SELFI: Evaluation of Techniques to Reduce Self-report Fatigue by Using Facial Expression of Emotion

Salma Mandi[1], Surjya Ghosh[2], Pradipta De[3], and Bivas Mitra[1]

[1] Indian Institute of Technology Kharagpur, Kharagpur, INDIA
salmamandi@kgpian.iitkgp.ac.in, bivas@cse.iitkgp.ac.in
[2] BITS Pilani Goa, INDIA
surjyag@goa.bits-pilani.ac.in
[3] Microsoft Corporation, USA
prade@microsoft.com

Abstract. This paper presents the SELFI framework which uses information from a range of indirect measures to reduce the burden on users of context-sensitive apps in the need to self-report their mental state. In this framework, we implement multiple combinations of facial emotion recognition tools (Amazon Rekognition, Google Vision, Microsoft Face), and feature reduction approaches to demonstrate the versatility of the framework in facial expression based emotion estimation. The evaluation of the framework involving 20 participants in a 28-week in-the-wild study reveals that the proposed framework can estimate emotion accurately using facial image (83% and 81% macro-F1 for valence and arousal, respectively), with an average reduction of 10% self-report burden. Moreover, we propose a solution to detect the performance drop of the model developed by $SELFI$, during runtime without the use of ground truth emotion, and we achieve accuracy improvements of 14%.

Keywords: Experience Sampling Method (ESM) · Emotion prediction · Self-report burden · Facial image processing.

1 Introduction

In recent years, context-sensitive mobile applications have become pervasive due to the penetration of smartphones in our daily life [34,6,29,24]. These context-aware applications dynamically adjust their behaviors depending on the present context (say, the current mental state of the user, the current computational and physical environment etc), so that the user can focus on their current activity. The core of context-sensitive applications is to accurately detect the context (such as location of use, collection of nearby people, emotion of the user, schedules for the day, etc) and adapt applications according to the changes in context [9,8,15]. Context-sensitive applications typically deploy supervised machine learning models, which are trained by correlating data collected from various sensors with the reported context information[19,41,17]. The availability of diverse

sensors facilitates the development of context-sensitive applications, however, collecting the ground truth context information is still a challenge.

Importantly, state-of-the-art context-sensitive applications mostly rely on manual efforts of users to collect self-reported context information [30,22,18]. For instance, consider a music recommendation system, which is capable of dynamically recommending music that adapts to the emotion of an individual. In order to develop such applications, researchers typically rely on manually collected emotion self-reports as ground truth context information relying on the Experience Sampling Methods [21,5]. However, as manual self-reporting is a burden and time-consuming for users, collecting context information from a long-term study is challenging as participants may respond arbitrarily, or drop out [26,27]. Therefore, efficient strategies to reduce the self-report burden while collecting the context information is essential to develop an effective context-sensitive application. In this paper, we aim to facilitate the users of the emotion aware context-sensitive applications, hence, we propose a mechanism to reduce the *human in the loop* while collecting ground truth emotion self-reports.

One promising approach for reducing the *human in the loop* in emotion self-report collection is to use the *alternative information sources* for inferring emotion, based on initial self-reports provided by the participant. Those inferred emotions will be subsequently considered as the ground truth self-reports (in place of directly asking the participants) to train the supervised models. A diverse set of indirect measures (physiological data, speech, facial expression, posts in Online Social Networks) is already known to carry a signature of human emotion [23,48,10,43]. For example, passively sensed heart rate data or skin conductance data using a wearable device can be used as an alternative information source to infer a user's emotion instead of asking for self-report labels [3,35]. Similarly, inferred emotion from the voice clips can be used as a substitute of emotion self-report [12]. Another promising alternative is to capture the facial image in place of directly collecting emotion self-report from the user, and subsequently infer the emotion label from the facial expression, which may work as a substitute of self-reported labels.

However, the applicability of these alternative sources of information to infer users' emotion labels poses a number of challenges. First of all, the performance and suitability of various alternative information sources widely vary depending on the context and the participants involved in the study. Notably, the reduction in self-reports leads to poor model training, which may significantly drop the emotion prediction performance of the model. Hence, a flexible toolbox is essential to conveniently explore the role of multiple alternative information sources and investigate their impact on reducing the self-report burden, with a trade-off with emotion prediction performance. The lack of such a toolbox creates a major bottleneck to examining the potential of various alternative sources of information in reducing the self-report burden. Second, recent advances in machine learning algorithms allow the extraction of a large set of features from these alternative information sources. However, identifying the most relevant features, which are effective in reducing the self-report burden for developing the emotion

prediction model, is a challenge. Deep learning models [39] may facilitate the automation of feature extraction, however, the scarcity of data from large scale field trials with users restricts the applicability of deep learning models. Third, the performance of the developed emotion prediction model may degrade over time, as the model gets obsolete due to the change in the context, environment, and behavioral pattern of the users. Hence, one needs to assess the quality of the estimates of users' emotional states on runtime and automatically *adapt* the framework, once it detects any degradation in the performance. Therefore, the development of an *adaptive framework*, which allows one to (i) explore various alternative sources of information, and (ii) to identify the relevant features from this sources of information, may facilitate the developers of the context-sensitive applications to reduce the burden of the participating users. This paper takes one step towards this direction.

In this paper, we propose *SELFI*, a framework which relies on alternative sources of information, to reduce the self-report burden of the users of emotion aware context-sensitive applications. In particular, we use facial expression as the alternative source of information for emotion estimation, since (i) facial expression is considered as a strong indicator of human emotion [19], (ii) images are easy to capture in a seamless manner, and (iii) a number of facial image recognition tools (Amazon Rekognition [46], Google Vision [2], and Microsoft Azure [1]) are commercially available. The framework consists of two major building blocks. The first building block (termed as *Facial image processing block*) takes the facial images (captured using a smartphone) as input and extracts a set of features. The second block (termed as *feature reduction tool*) takes the large feature set as input and decomposes it by adopting any suitable feature reduction method. Once the feature set is reduced, the correspondence between the facial image and the self-reports is established by training a machine learning model for inferring the emotion based on the facial image. The proposed framework provides flexibility to explore any information source (various facial image recognition tools, in this case) and any feature decomposition method (to select relevant features) for emotion estimation.

In order to evaluate the proposed framework, we have developed an Android application for capturing facial images and emotion self-reports. We conducted a 28-week field study of the framework with 20 participants, who recorded their facial expressions and instantaneous emotions throughout the day. The emotion self-report collection was guided by the Circumplex model of emotion, which suggests every emotion is a combination of valence and arousal [33]. Our experimental results demonstrate that the proposed framework estimates the valence and arousal with an average F1-score of 83% and 81% respectively, based on the facial images as an alternative information source. Side by side, *SELFI* reduces the volume of required self-reports up to 10% (on average).

