavior can also be characterized using the *persistence effect* of emotion. The existing literature highlights that once the user feels an emotion, the feeling persists for some duration (a couple of seconds up to several hours, or even longer) [17, 75]. Some emotion like sadness lasts the longest, while shame lasts the least [74]. As we captured the self-reports with minimum interval of two hours and a few emotions can persist over hours, persistence time indicates the continuation of previous emotion. Emotion persistence gives an intuition of how much time (denoted as *persistence time*), one user has stayed in a single emotion state. To compute persistence time of a particular emotion, we define the *elapsed time*, indicating the time-duration a user stays in the present emotion state, before providing the next self-report. We consider one emotion (say, emotion x) at a time and parse the sequence of observed self-reports of the user to find the blocks, where this emotion is reported. We compute the average elapsed time from each of these blocks. To compute the persistence time of that emotion (emotion x), we compute the average of the average elapsed times (obtained from different blocks). Precisely, if there are k different blocks for emotion x, with average elapsed time of i^{th} block as t_i for $1 \leq i \leq k$, the elapsed time for x is $T_x = \frac{\sum_{i=1}^{k} t_i}{k}$.

The above-mentioned process is repeated for all the emotions. For example, in Fig. 5, the *happy* state appears in two blocks - once for 2 times (column 1, 2), and the second one for 3 times (column 5 to 7). For the first occurrence, the elapsed times in *happy* states are 2 and 3 hours, and in the second occurrence, the elapsed times are 6, 4 and 2 hours, respectively. The average elapsed times from these two blocks are 2.5, and 4 hours respectively. Therefore, the persistence time of *happy* state is the average of 2.5, and 4 hours, i.e. 3.25 hours. In this way, we compute the emotion-wise persistence time of all the users.

4.1.3 Emotion Sequence Length. To quantify the self-reporting behavior, we also use *emotion sequence length* as an attribute. This feature captures the typical sequence length of a specific emotion self-report, i.e. once a user reports an emotion, how many times do they continue reporting the same emotion. To compute this, we parse the sequence of observed self-reports and identify the number of times a user has reported the same emotion at-a-stretch. In specific, if there are k different sequences for an emotion (say x), with the length of i^{th} sequence as d_i for $1 \leq i \leq k$, the sequence length for x is $D_x = \frac{\sum_{i=1}^{k} d_i}{k}$. For example, in Fig. 5, the chain of *happy* emotion appears twice - once for 2 times (column 1, 2), the second time for 3 times (column 5 to 7). We compute the average of these continuous sequence lengths and represent it as the sequence length for the *happy* state. As a result, the sequence length of *happy* state is average of 2, and 3 (i.e. 2.5). We repeat this procedure for all the users to compute the emotion sequence length for every emotion.

4.2 Measuring Self-reporting Behavior Similarity

Once the self-reporting behavior is quantified in terms of emotion-state transition, persistence effect and sequence length, we investigate if there is similarity in the self-reporting behavior across users. We compute the Pearson correlation coefficient between every two users' state-transition probabilities, persistence times, sequence lengths and show the heatmaps in Fig. 6a, 6b, 6c respectively.

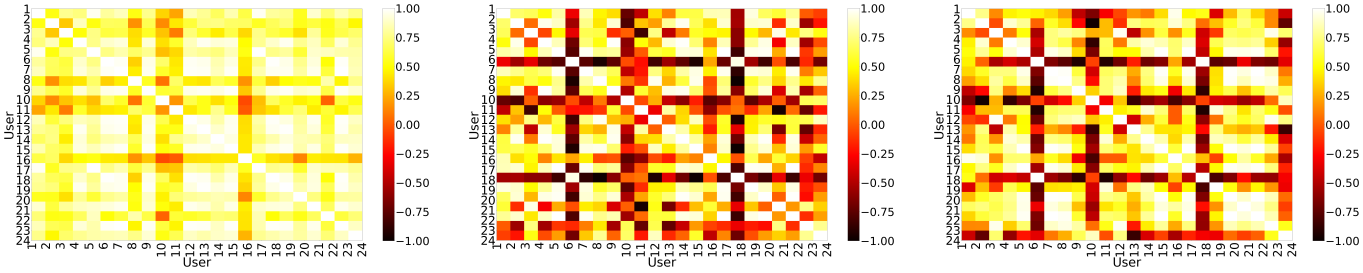
We observe that for each user, there exists a group of users exhibiting similar self-report transition probability, similar persistence effects and similar emotion sequence lengths. However, the number of these similar users may vary for every user. The *key takeaway* from these observations is that there exists a group of users with similar self-reporting behavior. Hence, the data among these users may be shared to train a model for estimating the (missing) self-reports of a user, who may have dropped from the ESM study.

4.3 Distinguishing Emotions with Self-reporting Characteristics

Once we could identify the similar users based on the self-reporting characteristics, we investigate if these self-reporting attributes could be used to distinguish different emotions (*happy, sad, stressed, relaxed*).

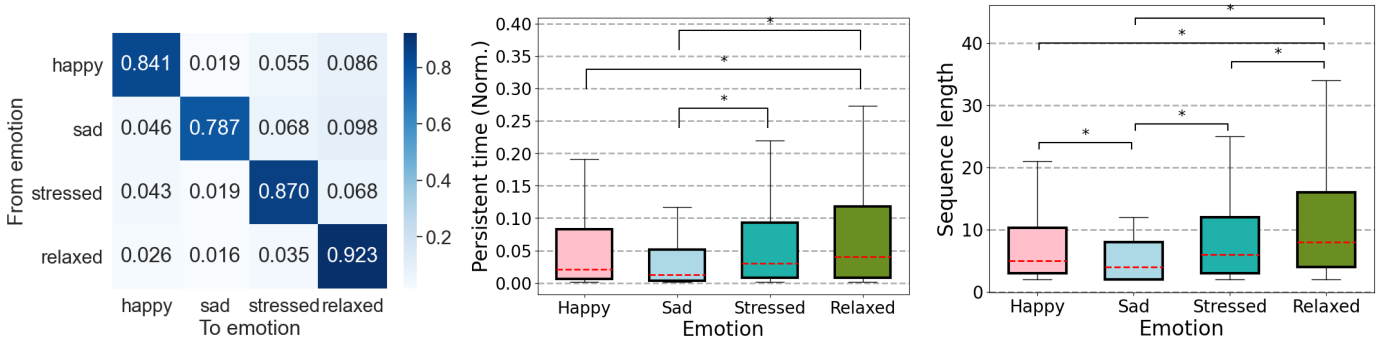
First, we investigate whether the self-report transition patterns (measured using emotion transition probabilities) are useful to detect the emotions. To investigate this, we computed the emotion transition probabilities (as per Eq. 1) in the aggregate data of the participants in Fig. 7a. We observe a heavy diagonal indicating that once a user is in a specific emotion, they are more likely to report the same emotion in the next self-report. This observation may help to determine the missing emotion self-reports.

Next, we investigate if another self-reporting characteristic i.e., persistence times vary across emotions. We grouped the persistence times values (after normalization) according to these emotions and compare them using Kruskal Wallis test [44] (Fig. 7b). The Kruskal Wallis test revealed a significant effect of emotion on



(a) Emotion-transition similarity (b) Persistence effect similarity (c) Sequence length similarity

Figure 6: Heat-maps showing the similarity in different self-reporting characteristics for every pair of users - (a) emotion transition similarities across users, (b) emotion persistence times similarity across users, and (c) emotion sequence length similarity across users.



(a) Emotion-transition probabilities (b) Persistence time comparison (c) Sequence length comparison

Figure 7: Comparison of the variations in the self-reporting characteristics across emotions. (a) the transition probabilities reveal that users are more likely to provide the same emotion self-report as the current one, (b) the persistence time (normalized) vary significantly (p<0.05) across emotions, (c) the emotion sequence lengths also vary significantly (p<0.05) across emotions.

persistence times ($\chi^2(3) = 13.17, p < 0.05$). A post-hoc test using Mann-Whitney tests with Bonferroni correction [45] showed the significant differences ($p < 0.05$) between following emotion pairs, *happy-relaxed*, *sad-stressed*, and *sad-relaxed*. Similarly, we performed the Kruskal Wallis test to identify the effect of emotions on the other self-reporting characteristic i.e., sequence length (Fig. 7c), which revelaed a significant effect with the following test statistics ($\{\chi^2(3) = 19.12, p < 0.05\}$). A post-hoc test using Mann-Whitney tests with Bonferroni correction showed the significant differences between every emotion pairs except the following one (happy-stressed). In summary, these self-reporting characteristics vary significantly across emotions, and therefore may be used to develop machine learning model for estimating the missing emotion self-reports.

