The Way it Makes you Feel Predicting Users' Engagement during Interviews with Biofeedback and Supervised Learning

Daniela Girardi*, Alessio Ferrari†, Nicole Novielli*, Paola Spoletini‡ Davide Fucci§, Thaide Huichapa‡

* University of Bari, Italy, † CNR-ISTI, Italy, ‡ Kennesaw State University, GA, USA, § BTH, Sweden

Email: daniela.girardi@uniba.it, alessio.ferrari@isti.cnr.it, nicole.novielli@uniba.it, pspoleti@kennesaw.edu
davide.fucci@bth.se, thuichap@students.kennesaw.edu

Abstract—**Capturing users' engagement is crucial for gathering feedback about the features of a software product. In a market-driven context, current approaches to collect and analyze users' feedback are based on techniques leveraging information extracted from product reviews and social media. These approaches are hardly applicable in bespoke software development, or in contexts in which one needs to gather information from specific users. In such cases, companies need to resort to face-to-face interviews to get feedback on their products. In this paper, we propose to utilize biofeedback to complement interviews with information about the engagement of the user on the discussed features and topics. We evaluate our approach by interviewing users while gathering their biometric data using an Empatica E4 wristband. Our results show that we can predict users' engagement by training supervised machine learning algorithms on the biometric data. The results of our work can be used to facilitate the prioritization of product features and to guide the interview based on users' engagement.**

I. Introduction

The central role of *users* in requirements engineering (RE), as well as the relationship between user involvement and product success, is widely acknowledged [1]. Keeping users in the loop and properly collecting their feedback supports the development of more usable products, leads to improved satisfaction [2] and understanding of requirements [3], and assists in maintaining long-term relationships with customers [4].

Users' feedback can be collected through different means. A recent stream of literature in *crowd RE* [5], [6] and *data-driven RE* [7], [8] focuses on gathering and analyzing feedback leveraging data analytics applied to users' opinions and product usage data. In the case of bespoken development (i.e., when customer- or domain-specific products' requirements need to be engineered), it is still common to follow traditional RE practices, such as prototyping, observations, usability testing, and focus groups [9]. Among these techniques, user interviews are one of the most widely used to gather requirements and feedback [10], [11]. Accordingly, the research community has focused on aspects that are related to interview success (and failure), such as the role of domain knowledge [11], [12], ambiguity [13], and typical mistakes of requirements analysts [14]. Currently, little attention is dedicated to the emotional aspects of interviews and, in particular, to users' *engagement*. Capturing users' engagement is crucial for gathering feedback about the features of a certain product, and have a better understanding of their preferences. The field of *affective RE* acknowledged the role of users' emotions and studied it extensively. Contributions include applications of sentiment analysis to app reviews [15], analysis of users' facial expressions [16], [17], the study of physiological reactions to ambiguity [18], and the augmentation of goal models with user emotions elicited through psychometric surveys [19].

In this paper, we aim to extend the body of knowledge in affective RE by studying users' emotions during interviews. We focus on *engagement*—i.e., the degree of positive or negative interest on a certain product-related aspect discussed in the interview. We perform a study with 31 participants taking part in a simulated interview during which we capture their biofeedback using an Empatica E4 wristband, and collect their self-assessed engagement. We compare different machine learning algorithms to predict users' engagement based exclusively on features extracted from biofeedback signals.

Our experiments show that topics related to *privacy*, *ethics* and *usage habits* tend to create more positive users' engagement. Furthermore, we show that engagement can be predicted in terms of valence and arousal [20] with an improvement in terms of F1-measure of 22% and 46%, respectively, when compared to a baseline.

This paper makes three contributions:

- One of the first studies on user engagement in requirements interviews, confirming the intuition that different types of engagement are experienced by users depending on the topic.
- A methodology, based on biometric features and machine learning, which can be applied to predict users' engagement during requirements interviews.
- A replication package[1] to enable other researchers to build on our results.

[1] https://doi.org/10.6084/m9.figshare.11864994

2332-6441/20/$31.00 ©2020 IEEE
DOI 10.1109/RE48521.2020.00016

The remainder of the paper is structured as follows. In Section II, we present background definitions of engagement and emotions, as well as related work in RE and software engineering. In Section III, we report our study design, whereas Section IV reports its results. We discuss the implications of our study in Section V and its limitations in Section VI. Finally, Section VII concludes the paper.

II. Background and Related Work

In this section, we present the background on affect modelling and emotion classification using biofeedback. Furthermore, we discuss relevant related work in the broader area of emotions in RE and use of biometrics in software engineering.

A. Engagement and Emotions

Affective states vary in their degree of stability, ranging from personality traits—i.e., long-standing, organized sets of characteristics of a person—to emotions—i.e., transient and typically episodic, dynamic, and structured events. Emotions involve perceptions, thoughts, feelings, bodily changes, and personal dispositions to experience them. Emotions are episodic and dynamic in that, over time, they can vary depending on several factors [21].

Several theories of emotions emerged in the last few decades. Specifically, cognitive models define emotions in terms of reactions to cognition. It is the case of the OCC model [22], which defines a taxonomy of emotions and identifies them as *valenced* (positive or negative) reactions to cognitive processes aimed at evaluating objects, events, and agents. Similarly, Lazarus describes nine negative (Anger, Anxiety, Guilt, Shame, Sadness, Envy, Jealousy, and Disgust) and six positive (Joy, Pride, Love, Relief, Hope, and Compassion) emotions, as well as their *appraisal* patterns. Positive emotions are triggered if the situation is congruent with one's goals; otherwise, negative emotions are triggered [23].

In line with these theories, we use emotions as a proxy for users' *engagement* during interviews. Our choice is further corroborated by previous empirical findings demonstrating how emotions can be leveraged for detecting engagement in speech-based analysis of conversations [24] or to detect students' motivation [25]. When evaluating the importance of a feature, the appraisal process of an individual is responsible for triggering an emotional reaction based on the perceived importance and relevance of a given aspect with respect to his/her goal, values, and desires.

Consistently with prior research on emotion awareness in software engineering [26]–[28], we use a dimensional representation of developers' emotions. Specifically, we adopt the Circumplex Model of Affect by Russel [20], which represents emotions according to two continuous dimensions—*valence* (from pleasant to unpleasant) and *arousal* (from activation to deactivation). Pleasant emotional states, such as happiness, are associated with *positive* valence, while unpleasant ones, such as sadness, are associated with *negative* valence. The arousal dimension captures the level of emotional activation, which ranges from inactive or *low* to active or *high*.

We expect to observe different forms of engagement in relation to valence and arousal: positive-high engagement (i.e., positive valence and high arousal) may occur when users discuss topics that they consider relevant and towards which they have a positive feeling, e.g., a feature users like and have an opinion they want to discuss about; negative-high engagement (i.e., negative valence and high arousal) may occur when topics are relevant but more controversial, such as a feature that users do not like, or a bug they find annoying. Low engagement may occur when the user does not have a strong opinion on the topic of the discussion, and is either calm (positive valence, low arousal) or bored by the conversation (negative valence, low arousal).

B. Biofeedback-based Classification of Emotions

Affective computing largely investigated emotion recognition from several physiological signals [29]–[33]. Previous research investigated the link between affective states and the electrical activity of the brain—e.g., using electroencephalogram (EEG) [32], [34]–[36], the electrical activity of the skin, or elecrodermal activity (EDA) [37], [38], the electrical activity of contracting muscles measured using electromyogram (EMG) [31], [33], [39], and the blood volume pulse (BVP) from which heart rate (HR) and its variability (HRV) are derived [29], [40].

Electrodermal activity (EDA) measures the electrical conductance of the skin due to the sweat glands activity. EDA correlates with the arousal dimension [41] and its Variation occur in presence of emotional arousal and cognitive workload. Hence, EDA has been employed to detect excitement, stress, interest, attention as well as anxiety and frustration [37], [38].

Heart-related metrics are also used in the literature. Blood volume pressure (BVP) is related to the changes in the volume of blood in vessels, while Heart Rate (HR) and its Variability (HRV) capture the rate of heart beats. Significant changes in the BVP are observed in response to increase cognitive and mental load [42]. Increases in HR occur when the body needs a higher blood supply, for example in presence of mental or physical stressors [43]. As such, heart-related metrics have been successfully employed for emotion detection [29], [40].