2 Background and Data Collection Apparatus

In this section, first we discuss the various ways of representing emotion. Next, we describe the development of the data collection application and procedure.

2.1 Emotion representation

Emotions are intense feelings caused by a specific event [11] persist for a short duration, whereas moods, in contrast, are feelings less intense than emotions that often do not depend on contextual stimuli [45] last longer than emotions. Notably, unlike moods, emotions tend to be more clearly revealed with facial expressions [16], which motivates us to rely on facial images as an alternative source of information for emotion estimation. In this paper, we followed the Circumplex model that presents emotion in a 2-dimensional plane (see Fig. 1). The x-axis called valence refers to the positive (pleasant) and negative (unpleasant) degree of emotion, while the y-axis called arousal refers to the degree to which an emotion is associated with high (activation) or low energy (deactivation). Thus any emotion can be described using valence and arousal dimensions mapped in one of the 4 quadrants. For instance, the emotion *happy* is high valence and high arousal state located in the 1st quadrant. We notice that four dominant emotions *happy*, *angry*, *sad*, and *relaxed* from four different quadrants of the Circumplex model are pretty discriminative Fig. 1. In the following, we develop an app (see Fig. 2) to collect valence and arousal states as emotion self-reports, directly from the users.

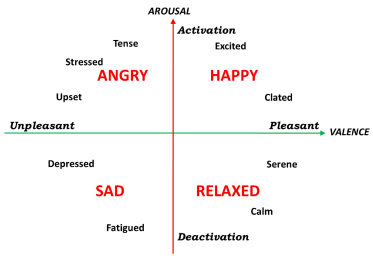


Fig. 1: Circumplex Emotion Model [28].

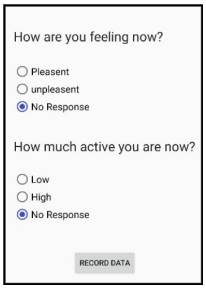


Fig. 2: Data collection UI.

2.2 UI development and self-report collection

We have developed a smartphone based Android application [4] to collect the data from the participants. This application enables us to collect two specific information, (a) self-reported emotion labels, and (b) facial images of the participants. Since capturing facial images may raise privacy concerns, we take utmost care to capture and process the facial images. The application checks the phone status (lock or unlock) at 2 hours intervals to probe a notification to record the data. If the screen is locked at the time of triggering the notification, the notification is held back and issued once the participant unlocks the phone. The interval value is chosen based on a study presented by Schmidt et al [36]. In response to the

[4] In `https://anonymous.4open.science/r/Image_collection_Upload_Dropbox-0565/` we provide the implementation of the data collection apparatus.

notification, the data collection UI records emotion labels and facial images of the participant. The process is illustrated in Fig. 3. The data collection UI consists of two modules, *self-report collection module* and *Facial Image collection module*.
(a) Self-report Collection module: This module collects the valence and arousal state from participants in the form of two questions (shown in Fig. 2). Participants are asked to select the suitable radio button to record their perceived emotion state and press the "Record data" to record the self-report. Notably, participants are also allowed to skip self-reporting by selecting the 'No Response' option, which is set as default. Once the participant responds to the notification, the notification probing time, emotion reporting time, and reported emotion states are saved in a file.
(b) Facial Image Collection Module: We have implemented the facial image collection component, which is embedded within the data collection UI. The responsibility of this module is to capture the facial images of the user at the moment of recording the emotion self-report, in a *privacy preserving* manner. The module triggers automatic collection of the facial image of the participant within one second of the probe notification. It captures three images in succession to record at least one image with the full frontal face. We record the date and time of the captured image, which facilitates us to map it with its corresponding emotion self-report.
The participants are instructed as part of the self-report collection to ensure that a selfie is captured. If the participant does not provide the self-report, then the photo is marked "No response". To ensure privacy, this application provides an *image review button* to review the images at any time before it is uploaded. Participants can select and delete the images with the delete button if they do not want to upload a specific image. If the participant deletes *all photos* captured with an emotion self-report, then the corresponding self-report label is also removed. This is important to note that participants can audit and remove pictures from phone storage, but capturing a photo is not allowed on their own.

3 Pilot Study

We conducted a pilot study to demonstrate the limitations of the facial image based emotion recognition tools. We have taken approval from the Institute Ethical Committee of IIT Kharagpur[5] to collect facial images and emotion self-reports from the participants.

3.1 Study focus group

We conducted a survey among 33 participants to judiciously select the volunteers, who are active with their smartphone engagements. We asked the participants to fill out a questionnaire through a Google form[6] to collect information

[5] The IRB approval number IIT/SRIC/SAO/2017.
[6] We advised the participants to rely on the *Digital Wellbeing and Parental Control* tool to respond to the smartphone usage related questionnaire.

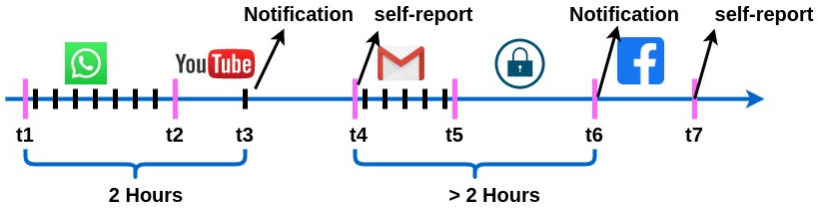


Fig. 3: Schematic showing the process of a participant recording self-report emotion label. For example, in the time interval between t_1 and t_3, the participant is engaged with different apps (Whatsapp, Youtube). A notification is received at t_3 as 2 hours are passed and it is answered at t_4. Similarly, more than 2 hours is elapsed from the last reported emotion and notification received at t_6 since the phone was locked.

regarding their daily smartphone usage, such as average duration of phone usage, types of applications usage and their duration, time of the day with peak usage of the phone, etc. Based on the collected data, we recruited 5 participants (2 males, 3 females; aged 17 to 35 years) from different job backgrounds (3 students, 1 teacher, and businessman), who are highly active with their smartphone usage, which ensures the collection of sufficient volume of data in a limited time period.

3.2 Study protocol

We instructed the participants to install the app on their devices to use it for 3 weeks. First, we familiarize the participants with the basic definition of emotion and apprise them about the functionality of the app. We instructed the participants to respond to the notification received to record their emotion self-report, whenever they strongly felt some emotion. After the pilot study, on average we have collected 100 emotion labels (on average 5 self-reports per day) and 165 facial images from each participant. We paid a token honorarium to all the volunteers.