5 MUSE FRAMEWORK

In this section, we discuss the architecture of the MUSE framework (see Fig. 8). We use Multi-task Learning (MTL) to estimate the emotion self-reports as it leverages well on the concept of generalization to return superior performance if the tasks are related [60]. In our case, learning a self-report prediction model for every user is considered a *task*. So, the MTL internally shares knowledge (training dara) among related tasks (users) and estimates the emotion self-reports for every user. Next, we discuss in detail the input to the framework and the self-report estimation model.

5.1 Input Data Representation

The inputs to the framework are a combination of emotion self-reporting characteristics (as discussed in Section 4.1) - (i) emotion state transition probabilities (ii) emotion persistence period (iii) emotion recurrence lengths. The emotion-transition probabilities indicate the likelihood of current emotion based on the previous one, whereas persistence effect indicates the time spent in a given emotion state. Both of these are combined to obtain an input vector (Influence$_e$ ($\mathbb{F}_e$)). The emotion recurrence length (Sequence_length ($\mathbb{L}$)) is considered as another input. These two features (Table 1) are the input to the self-report estimation model. We discuss the computation of these two inputs next.

5.1.1 Influence$_e$ ($\mathbb{F}_e$). Influence$_e$ ($\mathbb{F}_e$) captures the self-reporting pattern in terms of both emotion transition and emotion persistence. Precisely, it measures the influence of the previous emotion e' on

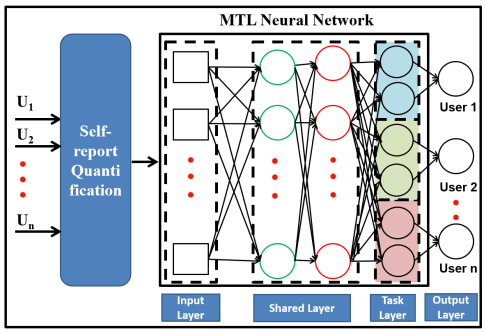


Figure 8: The architecture of the MUSE framework. The self-reports of all users are quantified and fed to the MTL network, which *implicitly* leverages user similarity and returns self-report prediction model for every user.

Parameter name	Parameter description
Influence$_e$ ($\mathbb{F}_e$)	Influence of self-report state e on current instance, where $e \in \{happy, sad, stressed, relaxed\}$
Sequence_length ($\mathbb{L}$)	Number of times the same self-report is recorded at a stretch as noted in current instance

Table 1: Inputs to the self-report estimation model

the current emotion e, where $e, e' \in \mathbb{E} = \{happy, sad, stressed, relaxed\}$. Based on the emotion transition behavior, it is possible to identify the most probable emotion e in the current emotion from the previous emotion e' (Section 4.1, Emotion State Transition). Moreover, a past emotion $e \in \mathbb{E}$ has an impact on the present one, based on the time elapsed between past self-report e and current self-report e (Section 4.1, Emotion Persistence Effect). Intuitively lower the elapsed time, higher will be the impact of the past self-report – e. Hence, the parameter $\mathbb{F}_e$ is designed to capture the influence of a previous self-report state $e' \in \mathbb{E}$ on current self-report e as the product of - (a) probability of the current emotion is determined as e based on the previous emotion e' and (b) normalized elapsed time since last observed emotion self-report, e. Mathematically, we express this as follows,

$$\mathbb{F}_e = p_e * (1 - \tau_e) \quad | \quad e \in \mathbb{E} \tag{2}$$

where p_e indicates the probability of the current emotion being determined as e based on the previous self report and τ_e indicates the normalized elapsed time from the last reported e state.

We apply discrete-time Markov Chain [66] to compute the probability (p_e) of the current emotion state e based on the previous self-report. Consider a 4−size vector (for *happy, sad, stressed, relaxed*) e_{n-1}, which represents the previous emotion self-report (at instance $(n-1)$) in terms of one-hot encoding [22]. We multiply e_{n-1} with the emotion transition matrix P (computed for every participating user following Eq. 1) to estimate the vector e_n, which denotes the probabilities of obtaining each emotion at the current (n^{th}) self-report instance. Mathematically, we express this as follows,

$$e_n = e_{n-1}.P \tag{3}$$

We select the value of p_e by looking into the corresponding position of emotion e at current self report vector e_n.

In order to compute the normalized elapsed time τ_e for emotion $e \in \mathbb{E}$, first we construct a 4-size vector, which holds the absolute elapsed time between the current time instance and the last self-reported emotion $e \in \mathbb{E}$. We update this vector with every new emotion self-report instance, as the elapsed time changes with every new reported emotion. Finally, we select τ_e by looking into the proper position of emotion e in this vector and normalize it by dividing the same with the highest value in the vector.

5.1.2 Sequence_length ($\mathbb{L}$). This parameter captures the number of times a user provides the *same* emotion self-report at a stretch, depicting the recurrence pattern in self-reporting. In order to compute, we check if the current emotion self-report is the same as the previous one, then the sequence length $\mathbb{L}$ is incremented, otherwise, it is reset to 0 (see Section 4.1, Fig. 5).

5.2 Self-report Estimation Model

We implement a neural network based Multi-Task Learning (MTL) [9] model for emotion self-report estimation. The model predicts the individual user's self-report, which is considered as a *task*. We attempt to construct the model by (a) *shared learning*; learning features of one user (one task) using related features from other similar users (shared tasks) and by (b) *task-specific learning*; in parallel, make some portion of the model task-specific to capture the personalized behavior (specific to a user). Notably, the shared learning alleviates the issue of training data scarcity for the personalized model.

The architecture of the proposed *Multi-task Learning Neural Network* (MTL-NN) model for self-report estimation is shown in Fig. 8. We implement two hidden layers to build the Neural Network. The initial layers are used to transform the input vector to learn the generic representation, while the final layers are used to obtain user-specific representation. The initial layers are *shared* across tasks to improve learning by using the training data samples of other related tasks (similar users). This data sharing allows the model to learn a generic representation by leveraging self-reporting characteristics similarity. For example, this layer learns by sharing information among those users, who have very high persistence time in the *sad* state or among those users, who frequently move to the *relaxed* state from the *sad* state.

On the other hand, the final *task-specific layers* allow learning user-specific representations. In this layer, the traits of the individual user are taken into consideration to estimate her emotion self-reports. This layer ignores inputs from other users (as shown in different colors in Fig. 8) by assigning them small weights. This task-specific layer uses the embedding obtained from the shared layers and adds user-specific customization to generate the final output. For example, this layer leverages the generic representation for a set of users exhibiting relatively high persistence time in the *sad* state, and then adopts customization for a specific user, who shows exclusively high persistence time in the *sad* state. As a result, personalized nature of self-reporting behaviour is captured using the generic representations obtained from the shared layers.

We apply the input features as noted in Table 1 to train the MTL-NN model. Since building a model for every user is considered as a separate task, while training the model, we aim to minimize the

total loss (L_{total}). Mathematically, we express it as follows,

$$L_{total} = \sum_{\forall u \in \mathbb{U}} \phi_u * L_u \tag{4}$$

where ϕ_u is the weight for user u and L_u is a user-specific loss function. We assign equal weight for every user's loss function (ϕ_u) for simplicity. We use categorical cross entropy (as there are four emotions) as the user specific loss function (L_u). The model takes multiple users' data at one time, however, during training a single batch contains data from one user at a time. This training approach is inspired by work for mood prediction using personalized MTL strategy [29]. To take care of the ordering effect in a batch, shuffling is done. This training data organization approach helps to predict the emotion self report of that user only, and as a result, the errors made during prediction are backpropagated to learn the correct weights in every layer (shared and task-specific) of that user. In this way, by selecting each user and continuously updating the task-specific and shared weights, the MTL-NN model learns the generic and task specific representation for every user. Notably, MTL itself works as a regularization tool to avoid overfitting [60]; additionally, we apply dropout as the standard approach for regularization in neural network.

We used the following neural network configuration. The network consisted of 3 hidden layers. The number of nodes are 5, 8, and 8 respectively. Each of the layers used the ReLU activation function. The task-specific layer used softmax as the activation function (as there are four emotions). The outputs are the probabilities of the next emotion state. One-hot encoding is done on the probability values to find out the next state. The total loss is summed across all the tasks (Eq. 4), where individual loss functions are the categorical cross entropy loss functions. During training, first, we compute the transition matrix based on the observed emotion sequence. In case of training, as we know the current state, we directly use the corresponding one-hot encoding of the emotion for influence vector calculation. During testing, we predict one instance at a time. In specific, we estimated the current emotion. Once we estimate the current emotion, we multiply the corresponding one-hot vector with the transition matrix to obtain the influence vector (applying Markov Chain). Also, if the estimated current emotion matches with the previous emotion, we increase the sequence length by one. These values are used as input to the network for estimating the next state.

6 EVALUATION

In this section, first, we describe the experimental setup and then evaluate the performance of the MUSE framework.