In a recent study, Girardi et al. [28] identify a minimum set of sensors including EDA, BVP, and HR for valence and arousal classification. To collect such biometrics, they use the Empatica E4 wristband and detect developers' emotions during software development tasks. They found that the performance obtained using only the wristband are comparable to the one obtained using an EEG helmet together with the wristband. Accordingly, in this study we use EDA, BVP, and HR collected using Empatica E4, a noninvasive device that participants can comfortably wear during interviews (see Section III-C).

C. Sentiment and Emotions in Requirements Engineering

Researchers recognize the importance of considering users' emotions in RE activities [44].

Data such as stakeholders' communication traces and feedback (e.g., tweets and app reviews) are collected and analyzed once a software product is in use (e.g., sentiment extracted from reviews on the current version of an app is analyzed to prioritize new features). Studies in this area focus on the application of natural language processing to textual artefacts. For example, Guzman et al. [45] used sentiment analysis on a large dataset of 10M tweets about 30 different software applications. They found that tweets are mostly neutral (85%), whereas negative emotions correlates with complaints and positive with praises about existing features. Martens and Maalej [46] applied sentiment analysis to 7M app reviews over 245 free and paid apps. They found a correlation between users' sentiment and app category and a moderate correlation between the rating (e.g., 1–5 stars) and sentiment.

Users' emotions extracted from app stores reviews have been also used to evaluate single app features (i.e., [15], [47], [48]). Sentiment information extracted from a textual source provides features for machine learning approaches developed to support RE tasks. For example, Maalej and Nabil [49] proposed a method that uses sentiment scores to classify app reviews into bug reports or feature requests to help stakeholder dealing with large amount of feedback. Kurtanović and Maalej [50] use sentiment scores to investigate how users argue and justify their decisions in Amazon App Store reviews. Other uses of sentiment analysis in RE include the prediction of tickets escalation in customer support systems [51].

Finally, emotions are considered in early-stage RE activities, such as elicitation and modelling. For example, Colomo-Palacios et al. [52] asked users to rank requirements according to Russel's Valence-Arousal theory. Such information is then used to improve the resolution of conflicting requirements. Other researchers used information regarding users' emotions gathered through psychometrics (e.g., surveys) to augment traditional requirements goal modelling approaches [19], [53] and artefacts, such as user stories [54].

D. Biofeedback in Software Engineering and RE

A recent research trend emerged to study the use of biometric sensors for recognition of cognitive and affective states of software developers. Fucci et al. [55] use EEG, EDA, HR, and BVP to distinguish between code and prose comprehension tasks. Fritz et al. [56] use EEG, BVP, and eye tracker to measure the difficulty of programming tasks and prevent the introduction of bugs. In a follow up work, the same set of sensors is used to classify emotional valence during programming tasks [26]. Girardi et al. [28] replicate previous findings by Müller and Fritz [26] regarding the use of non-invasive sensors for valence classification during software development tasks. Furthermore, they also address the classification of emotional arousal. Combining EDA, HR, HRV allows predicting developers' interruptibility [57] and identifying code quality concerns [58].

Biofeedback has been used also in RE, mainly to capture users' emotions *while* using an app. For example, Scherr et al. [59] and Mennig et al. [17] uses the mobile phone cameras to recognize facial muscle movements and associate them to the emotions users experience when using different features of an app. This methodology has been recently proposed and applied to user validation of new requirements [16] and to the identification of usability issues [60] with minimal privacy concerns [61]. Specifically focused on requirements elicitation interviews is the proposal of Spoletini et al. [18]. Their work focuses on ambiguity and it is at the research preview stage (i.e., no experiments have been published).

With respect to works using biofeedback in software engineering and RE, our study is among the first ones to specifically focus on users' interviews rather than product usage or development tasks. Previous work focusing on users during product interaction (e.g., [17], [59]) can detect the engagement experienced *while* using the software features. We aim to detect users' engagement about certain features when users reflect on the features and speak about them, thus capturing a different moment—a verbalized, more rational one—of the relationship between the user and the product. Furthermore, in interviews we can consider *what if* scenarios (e.g., financial and privacy-related questions in Table I), which is not possible when performing observations without interacting with users.

III. Research Design

A. Research Questions

The main goal of this study is to understand to what extent we can use biofeedback devices to predict users' engagement during interviews. Accordingly, we formulate the following research questions (RQs).

- **RQ1:** *What range of engagement do users report during an interview?* With this question, we aim at gaining a preliminary understanding of the ranges of the engagement-related data obtained from users. Specifically, we want to understand which are the variations in terms of engagement reported by users when providing opinions about a certain product. To that end, we interview Facebook users[2], asking their opinion about the platform. After the interview, we ask them to report their engagement for each of the different questions.
- **RQ2:** *To what extent can we predict users' engagement using biofeedback measurements and supervised classifiers?* With this question, we aim to understand whether it is possible to automatically recognize engagement. More specifically, we aim at assessing to what extent we can recognize emotional valence and arousal—i.e., the two dimensions we use for the operationalization of engagement. To achieve this goal, we evaluate and compare different supervised machine learning classifiers. During the interviews with users, we acquire their raw biometrics. We use features extracted from such signals, and consider intervals of reported engagement as classes to be predicted.

B. Study Participants

We recruited 31 participants among the students of Kennesaw State University with an opportunistic sampling. The

[2] Although our study is not primarily oriented to consumers' products, selecting Facebook as discussion topic facilitates the selection of participants.

participation was not restricted by major or academic level, but the only main requirement was to be an active Facebook user (access to Facebook at least once per day, self-declared), as the user interview questions dealt with this social network. More than 90% of the participants were undergraduate students divided in 11 majors. To account for differences in biometrics due to physiological aspects [62], we try to have a pool of participants as varied as possible by including multiple ethnic groups and both female and male subjects. Specifically, approximately 65% of the participants were male, and their age varied between 18 and 34 with both median and average equal to 22. Participants were either native speakers or proficient in English. The majority (58%) were white/Caucasian, 23% black/African American, 13% Hispanic/Latino, and the remaining 6% was Asian/Pacific islander. During the data analysis, we removed 10 participants because either the collected data were incomplete or the available information were not considered reliable (e.g., they provided the same response to all the questions in the surveys). Of the remaining 21 participants, approximately 67% were male with the following racial/ethnicity distribution, 67% white/Caucasian, 14% black/African American, 14% Hispanic/Latino, and 5% Asian/Pacific islander. Participants received a monetary incentive of $25 for up of one hour of their time. The study was approved by the Kennesaw State University review board (study 16-068).

C. Device and Signals

The device we use to acquired the biofeedback is the Empatica E4[3] wristband. We selected it as it is used in several studies in affective computing [43] as well as in the field of software engineering [26], [55]). Furthermore, recent research identified a minimal set of biometrics for reliable valence and arousal detection, consisting in the EDA, BVP, and HR measured by the E4 wristband [28]. Using the Empatica E4, we collected the following signals:

- **Electrodermal Activity**: EDA can be evaluated based on measures of skin resistance. Empatica E4 achieves this by passing a small amount of current between two electrodes in contact with the skin, and measuring electrical conductance (inverse of resistance) across the skin. EDA is considered a biomarker of individual characteristics of emotional responsiveness and, in particular, it tends to vary based on attentive, affective, and emotional processes [63].
- **Blood Volume Pulse**: BVP is measured by Empatica E4 through a photoplethysmography (PPG)—an optical sensor that senses changes in light absorption density of the skin and tissue when illuminated with green and red lights [64], [65].
- **Heart Rate**: HR is measured by Empatica E4 based on elaboration of the BVP signal with a proprietary algorithm.

D. Supervised Learning Algorithms

We address the problem of predicting user engagement (RQ2) using machine learning. In line with previous research on biometrics [26], [28], [31], [55], we chose popular algorithms—i.e., Naive Bayes (nb), K-Nearest Neighbor (knn), C4.5-like trees (J48), SVM with linear kernel (svm), Multilayer Perceptron for neural network (mlp), and Random Forest (rf).