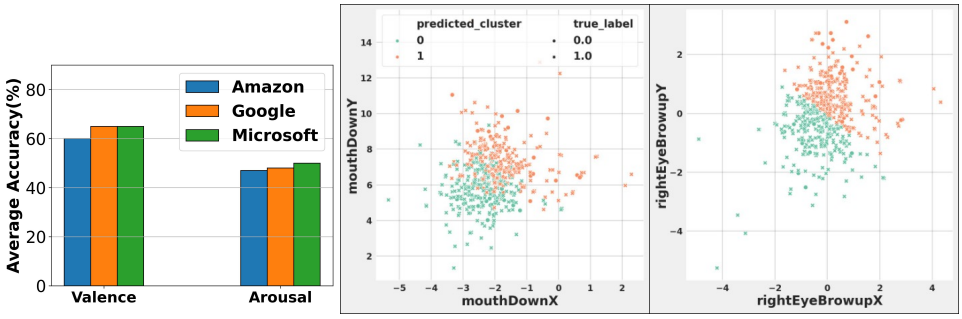
3.3 Challenges of using facial images

(a) Limitations of commercial image processing tools: First, we demonstrate the limitations of the state of the art facial image processing tools such as Amazon Rekognition [46], Google Vision [2], and Microsoft Azure [1] to predict emotion from the collected facial images. These tools return a set of discrete emotions from the facial images, along with their respective confidence scores. We consider the emotion with the highest confidence score as the predicted emotion. We map the predicted emotion of each participant in the *valence* and *arousal* scale of the Circumplex model. In Fig. 4a, we compare this predicted emotion with the ground truth emotion label. We observe that, on average, the

prediction accuracy is 63% and 48% for valence and arousal, respectively, which demonstrates the limitations of the aforementioned pre-trained tools.

It has been shown in the literature that specific facial landmarks are responsible for expressing a particular emotion [13]. For example, points in the cheek and lip region express emotion *happy*; on the contrary, the eyebrow and lip corner mainly reveal emotion *sad*. Therefore we focus on two individual landmarks, say the right eyebrow and the mouth position, to investigate if they are capable to correctly discriminate the (high and low) valance states. We extract the 2D landmark features (right eyebrow, mouth position) from the facial images using commercial image processing tools. Next, we apply k-means clustering (with $k = 2$) to cluster all the images in two classes based on those extracted landmark features. In Fig. 4b, we present a scatter plot, where the coordinate of each point represents the 2D landmark features obtained from the facial image, the color of the point (brown or green) represents the predicted valance state (high or low), and the shape of the point (star or disk) represents the ground truth emotion label collected from the participants. We observe that albeit the facial landmark features exhibit some potential to classify the high and low valance states, however, their accuracy does not cross 50% (for both right eyebrow and mouth position features), revealing the limitations of the *individual* landmark features.

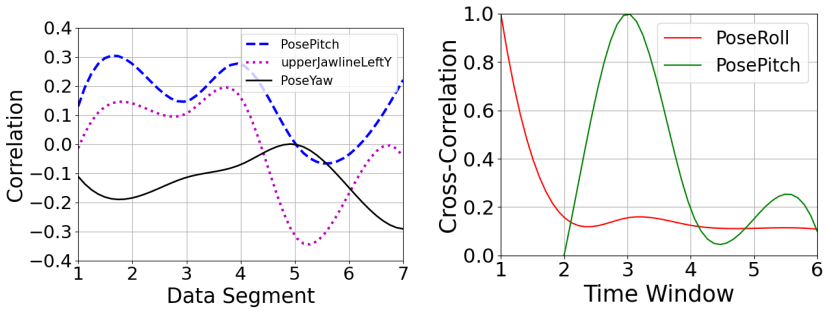
The aforementioned experiments point to the necessity of developing a *personalized* toolbox, as we cannot directly rely on the generalized pre-trained tools for automatic emotion recognition. Albeit these commercial tools provide a large set of facial landmark features to represent a facial expression, however, *individually* those landmark features do not exhibit elegance in classifying the valance states.



(a) (b)

Fig. 4: (a) Performance of three facial image processing tools for emotion prediction, (b) Scatter plot showing the role of the facial landmark features (mouth down, right eyebrow) in classifying the emotion label. The coordinate of each point represents the landmark features, the color of the point (brown or green) represents the k-means based valance state (high or low), and the shape of the point (star or disk) represents the ground truth emotion label.

(b) Variation of facial features over time: We consider the data collected from a specific user U_1 and extract features from facial images using different image analysis tools. We split the data (both emotion labels and the features) into fixed size time windows. We take the data segment of the first window and calculate the biserial correlation [20] between every facial feature and ground truth emotion labels. In Fig 5a, we show the top correlated feature in the second, fifth, and seventh time windows. Interestingly, we observe that the top correlated facial feature "PosePitch" in the second window drops in the fifth and seventh time window, whereas the other facial feature "upperJawlineLeftY" and "PoseYaw" appears at the top in the fifth and seventh time windows, respectively. This observation indicates that facial features, which are highly correlated with self-reported emotion in one time window, may exhibit low correlation in the subsequent time windows. Further, we compute the cross-correlation between the top correlated feature (PoseRoll) of one time window, with the top correlated feature (PosePitch) of the subsequent window. Fig. 5b depicts that the cross-correlation between the pair of top features in two subsequent windows is pretty low, which points to the significant changes of facial features across various time windows. This study shows that the critical facial features, which may play a key role in developing emotion estimation models, vary across different time windows. Hence, one time development and training of the emotion estimation model may not suffice, as the model may get obsolete over time.



(a) (b)

Fig. 5: (a) Correlation score for the top correlated feature in one time window decreases gradually in the subsequent time window, (b) Cross-correlation score between the top correlated feature from different time window reduces with the higher time window indicates the changes of the top correlated feature.

4 Methodology

In this section, first we illustrate the problem addressed in this paper, followed by the implementation of the proposed framework *SELFI*.

4.1 Problem statement

Consider a scenario where we wish to predict the emotion label $e_{(n+1)}$ of a user at time instance $(n+1)$, from the past n self-reported emotion labels. This predicted emotion label $e_{(n+1)}$ will be later considered as $(n+1)^{th}$ self-report of the user. Here the emotion label e_i represents the (high or low) valence and arousal state of the user at time instance i. Multiple models [47] have been proposed in the literature to estimate the emotion $e_{(n+1)}$ from the past n labels, incurring the burden of self-report fatigue n. The objective of this paper is to develop a framework $SELFI$, which only collects initial self-report emotion labels $n' < n$ directly from the users, and relies on the facial image $f_{(n+1)}$ collected at time instance $(n+1)$ (as alternative information), to correctly estimate the emotion label $e_{(n+1)}$ at time instance $(n+1)$. Hence, the proposed framework $SELFI$ aims to reduce the self-report burden by $(n-n')$ with correctly estimating the emotion labels for the time period $(n-n')$.