6.1 Experiment Setup

We split the data into a 60 − 40% ratio for every user, where the initial 60% is used to train the model and the remaining 40% is used for testing. We train the MTL model using the initial 60% data combined from each user and test it on the remaining 40% data of one user at a time. This setup simulates the condition where a participant drops from the study in between (during the study, the participant provided 60% self-reports and the missing 40% emotion self-reports need to be estimated.)

Hyperparameter tuning: To select the optimal values of the hyperparameters, we perform a grid search. We try with the following (i) batch sizes (8, 16, 24), (ii) epochs (20, 30, 40, 50) and (ii) dropout rates (0.15, 0.20, 0.25, 0.25) to train the model. It is observed that for the batch size of 8, the epoch of 40, and dropout of 0.15, the best classification performance is obtained. Hence, we fix these values as the model hyperparameters.

6.1.1 Performance Metric. We use AUCROC (Area under the Receiver Operating Characteristic curve) as the performance metric to measure the performance of the self-report estimation model. We rely on this metric, as it is typically used for an unbalanced dataset, which is also the case for us. We report the weighted average of AUCROC for every user. We calculate the weighted average of AUCROC (auc_{wt}) from four different emotion self-reports as per Eq. 5, where f_i, and auc_i indicate the fraction of samples and AUCROC for emotion i respectively.

$$\begin{aligned} auc_{wt} &= \textstyle\sum_{\forall i \in \mathbb{B}} f_i \times auc_i, \text{where } \mathbb{E} \\ &= \{happy, sad, stressed, relaxed\} \end{aligned} \tag{5}$$

We decide to use the weighted average so that the sample imbalance does not inadvertently impact the overall classification performance. At the same time, we report the mean and std. deviation of AUCROC per emotion to get a fair idea how accurately each emotion is identified.

6.1.2 Baselines. We compare the performance of MUSE with the following baselines. For all the baselines, we use the initial 60% self-reports of the user for training and the remaining 40% self-reports for testing.

- **Most Represented Emotion Model (MRE):** In our dataset, most of the users have reported *relaxed* emotion as the most frequent one (see Fig. 3b). Hence, we implement this personalized baseline model, which always predicts the most frequently represented emotion of a user. We identify the most represented emotion from the training data of the user and predict it as the estimated self-report in the testing phase.
- **Top-2 Most Represented Emotion Model (MRE-2):** This is a variant of MRE, where we identify the top two frequently occurring emotions present in every user's training data. In the testing phase, the model declares an agreement whenever one of these two states is found in the test emotion label, otherwise a mismatch.
- **Sequential Model (SEQ):** We develop a personalized LSTM (Long short-term memory) [25] based model for this baseline. To construct the input to the SEQ model, we look into the self-reports used in the past 24 hours and use them as the input. As the self-reports are collected with a minimum interval of 2 hours, we use a sequence size of 12 (i.e., previous 12 time steps). However, if there are not 12 self-reports in the previous 24 hours, we pad it with the most frequently reported emotion self-report within the 24 hours. This baseline enables us to assess, whether we can estimate the missing self-reports based on only the previous self-report sequence.
- **Single-Task Learning Model (STL):** We develop a personalized DNN (deep neural network) model, where we do not use self

reports from any other users. It has the same network configuration like MUSE and it uses the same set of features (Table 1) as used in the MUSE framework based on the specific user only. This model enables us to assess if we can estimate the missing self-reports from observed *personal* self-reports only, without relying on the other users.

- **All-User Model (AU):** The proposed MTL model leverages the similarity in self-reporting behavior among users to estimate the missing self-reports. This baseline, on the contrary, overlooks the aspect of self-reporting similarity. To construct this model for a user, we aggregate the observed data (initial 60%) and the self-reports from all other users. It extracts the same of features (Table 1) from this aggregated data and trains a DNN model (using the same configuration like MUSE) for estimating the missing self-reports. Comparing performance with this baseline signifies the importance of sharing training data among similar users as performed in the MUSE framework using MTL.
- **Markov Chain Model (MKC):** We develop a personalized Markov Chain model as a baseline. In this model, for every user based on the observed self-reports (initial 60%), we construct a transition matrix. This is used to estimate the unobserved emotion self-reports, where the current emotion self-report is multiplied with the transition matrix to obtain the next self-report. This model helps to assess the role of self-report transition pattern.
- **Similar User Clustering (CST):** We also develop a baseline by combining data from the similar users using a clustering approach (similar to approach taken in [34]). In specific, for every user, based on the observed self-reports (initial 60%), we compute the transition similarity vector and then run a k-means clustering algorithm on other users' state-transition similarity vectors. To identify the optimal number of similar users, we vary the value of k and select the one having highest silhoutte score [63]. Once, the similar users are identified, data from these users are combined to train a feed forward network. It has the same network configuration like MUSE and it takes the same features (like MUSE) as input. The output layer produces a 4-dimensional output corresponding to every emotion self-report state. We apply softmax activation with cross-entropy loss for classification.
- **Tree-based Model (TBM):** We develop a personalized model using Random Forest (using 100 trees) for this baseline. In this model, for every user, based on the observed self-reports (initial 60%), we construct the self-report estimation model using the same features as outlined in Table 1. This is used to estimate the unobserved emotion self-reports. Comparison with the this model helps to assess the superiority of the MUSE framework over tree-based models.
- **Collaborative Filtering based Model (CFM):** In this baseline, first, we create the emotion transition matrix of every user observing the initial 60% emotion self-reports. We apply the k-means clustering on this profile data (transition matrix) to identify the similar users. To identify the optimal number of clusters, we vary the number of clusters (k) and select the number of cluster, which returns the highest value of the silhouette score [63]. Once we identify the similar users, we train separate personalized model for each user. Later, to make prediction for a user (say u), we input the features (as noted in Table 1) to each of the models of the users belonging to the same cluster as u and apply majority voting to obtain the final output. This approach is similar to collaborative filtering adopted by Xu et. al [79].

6.2 MUSE's Performance: Self-report Estimation on Homogeneous Dataset

In this section, first we compare the self-report estimation performance of MUSE with the baselines (on the Homogeneous dataset). Later, we report the user-wise and emotion-wise estimation performance.

6.2.1 Comparison with Baselines. We present the comparison of self-report estimation between MUSE and the baselines in Fig. 9. **MUSE** achieves an average AUCROC of 84% (std dev. 18%) by outperforming all the baselines. The most represented emotion (**MRE**) baseline exhibits significantly poor average AUCROC of 58% (std dev. 24%), whereas, the model based on the top two most representative emotions (**MRE-2**) shows an average AUCROC of 73% (std dev. 21%). The **SEQ** model and the **STL** model also perform poorly with an average AUCROC of 58% (std dev. 19%), and 61% (std dev. 16%). Notably, the **AU**, **MKC**, **CST**, **TBM**, and **CFM** baselines show relatively better performance, with average AUCROC of 75% (std dev. 20%), 79% (std dev. 11%), 78% (std dev. 21%), 74% (std dev. 22%), and 78% (std dev. 15%) respectively.

These observations demonstrate that always predicting the most frequent (or top-two frequent) self-report (**MRE** model variations) as the outcome is not a good choice. The poor performance of the **SEQ** baseline can be attributed to the scarcity in data volume, which does not allow the model to learn the variations in the self-report sequence. However, relatively better performance is obtained for the tree-based model (**TBM**). The usage of only personal self-reports (**STL** baseline) to predict future labels does not work well since there may not be enough training samples. On the contrary, aggregating self-reports across users helps to improves prediction performance as observed in the **AU** baseline. The performance improves if data from similar users are aggregated as observed in the **CST** baseline or by adopting the collaborative filtering strategy (**CFM**) or if the emotion transition patterns are leveraged as observed in the **MKC** model. Both of these (aggregating data among similar users, leveraging emotion transition pattern) captured in the MUSE framework by leveraging data from the similar users based on the self-reporting behavior (performed implicitly using the MTL), which help to obtain superior performance to estimate missing self-reports.

6.2.2 Self-report Estimation Performance. We show the user-wise and emotion-wise self-report performance of MUSE in Fig. 10. We obtain an average user-wise AUCROC of 84%, while the AUCROC is more than 60% for 83% of the participants (see Fig. 10a). Notably, for a few users (1, 3, 5, 20, and 22) in the test data, only two emotions are present and each instance of these emotions is identified correctly, boosting their auc_{wt} to 100%. On the other hand, users (11, 17) observed anomalies (major difference) in their self-reporting behavior in the test phase, dropping their AUCROC.