E. Experimental Protocol and Data Collection

Three main roles are involved in the experiment: *interviewer*, *user*, and *observer*. The interviewer leads the experiment by asking questions to the user, while the observer tracks the interview by annotating timestamps of each question, monitoring the output of the wristband, and annotating general observations on the interview and behaviour of the user.

The experimental protocol consists of four phases (i) device calibration and emotion triggering, (ii) user's interview, (iii) self-assessment questionnaire, and (iv) wrap-up.

a) Device calibration and emotion triggering: In line with previous research [26], [28] we run a preliminary step for device calibration and emotion elicitation. The purpose of this phase is threefold. First, we want to check the correct acquisition of the signal by letting the wristband record the raw signals for all biometric sensors under the experimenter scrutiny. Second, the collected data will be needed to adjust the scores obtained during the self-assessment questionnaire (see Sect. III-F). Third, we want the participants to get acquainted with the emotion self-report task. Accordingly, we run a short emotion elicitation task using a set of emotion-triggering pictures. Each participant watches a slideshow of 35 pictures. Each picture is displayed for 10 seconds, with intervals of five seconds between them to allow the user to relax. The whole slideshow lasts for nine minutes. During the first and last three minutes, calming pictures are shown to induce a neutral emotional state, while during the central 3 minutes the user sees pictures aimed at triggering negative and positive emotions. The pictures have been selected from the Geneva database [66] previously used in software engineering studies by Müller and Fritz [26]. The user is then asked to fill a form to report the degree of arousal and valence they associated to the pictures on a visual scale from 0 to 100. As done in previous work [26], for each picture, the user is asked two questions, 1) You are judging this image as 0 = Very Negative; 50 = Neutral; 100 = Very Positive; 2) Confronted with this image you are feeling 0 = Relaxed, 50 = Neutral, 100 = Stimulated.

b) User's Interview: A trained interviewer conducts the interview with each user. The interview script consists of 38 questions concerning the Facebook platform. Questions are grouped into seven topics—i.e., *usage habits*, *privacy*, *procedures*, *relationships*, *information*, *money*, and *ethics*. The questions are reported in Table I. For each topic, we include multiple questions, to allow users sufficient time to get immersed in the topic, and have more stable biofeedback parameters in relation to the topic. Questions related to topics we expect to raise more engagement, (i.e., privacy, relationship, money, and

[3] https://www.empatica.com/research/e4/

USAGE HABITS
1. Do you use the Facebook chat function? 2. (If yes to 1) Who are the people you talk to most frequently using the Facebook chat? (If no to 1) Do you use any other chat applications? 3. How many hours do you use Facebook per day? 4. When you check Facebook, what is the average length of time you spend per session? 5. Is Facebook your primary source of social media? (If yes, why? If no, what other social media you use more often? Why is it superior?)
PRIVACY
6. If someone shared a photo of you in an embarrassing, incriminating, or shameful situation, how would you react? (Do you think Facebook has a responsibility to prevent it from happening? Should they be allowed to remove the photo on your behalf?) 7. If someone tagged you in a post which contained topics you are not comfortable sharing on Facebook (e.g., your political view, sexual preference, . . .), how would you react? (Do you think Facebook has a responsibility to prevent it from happening?) 8. How would you feel knowing that someone (e.g. your SO) accessed your profile and searched it? 9. Imagine Facebook begins using profile information to generate ad content. Would you be okay with this? (why?) 10. In relation to Facebook, what is private information?
PROCEDURE
11. Can you explain me how to add a new friend on Facebook? 12. Can you explain me how to find Facebook pages that match your interest? 13. Can you explain to me how to block a person on Facebook?
RELATIONSHIP
14. Are you connected on Facebook with members of your family? (If so, do you interact with them using Facebook? If not, why?) 15. Have you ever had a family member (even of your extended family) delete you from his/her friend list? (If so why?) 16. Have you ever wanted to delete or deleted a family member (even of the extended family) from your set of friends? (If so why?) 17. Have you ever used Facebook to begin a long-distance relationship with someone you could not realistically meet? (If so, tell us about it.) 18. Have you ever considered ending a friendship/relationship over their Facebook behavior? (What did they do to make you consider this?)
USAGE HABITS
19. Do you use Facebook using the mobile app or your PC? 20. Do you post regularly on the dashboard? 21. Do you click on posts that link to other websites?
PROCEDURE
22. Can you explain to me how to set the privacy settings? 23. Can you explain to me how to change the password?
MONEY
24. Would you agree to pay a subscription to use Facebook? If yes, how much would you consider a reasonable amount to pay? (If not, why?) 25. If the application for PC available from your browser was free, but the mobile app was not. Would you pay for it? 26. Suppose that the free access to Facebook was limited in time, information you can access or which version of the app you can use. Which of these functionalities would have to be excluded from the free version for you to be interested in the subscription? Why that Specific one? 27. If Facebook would pay you in exchange for you performing tasks or taking surveys, would you be interested in them? (If yes, for how much? If the tasks could be considered unethical, would you still do it?) 28. Suppose Facebook will become a subscription service starting from tomorrow and you decide not to pay. What should Facebook do with your profile and data?
INFORMATION
29. When you read something that you find interesting, do you share it?(What motivates you to share it? Are you likely to share something without reading it?) 30. Is the information on Facebook more or less reliable than other sources? (For what reason?) 31. What is inappropriate information for Facebook? (Is there any information that should never reach Facebook? Should Facebook be used as a news source?)
PROCEDURE
32. Can you explain to me how create a post and tag someone into it? 33. Can you explain to me how to find friends that have no mutual friends?
ETHICS
34. FB censures some photos and posts if their content is signaled as inappropriate. Do you think this is correct? Where should the line be drawn between censure and freedom? 35. Recently FB has censored pictures of women breastfeeding even if the breast was not visible? Why do you think they do this? Should they be allowed to? 36. Recently FB workers admitted to routinely suppressing conservative content, do you feel they did anything wrong? (Why or why not?) 37. Should FB play a role in limiting/removing hate speech from the site? Is it ethical if they do? 38. Terrorist groups are known to have very active social media presences. Suppose Facebook began submitting information from all profiles to the government for help in tracking these groups. Would you be okay with this? Why?

TABLE I: List of questions asked during the Interview Phase

ethics) are separated by questions on topics that are expected to reduce user engagement (i.e., usage habits, procedures, and information). The lower degree of engagement for the latter topics was assessed during preliminary experiments in which the questions were drafted and finalised[4]. During the interview, the wristband records the biofeedback parameters while the observer annotates the timestamp of each question. We use this information to align the sensor data with the questions. Based on a preliminary run, each interview was estimated to last for about 20 minutes.

c) Self-assessment Questionnaire: For each question in the interview script (i.e., Q_i), the interviewer asks the participant to report their involvement using two 10-point rating scale items: ($q_A(Q_i)$) How much did you feel involved with this topic? (1 = Not at all involved; 10 = Extremely involved); ($q_V(Q_i)$) How would you rate the quality of your involvement? (1 = Extremely negative; 10 = Extremely positive). These two questions aim at measuring the engagement of the user in terms *arousal* (q_A) and *valence* (q_V). The participants' answers to these questions represent our gold standard for the machine-learning study (see Section III-F).

d) Wrap-up: The observer downloads and stores the wristband data as well as the questionnaires filled by the participant. The wristband memory is then erased to allow further recording sessions.

F. Data Collection, Pre-processing and Feature Extraction

The data from the interview questionnaire are used to produce the gold standard—i.e., the labels for valence and arousal to be predicted.

We define *positive*, *negative*, and *neutral* labels for valence, and *high*, *low*, and *neutral* labels for arousal. We discretize the scores in the rating scale following an approach utilized in previous research [26], [28]. First, we adjust the valence and arousal scores based on the mean values reported while watching the emotion-triggering pictures (see Section III-E0a). This step is necessary to take into account fluctuations due individual differences in the interpretation of the scales in the interview questionnaire. Then, we perform a discretization of the values into the three categories (i.e., labels) for each dimension using k-means clustering.[5]

To synchronize the measurement of the biometric signals with the self-assessment, we (1) save the timestamp corresponding to the interviewer asking question Q_i (i.e., $timestamp(Q_i)$), (2) calculate the timestamp associated to the next question Q_{i+1} ($timestamp(Q_{i+1})$), and (3) select each signal samples recorded between $timestamp(Q_i)$ and $timestamp(Q_{i+1})$.