4.2 Development of *SELFI*

The overview of the proposed framework $SELFI$ [7] is presented in Fig. 6. This framework relies on two different input blocks. In the first block, (a) **Emotion self-report processing**, we collect the past self-reported emotion labels (e_i, $i \in [1, n']$), provided by the user and subsequently compute the handcrafted features (i) Influence and (ii) Emotion Sequence Length. The second block (b) **Facial image processing**, we implement this block in two steps, (i) first, we fed the facial image $f_{(n+1)}$ collected at time instance $(n+1)$ to the image analysis tool (say, Amazon Rekognition[46], Google Vision[2], and Microsoft Azure[1]), which extracts all facial landmarks as facial features. (ii) In the second step, those features are fed to the feature reduction tool (say, Kernel Principal Component Analysis ($KPCA$) and Kernel Discriminant Analysis (KDA), which reduces the dimension of features. Finally, we use the extracted (a) self-reported features and (b) facial features to develop a machine learning model, which predicts the emotion label $e_{(n+1)}$ at time instance $(n+1)$. In this section, we describe in detail the various building blocks of $SELFI$. We develop two separate prediction models for $SELFI$; one to estimate the valence state and another one to estimate the arousal state of the user.

Emotion self-report processing: We compute the following two self-report features from the past n' emotion labels e_i (say, valence) $\in \{1, 0\}$ collected from the user.

(a) Influence (F_{e_i}): This feature measures the influence of the current self-report emotion e_i on the next self-report $e_{(i+1)}$, where $e_i, e_{(i+1)} \in \{1, 0\}$. In order to compute *Influence*, first we define a 2×2 state-transition matrix P, where each element of $P = \{p_{\text{ij}}\}$, $\forall i, j \in \{1, 0\}$ denotes the state transition probability from emotion label e_i to label e_j, where $e_i, e_j \in \{1, 0\}$. Moreover, a current emotion label e_i has an impact on the next state based on the time elapsed between

[7] In `https://anonymous.4open.science/r/SELFI-77A3/` we provide the implementation of the SELFI framework, with a toy dataset.

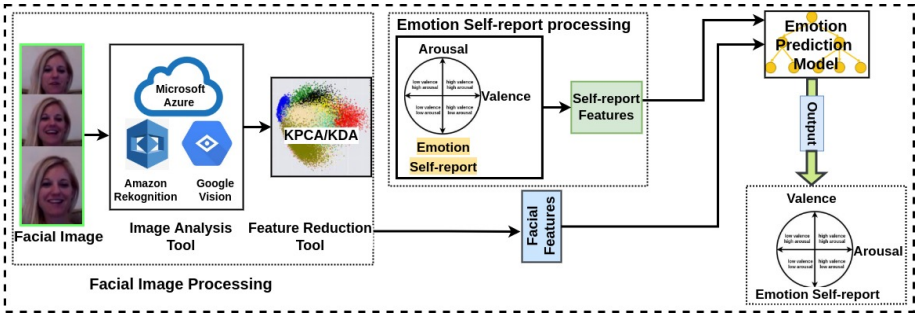


Fig. 6: The overall architecture of the emotion estimation framework $SELFI$.

the current self-report e_i and the next self-report $e_{(i+1)}$. We define $\tau_{e_{(i+1)}}$ as a normalized elapsed time since last observed self-report $e_{(i+1)}$. Hence, the feature *Influence* F_{e_i} is designed to capture the influence of a current self-report state e_i on the next self-report $e_{(i+1)}$ as $F_{e_{(i+1)}} = p_{e_i e_{(i+1)}} \times (1 - \tau_{e_{(i+1)}})$, where $p_{e_i e_{(i+1)}}$ indicates the probability of the next emotion state determined as $e_{(i+1)}$ based on the current self-report e_i and $(1 - \tau_{e_{(i+1)}})$ indicates a weight of the current emotion state to the next emotion state in terms of the elapsed time.

(b) Emotion Sequence Length (L_{e_i})***:*** This feature captures the typical sequence length of a specific emotion self-report label e_i, reflecting once a user reports an emotion label e_i, how many times in a row, the user repeats the same emotion at-a-stretch.

Facial image processing: We implement this block following two steps.

(a) Facial Feature Extraction: We rely on the various facial image analysis tools (such as, Amazon Rekognition[46], Google Vision[2], and Microsoft Azure[1] etc) to extract features from facial images. The basic principle of those tools is to first detect the faces in the image and return a bounding box demarcating the face in the image. Next, it identifies the various facial landmarks, such as the position of the eyebrow, pupil, mouth, nose, chin, etc as attributes (e.g., leftEyeBrowLeft, rightPupil, mouthUp, noseLeft, chinBottom, etc). These tools provide the location of each landmark attribute on the face in terms of two or three dimension coordinates. Apart from landmarks, those tools also return some other attributes from the images, such as age range, gender, image quality, as well as some binary attributes such as the presence of beard on the face, the presence of eyeglasses, etc. However, in our model construction, we only consider the extracted facial landmarks and their respective attribute coordinates as the facial features. Precisely, *Amazon Rekognition*, *Microsoft Azure* and *Google Vision* provide us 60, 54 and 96 facial features, respectively from each facial image. We stress on the fact that in the development of $SELFI$, one may feel free to choose any commercially available facial image analysis tools to implement the *Facial image processing* block for extracting the facial features.

(b) Feature Reduction: Since facial image analysis tools extract a large number of facial features from images, it is necessary to apply feature reduction to maintain a sufficient density of the samples per feature. This reduces the risk of model overfitting, developed based on facial features. This issue is specifically critical in our context, as we aim to collect fewer self-report emotion labels from each user, limiting the volume of available data samples. We implement two specific feature reduction techniques such as (i) Kernel Principal Component Analysis (*KPCA*) [37] considers the correlation between independent features and (ii) Kernel Discriminant Analysis (*KDA*) [32] considers the correlation between independent and dependent features, both of which are set to return a *single facial feature* after reduction. We implement *KPCA* with Gaussian Radial Basis Function (RBF) kernel [31], select *arpack* as eigensolver, and set one as the number of principal components to be returned. In *KDA*, we implement the same kernel function as used for *KPCA*. We stress on the fact that in the implementation of *Facial image processing* block, one may feel free to (i) choose a suitable feature reduction method and (ii) decide the dimension of the reduced features.

Emotion prediction model: Using the aforesaid (i) self-report features, computed from the past n' collected emotion labels and (ii) facial features, computed from the facial image captured for all time instance n', we train a Random Forest (RF) model to predict the emotion $e_{(n'+1)}$ at time instance $(n'+1)$. We develop two personalized models for each user to predict valence and arousal separately; each model implements a two state classifier to predict the state (high, low) of the emotion (valence and arousal). We implement 50 decision trees for the RF model, with no specific value chosen for the tree's maximum depth (hence depth is unlimited). This is important to note that state-of-the-art packages, such as ATOM[25], H2O[14], can be utilized to implement this block, enabling flexibility to explore the suitable ML models.