We show the emotion-wise AUCROC in Fig. 10b. It is observed that *relaxed* is the most accurately identified emotion (mean AUCROC - 82%), followed by the *stressed* emotion (mean AUCROC -

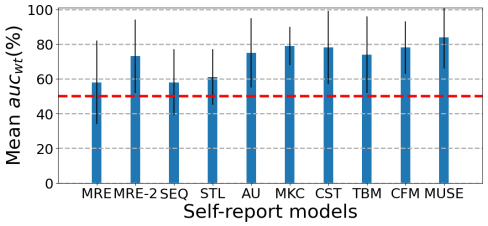
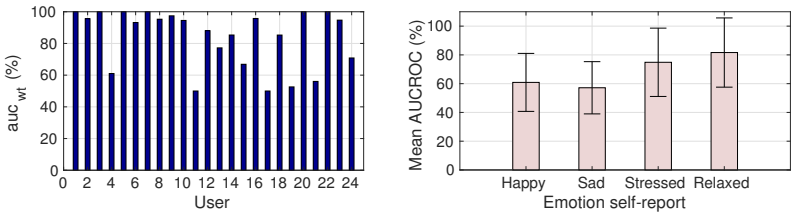
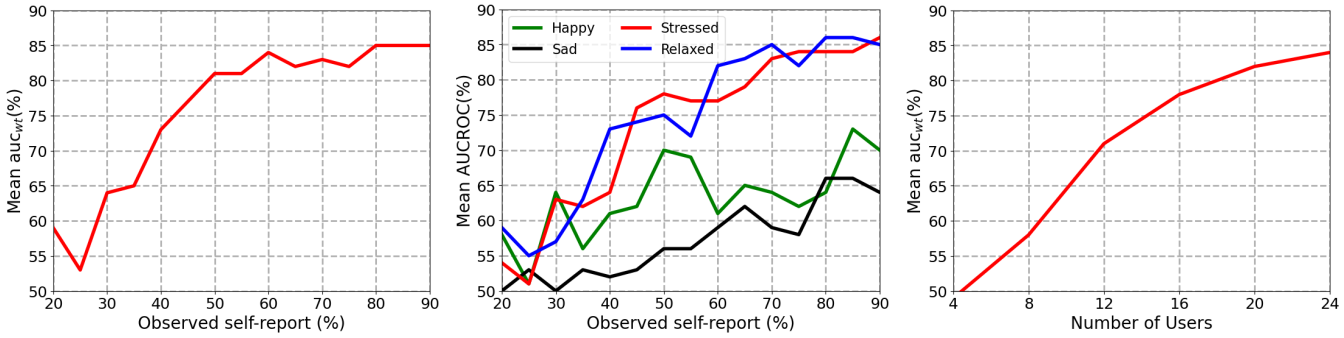


Figure 9: MUSE's emotion self-report estimation performance on Homogeneous dataset - comparison with baseline AUCROC. Error bar indicates std. dev. The red (dashed) line indicates 50% AUCROC.



(a) User-wise AUCROC **(a) Emotion-wise mean AUCROC**

Figure 10: MUSE's self-report estimation performance on the Homogeneous dataset - (a) user-wise AUCROC (b) emotion-wise AUCROC. Error bar indicates std. dev.



(a) User-wise mean AUCROC **(b) Emotion-wise mean AUCROC** **(c) User-wise mean AUCROC**

Figure 11: Change in self-report estimation performance with varying amount of observed self-reports and varying number of users. (a) Mean AUCROC gradually improves with increasing self-reports (b) Improvement in mean AUCROC is more pronounced for *stressed, relaxed* emotion (c) Mean AUCROC improves with increasing number of users

75%). The distribution of emotion self-reports (Fig. 3b) reveals that the relaxed emotion is the most frequent one, therefore, the model can learn more effectively as the number of data points are larger. As a result, the model returns the best performance for the relaxed emotion.

6.3 MUSE's Performance: Influence of Self-report Volume and Number of Users

We investigate the amount of self-reports to be observed so that the missing self-reports can be estimated reliably. This helps to understand how many self-reports from a user needs to be captured (before the participant drops out) so that MUSE can determine the missing emotion self-reports. To investigate this, we vary the number of observed self-report from 20% to 60% with an increment of 5% in each iteration. In each iteration, we train the model for the observed self-reports (i.e., 20%, 25%, ...60%) and test with the unobserved self-reports (i.e., 80%, 75%, ...40%). Notably, as MUSE needs to estimate the missing self-reports, for x% observed self-reports, it needs to estimate (100-x)% self-reports.

We show the variation in user-wise and emotion-wise AUCROC with increasing number of self-reports in Fig. 11. We observe that user-wise AUCROC improves with increasing number of self-reports initially and then it stabilizes (Fig. 11a). This is intuitive (as the amount of training data increases with increasing number of self-reports and once sufficient self-reports are collected, the performance stabilizes). We also note that for estimating the (missing) self-reports with a good AUCROC (AUCROC of $\approx$ 80%), we need to observe $\approx$ 50% self-reports (i.e., if a participant drops out even before providing half of the required self-reports, it becomes challenging to estimate the missing self-reports with a high AUCROC). Similarly, the emotion-wise AUCROC also improves with increasing self-reports (Fig. 11b) , however, the improvement is more pronounced for the *stressed, relaxed* emotions as these two emotions have a comparatively large number of samples (Section 3.3).

We investigate deeper to find out the influence of number of users on self-report estimation performance of MUSE. We vary the number of users from 4 to 24 (with an increment of 4) for a fixed amount of observed self-report (60%) and plot the user-wise AUCROC in Fig. 11c. We observe that as the number of users increases the AUCROC also improves. This can be attributed to the presence of more similar users in a larger user pool, which allows to accumulate more training data and therefore, the overall user-wise AUCROC improves.

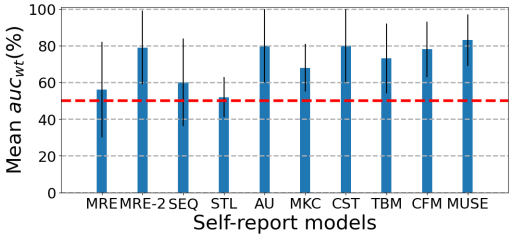
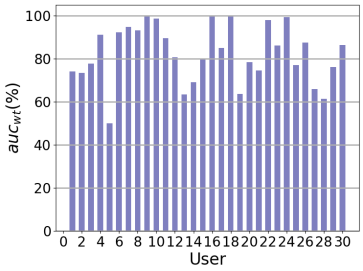


Figure 12: Comparing prediction performance with different baseline models on the Heterogeneous dataset. Error bar indicates the standard deviation. The red (dashed) line indicates 50% AUCROC.



(a) User-wise AUCROC

(a) Emotion-wise mean AUCROC

Figure 13: MUSE's self-report estimation performance on the Heterogeneous dataset - (a) user-wise AUCROC (b) emotion-wise AUCROC. Error bar indicates std. dev.

6.4 Relationship between Response Rate and Self-report Estimation Performance

In this section, we investigate the relationship between emotion self-report estimation performance and response rate. In an ESM study, a high response rate is desired [3]. Therefore, we needed to investigate if MUSE is biased towards the high responders (i.e., whether it can estimate self-reports only for the highly responsive participants). To investigate further, we compute the Pearson correlation coefficient between the participants' response rate (as defined in Section 3.3) and user-wise AUCROC of the framework. We observe a correlation value of 0.177, which indicates no strong correlation between response rate and AUCROC; thus MUSE does not favor the highly responsive participants.

7 USER STUDY II: HETEROGENEOUS POPULATION

In this section, we discuss the user study and the findings obtained by evaluating MUSE on a diverse population, comprising participants with various age groups, professional background, gender, and location.

7.1 Study Participants and Study Procedure

We carried out the 8-week in-the-wild study involving 30 participants (16 males, 14 females) with diverse profiles. We recruited the participants adopting an word-of-mouth approach [76]. The participants were in the age range of 21 – 55 years. The age distribution of the participants were as follows - 21 to 25 years (10%), 26 to 30 years (30%), 31 to 35 years (30%), 36 to 40 years (13%), 41 to 45 years (7%), 46 to 50 years (7%), and 51 to 55 years (3%). The participants took part in the study from 10 different cities in three different countries: Germany (2 cities), the Netherlands (2 cities), and India (6 cities). The participants were from a diverse professional background, such as Information Technology professional (20%), homemaker (13%), professor (10%), businessman (10%), school teacher (10%), administration (10%), manager (6.7%), graduate student (6.7%), unemployed (6.7%), researcher (3.3%), and nurse (3.3%). Each of these participants were awarded a gift voucher worth 10 USD. Notably, this study population profile is strikingly different from the Homogeneous dataset (Section 3.3), where the population profile was mostly homogeneous.

We installed the Android application (described in Section 3.1) on the smartphones of the participants. They were provided with the same guidelines as outlined in Section 3.2 to conduct the study. The participants were informed that they would receive a survey pop-up (Fig. 2a) throughout the day, where they require to record their emotions (*happy, sad, stressed, relaxed*). They were also instructed to select the *No Response* option if they would like to skip the self-reporting.