For each interview question Q_i, we have a set of signal samples (for EDA, BVP and HR) within the time interval associated to Q_i, and two labels, one representing the arousal ($q_A(Q_i)$) and the other representing the valence ($q_V(Q_i)$) according to the self-assessment questionnaire. The labels are used to form the gold standard to be predicted by the algorithms based on features extracted from the signal samples.

We normalize the signals collected during the entire duration of the experiment to each participant's baseline using $Z-$score [26]. To maximize the signal information and reduce noise caused by movements, we apply multiple filtering techniques. Regarding BVP, we extract frequency bands using a band-pass filter algorithm at different intervals [29]. The EDA signal consists of a tonic component (i.e., the level of electrical conductivity of the skin) and a phasic one representing phasic changes in electrical conductivity or skin conductance response (SCR) [67]. We extract the two components using the cvxEDA algorithm [68].

TABLE II: Machine learning features grouped by physiological signal.

Signal	Features
EDA	- mean tonic
	- phasic AUC
	- phasic min, max, mean, sum peaks amplitudes
BVP	- min, max, sum peaks amplitudes
	- mean peak amplitude (diff. between baseline and task)
HR	- mean, sd. deviation (diff. between baseline and task)

After signals pre-processing, we extracted the features presented in Table II, which we use to train our classifiers. We select features based on previous studies using the same signals [26], [28], [55].

G. Analysis Procedure

The analysis procedure aims at answering the two RQs, as detailed in the following.

a) RQ1 (type of engagement and measurements): We first measure the range of engagement in terms of arousal and valence, based on the results of the self-assessment questionnaire. This allows us to understand which are the most engaging topics according to the users, and to what extent engagement varies during the interview. We collected descriptive data and provide qualitative considerations.

b) RQ2 (supervised learning): For each user, we use the biometrics gathered in the user's interview phase as input features for the different classifiers listed in Sect. III-D.

In line with previous research [26], [28], we target a binary classification task using machine learning. In particular, we distinguish between *positive* and *negative* valence and *high* and *low* arousal. As such, we exclude the *neutral* label from the gold standard and focus on more polarised values. Although this reduces our dataset, it also facilitates the separation between clearly distinguished emotional states[6]. Table III reports the gold standard dataset with valence and arousal distribution.

We evaluate our classifiers in the *Hold-out* setting. Therefore, we split the gold standard into train (70%) and test (30%) sets using the stratified sampling strategy implemented in the

[4] During the experiments reported in this paper, we saw that *usage habits* was associated with higher engagement, instead. Discussion on this aspect is reported in Sect. IV.

[5] We use the k-means implementation in by the `arules` R package

[6] Preliminary experiments were performed considering a 3-label problem, but the number of vectors resulted too small to achieve acceptable results.

Arousal			*Valence*		
High	**Low**	Neut.	**Positive**	**Negative**	Neut.
245 (66%)	191 (44%)	340	345 (79%)	89 (21%)	342

TABLE III: Label distribution in the gold standard.

R caret package [69]. We search for the optimal hyperparameters [70], [71] using leave-one-out cross validation—i.e., the recommended approach for small training sets [72] such as ours. The resulting model is then evaluated on the hold-out test set to assess its performance on unseen data and avoid overfitting. We repeat this process 10 times to further increase the validity of the results. The performance is then evaluated by computing the mean for precision, recall, F-measure, and accuracy over the different runs. This setting is directly comparable to the one implemented by Müller and Fritz [26] and by Girardi et al. [28], which includes data from the same subject in both training and test sets.

IV. Execution and Results

The data were initially gathered from 31 participants. Interviews lasted 18 minutes on average. We discarded the data from those subjects for which data were largely incomplete, or that appeared to have a low degree of standard deviation (i.e., lower than 1) in their labels of valence and arousal. Indeed, although these subjects may in principle have had little variations in their actual emotions, they can be considered outliers with respect of the rest of the subjects. As data are treated in aggregate form, and given the limited number of data points, including these outliers could have introduced undesired noise. We also discarded data whenever some inconsistency was observed through the different pre-processing steps, as, e.g., timestamps not plausible.

At the end of this process, we produced the feature vectors and associated labels for valence and arousal (776 vectors in total from 21 subjects). The scatter plot for the two dimensions is reported in Fig. 1. The normalised range of the labels, evaluated by means of k-means clustering as explained in Sect. III-G, is as follows. For valence we have: [-4.94,-1.03) *negative*; [-1.03,2.52) *neutral*; [2.52,5.31] *positive*. For arousal we have: [-4.8,0.308) *low*; [0.308,3.57) *neutral*; [3.57,7] *high*. These vectors and labels are used to compute the statistics useful to answer RQ1 (see Table III).

As our goal for RQ2 is to discriminate between high (positive) and low (negative) arousal (valence), we removed all the items for which the label resulted *neutral* for the dimensions, based on the participant's answers. Therefore, our gold standard includes only the vectors labelled as high (positive) or low (negative) and we model our problem as a binary classification task. Below, we report the results of the analysis and we answer the RQs.

A. RQ1: What is the range of reported engagement and biofeedback measurements of a user during an interview?

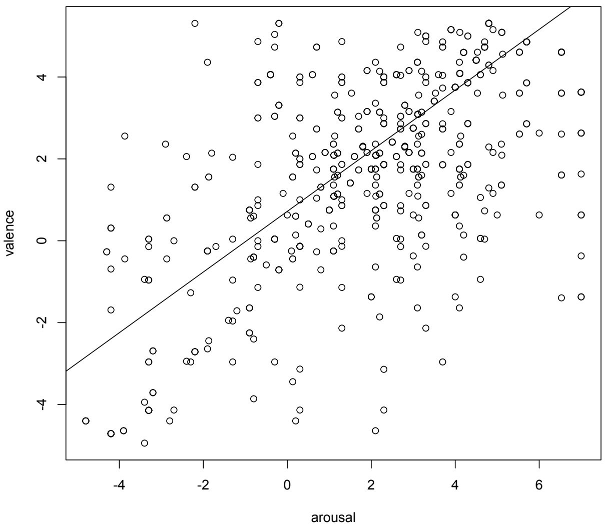

Fig. 1: Scatter plots of normalised valence and arousal according to the self-assessment questionnaire.

	Valence	Valence (norm)	Arousal	Arousal (norm)
Average	7.23	1.90	7.06	2.13
Minimum	1	-4.94	1	-4.8
Maximum	10	5.31	10	7
Std. Dev.	1.47	1.58	2.17	2.17

TABLE IV: Descriptive statistics of the reported engagement.

Table IV reports the ranges of valence and arousal, according to the self-assessment questionnaire. We report both original values and normalised ones (“norm”, in the table). We see that, overall, users tend to give high scores both for arousal and valence (both averages are above 7), indicating that the interview is generally perceived as *positively engaging*. Although they used the whole 1 to 10 scale for both dimensions, indicating that the interview appeared to cover the whole range of emotions, we see that the standard deviation is not particularly large, especially for valence. Indeed, considering the more intuitive 1-10 scale, the value of standard deviation (Std. Dev. in Table IV) indicates that around 68% of the subjects gave score in [6-9] for valence, and in [5-9] for arousal. This indicates that subjects tended to report scores around the average, and that apparently most of the interview triggered a similar level of engagement.

To gain more insight, it is useful to look at the reported engagement for each question[7]. Figure 2 reports the box plots for valence and arousal for each question, divided by question group. We see that questions related to *privacy*, *ethics* and *usage habits* tend to create more (positive) arousal in average, while questions related to *procedures* are associated to more neutral values of arousal and valence (i.e., closer to 0 in the plot). Interestingly, questions related to *relationships* show the largest variation in terms of arousal and valence (the box-plot appears larger), indicating that this is a sensitive

[7] The statistics in this case consider solely those subjects that responded to all questions, i.e., 10 in total

topic for the users, leading to more polarised scores in terms of emotional dimensions. The maximum average valence, instead, is observed for questions related to *ethics*.