Estimating emotion at time $(n+1)$ from initial n' self-reports: From the collected initial n' self-reports and the captured facial image $f_{(n'+1)}$ at time instance $(n'+1)$, $SELFI$ predicts the emotion label $e_{(n'+1)}$ for time instance $(n'+1)$. The predicted emotion label $e_{(n'+1)}$ works as a substitute of the self-report for time instance $(n'+1)$. Extending this principle, $SELFI$ relies on the past n self-reports and the facial image $f_{(n+1)}$ at time instance $(n+1)$, to estimate the emotion label $e_{(n+1)}$ at time instance $(n+1)$. Here we obtain the past n self-reports as the (i) initial n' self-reports, collected directly from the users, and (ii) the remaining $(n-n')$ emotion labels recursively predicted by the model, thus reducing the self-report burden by $(n-n')$.

4.3 Retraining and runtime adaptation of *SELFI*

The developed model may get stale over a period of time, due to the changes in the feature pattern, which may result in a drop in accuracy in the estimated emotion labels. During runtime, we first check if retraining is needed after a certain time interval, and if it is needed, we record the ground truth emotion self-reports directly from the users and retrain the model. Initially, we train

the model with a n number of data samples collected over the first few days, where the value of n is decided based on the study in the development phase. On these training samples, we calculate the K-L divergence score [7] T_k between the features labeled as high and low valance (arousal) states. Next, in the running phase, we fix the time instances at which the model estimates the emotion labels in a sequence. After each prediction phase, we calculate the K-L divergence score P_H between the features predicted as high valance (arousal) state in the prediction phase and the features labeled as high valance (arousal) state in the training data. Likewise, a similar score P_L is calculated for features with low valance (arousal) states. Subsequently, we take an average of P_L and P_H, denoted as P_{avg}, which indicates, on average, the deviation between the test and the training data distribution. Finally, we compare P_{avg} against T_k, where higher P_{avg} indicates the misclassification in valance (arousal) estimation and subsequently calls for retraining the model with the new self-reported emotion labels, directly collected from the users.

5 Experimental Setup

In this section, we describe the field study procedure including the collected dataset, and explain the experiment procedure in detail.

5.1 Field study

Survey focus group: Initially, we recruited 33 participants (24 Female, 9 Male) aged between 17 to 60 years from different work backgrounds (such as office executive, student, teacher, nurse, businessman, retired person, etc) via an offline snowball recruiting method [38] maintaining work background, age, gender wise diversity. In addition, our participants were chosen from a diverse range of cities such as tier-II cities (3), tier-I urban agglomeration (7), urban agglomeration (18), and metropolitan cities (5), etc. We asked participants to install the app (described in Sec. 2) on their smartphones, allow the app to access the smartphone camera, and instructed them to use it for 28 weeks to record their emotion states. We notice that the participants use a variety of Android-based smartphones (such as Samsung, Redmi, Realme, Oppo, etc) with different configurations. During the study, we came across several challenges such as participants leaving the study in the middle (four participants), the data collection app stopped working due to some issues with their phone model (six participants), certain participants recording data rarely (three participants recorded less than 100 emotion labels), etc. Finally, we collected data from 20 participants (12 female, 8 male).

Instruction to the focus group: Once the participants installed the app on their devices, we advised them to engage normally with their phones. We have taken consent from the participants for capturing the facial images using the app and asked them to fill up a registration form recording their name, age,

gender, occupation, and demographic information. We apprise the participants that sometimes they may receive a notification to record their emotion self-report, once they unlock their phone. Participants are instructed to respond to the notification *only if* they feel a strong emotion at that moment (otherwise they may skip the notification). We also advised participants to record *No response* label if they do not wish to record the emotion self-report. Moreover, we informed the participants that they can review the recorded images anytime using the *image review button*. This makes sure that we capture emotion, not mood.

Field study dataset: On average, we have collected 350 emotion labels and 600 facial images from each participant. Next, we preprocess the collected data as follows, (a) First, we manually remove all the images with no facial impression (for example, participants with face masks, blank images) and the images which are difficult to visualize (say, too dark or too bright images). (b) We remove the entries with *No response* labels, as they do not reveal any emotion. Finally, after data preprocessing, on average we obtain 310 emotion labels and 500 facial images from each participant, on which we conduct the evaluation experiments. Valence and Arousal in both datasets, we maintain a 60:40 ratio between high and low class.

5.2 Evaluation procedure

We evaluate the framework using the nested cross-validation method [44], as the traditional cross-validation approach is not suitable due to the presence of temporal dependency in time series data. In every fold (or iteration) of this cross-validation approach, the temporal dependency is maintained i.e., the training portion does not include data from the future segment. In each iteration, we use 80% of data (first 60% for training and next 20% for testing). In the first iteration, an initial 60% is used for training, and the next 20% is used for testing, while in the second iteration, we discard the initial 20%, use the next 60% for training, and the last 20% for testing. This approach ensures that at every fold, we perform the training and testing on an equal amount of data and at the same time, future data is not used for training. In order to evaluate $SELFI$, we measure the macro f1-score from every iteration and compute the average values of two iterations to obtain the performance metric.

5.3 Baseline algorithms

We implement the following baseline algorithms to evaluate the performance of $SELFI$.
(a) Feature Based Emotion Model (FBEM): We implement a variation of $SELFI$ as *FBEM*, which aims to estimate the emotion label $e_{(n+1)}$ of a user at time instance $(n+1)$, *only* relying on the 'Facial image processing' block of the $SELFI$ framework (and ignoring the past emotion self-reports). We apply the image analysis tools of the facial image processing block to obtain the features from the facial image $f_{(n+1)}$ captured at time instance $(n+1)$. Subsequently,

we compute the correlation between each facial feature and the self-reported emotion label, and select the top two features with the highest correlation. Finally, we build the emotion prediction model based on these top facial features to predict the emotion label $e_{(n+1)}$.

(b) Self-reported Emotion Model (SREM): We implement a variation of $SELFI$ as *SREM*, which aims to estimate the emotion label $e_{(n+1)}$ of a user at time instance $(n+1)$, *only* relying on self-report emotion labels in the past n time instances [42]. Precisely, by implementing this *SREM*, we conducted an ablation study of $SELFI$, dropping the 'Facial image processing' block from the framework pipeline.

(c) Past-Current Data Based Emotion Model (PCDEM): We implement another variation of $SELFI$ as *PCDEM*, which aims to estimate the emotion label $e_{(n+1)}$ of a user at time instance $(n+1)$, relying on all the collected self-report emotion labels in the past n time instances and the facial image collected at time instance n and $(n+1)$, both [40]. We develop this baseline to jointly observe the influence of the current and the previous facial image on the prediction of emotion labels for the current instance.