7.2 Heterogeneous Dataset Description

During this study, we collected 7314 emotion self-reports (mean: 243.8, std. dev: 164.76) from the participants. Each participant recorded at least 100 self-reports. Regarding the distribution of different emotion self-reports, overall we observed 11% *happy*, 9% *sad*, 29% *stressed*, and 50% *relaxed* emotion self-reports.

7.3 MUSE's Performance: Self-report Estimation on Heterogeneous Dataset

We use the same experimental setup as described in Section 6.1, to evaluate the performance of the MUSE on this newly collected dataset. For every user, we split the data into a 60 – 40% ratio, where the initial 60% self-reports are used to train the model and the remaining 40% self-reports are used for testing. We fix the model hyperparameters following the grid search, as described in Section 6.1.

7.3.1 Comparison with Baselines. We compare the performance of the MUSE with the baselines on the Heterogeneous dataset in this section. Like Homogeneous dataset, in this case also, we observe that MUSE outperforms all the baselines (average AUCROC of 82%, std. dev 14%) although some baselines (**CST, AU, TBM, CFM**) return good performance (Fig. 12). However, as MUSE leverages the self-reporting similarity across users using MTL, it returns the best performance.

7.3.2 Self-report Estimation Performance. We present the emotion self-report estimation performance of the MUSE framework in Fig. 13. We show the user-wise AUCROC of the MUSE framework in Fig. 13a,. We obtain an average user-wise auc_{wt} of 82% (std.

dev 14%), where 63% of the participants achieve auc_{wt} more than 80%. Emotion-wise AUCROC is reported in Fig. 13b, where we observe that *relaxed* state is identified with the highest AUCROC (mean AUCROC 85%), followed by *stressed, happy,* and *sad* state respectively. The results obtained on this Heterogeneous dataset are consistent with the findings on the Homogeneous dataset, as reported in section 6.2. For instance, the emotions present in a comparatively large number (say, *stressed, relaxed*) are estimated more accurately compared to other emotions (say, *happy, sad*).

These findings demonstrate that MUSE performs equally well on a heterogeneous population, comprising participants from various age groups, professional backgrounds, and locations; thus applicability of the proposed approach while dealing with diverse group of participants in an ESM study.

7.4 MUSE's Performance: Transferability across Different Study Population

We next evaluate the transferability of the MUSE framework to assess how well it can generalize across different participant profile for estimating the same emotion self-reports. In specific, we train the proposed model on the Heterogeneous dataset (Section 7.2) and test it on the Homogeneous dataset (Section 3.3). To perform the evaluation, we select one user at a time (say m) from the Homogeneous dataset (test), include the user's (m) initial 60% self-reports along with the self-report details of the users of the Heterogeneous dataset to train the model, and estimated the user's (m) remaining 40% emotion self-reports using the constructed model. This approach enables us to evaluate the transferability (on a different population - Homogeneous dataset) of the proposed model, which is constructed using a separate, relatively diverse population (Heterogeneous dataset). We observe an average user-wise AUCROC of 89% (std dev. 10%). This result is better than the earlier findings, where the self-report estimation performance on the Homogeneous dataset is 84% (Fig. 10), and on the Heterogeneous dataset is 82% (Fig. 13). This improvement suggests that by training the MTL model on a larger, diverse dataset, superior performance can be obtained; thus demonstrating the generalizability of the MUSE framework across different group of study population.

8 USE CASE OF MUSE: SMARTPHONE KEYBOARD INTERACTION BASED EMOTION DETECTION

In this section, we demonstrate the utility of MUSE for smartphone-keyboard interaction-based emotion detection. We selected smartphone-keyboard interaction-based emotion detection as the downstream task for the following reasons: (a) due to the overwhelming usage of various IM (instant messaging) applications (where frequently emotions are expressed), keyboard interaction has become an useful modality for emotion inference [58, 77], (b) due to the ubiquity of smartphone, it allows in-situ sampling of human behavior (in our case emotion self-report) efficiently [3, 52].

8.1 User Study and Dataset

We performed a 2-week in-the-wild study involving 14 participants (10 M, 4 F). The participants were rewarded with a gift voucher worth 5 USD. We obtained the IRB approval from the institute prior to the study. We developed an Android keyboard app, which keeps track of typing interactions (not actual text) and collects emotion self-reports (*happy, sad, stressed, relaxed*). We discuss the keyboard interaction tracking and self-report collection process next.

In this user study, we track the user's typing session, which is defined as the time spent at-a-stretch in a single app (before changing to the next one). To track the typing interactions, we developed an Android smartphone keyboard (as shown in Fig. 14a) using the Android IME (Input Method Editor) functionality[1]. We capture the timestamp of every key press event in a session. This allows us to measure typing features like session duration, typing speed, and session length. Additionally, we track if the backspace (or delete) key was pressed in a typing session. This allows measuring the typing mistake rate. Furthermore, we keep a track of special characters inserted by the user. We consider any non-alphanumeric character (except backspace and delete) inputted in a session as a special character.

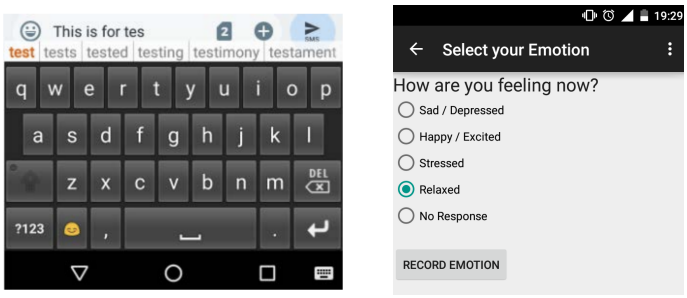
We also collect the user's current emotion as soon as a typing session is over (i.e., typing is done in an app and the user changes the application) via the self-report collection UI (Fig. 14b). The user reports one of the four emotion choices as shown in Fig. 14b. We use the same interface (as used for self-report collection, Section 3). We collected the same emotions as proposed in the MUSE framework (Section 5).

The participants installed the Android keyboard app on their personal phone, and used it for their regular typing activities. The users were instructed that they would receive a pop-up based on typing activities done throughout the day. In the pop-up, they had to report the current emotion. They were also instructed to select the *No response* option if they would like to avoid self-reporting at the given point of time. We collected 2115 typing sessions from the study. On average, there are 151.07 sessions (SD: 122.08) per user.

8.2 Emotion Inference Modeling

We use a Random Forest based emotion classification model. The model is personalized because typing interactions vary across individual [14, 19]. The model uses five features (typing speed, error rate, special character usage rate, session duration, session length) to determine the user's emotion during a typing session. These five features have been calculated as follows: (f_1): typing speed - The average elapsed time between two consecutive typing events in a session is considered as the typing speed. (f_2): error rate - The fraction of characters deleted in a session is considered as the error rate. (f_3): session duration - The total elapsed time in a session is the session duration. (f_4): session length - The number of characters typed in a session is considered as the session length. (f_5): special character usage rate - The fraction of special characters in a session is considered the special character usage rate. We use these five features as these are effective for smartphone keyboard based emotion inference [18, 19]. We used the weighted F-score as the classification performance metric.

[1] http://tinyurl.com/y54c69yf



(a) App keyboard (b) Self-reporting UI

Figure 14: Experiment apparatus used in the user study - (a) app keyboard (b) emotion self-report collection UI.

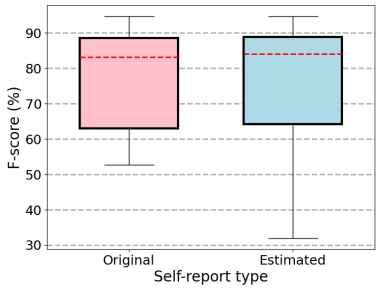


Figure 15: Comparing emotion detection performance using the original self-reports and the estimated elf-reports. No significant difference is observed in the user-wise F-score using Mann-Whitney U-test.

8.3 Evaluation

To perform the downstream task of emotion inference, we train a machine learning model combining keyboard interaction features and emotion self-reports. We train the model in two ways. First, we train the model using initial 80% of the actual self-reports (as provided by the users), and test the model using the remaining self-reports (20%). Second, to simulate the missing self-reports from a user, we adopted the following approach. We used the initial 40% self-reports to estimate the next 40% self-reports using MUSE. Now, this 80% (40% actual, and 40% estimated using MUSE) self-reports are used to train the model, and the remaining 20% self-reports (the last 20% of the actual) are used for testing. This evaluation approach ensures that the model is tested on the same dataset, while trained in different ways (once with original self-reports, and once with the estimated self-reports).