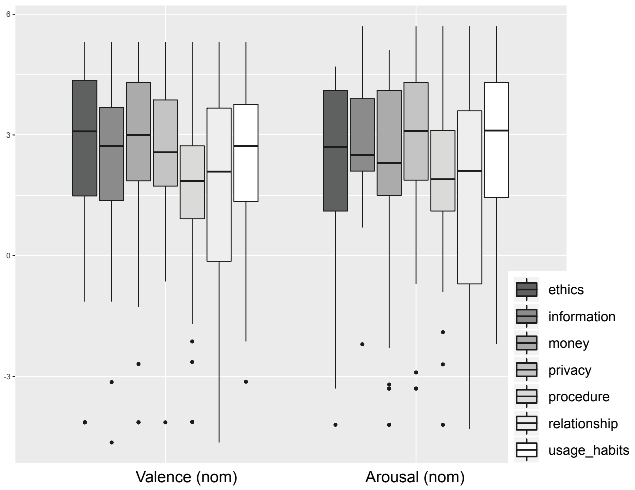

Fig. 2: Box plots of valence and arousal for each group of questions, according to the self-assessment questionnaire.

	Valence			*Arousal*		
algorithm	Prec	Rec	F1	Prec	Rec	F1
C5.0	0.400	0.500	0.440	0.594	0.575	0.615
J48	0.451	0.508	0.455	0.621	0.613	0.632
knn	0.498	0.504	0.460	0.610	0.609	0.614
mlp	0.582	0.543	0.526	0.555	0.547	0.568
nb	0.506	0.514	0.495	0.578	0.574	0.588
rf	**0.723**	**0.564**	**0.566**	**0.671**	**0.663**	**0.663**
svmLinear	0.400	0.499	0.440	0.502	0.364	0.563

TABLE V: Comparison of the performance of the different supervised learning algorithms

B. RQ2: To what extent can we predict users' reported engagement using biofeedback measurements and supervised classifiers?

	Precision	**Recall**	**F1**	**Accuracy**
		Valence		
Random Forest	0.72	0.56	0.57	0.81
Baseline	0.40	0.50	0.44	0.79
Improvement	0.33 (45%)	0.06 (11%)	0.12 (22%)	0.02 (2%)
		Arousal		
Random Forest	0.67	0.66	0.66	0.67
Baseline	0.28	0.50	0.36	0.56
Improvement	0.38 (58%)	0.16 (25%)	0.30 (46%)	0.11 (17%)

TABLE VI: Best valence and arousal classifier's performance and comparison with majority class baseline classifier. Improvement over the baseline is also shown.

In table V we report the performance of the different classifiers together with their precision, recall, and F1-measure. Specifically, for each metric we report the mean over the ten runs of the Hold-out train-test procedure, i.e. the macro-average. This choice is in line with consolidated recommendations from literature on classification tasks using machine learning [73]. Specifically, using macro-averaging is recommended with unbalanced data as ours, as it emphasizes the ability of a classifier to behave well also on categories with fewer positive training instances. We see that the Random Forest algorithm (rf) achieves the best performance across all the measures considered.

In Table VI, we report the result of the rf algorithm, and we compare it with a baseline. Following previous research on sensor-based emotion recognition in software development [28], we select as baseline the trivial classifier always predicting the majority class, that is *high* for arousal and *positive* for valence. For the sake of completeness, we also report accuracy even if its usage is not recommended in presence of unbalanced data as ours.

For valence, the Random Forest classifier distinguishes between negative and positive emotions with an F1 of 0.57, thus obtaining an increment of 22% with respect to the baseline. Furthermore, we observe an improvement in precision of 45% (from 0.40 of the baseline to 0.72 of random forest) and 11% in recall (from 0.50 to 0.56). These results indicate that the classifiers' behavior is substantially better than the baseline classifier that always predicts the positive class. Furthermore, it confirms the inadequacy of accuracy as a metric for assessing performance in supervised learning for imbalanced data.

As for arousal, we observe a better performance. Our classifier distinguishes between high and low activation with an F1 of 0.66, representing an improvement of 46% over the baseline (0.36). Again, the classifier substantially outperforms the baseline with an improvement of 58% for precision (from 0.28 to 0.67) and 25% for recall (from 0.50 to 0.66).

Looking at the confusion matrix of the best train-test round (see Table VII), we observe that the main reason for misclassification is due to the classifier bias towards the majority class for both valence and arousal. In fact, both classifiers tend to predict more often the *positive* label for valence and the *high* label for arousal, thus lowering the recall of the *negative* and *low* class and the precision for the the *positive* and *high* classes.

V. Discussion

The main take-away messages of this study are 1) users' interviews are activities that can trigger positive engagement in the involved users, 2) different levels of engagement are experienced depending on the topic of the question, with topics such as *privacy*, *ethics* and *usage habits* leading to higher engagement, and *relationships* leading to larger variations of engagement, 3) by combining biometric features into vectors and by training a Random Forest algorithm, it is possible to predict the engagement in a way that outperforms a majority-class baseline. We discuss our results in relation to existing literature and outline possible applications of our results.

Engagement and topics The results of RQ1 indicate that users experienced different levels of engagement with respect to the question topic. Specifically, our participants reported a positive attitude when discussing privacy, ethics, and usage habits. Concerning privacy and ethics, these topics were

Arousal							
Confusion matrix				**Performance**			
		Predicted Label		*Label*	*Precision*	*Recall*	*F1*
		High	Low	High	0.75	0.74	0.74
Gold Label	High	54	19	Low	0.67	0.68	0.68
	Low	18	39	macroAvg	0.71	0.71	0.71
Valence							
Confusion matrix				**Performance**			
		Predicted Label		*Label*	*Precision*	*Recall*	*F1*
		Negative	Positive	Negative	0.55	0.23	0.32
Gold Label	Negative	6	20	Positive	0.83	0.95	0.89
	Positive	5	98	macroAvg	0.69	0.59	0.61

TABLE VII: Confusion matrices and performance by class for the best train-test round with Random Forest for Arousal and Valence.

selected on purpose to trigger higher engagement. Given the raising interest in these two fields, especially in relation to Facebook and online communities in general (e.g., Trice et al. [74]) the obtained results are not surprising. Concerning *usage habits*, we expected to see lower values of arousal. As questions regarding usage habits were asked at the beginning, the high arousal observed may be resulting from the excitement of the new experience. However, we observed that question 19, also about usage habits but asked later, was the one with the highest average arousal (3.6 in normalised values, while the average for all questions regarding usage habits is 2.8) and valence (3.2 vs 2.5)[8]. Therefore, we argue that speaking about usage habits triggers positive engagement. This indicates that users generally like the platform and are interested in speaking about their habitual relation with it. Qualitative analysis of the audio of the actual answers, not performed in this study, can further clarify these aspects. Overall, these results show that 1) users' interviews elicit emotions and engagement, with varying degrees of reactions depending on the topic; and 2) some topics are perceived as more engaging than others.

Performance comparison with related studies In this study, we adopt emotions as a proxy for engagement (see Sect. II). Specifically, we operationalize emotions along the valence and arousal dimensions which we recognize using biometrics. In particular, using machine learning, we are able to distinguish between positive and negative valence and high and low arousal with a performance that is comparable to the one obtained by previous studies [26], [28]. A direct comparison is possible with the results of the empirical study by Girardi et al. [28] as we use the same device (i.e., Empatica E4 wristband) to replicate their sensor setting including EDA, BVP, and HR. Our macro-average F1 for arousal (0.66) and valence (0.57) is comparable to the one they obtain using only the Empatica—i.e., 0.55 for arousal and 0.59 for valence. They report a slightly better performance when including also the EEG helmet (F1 = .59 for aousal and F1 = 0.60 for valence). Müeller and Fritz [26] report an accuracy of 0.71 for valence, using a combination of features based on EEG, HR, and pupil size captured by an eye-tracker.

Our approach shows better performance for arousal than for valence. This can be due to the link between emotional arousal and the biometrics collected by Empatica E4. Previous work suggests that changes in the EEG spectrum indicate the overall levels of arousal or alertness [34] as well as pleasantness of the emotion stimulus [35]. Soleymani et al. [32] found that high-frequencies sensed from electrodes positioned on the frontal, parietal, and occipital lobes have high correlation with valence. Accordingly, further replications should include EEG sensors to investigate the extent to which such signals can improve emotion classification performance in similar settings.