6 Evaluation of $SELFI$ Framework

In this section, we evaluate the emotion prediction performance and reduction in the self-report burden of $SELFI$. We demonstrate the versatility of the $SELFI$ framework across various image analysis toolkits.

6.1 Emotion prediction performance

In Table 1, we evaluate and compare the performance of $SELFI$ framework with baseline algorithms, in the light of correctly estimating the emotion label $e_{(n+1)}$ at time instance $(n+1)$. In this evaluation, for all the algorithms, we have collected the *identical volume* of self-report emotion labels directly from the participants in the past n time instances to train the respective models.

Valence estimation: In Table 1, we observe that for all the model variants, $SELFI$ framework consistently outperforms the baseline algorithms in the valence dimension. The macro F1-score of $SELFI$ and the *SREM* algorithm is 83% and 77%, respectively, which justifies the utility of the 'Facial image processing' block of $SELFI$. Poor performance of *FBEM* demonstrates the important role played by the past emotion self-reports of the $SELFI$ framework. On the other hand, the baseline model *PCDEM* exhibits marginally inferior performance compared to $SELFI$, since the collected facial image at the previous time instance introduces noise in the model performance. Among all the variants of $SELFI$, we obtain the best performance for the {*Microsoft Azure and KPCA*} combination (macro F1: 83%). Nevertheless, for the other two facial emotion recognition toolkit (Amazon, Google) and feature reduction technique (*KPCA*, *KDA*) combination also, the $SELFI$ exhibit a decent performance (macro F1) ranging from 80% to 82%.

	Facial Image					
Features	**Amazon**		**Google**		**Microsoft**	
Emotion Class	Valence	Arousal	Valence	Arousal	Valence	Arousal
FBEM	60	54	60	57	62	60
SREM	77	77	77	77	77	77
PCDEM	80	79	80	77	81	80
KPCA SELFI	**82**	80	**82**	80	**83**	**81**
KDA SELFI	81	79	80	77	81	78

Table 1: Reporting F1-score (%) of valence and arousal for information source Facial Image. We report F1-score for different features for 3 image analysis tools (*Amazon, Google, Microsoft*) across all baselines and $SELFI$ framework. We highlight the highest accuracy score achieved by model $SELFI$.

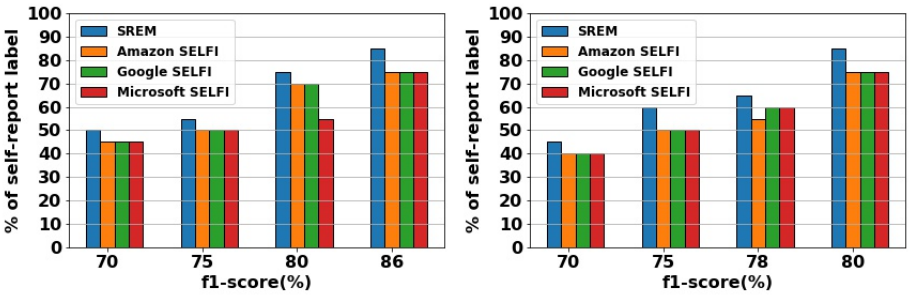
Arousal estimation: In the arousal dimension too, Table 1 demonstrates that $SELFI$ outperforms the baseline algorithm for all the variants. $SELFI$ achieves the mean F1-score of 80% for Amazon, Google and Microsoft toolkit (with both *KPCA* and *KDA* based model), whereas baseline models *FBEM*, *SREM* and *PCDEM* achieve the mean F1-score of 57%, 77% and 80% respectively. Notably, among all the $SELFI$ variants, the combination {*Microsoft Azure and KPCA*} returns the best arousal detection performance (macro F1: 81%).

In summary, $SELFI$ framework outperforms all the baseline algorithms across all model variants and emotion dimensions (valence, arousal). Since *SREM* baseline model performs a little better compared to the *FBEM*, in the rest of the paper we compare the performance of $SELFI$ with the *SREM* model only. Notably, the combination of Microsoft Azure toolkit and *KPCA* in the case of facial image exhibits superior performance for the $SELFI$ framework. The improvement in emotion prediction performance of $SELFI$ opens up the possibility of reducing the self-report burden, which we explore in detail in the following section.

6.2 Reduction in self-report burden

In this section, we measure the volume of emotion self-reports required to train the $SELFI$ framework, which can achieve the identical emotion prediction performance as the baseline model *SREM* (see Fig. 7). In this experiment, we implement the $SELFI$ framework with various image processing tools (Amazon Rekognition, Google Vision, and Microsoft Azure) and the *KPCA* feature reduction technique (since it exhibits the best performance). In Fig. 7a and Fig. 7b, the x-axis represents the model performance in terms of F1-score (%) and the y-axis represents the volume (%) of emotion self-report labels required to achieve that respective emotion prediction performance (for both valence and arousal). Overall, we observe that $SELFI$ requires fewer volume of self-report labels compared to the baseline algorithm, to achieve a similar emotion prediction performance. Precisely, in the case of valence, $SELFI$ reduces the self-report

burden by 10% for Microsoft Azure and by 6% for Amazon and Google Vision toolkits, compared to the baseline algorithm. Similarly, in the case of arousal, $SELFI$ reduces the self-report burden by 8% for Amazon Rekognition, and by 6% for Google Vision and Microsoft Azure.



(a) Self-report volume in valence (b) Self-report volume in arousal

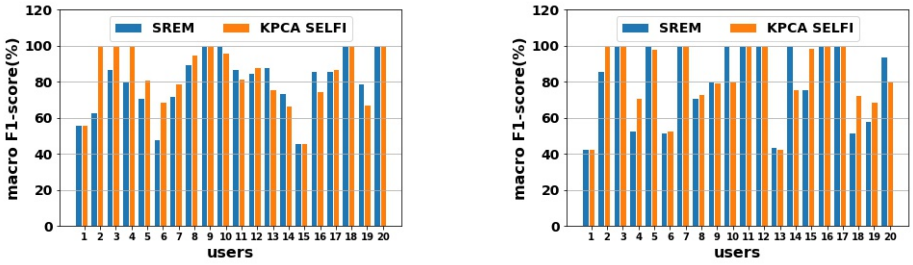
Fig. 7: $SELFI$ reduces the self-report burden.

6.3 Versatility of the *SELFI* framework

In Fig. 8, we demonstrate the versatility of the $SELFI$ framework, which allows one to plug and play with various image analysis tools (Amazon Rekognition, Google Vision, and Microsoft Azure) to implement the framework. This is comforting for us to observe that, for a majority of participants, the $SELFI$ framework outperforms the baseline model *SREM* across all the facial image analysis tools. This emphasizes the role of facial expression as an alternative information source to supplement the past self-reported emotion labels to improve the prediction performance, which in turn, paves the way to reduce the self-report burden. User centric close inspection reveals that *Microsoft Azure* based $SELFI$ framework achieves the best performance across all the image analysis tools, where 9 participants outperformed the *Baseline*. We observe that for 6 participants, the performance of $SELFI$ framework remains same as the *Baseline* algorithm, indicating that facial features do not provide any additional benefit for those participants. This performance is followed by *Amazon Rekognition* and *Google Vision* toolkit, which exhibits superior performance for 8 participants and 7 participants, respectively. In a nutshell, *Microsoft Azure* exhibits 4% performance improvement over the *SREM* model, closely followed by *Amazon Rekognition* model, which exhibits 3% improvement.