We compare the user-wise F-score using both these approaches in Fig. 15. We observe that in the first case (the model is trained using original self-reports), the average F-score is 76.8% (SD: 15.1%), whereas in the second case (the model is trained using estimated self-reports for the missing ones) the average F-score is 74.1% (SD: 21.5%). This implies that if estimated self-reports are used for the missing self-reports, there is not much difference in the performance of the downstream task. Since the obtained F-score values are not normally distributed ($p < 0.05$ with Shapiro-Wilk test [81]), we perform the paired Mann-Whitney U-test. However, we did not observe a significant effect of self-report types on the F-score values (U = 96.0, p = 0.47). These findings indicate that if the missing self-reports are estimated using MUSE and those are used as ground truth for the downstream task, we obtain similar performance (as obtained using original self-reports) for the downstream task.

9 DISCUSSION

The experimental findings demonstrates that the MUSE framework can estimate the missing emotion self-reports of dropout participants, thus assisting the researchers in HCI community to deal with data collection challenges in an ESM study. However, deploying the proposed framework for emotion self-report collection studies or other ESM studies needs to consider a few aspects, which we discuss next. We also highlight the limitations of the framework.

9.1 Implications of the Findings

The major implication from the current findings is that MUSE reduces the data collection overhead of the researchers (or study designers), who face the challenge of limited data because of unplanned dropout of users in an ESM study. If the users have reported sufficient number of self-reports before dropping out, the model can estimate the missing self-reports accurately. Another key takeaway from the study is that MUSE performs efficiently for same task (i.e., estimates the same missing emotion self-reports) across different study population if the training data is large and diverse (as observed in Section 7.4). While this confirms the generalizability of the framework across different study population, the generalizability in other aspects (e.g., different emotion self-report estimation, different ESM study) needs to be considered. Finally, the utility of MUSE in missing self-report estimation for smartphone keyboard interaction based emotion detection (Section 8) underscores the effectiveness of proposed approach for downstream tasks.

9.2 Generalizability of the MUSE Framework

In this paper, we demonstrated that MUSE can estimate four emotion self-reports (*happy, sad, stressed, relaxed*) collected using the UI (Fig. 2a) in an ESM study. However, whether the same approach can be used for (a) other emotions (b) other ESM studies and (c) large diverse population is discussed next.

First, we do not foresee any major challenge while extending the proposed approach to more number of emotion choices (beyond just 4 emotions), multiple-choice questions, etc. Additionally, it is possible to implement other scales like Self-assessment Manikin (SAM) [5], Ekman's six basic emotion model [13], Plutchik's emotion wheel [57] in the self-report study design. However, if the number of emotions increases, the emotion transition matrix will also increase in size and can result into a sparse transition matrix. To address this, we may need to collect more number of self-reports of different emotions. The key idea of applying MUSE for other emotion self-report studies would be to quantify the emotion self-reporting pattern and use them as input to the framework.

Second, we also envision that the proposed framework can be extended to other ESM studies (beyond emotion self-reports). The crux of the MUSE framework lies in modeling the emotion state

transitions, which we observe from the sequence of obtained self-reports. For any ESM study (in another domain), if the experimenter can figure out the underlying transition pattern among different self-reports (for that domain), the concepts of this framework can be applied. For example, a fitness routine related ESM study, which recommends future exercises to a player, could adopt the proposed framework, since fitness exercises follow a regular transition pattern (as people work-out with different body-parts at a certain intervals).

Finally, the current findings on a reasonably large and diverse profile of participants demonstrate the efficacy of MUSE (Section 7). However, recent findings highlight that sensor-based complex models may not generalize well in longitudinal studies across country, and participant profile [46, 80]. From that perspective, as the proposed approach does not use any sensor data and relies only on self-reporting characteristics, identifying similar users is easier and therefore, can be effective in a large and diverse population.

9.3 Deployment Considerations

We discuss several factors that need to be considered for deploying MUSE in an ESM study. First, it is challenging to understand apriori how many self-reports are required to be collected from a user to estimate the missing emotion self-reports accurately. Ideally, MUSE should estimate the missing self-reports accurately by observing few self-reports so that even if the participant drops (from the study) early, the missing self-reports can be estimated. We demonstrated that for the Homogeneous dataset, MUSE can estimate the missing self-reports with an accuracy of 80 - 84% by observing 50 - 60% self-reports (Fig. 11a). For example, although we collected 100 self-reports in the user study (Section 3), in reality using half of these, MUSE can estimate the remaining self-reports. This implies that even if a user drops out after providing self-reports till half the duration of the original one, the proposed approach can estimate the self-reports. Nevertheless, this is empirically derived based on this dataset, which in reality can be further complicated by user self-report behavior, temporal and external events, and anomalies. Also, as we observed MUSE performs well with relatively larger number of participants than the usual sample size of HCI studies [7], the proposed approach should be deployable for typical user base in ESM studies (e.g., hundreds of users). If the framework is fed with data from more users, the model gets the opportunity to learn more effectively. However, in case of significantly large number of users (e.g., thousands of users), the multi-task learning model may encounter some delay during training as we consider self-report estimation for every user a separate task. Therefore, it is recommended to test the model performance before deploying it for a significantly larger sample size.

Second, another challenge is to decide whether (a) to bootstrap the model with a small number of self-reports and then retrain at some interval *or* (b) to train the model once with sufficient number of self-reports. Retraining may be useful for a long-running study, as self-reporting behavior of a participant may change over time due to external effects, environment, contextual fluctuations etc, [82] subsequently the trained models get outdated [30]. An effective way to build a reliable model could be to probe users intermittently, collect the intermediate self-reports and retrain the model based on newly recorded samples. However, this poses the challenge of automatic identification of the retraining points.

Finally, MUSE leverages the self-reporting behavior similarity, which we expressed in terms of emotion-transition, emotion persistence time and emotion recurrence length. However, self-reporting similarities may be computed from the various other modalities. For example, identifying a group of users, who reacts similarly to a specific emotion stimulus (e.g., watches media content more in a *relaxed* emotion) may be considered as similar. Moreover, participants sharing similar profile (say attending same class, staying in same dormitory etc), may also exhibit a similarity in their self-reporting behavior. However, such approaches may require additional sensor details and usage logs, which may raise privacy concerns.

9.4 Limitations

We acknowledge that as we are estimating the missing self-reports (once a participant drops out) there may be some deviations than if the participant would have continued and provided the actual self-reports. The true self-reports are influenced by several temporal, contextual, and external factors, which are extremely challenging to factor in a machine learning model. Additionally, as MUSE estimates the self-reports once a participants drops out (or completely stops responding), the current implementation cannot be used for estimating intermittent missing self-reports from a participant, who responds for a few days and then becomes silent and so on. In future, we will extend MUSE to handle a situation where the participant responds initially, and then fall silent for a certain period of time, before becoming active again.

10 CONCLUSION

In this paper, we propose MUSE, a multi-task learning (MTL) framework to estimate the missing self-reports of dropout participants in an ESM study. This framework will allow researchers (or study designers) to deal with the unplanned withdrawal of the participants and saves them from running multiple studies to collect required amount of self-reports. To estimate the missing emotion self-reports, MUSE leverages the similarity in the self-reporting behavior among the participants. In specific, MUSE proposes an approach to (a) quantitatively express the self-reporting pattern of every user in terms of emotion transition probabilities, emotion persistence times, and emotion recurrence length (b) share data among similar users using the MTL network to estimate the missing emotion self-reports of a dropout participant. We evaluated MUSE by conducting two in-the-wild studies (N1 = 24, N2 = 30) of duration 6-week and 8-week respectively. The participants self-reported four different emotions (*happy, sad, stressed, relaxed*) during this period. The evaluation of MUSE on these datasets reveals that it can estimate the missing self-reports with an average AUCROC of 84% (Study I) and 82% (Study II). Furthermore, the evaluation of MUSE on a smartphone keyboard based emotion inference scenario highlights that using estimated self-reports, similar emotion inference performance (like original self-reports) is obtained. These findings demonstrate the possibility of reducing the data collection overhead for study designers by estimating the missing emotion self-reports from dropout participants in an ESM study.

ACKNOWLEDGMENTS

This research has been supported by the SURE grant (SUR/2022/001965) of SERB (Science and Engineering Research Board) of the Department of Science & Technology (DST), Government of India, the EHS grant (VN/BK/18-19/AUG/65) of TCS Research Lab, and the ACG grant (GOA/ACG/2022-2023/Oct/11) of BITS Pilani Goa. We also thank the anonymous reviewers for their insightful feedback to improve the quality of the work.