Compared to ours, studies in affective computing report better performance—e.g., accuracy of 0.97 for arousal [75]–[77] and 0.91 for valence [39]. However, such studies rely on high-definition EEG helmets [75]–[77] and facial electrodes for EMG [39] which are *invasive* and cannot be used outside laboratory settings—e.g., during real interviews with users.

Looking at the confusion matrices (see Table VII), we observe a drop in performance due to misclassification of cases from the minority class as belonging to the majority class. This is prevalent for valence as it has the most unbalanced distribution of positive/negative labels. This evidence suggests the needs of further replications to assess the validity of our findings with a richer, more balanced dataset which should include new subjects and a larger set of questions to trigger low-valence states.

Applications. Direct applications of our results are not straightforward as the study is oriented to have a first understanding of engagement in user interviews and on the potential usage of biofeedback devices in this context. However, we argue that our results can be useful to better investigate possible discrepancies between user engagement as sensed by the wristband and reported relevance of features, to facilitate requirements prioritization tasks. Furthermore, the usage of these technologies could be extended to identify the engagement of the user *on-the-fly*—i.e., during the interview—to guide analyst steering the flow of the interview. Similarly, we can support requirements elicitation interviews to improve the analyst ability to create a trustworthy relationship wit the customer, and improve the quality of the interview and the collected data [78]. In this context, it it relevant to extend the work to identify the customer's frustration, which often corresponds to the first step to create mistrust in the analyst [78]. Frustration can be detected using biofeedback by analyzing the changes in the heart-rate, temperature, and

[8] Results for each individual question not shown in the paper.

other vitals [79]–[82] and used to warn the analyst. In line with the proposal of Spoletini et al. [18], biofeedback can also be combined with those acoustic properties of speech that indicate emotional differences (i.e., *emotional prosody* [83]) to further evaluate the current status of the user or customer during interviews [84]. Vocal cues could be integrated in the analysis to increase the reliability of our approach.

VI. Threats to Validity

In this section, we discuss the main limitations of our study and report how do we address them.

External validity - Given the limited amount of subjects and data points, we cannot claim a large generalization power of our results. However, our study participants include multiple ethnic groups and both female and male subjects (see Sect. III-B), although with some imbalance, to account for biometric differentiation due to physiological aspects [62]. Further replications with a confirmatory design should engage more participants, and consider more balance between ethnicity, culture, age and gender to account for the differences in emotional reactions due to these aspects.

Conclusion validity - The validity of our conclusions relies on the robustness of the machine learning models. To mitigate any threat arising from having a small dataset, we ran several algorithms addressing the same classification task. In all runs, we performed hyperparameters tuning as recommended by state-of-the-art research. Following consolidated guidelines for machine learning, we split our data in two train-test subsets. The training is performed using cross validation and the final model performance is assessed on the hold-out test set. The entire process is repeated ten times for each algorithm, to further increase the validity of the study.

Construct validity - Threats to construct validity refer to the reliability of the operationalization of the problem. Our study may suffer to threats to construct validity in capturing emotions with self-report. To address the problem of potential unreliability of the self-reported rating, we performed data quality assurance and did not consider participants who provided always the same score or scores with overall low standard deviation. Another threat is concerned with the selection of Facebook as main argument of the interview. This was driven by the need to balance between the choice of a representative product and the ease of participants' sampling. Associated threats cannot be entirely ruled out. However, we arguably believe that the designed interview script is sufficiently representative of typical users' interviews in terms of triggered engagement.

Internal validity - Threats to internal validity regard any confounding factors that can influence the results of a study. We collected data in a laboratory setting. Factors existing in our settings, such as the presence of the experimenter, can influence the emotional status as the participants may feel they are being observed. Furthermore, self-assessment questionnaires were filled *immediately after* the interview. This choice was driven to the need to preserve a realistic interview context. However, with this design, the engagement is recalled by the subject and not reported in the moment in which it emerged. Therefore, discrepancies may occur between the feeling of engagement and its rationally processed memory. Similarly, to maintain a realistic settings, we did not perform pre-interviews to assess the participants' mood (i.e., the presence of a long-lasting emotion). We acknowledge that an emotionally-charged event in the life of a participant, either sad or happy, before the interview took place can impact the results. Furthermore, the Hawthorne effect might occur in human studies, that is changes in participants' behavior due to their awareness of being. Establishing a trust-based rapport with the participants in a relaxed setting is crucial to mitigate these threats. Thus, we invited the participant to wear the wristband when entering in the room, before the actual interview started, in order to get acquainted with the device.

VII. Conclusion and Future Work

This paper presents the first study about engagement prediction in user interviews. In particular, we show that, through the usage of biofeedback measurements acquired through a wristband and the application of supervised machine learning algorithms, it is possible to predict the positive or negative engagement of a user during an interview about a product. The approach can be extended to large scale scenarios, for example for A/B testing, when low-cost devices will be available to acquire the considered measurements. The study is exploratory in nature, and application of our results require further investigation, especially concerning the acceptance of the non-intrusive, yet potentially undesired, biofeedback device. Among the future works, we plan to: (a) replicate the experiment with a larger and more representative sample of participants; (b) complement our analysis with the usage of other emotion-revealing signals considered in other studies, such as facial expressions captured through cameras [32], voice recording [18], and electroencephalographic (EEG) activity data [26], [28]; (c) apply the study protocol to requirements elicitation interviews for novel products to be developed; (d) investigate requirements analyst's emotions in relation with users' emotions during interviews, to explore the emotional dialogue that occurs between the two of them; (e) investigate and compare the emotional footprint of different software-related tasks. This can be done for example by looking at the difference between physiological signals of the multiple actors of the development process across different phases, such as of development, elicitation, testing, *etc.* Overall, we believe that the current work, with its promising results, establishes the basis for further research on emotions during the many human-intensive activities of system development.

Acknowledgment

This work was partially supported by the KKS foundation through the S.E.R.T. Research Profile project at Blekinge Institute of Technology and by the National Science Foundation under grant CCF-1718377.

REFERENCES

[1] M. Bano and D. Zowghi, "A systematic review on the relationship between user involvement and system success," *Information and Software Technology*, vol. 58, pp. 148–169, 2015.

[2] Z. Bakalova and M. Daneva, "A comparative case study on clients participation in a'traditional'and in an agile software company," in *Proc. of the 12th Int. Conf. on product focused software development and process improvement*, 2011, pp. 74–80.

[3] G. K. Hanssen and T. E. Fægri, "Agile customer engagement: a longitudinal qualitative case study," in *Proc. of the 2006 ACM/IEEE Int. symposium on Empirical software engineering*, 2006, pp. 164–173.

[4] J. Heiskari and L. Lehtola, "Investigating the state of user involvement in practice," in *2009 16th Asia-Pacific Software Engineering Conf.* IEEE, 2009, pp. 433–440.

[5] P. K. Murukannaiah, N. Ajmeri, and M. P. Singh, "Acquiring creative requirements from the crowd: Understanding the influences of personality and creative potential in crowd re," in *2016 IEEE 24th Int. Requirements Engineering Conf. (RE)*. IEEE, 2016, pp. 176–185.

[6] E. C. Groen, N. Seyff, R. Ali, F. Dalpiaz, J. Doerr, E. Guzman, M. Hosseini, J. Marco, M. Oriol, A. Perini *et al.*, "The crowd in requirements engineering: The landscape and challenges," *IEEE software*, vol. 34, no. 2, pp. 44–52, 2017.

[7] W. Maalej, M. Nayebi, T. Johann, and G. Ruhe, "Toward data-driven requirements engineering," *IEEE Software*, vol. 33, no. 1, pp. 48–54, 2015.

[8] G. Williams and A. Mahmoud, "Mining twitter feeds for software user requirements," in *2017 IEEE 25th Int. Requirements Engineering Conf. (RE)*. IEEE, 2017, pp. 1–10.

[9] D. Zowghi and C. Coulin, "Requirements elicitation: A survey of techniques, approaches, and tools," in *Engineering and managing software requirements*. Springer, 2005, pp. 19–46.

[10] A. Davis, O. Dieste, A. Hickey, N. Juristo, and A. M. Moreno, "Effectiveness of requirements elicitation techniques: Empirical results derived from a systematic review," in *14th IEEE Int. Requirements Engineering Conf. (RE'06)*. IEEE, 2006, pp. 179–188.