Finally, we identify 5 participants, whose performance drops compared to the *SREM* model across all the facial image analysis tools. Delving deep, we observe that *permutation importance* [4] (depicting feature importance) of facial features is negative for all these 5 participants, indicating that facial features

play a detrimental role in predicting the emotion for those participants. Manual inspection of images reveals that indeed the facial expression of those participants *visibly* does not reflect their self-reported emotion, resulting in a drop in the model performance. This observation opens up the possibility to identify the



(a) Microsoft valence performance

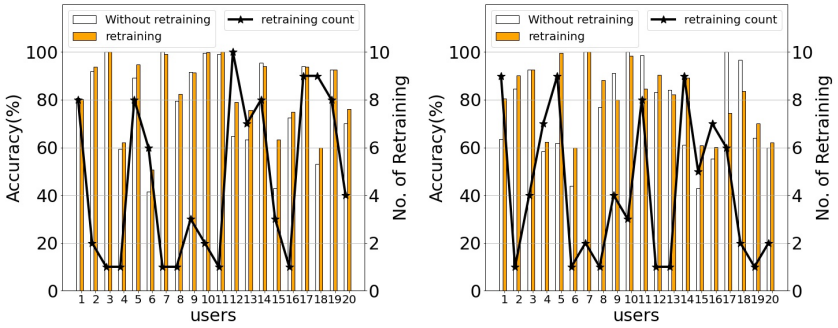
(b) Microsoft arousal performance

Fig. 8: User-centric emotion estimation performance of the best model of $SELFI$ framework.

pertinent users apriori, for whom the $SELFI$ framework is most suitable. Before applying $SELFI$, one may calculate point biserial correlation [20] between the facial features and self-reported emotion label for each user. If the facial features correlate well with the emotion self-reports, then we consider that user as *pertinent*, hence applying the $SELFI$ framework only on the *pertinent* users. Since computing correlations from numerous facial features, obtained from the facial analysis tools, may be expensive, we use the *feature reduction* step (sec 4) to select the most relevant features to compute the correlations. While applying $SELFI$ only on the *pertinent* users (15 participants in our dataset), we observe on average an appreciable 8% and 12% performance improvement over baselines, for valence and arousal respectively. However, the reduction in the self-report burden remains almost identical to Sec VI-C, on average 10% and 8% for valence and arousal respectively.

6.4 Runtime evaluation

In this section, we evaluate the runtime performance of the model developed using $SELFI$ with and without applying the retraining point detection method. We take 50% and 5% of the size of the whole dataset as the training and prediction window size since it is observed that this window size is enough to have almost balanced data. Given these settings, the number of prediction phases becomes 10. We summarize the best results in figure 9, showing two bars for each user that indicate accuracy with and without retraining. In the facial image based $SELFI$ model, we achieve the best performance for *Microsoft Azure* and *KPCA* for both the case valence and arousal. It is observed that the model accuracy improved significantly for applying retraining point detection for 12 users.



(a) Microsoft valence performance (b) Microsoft arousal performance

Fig. 9: User-wise runtime emotion estimation performance with and without retraining detection method and count of retraining instances.

On average, we achieve a 14% accuracy increment over 5 retraining instances for valence and arousal.

7 Conclusion and Discussion

The major contribution of this paper is to provide a generic platform, where one can explore various alternative information sources to train supervised models for emotion prediction, with *reduced self-report burden*. Our experiments have shown that *SELFI* framework exhibits 4% performance improvement in emotion prediction (with 83% & 81% macro-F1 in valence and arousal, respectively), compared to the baseline model *SREM*, which solely relies on the past emotion self-reports. This manifests the role of facial expression as an alternative information source, to replace the self-reported emotions directly collected from the participants. Evidently, *SELFI* facilitates us to achieve 10% reduction in self-report burden, compared to the *Baseline* algorithm.

Importantly, the elegance of $SELFI$ comes from its flexibility, which allows one to plug in various image analysis tools (say, Amazon Rekognition, Google Vision, and Microsoft Azure) and feature reduction techniques (say, *KPCA*, *KDA* etc) to develop the framework. Considering the (un)reliability of facial expression as an indicator of emotion, one may suitably explore the plausibility of various other mediums (such as audio, video, IMU sensors etc) as the alternative source of information. Moreover, in order to assess the potential benefit of $SELFI$ *in practice*, one may develop context-sensitive applications, which solely rely on the emotions estimated from *SELFI* to train the supervised models (in place of directly collecting labels from the users). The performance of those context-sensitive applications will essentially manifest the effectiveness of *SELFI* to reduce the self-report burden in practice.

References

1. Face api - facial recognition software — microsoft azure (2021), https://azure.microsoft.com/en-in/overview/what-is-azure/, [Online; accessed 29-December-2021]
2. Vision api - image content analysis — google cloud. (2021), https://cloud.google.com/vision/, [Online; accessed 29-December-2021]
3. Agrafioti, F., Hatzinakos, D., Anderson, A.K.: Ecg pattern analysis for emotion detection. IEEE Transactions on affective computing **3**(1), 102–115 (2011)
4. Altmann, A., Toloşi, L., Sander, O., Lengauer, T.: Permutation importance: a corrected feature importance measure. Bioinformatics **26**(10), 1340–1347 (2010)
5. Arshad, R., Baig, M.A., Tariq, M., Shahid, S.: Acceptability of persuasive prompts to induce behavioral change in people suffering from depression. In: Human-Computer Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part IV 17. pp. 120–139. Springer (2019)
6. Asim, Y., Azam, M.A., Ehatisham-ul Haq, M., Naeem, U., Khalid, A.: Context-aware human activity recognition (cahar) in-the-wild using smartphone accelerometer. IEEE Sensors Journal **20**(8), 4361–4371 (2020)
7. Bouhlel, N., Dziri, A.: Kullback–leibler divergence between multivariate generalized gaussian distributions. IEEE Signal Processing Letters **26**(7), 1021–1025 (2019). https://doi.org/10.1109/LSP.2019.2915000
8. Cao, L., Wang, Y., Zhang, B., Jin, Q., Vasilakos, A.V.: Gchar: An efficient group-based context—aware human activity recognition on smartphone. Journal of Parallel and Distributed Computing **118**, 67–80 (2018)
9. Chitkara, S., Gothoskar, N., Harish, S., Hong, J.I., Agarwal, Y.: Does this app really need my location? context-aware privacy management for smartphones. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **1**(3), 1–22 (2017)
10. Diamantini, C., Mircoli, A., Potena, D., Storti, E.: Automatic annotation of corpora for emotion recognition through facial expressions analysis. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 5650–5657. IEEE (2021)
11. Frijda, N.H.: Moods, emotion episodes, and emotions. (1993)
12. Furey, E., Blue, J.: Alexa, emotions, privacy and gdpr. In: Proceedings of the 32nd International BCS Human Computer Interaction Conference 32. pp. 1–5 (2018)
13. Gund, M., Bharadwaj, A.R., Nwogu, I.: Interpretable emotion classification using temporal convolutional models. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 6367–6374. IEEE (2021)
14. H2O.ai: H2O: Scalable Machine Learning Platform (2020), https://github.com/h2oai/h2o-3, version 3.30.0.6
15. Huang, Y.N., Zhao, S., Rivera, M.L., Hong, J.I., Kraut, R.E.: Predicting well-being using short ecological momentary audio recordings. In: Extended abstracts of the 2021 CHI conference on human factors in computing systems. pp. 1–7 (2021)
16. Hume, D.: Emotions and moods. Organizational behavior (258-297) (2012)
17. Khalil, R.A., Jones, E., Babar, M.I., Jan, T., Zafar, M.H., Alhussain, T.: Speech emotion recognition using deep learning techniques: A review. IEEE Access **7**, 117327–117345 (2019)
18. Khwaja, M., Matic, A.: Personality is revealed during weekends: Towards data minimisation for smartphone based personality classification. In: Human-Computer

Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part III 17. pp. 551–560. Springer (2019)
19. Ko, B.C.: A brief review of facial emotion recognition based on visual information. sensors **18**(2), 401 (2018)
20. Kornbrot, D.: Point biserial correlation. Wiley StatsRef: Statistics Reference Online (2014)
21. Larson, R., Csikszentmihalyi, M.: The experience sampling method. In: Flow and the foundations of positive psychology, pp. 21–34. Springer (2014)
22. Lim, J., Jeong, C.Y., Lim, J.M., Chung, S., Kim, G., Noh, K.J., Jeong, H.: Assessing sleep quality using mobile emas: opportunities, practical consideration, and challenges. IEEE Access **10**, 2063–2076 (2022)
23. Liu, W., Zhang, L., Tao, D., Cheng, J.: Reinforcement online learning for emotion prediction by using physiological signals. Pattern Recognition Letters **107**, 123–130 (2018). https://doi.org/https://doi.org/10.1016/j.patrec.2017.06.004, `https://www.sciencedirect.com/science/article/pii/S0167865517302003`, video Surveillance-oriented Biometrics
24. Mandi, S., Ghosh, S., De, P., Mitra, B.: Emotion detection from smartphone keyboard interactions: role of temporal vs spectral features. In: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing. pp. 677–680 (2022)
25. Mavs: ATOM: A Python package for fast exploration of machine learning pipelines (2019), `https://tvdboom.github.io/ATOM/`, aTOM version 2.0.3
26. Mehrotra, A., Vermeulen, J., Pejovic, V., Musolesi, M.: 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. pp. 723–732 (2015)
27. Pejovic, V., Musolesi, M.: Interruptme: designing intelligent prompting mechanisms for pervasive applications. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. pp. 897–908 (2014)
28. Posner, J., Russell, J.A., Peterson, B.S.: The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and psychopathology **17**(3), 715–734 (2005)
29. Qi, W., Su, H., Aliverti, A.: A smartphone-based adaptive recognition and real-time monitoring system for human activities. IEEE Transactions on Human-Machine Systems **50**(5), 414–423 (2020)
30. Rabbi, M., Li, K., Yan, H.Y., Hall, K., Klasnja, P., Murphy, S.: Revibe: a context-assisted evening recall approach to improve self-report adherence. Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies **3**(4), 1–27 (2019)
31. Rasmussen, C.E.: Gaussian processes in machine learning. In: Summer school on machine learning. pp. 63–71. Springer (2003)
32. Roth, V., Steinhage, V.: Nonlinear discriminant analysis using kernel functions. In: NIPS. vol. 12, pp. 568–574 (1999)
33. Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology **39**(6), 1161–1178 (1980)
34. Sarker, I.H., Abushark, Y.B., Khan, A.I.: Contextpca: Predicting context-aware smartphone apps usage based on machine learning techniques. Symmetry **12**(4), 499 (2020)
35. Schmidt, P., Reiss, A., Dürichen, R., Laerhoven, K.V.: Wearable-based affect recognition—a review. Sensors **19**(19), 4079 (2019)

36. Schmidt, P., Reiss, A., Dürichen, R., Van Laerhoven, K.: Labelling affective states" in the wild" practical guidelines and lessons learned. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. pp. 654–659 (2018)
37. Schölkopf, B., Smola, A., Müller, K.R.: Kernel principal component analysis. In: International conference on artificial neural networks. pp. 583–588. Springer (1997)
38. Sedgwick, P.: Snowball sampling. Bmj **347** (2013)
39. Sepas-Moghaddam, A., Etemad, A., Correia, P.L., Pereira, F.: A deep framework for facial emotion recognition using light field images. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 1–7 (2019). https://doi.org/10.1109/ACII.2019.8925445
40. Shahriar, S., Kim, Y.: Audio-visual emotion forecasting: Characterizing and predicting future emotion using deep learning. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). pp. 1–7. IEEE (2019)
41. Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., Xu, X., Yang, X.: A review of emotion recognition using physiological signals. Sensors **18**(7), 2074 (2018)
42. Suhara, Y., Xu, Y., Pentland, A.: Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In: Proceedings of the 26th International Conference on World Wide Web. pp. 715–724 (2017)
43. Tashtoush, Y.M., Orabi, D.A.A.A.: Tweets emotion prediction by using fuzzy logic system. In: 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). pp. 83–90. IEEE (2019)
44. Varma, S., Simon, R.: Bias in error estimation when using cross-validation for model selection. BMC bioinformatics **7**(1), 1–8 (2006)
45. Weiss, H.M., Cropanzano, R.: Affective events theory. Research in organizational behavior **18**(1), 1–74 (1996)
46. Wikipedia contributors: Amazon rekognition — Wikipedia, the free encyclopedia (2021), `https://en.wikipedia.org/w/index.php?title=Amazon_Rekognition&oldid=1024901190`, [Online; accessed 29-December-2021]
47. Zhang, X., Li, W., Chen, X., Lu, S.: Moodexplorer: Towards compound emotion detection via smartphone sensing. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **1**(4), 1–30 (2018)
48. Zhang, Z., Wu, B., Schuller, B.: Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6705–6709. IEEE (2019)