REFERENCES

[1] Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2019. Automatic Labeling Affective Scenes in Spoken Conversations. In *Cognitive Infocommunications, Theory and Applications*. Springer International Publishing, Cham, 109–130. https://doi.org/10.1007/978-3-319-95996-2_6

[2] Wageesha Bangamuarachchi, Anju Chamantha, Lakmal Meegahapola, Haeeun Kim, Salvador Ruiz-Correa, Indika Perera, and Daniel Gatica-Perez. 2023. Inferring Mood-While-Eating with Smartphone Sensing and Community-Based Model Personalization. arXiv:2306.00723 [cs.HC]

[3] Niels Van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The Experience Sampling Method on Mobile Devices. *ACM Computing Surveys (CSUR)* 50, 6 (2017), 93.

[4] Allan Berrocal and Katarzyna Wac. 2018. Peer-vasive Computing: Leveraging Peers to Enhance the Accuracy of Self-Reports in Mobile Human Studies. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (Singapore, Singapore) *(UbiComp '18)*. Association for Computing Machinery, New York, NY, USA, 600–605. https://doi.org/10.1145/3267305.3267542

[5] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.

[6] John Brebner. 1990. Personality factors in stress and anxiety. *Cross-cultural anxiety* 4 (1990), 11–19.

[7] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[8] 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* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 1293–1304. https://doi.org/10.1145/2750858.2805845

[9] Rich Caruana. 1997. Multitask learning. *Machine learning* 28, 1 (1997), 41–75.

[10] Yu-Lin Chang, Yung-Ju Chang, and Chih-Ya Shen. 2019. She is in a Bad Mood Now: Leveraging Peers to Increase Data Quantity via a Chatbot-Based ESM. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan) *(MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 58, 6 pages. https://doi.org/10.1145/3338286.3344406

[11] Lee Anna Clark, David Watson, and Susan Mineka. 1994. Temperament, personality, and the mood and anxiety disorders. *Journal of abnormal psychology* 103, 1 (1994), 103.

[12] Tamlin S Conner, Howard Tennen, William Fleeson, and Lisa Feldman Barrett. 2009. Experience sampling methods: A modern idiographic approach to personality research. *Social and personality psychology compass* 3, 3 (2009), 292–313.

[13] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.

[14] Clayton Epp, Michael Lippold, and Regan L. Mandryk. 2011. Identifying emotional states using keystroke dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (, Vancouver, BC, Canada,) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 715–724. https://doi.org/10.1145/1978942.1979046

[15] Ramin Fallahzadeh and Hassan Ghasemzadeh. 2017. Personalization without user interruption: boosting activity recognition in new subjects using unlabeled data. In *Proceedings of the 8th International Conference on Cyber-Physical Systems* (Pittsburgh, Pennsylvania) *(ICCPS '17)*. Association for Computing Machinery, New York, NY, USA, 293–302. https://doi.org/10.1145/3055004.3055015

[16] Rosta Farzan, Joan M. DiMicco, David R. Millen, Casey Dugan, Werner Geyer, and Elizabeth A. Brownholtz. 2008. Results from deploying a participation incentive mechanism within the enterprise. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 563–572. https://doi.org/10.1145/1357054.1357145

[17] Nico H Frijda, Batja Mesquita, Joep Sonnemans, Stephanie Van Goozen, and KT Strongman. 1991. The duration of affective phenomena or emotions, sentiments and passions. *International Review of Studies on Emotion* 1 (1991), 187–225.

[18] Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2017. Evaluating effectiveness of smartphone typing as an indicator of user emotion. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, San Antonio, TX, USA, 146–151. https://doi.org/10.1109/ACII.2017.8273592

[19] Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2017. TapSense: combining self-report patterns and typing characteristics for smartphone based emotion detection. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 2, 12 pages.

[20] Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2019. Designing an experience sampling method for smartphone based emotion detection. *IEEE Transactions on Affective Computing* 12, 4 (2019), 913–927.

[21] Surjya Ghosh, Bivas Mitra, and Pradipta De. 2020. Towards improving emotion self-report collection using self-reflection. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3383019

[22] Sarah Harris and David Harris. 2015. *Digital design and computer architecture: arm edition*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[23] Muhammad Asif Hasan, Nurul Fazmidar Mohd Noor, Siti Soraya Binti Abdul Rahman, and Mohammad Mustaneer Rahman. 2020. The transition from intelligent to affective tutoring system: a review and open issues. *IEEE Access* 8 (2020), 204612–204638.

[24] Joel M Hektner, Jennifer A Schmidt, and Mihaly Csikszentmihalyi. 2007. *Experience sampling method: Measuring the quality of everyday life*. SAGE Publications, Inc., Thousand Oaks, CA. https://doi.org/10.4135/9781412984201

[25] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[26] Wassily Hoeffding. 1992. *A class of statistics with asymptotically normal distribution*. Springer, New York, NY, 308–334. https://doi.org/10.1007/978-1-4612-0919-5_20

[27] Gary Hsieh, Ian Li, Anind Dey, Jodi Forlizzi, and Scott E Hudson. 2008. Using visualizations to increase compliance in experience sampling. In *Proceedings of the 10th International Conference on Ubiquitous Computing*. Association for Computing Machinery, New York, NY, USA, 164–167. https://doi.org/10.1145/1409635.1409657

[28] Maxwell L. Hutchinson, Erin Antono, Brenna M. Gibbons, Sean Paradiso, Julia Ling, and Bryce Meredig. 2017. Overcoming data scarcity with transfer learning. arXiv:1711.05099 [cs.LG]

[29] Natasha Jaques, Sara Taylor, Akane Sano, Rosalind Picard, et al. 2017. Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation. In *Proceedings of IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*. PMLR, Melbourne, Australia, 17–33.

[30] Matthew Keally, Gang Zhou, Guoliang Xing, Jianxin Wu, and Andrew Pyles. 2011. Pbn: towards practical activity recognition using smartphone-based body sensor networks. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*. Association for Computing Machinery, New York, NY, USA, 246–259. https://doi.org/10.1145/2070942.2070968

[31] Md Abdullah Al Hafiz Khan and Nirmalya Roy. 2018. Untran: Recognizing unseen activities with unlabeled data using transfer learning. In *2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, Orlando, FL, USA, 37–47.

[32] Richard N Landers, Kristina N Bauer, and Rachel C Callan. 2017. Gamification of task performance with leaderboards: A goal setting experiment. *Computers in Human Behavior* 71 (2017), 508–515.

[33] Nicholas D Lane, Ye Xu, Hong Lu, Shaohan Hu, Tanzeem Choudhury, Andrew T Campbell, and Feng Zhao. 2011. Enabling large-scale human activity inference on smartphones using community similarity networks (csn). In *Proceedings of the 13th international conference on Ubiquitous computing*. Association for Computing Machinery, New York, NY, USA, 355–364. https://doi.org/10.1145/2030112.2030160

[34] Nicholas D. Lane, Ye Xu, Hong Lu, Shaohan Hu, Tanzeem Choudhury, Andrew T. Campbell, and Feng Zhao. 2014. Community Similarity Networks. *Personal Ubiquitous Comput.* 18, 2 (feb 2014), 355–368. https://doi.org/10.1007/s00779-013-0655-1

[35] Reed Larson and Mihaly Csikszentmihalyi. 2014. *The experience sampling method*. Springer Netherlands, Dordrecht, 21–34. https://doi.org/10.1007/978-94-017-9088-8_2

[36] Xiang Li, Ben Aldridge, Lucia Ballerini, Robert Fisher, and Jonathan Rees. 2009. *Depth data improves skin lesion segmentation*. Springer, Berlin, Heidelberg, 1100–1107. https://doi.org/10.1007/978-3-642-04271-3_133

[37] Xiang Li, Ben Aldridge, Robert Fisher, and Jonathan Rees. 2011. Estimating the ground truth from multiple individual segmentations incorporating prior pattern analysis with application to skin lesion segmentation. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, Chicago, IL, USA, 1438–1441. https://doi.org/10.1109/ISBI.2011.5872670

[38] Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. 2013. Moodscope: Building a mood sensor from smartphone usage patterns. In *ACM Mobisys*. Association for Computing Machinery, New York, NY, USA, 389–402. https://doi.org/10.1145/2462456.2464449

[39] Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. 2023. Large Language Models are Few-Shot Health Learners. arXiv:2305.15525 [cs.CL]

[40] Daniel Lopez-Martinez and Rosalind Picard. 2017. Multi-task neural networks for personalized pain recognition from physiological signals. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, San Antonio, TX, USA, 181–184.

[41] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T Chittaranjan, Andrew T Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. 2012. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. Association for Computing Machinery, New York, NY, USA, 351–360. https://doi.org/10.1145/2370216.2370270

[42] Peter Lynn. 2001. The impact of incentives on response rates to personal interview surveys: Role and perceptions of interviewers. *International Journal of Public Opinion Research* 13, 3 (2001), 326–336.