[11] I. Hadar, P. Soffer, and K. Kenzi, "The role of domain knowledge in requirements elicitation via interviews: an exploratory study," *Requirements Engineering*, vol. 19, no. 2, pp. 143–159, 2014.

[12] A. M. Aranda, O. Dieste, and N. Juristo, "Effect of domain knowledge on elicitation effectiveness: an internally replicated controlled experiment," *IEEE Transactions on Software Engineering*, vol. 42, no. 5, pp. 427–451, 2015.

[13] A. Ferrari, P. Spoletini, and S. Gnesi, "Ambiguity and tacit knowledge in requirements elicitation interviews," *Requirements Engineering*, vol. 21, no. 3, pp. 333–355, 2016.

[14] M. Bano, D. Zowghi, A. Ferrari, P. Spoletini, and B. Donati, "Teaching requirements elicitation interviews: an empirical study of learning from mistakes," *Requirements Engineering*, vol. 24, no. 3, pp. 259–289, 2019.

[15] E. Guzman and W. Maalej, "How do users like this feature? a fine grained sentiment analysis of app reviews," in *2014 IEEE 22nd Int. requirements engineering Conf. (RE)*. IEEE, 2014, pp. 153–162.

[16] S. A. Scherr, C. Kammler, and F. Elberzhager, "Detecting user emotions with the true-depth camera to support mobile app quality assurance," in *2019 45th Euromicro Conf. on Software Engineering and Advanced Applications (SEAA)*. IEEE, 2019, pp. 169–173.

[17] P. Mennig, S. A. Scherr, and F. Elberzhager, "Supporting rapid product changes through emotional tracking," in *2019 IEEE/ACM 4th Int. Workshop on Emotion Awareness in Software Engineering (SEmotion)*. IEEE, 2019, pp. 8–12.

[18] P. Spoletini, C. Brock, R. Shahwar, and A. Ferrari, "Empowering requirements elicitation interviews with vocal and biofeedback analysis," in *2016 IEEE 24th Int. Requirements Engineering Conf. (RE)*. IEEE, 2016, pp. 371–376.

[19] K. Taveter, L. Sterling, S. Pedell, R. Burrows, and E. M. Taveter, "A method for eliciting and representing emotional requirements: Two case studies in e-healthcare," in *2019 IEEE 27th Int. Requirements Engineering Conf. Workshops (REW)*. IEEE, 2019, pp. 100–105.

[20] J. Russell, "Culture and the categorization of emotions," *Psychological Bulletin*, vol. 110 (3), pp. 426–450, 1991.

[21] R. Cowie, N. Sussman, and A. Ben-Ze'ev, "Emotion: Concepts and definitions," in *Emotion-oriented systems*. Springer, 2011, pp. 9–30.

[22] A. Ortony, G. Clore, and A. Collins, *The Cognitive Structure of Emotion*, 01 1988, vol. 18.

[23] R. S. Lazarus, *Emotion and Adaptation*. Oxford University Press, 1991.

[24] C. Yu, P. M. Aoki, and A. Woodruff, "Detecting user engagement in everyday conversations," in *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004*. ISCA, 2004. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2004/i04_1329.html

[25] A. Barhenke, A. L. Miller, E. Brown, R. Seifer, and S. Dickstein, "Observed emotional and behavioral indicators of motivation predict school readiness in head start graduates," *Early Childhood Research Quarterly*, vol. 26, no. 4, pp. 430–441, 2011.

[26] S. C. Müller and T. Fritz, "Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress," in *37th IEEE/ACM Int. Conf. on Software Engineering, ICSE 2015, Florence, Italy, May 16-24, 2015*, 2015, pp. 688–699.

[27] D. Graziotin, X. Wang, and P. Abrahamsson, "Do feelings matter? on the correlation of affects and the self-assessed productivity in software engineering," *Journal of Software: Evolution and Process*, vol. 27, no. 7, pp. 467–487, 2015. [Online]. Available: https://doi.org/10.1002/smr.1673

[28] D. Girardi, N. Novielli, D. Fucci, and F. Lanubile, "Recognizing Developers' Emotions while Programming," in *42nd Int. Conf. on Software Engineering (ICSE '20), May 23–29, 2020, Seoul, Republic of Korea*, 2020. [Online]. Available: https://doi.org/10.1145/3377811.3380374

[29] F. Canento, A. Fred, H. Silva, H. Gamboa, and A. Lourenço, "Multimodal biosignal sensor data handling for emotion recognition," in *SENSORS*. IEEE, 2011, pp. 647–650.

[30] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008. [Online]. Available: https://doi.org/10.1109/TPAMI.2008.26

[31] S. Koelstra, C. Mühl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis using physiological signals," *IEEE Transaction on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012. [Online]. Available: https://doi.org/10.1109/T-AFFC.2011.15

[32] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Transaction on Affective Computing*, vol. 7, no. 1, pp. 17–28, 2016. [Online]. Available: https://doi.org/10.1109/TAFFC.2015.2436926

[33] D. Girardi, F. Lanubile, and N. Novielli, "Emotion detection using noninvasive low cost sensors," in *Seventh Int. Conf. on Affective Computing and Intelligent Interaction, ACII 2017, San Antonio, TX, USA, October 23-26, 2017*, 2017, pp. 125–130.

[34] A. E. Kramer, *Physiological Metrics of Mental Workload: A Review of Recent Progress*, D. T. I. Center, Ed., 06 1990.

[35] B. Reuderink, C. Mühl, and M. Poel, "Valence, arousal and dominance in the eeg during game play," *Int. Journal of Autonomous and Adaptive Communications Systems*, vol. 6, no. 1, pp. 45–62, 2013. [Online]. Available: http://dx.doi.org/10.1504/IJAACS.2013.050691

[36] M. Li and B.-L. Lu, "Emotion classification based on gamma-band eeg," in *2009 Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, Sep. 2009, pp. 1223–1226.

[37] W. Burleson and R. W. Picard, "Affective agents: Sustaining motivation to learn through failure and state of "stuck"," in *Social and Emotional Intelligence in Learning Environments Workshop.*, 8 2004.

[38] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *Int. Journal Human-Computer Studies*, vol. 65, no. 8, pp. 724–736, Aug. 2007.

[39] P. A. Nogueira, R. A. Rodrigues, E. C. Oliveira, and L. E. Nacke, "A hybrid approach at emotional state detection: Merging theoretical models of emotion with data-driven statistical classifiers," in *2013 IEEE/WIC/ACM Int. Conf. on Intelligent Agent Technology, IAT 2013*. IEEE, 2013, pp. 253–260.

[40] J. Scheirer, R. Fernandez, J. Klein, and R. W. Picard, "Frustrating the user on purpose: a step toward building an affective computer," *Interacting with Computers*, vol. 14, pp. 93–118, 2002. [Online]. Available: https://ieeexplore.ieee.org/document/8160759

[41] P. J. Lang and M. Bradley, "The int. affective picture system (iaps) in the study of emotion and attention," in *Handbook of Emotion Elicitation and Attention*, J. A. Coan and J. J. B. Allen, Eds. Oxford University Press, 2007, ch. 2, pp. 29–46.

[42] A. Kushki, J. Fairley, S. Merja, G. King, and T. Chau, "Comparison of blood volume pulse and skin conductance responses to mental and af-

fective stimuli at different anatomical sites," *Physiological measurement*, vol. 32, no. 10, p. 1529, 2011.

[43] S. Greene, H. Thapliyal, and A. Caban-Holt, "A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health," *IEEE Consumer Electronics Magazine*, vol. 5, no. 4, pp. 44–56, 2016.

[44] A. Sutcliffe, "Emotional requirements engineering," in *2011 IEEE 19th Int. Requirements Engineering Conf.* IEEE, 2011, pp. 321–322.

[45] E. Guzman, R. Alkadhi, and N. Seyff, "An exploratory study of twitter messages about software applications," *Requirements Engineering*, vol. 22, no. 3, pp. 387–412, 2017.

[46] D. Martens and W. Maalej, "Release early, release often, and watch your users' emotions: Lessons from emotional patterns," *IEEE Software*, vol. 36, no. 5, pp. 32–37, 2019.

[47] T. Johann, C. Stanik, and W. Maalej, "Safe: A simple approach for feature extraction from app descriptions and app reviews," in *2017 IEEE 25th Int. Requirements Engineering Conf. (RE)*. IEEE, 2017, pp. 21–30.

[48] F. A. Shah, K. Sirts, and D. Pfahl, "Using app reviews for competitive analysis: tool support," in *Proc. of the 3rd ACM SIGSOFT Int. Workshop on App Market Analytics*, 2019, pp. 40–46.

[49] W. Maalej and H. Nabil, "Bug report, feature request, or simply praise? on automatically classifying app reviews," in *2015 IEEE 23rd Int. requirements engineering Conf. (RE)*. IEEE, 2015, pp. 116–125.

[50] Z. Kurtanović and W. Maalej, "Mining user rationale from software reviews," in *2017 IEEE 25th Int. Requirements Engineering Conf. (RE)*. IEEE, 2017, pp. 61–70.

[51] C. Werner, Z. S. Li, and D. Damian, "Can a machine learn through customer sentiment?: A cost-aware approach to predict support ticket escalations," *IEEE Software*, vol. 36, no. 5, pp. 38–45, 2019.

[52] R. Colomo-Palacios, C. Casado-Lumbreras, P. Soto-Acosta, and Á. García-Crespo, "Using the affect grid to measure emotions in software requirements engineering," 2011.

[53] T. Miller, S. Pedell, A. A. Lopez-Lorca, A. Mendoza, L. Sterling, and A. Keirnan, "Emotion-led modelling for people-oriented requirements engineering: the case study of emergency systems," *Journal of Systems and Software*, vol. 105, pp. 54–71, 2015.

[54] P. Kamthan and N. Shahmir, "Effective user stories are affective," in *Int. Conf. on Ubiquitous Computing and Ambient Intelligence*. Springer, 2017, pp. 605–611.

[55] D. Fucci, D. Girardi, N. Novielli, L. Quaranta, and F. Lanubile, "A replication study on code comprehension and expertise using lightweight biometric sensors," in *Proc. of the 27th Int. Conf. on Program Comprehension, ICPC 2019, Montreal, QC, Canada, May 25-31, 2019*, 2019, pp. 311–322. [Online]. Available: https://dl.acm.org/citation.cfm?id=3339126

[56] T. Fritz, A. Begel, S. C. Müller, S. Yigit-Elliott, and M. Züger, "Using psycho-physiological measures to assess task difficulty in software development," in *36th Int. Conf. on Software Engineering, ICSE '14, Hyderabad, India - May 31 - 7 June, 2014*, 2014, pp. 402–413. [Online]. Available: https://doi.org/10.1145/2568225.2568266

[57] M. Züger, S. C. Müller, A. N. Meyer, and T. Fritz, "Sensing interruptibility in the office: A field study on the use of biometric and computer interaction sensors," in *Proc. of the 2018 CHI Conf. on Human Factors in Computing Systems, (CHI 2018)*, 2018, p. 591. [Online]. Available: https://doi.org/10.1145/3173574.3174165

[58] S. C. Müller and T. Fritz, "Using (bio)metrics to predict code quality online," in *Proc. of the 38th Int. Conf. on Software Engineering, ICSE 2016, Austin, TX, USA, May 14-22, 2016*, 2016, pp. 452–463. [Online]. Available: https://doi.org/10.1145/2884781.2884803

[59] S. A. Scherr, P. Mennig, C. Kammler, and F. Elberzhager, "On the road to enriching the app improvement process with emotions," in *2019 IEEE 27th Int. Requirements Engineering Conf. Workshops (REW)*. IEEE, 2019, pp. 84–91.

[60] J. O. Johanssen, J. P. Bernius, and B. Bruegge, "Toward usability problem identification based on user emotions derived from facial expressions," in *2019 IEEE/ACM 4th Int. Workshop on Emotion Awareness in Software Engineering (SEmotion)*. IEEE, 2019, pp. 1–7.

[61] M. Stade, S. A. Scherr, P. Mennig, F. Elberzhager, and N. Seyff, "Don't worry, be happy–exploring users' emotions during app usage for requirements engineering," in *2019 IEEE 27th Int. Requirements Engineering Conf. (RE)*. IEEE, 2019, pp. 375–380.

[62] B. Bent, B. A. Goldstein, W. A. Kibbe, and J. P. Dunn, "Investigating sources of inaccuracy in wearable optical heart rate sensors," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–9, 2020.

[63] H. Critchley and Y. Nagai, *Electrodermal Activity (EDA)*. New York, NY: Springer New York, 2013, pp. 666–669.

[64] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological measurement*, vol. 28, no. 3, 2007.

[65] J. E. Sinex, "Pulse oximetry: principles and limitations," *The American journal of emergency medicine*, vol. 17, no. 1, pp. 59–66, 1999.

[66] E. S. Dan-Glauser and K. R. Scherer, "The geneva affective picture database (gaped): a new 730-picture database focusing on valence and normative significance," *Behavior research methods*, vol. 43, no. 2, p. 468, 2011.

[67] J. J. Braithwaite, D. G. Watson, R. Jones, and M. Rowe, "A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments," University of Birmingham, UK, University of Birmingham, UK, Tech. Rep., 2015.

[68] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 797–804, April 2016.

[69] M. Kuhn, "The caret package," http://topepo.github.io/caret/index.html, 2009.

[70] C. Tantithamthavorn, S. McIntosh, A. E. Hassan, and K. Matsumoto, "Automated parameter optimization of classification techniques for defect prediction models," in *Proc. of the 38th Int. Conf. on Software Engineering, ICSE 2016, Austin, TX, USA, May 14-22, 2016*, 2016, pp. 321–332.

[71] ——, "The impact of automated parameter optimization on defect prediction models," *IEEE Transactions on Software Engineering*, vol. 45, no. 7, pp. 683–711, July 2019.

[72] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," *CoRR*, vol. abs/1811.12808, 2018. [Online]. Available: http://arxiv.org/abs/1811.12808

[73] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, pp. 1–47, 04 2001.

[74] M. Trice, L. Potts, and R. Small, "Values versus rules in social media communities: How platforms generate amorality on reddit and facebook," in *Digital Ethics*. Routledge, 2019, pp. 33–50.

[75] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos (extended abstract)," in *2015 Int. Conf. on Affective Computing and Intelligent Interaction, ACII 2015, Xi'an, China, September 21-24, 2015*, Sep. 2015, pp. 491–497.

[76] M. Chen, J. Han, L. Guo, J. Wang, and I. Patras, "Identifying valence and arousal levels via connectivity between eeg channels," in *Proc. of the 2015 Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*, ser. ACII '15. USA: IEEE Computer Society, 2015, pp. 63–69.

[77] H. F. García, M. A. Álvarez, and Á. Á. Orozco, "Gaussian process dynamical models for multimodal affect recognition," in *38th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society, EMBC 2016*, Aug 2016, pp. 850–853.

[78] A. Distanont, H. Haapasalo, M. Vaananen, and J. Lehto, "The engagement between knowledge transfer and requirements engineering," *IJKL*, vol. 1, no. 2, pp. 131–156, 2012.

[79] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion recognition using bio-sensors: First steps towards an automatic system," in *ADS'04*. Springer, 2004, pp. 36–48.

[80] J. Wagner, J. Kim, and E. Andre, "From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification," in *ICME 2005*. IEEE, 2005.

[81] R. Mandryk, K. Inkpen, and T. Calvert, "Using psychophysiological techniques to measure user experience with entertainment technologies," vol. 25, no. 2, pp. 141–158, 2006.

[82] J. Scheirer, R. Fernandez, J. Klein, and R. Picard, "Frustrating the user on purpose: a step toward building an affective computer," *Interacting with Computers*, Jan. 2002.

[83] T. W. Buchanan, K. Lutz, S. Mirzazade, K. Specht, N. J. Shah, K. Zilles, and L. Jäncke, "Recognition of emotional prosody and verbal components of spoken language: an fmri study," *Cognitive Brain Research*, vol. 9, no. 3, pp. 227–238, 2000.

[84] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, no. 1-2, pp. 227–256, 2003.