[43] Iris B Mauss and Michael D Robinson. 2009. Measures of emotion: A review. *Cognition and emotion* 23, 2 (2009), 209–237. https://doi.org/10.1080/02699930802204677

[44] Patrick E McKight and Julius Najab. 2010. Kruskal-wallis test. , 1 pages.

[45] Patrick E McKnight and Julius Najab. 2010. Mann-Whitney U Test. , 1 pages. https://doi.org/10.1002/9780470479216.corpsy0524

[46] Lakmal Meegahapola, William Droz, Peter Kun, Amalia De Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, et al. 2023. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–32.

[47] Lakmal Meegahapola, Salvador Ruiz-Correa, Viridiana del Carmen Robledo-Valero, Emilio Ernesto Hernandez-Huerfano, Leonardo Alvarez-Rivera, Ronald Chenu-Abente, and Daniel Gatica-Perez. 2021. One more bite? Inferring food consumption level of college students using smartphone sensing and self-reports. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–28.

[48] Abhinav Mehrotra, Jo Vermeulen, Veljko Pejovic, and Mirco Musolesi. 2015. Ask, but don't interrupt: the case for interruptibility-aware mobile experience sampling. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. Association for Computing Machinery, New York, NY, USA, 723–732. https://doi.org/10.1145/2800835.2804397

[49] Prasanth Murali, Javier Hernandez, Daniel McDuff, Kael Rowan, Jina Suh, and Mary Czerwinski. 2021. Affectivespotlight: Facilitating the communication of affective responses from audience members during online presentations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445235

[50] Mohamed Musthag, Andrew Raij, Deepak Ganesan, Santosh Kumar, and Saul Shiffman. 2011. Exploring micro-incentive strategies for participant compensation in high-burden studies. In *Proceedings of the 13th international conference on Ubiquitous computing*. Association for Computing Machinery, New York, NY, USA, 435–444. https://doi.org/10.1145/2030112.2030170

[51] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.

[52] Veljko Pejovic, Neal Lathia, Cecilia Mascolo, and Mirco Musolesi. 2016. Mobile-based experience sampling for behaviour research. In *Emotions and Personality in Personalized Services*. Springer International Publishing, Cham, 141–161.

[53] Veljko Pejovic and Mirco Musolesi. 2014. InterruptMe: designing intelligent prompting mechanisms for pervasive applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. Association for Computing Machinery, New York, NY, USA, 897–908. https://doi.org/10.1145/2632048.2632062

[54] Le Vy Phan, Nick Modersitzki, Kim K Gloystein, and Sandrine Müller. 2022. Mobile Sensing Around the Globe: Considerations for Cross-Cultural Research. https://doi.org/10.31234/osf.io/q8c7y

[55] Stefano Piana, Alessandra Staglianò, Francesca Odone, and Antonio Camurri. 2016. Adaptive body gesture representation for automatic emotion recognition. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 6, 1 (2016), 1–31.

[56] Martin Pielot, Tilman Dingler, Jose San Pedro, and Nuria Oliver. 2015. When attention is not scarce-detecting boredom from mobile phone usage. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. Association for Computing Machinery, New York, NY, USA, 825–836. https://doi.org/10.1145/2750858.2804252

[57] Robert Plutchik. 1980. *A general psychoevolutionary theory of emotion*. Elsevier, Atlanta, 3–33. https://doi.org/10.1016/b978-0-12-558701-3.50007-7

[58] M Prajwal, Ayush Raj, Sougata Sen, Snehanshu Saha, and Surjya Ghosh. 2023. Towards Efficient Emotion Self-report Collection Using Human-AI Collaboration: A Case Study on Smartphone Keyboard Interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 2 (2023), 1–23.

[59] Andrew Raij, Animikh Ghosh, Santosh Kumar, and Mani Srivastava. 2011. Privacy risks emerging from the adoption of innocuous wearable sensors in the mobile environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 11–20. https://doi.org/10.1145/1978942.1978945

[60] Michael T Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G Dietterich. 2005. To transfer or not to transfer. In *NIPS 2005 workshop on transfer learning*, Vol. 898. MIT Press, Vancouver British Columbia Canada, 1–4.

[61] James A Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178.

[62] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. 2019. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–30.

[63] Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, Sydney, NSW, Australia, 747–748.

[64] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. Association for Computing Machinery, New York, NY, USA, 614–622. https://doi.org/10.1145/1401890.1401965

[65] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, USA, 254–263.

[66] William J. Stewart. 1995. *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, Princeton. https://doi.org/10.1515/9780691223384

[67] Arthur A Stone, Ronald C Kessler, and Jennifer A Haythomthwatte. 1991. Measuring daily events and experiences: Decisions for the researcher. *Journal of personality* 59, 3 (1991), 575–607.

[68] Boyuan Sun, Qiang Ma, Shanfeng Zhang, Kebin Liu, and Yunhao Liu. 2015. iSelf: Towards cold-start emotion labeling using transfer learning with smartphones. In *2015 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, Hong Kong, China, 1203–1211.

[69] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, Soren Brage, Nick Wareham, and Cecilia Mascolo. 2021. Selfhar: Improving human activity recognition through self-training with unlabeled data. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5, 1 (2021), 1–30.

[70] Sara Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2020. Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE Transactions on Affective Computing* 11, 2 (2020), 200–213.

[71] Mark A Thornton and Diana I Tamir. 2017. Mental models accurately predict emotion transitions. *Proceedings of the National Academy of Sciences* 114, 23 (2017), 5982–5987.

[72] Liam D Turner, Stuart M Allen, and Roger M Whitaker. 2015. Push or delay? decomposing smartphone notification response behaviour. In *Human Behavior Understanding: 6th International Workshop, HBU 2015, Osaka, Japan, September 8, 2015, Proceedings*. Springer, Cham, 69–83.

[73] Niels Van Berkel, Jorge Goncalves, Simo Hosio, and Vassilis Kostakos. 2017. Gamification of mobile experience sampling improves data quality and quantity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–21.

[74] Philippe Verduyn, Ellen Delvaux, Hermina Van Coillie, Francis Tuerlinckx, and Iven Van Mechelen. 2009. Predicting the duration of emotional experience: two experience sampling studies. *Emotion* 9, 1 (2009), 83.

[75] Philippe Verduyn and Saskia Lavrijsen. 2015. Which emotions last longest and why: The role of event importance and rumination. *Motivation and Emotion* 39, 1 (2015), 119–127.

[76] Yufeng Wang, Wei Dai, Bo Zhang, Jianhua Ma, and Athanasios V Vasilakos. 2017. Word of mouth mobile crowdsourcing: Increasing awareness of physical, cyber, and social interactions. *IEEE MultiMedia* 24, 4 (2017), 26–37.

[77] Sophie F Waterloo, Susanne E Baumgartner, Jochen Peter, and Patti M Valkenburg. 2018. Norms of online expressions of emotion: Comparing Facebook, Twitter, Instagram, and WhatsApp. *New media & society* 20, 5 (2018), 1813–1831.

[78] Rui Xia and Yang Liu. 2015. A multi-task learning framework for emotion recognition using 2D continuous space. *IEEE Transactions on Affective Computing* 8, 1 (2015), 3–14.

[79] Xuhai Xu, Prerna Chikersal, Janine M Dutcher, Yasaman S Sefidgar, Woosuk Seo, Michael J Tumminia, Daniella K Villalba, Sheldon Cohen, Kasey G Creswell, J David Creswell, et al. 2021. Leveraging collaborative-filtering for personalized behavior modeling: a case study of depression detection among college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–27.

[80] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S Kuehn, Jeremy F Huckins, Margaret E Morris, et al. 2023. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–34.

[81] Bee Wah Yap and Chiaw Hock Sim. 2011. Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation* 81, 12 (2011), 2141–2155.

[82] Gerald Young and Gaurav Suri. 2019. Emotion regulation choice: A broad examination of external factors. *Cognition and Emotion* 34, 2 (2019), 242–261.

[83] Xiaojing Yuan, Ning Situ, and George Zouridakis. 2009. A narrow band graph partitioning method for skin lesion segmentation. *Pattern Recognition* 42, 6 (2009), 1017–1028.

[84] Zhen Yue, Eden Litt, Carrie J Cai, Jeff Stern, Kathy K Baxter, Zhiwei Guan, Nikhil Sharma, and Guangqiang Zhang. 2014. Photographing information needs: the role of photos in experience sampling method-style research. In *Proceedings of the sigchi conference on human factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1545–1554. https://doi.org/10.1145/2556288.2557192

[85] Omar F Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, USA, 1220–1229.

[86] A. Zenonos, A. Khan, G. Kalogridis, S. Vatsikas, T. Lewis, and M. Sooriyabandara. 2016. HealthyOffice: Mood recognition at work using smartphones and wearable sensors. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. IEEE, Sydney, NSW, Australia, 1–6. https://doi.org/10.1109/PERCOMW.2016.7457166