Noname manuscript No.
(will be inserted by the editor)

Using Voice and Biofeedback to Predict User Engagement during Requirements Interviews

Alessio Ferrari · Thaide Huichapa · Paola Spoletini · Nicole Novielli · Davide Fucci · Daniela Girardi

Received: date / Accepted: date

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 biometric data, in terms of physiological and voice features, to complement interviews with information about the engagement of the user on the discussed product-relevant topics. We evaluate our approach by interviewing users while gathering their physiological data (i.e., *biofeedback*) using an Empatica E4 wristband, and capturing their voice through the default audio-recorder of a common laptop. Our results

Alessio Ferrari
CNR-ISTI, Pisa, Italy
E-mail: alessio.ferrari@isti.cnr.it

Thaide Huichapa
Kennesaw State University (KSU), GA, USA
E-mail: thuichap@students.kennesaw.edu

Paola Spoletini
Kennesaw State University (KSU), GA, USA
E-mail: pspoleti@kennesaw.edu

Nicole Novielli
University of Bari, Bari, Italy
E-mail: nicole.novielli@uniba.it

Davide Fucci
Blekinge Tekniska Högskola (BTH), Karlskrona, Sweden
E-mail: davide.fucci@bth.se

Daniela Girardi
University of Bari, Bari, Italy
E-mail: daniela.girardi@uniba.it

show that we can predict users' engagement by training supervised machine learning algorithms on biometric data, and that voice features alone can be sufficiently effective. The performance of the prediction algorithms are maximised when pre-processing the training data with the synthetic minority oversampling technique (SMOTE). The results of our work suggest that biofeedback and voice analysis can be used to facilitate prioritization of requirements oriented to product improvement, and to steer the interview based on users' engagement. Furthermore, the usage of voice features can be particularly helpful for emotion-aware requirements elicitation in remote communication, either performed by human analysts or voice-based chatbots.

Keywords Software Engineering · Requirements Engineering · Biofeedback · Voice Analysis · Machine Learning · Requirements Elicitation · Interviews · Chatbots

1 Introduction

The development of novel software products, as well as the improvement of existing ones, can deeply benefit from the involvement of users in requirements engineering (RE) activities (Bano and Zowghi, 2015). Getting feedback from the user base has been recognised to lead to increased usability, improved satisfaction (Bakalova and Daneva, 2011), better understanding of requirements (Hanssen and Fægri, 2006), and creation of long-term relationships with customers (Heiskari and Lehtola, 2009).

User feedback can take implicit and explicit forms, and different means are available to collect this information. In particular, data analytics applied to user opinions and to usage data has seen an increasing interest in the last years, leading to the birth of RE sub-fields such as *crowd RE* (Murukannaiah et al., 2016; Groen et al., 2017) and *data-driven RE* (Maalej et al., 2015; Williams and Mahmoud, 2017). In the case of bespoken development (i.e., when customer- or domain-specific products' requirements need to be engineered), it is still common to resort to traditional RE practices, such as prototyping, observations, usability testing, and focus groups (Zowghi and Coulin, 2005). Among the classical techniques, user interviews are one of the most commonly used to gather requirements and feedback (Fernández et al., 2017; Davis et al., 2006; Hadar et al., 2014). Several aspects have been observed to influence the success and failure of interviews, such as the domain knowledge of the requirements analyst (Hadar et al., 2014; Aranda et al., 2015), ambiguity in communication (Ferrari et al., 2016a), and typical mistakes such as not providing a wrap-up summary at the end of the interview session, or not creating rapport with the interviewee (Bano et al., 2019).

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*

recognised the role of users' emotions and studied it extensively. Contributions include applications of sentiment analysis to app reviews (Guzman and Maalej, 2014; Kurtanović and Maalej, 2018), analysis of users' facial expressions (Scherr et al., 2019a; Mennig et al., 2019), the study of physiological reactions to ambiguity (Spoletini et al., 2016), and the augmentation of goal models with user emotions elicited through psychometric surveys (Taveter et al., 2019).

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, we record their voice through a common laptop recorder, and collect their self-assessed engagement. We compare different machine learning algorithms to predict users' engagement based on features extracted from biofeedback and voice 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 (Russell, 1991) with an F1-measure of 98% and 97%, considering solely biofeedback signals. When using voice signals alone, the performance in terms of F1-measure increases to 100% and 97%, showing the voice features alone can still be strongly predictive of users' engagement. The combination of biofeedback and voice features leads to the best performance for both valence and arousal (F1-measure of 100%).

This paper makes two main contributions:

- A methodology, based on the use of machine learning and biometric features, including physiology- and voice-related metrics, which can be applied to predict users' engagement during requirements interviews.
- A replication package[1] to enable other researchers to build on our results.

This paper builds upon a previous conference contribution by the same authors (Girardi et al., 2020a), in which only the biofeedback signals were used for prediction. The current paper repeats and expands the experiments from Girardi et al. (2020a). In particular, we introduce additional biometric features, based on voice signals, as well as additional data preparation options—namely standard scaling, oversampling and data imputation—that allow us to radically improve the previous results, even with voice features alone. This is a particularly relevant result, as using voice analysis can have multiple benefits:

1. It dramatically decreases the overall cost of the proposed methodology, which is due to the use of specific biofeedback devices (i.e., Empatica E4 wristbands);

[1] https://github.com/alessioferrari/VoiceBiofeedEmo

2. It can be effectively applied during remotely performed interviews—a common scenario nowadays, especially due to the COVID-19 pandemic—as voice is remotely transmitted as main part of the conversation, while biofeedback need to be pre-processed locally, and its transmission requires additional data transfer;
3. It scales the approach to a larger number of users, possibly interviewed in a semi-automated way by artificial agents;
4. It mitigates potential issues related to the acceptance of the non-intrusive, yet potentially undesired, biofeedback device.

The remainder of the paper is structured as follows. In Section 2, we present background definitions of engagement and emotions, as well as related work in RE and software engineering. In Section 3, we report our study design, whereas Section 4 reports its results. We discuss the implications of our study in Section 5 and its limitations in Section 6. Finally, Section 7 concludes the paper.

2 Background and Related Work

In this section, we first clarify the relationship between emotion modelling and engagement (Sect. 2.1). Then, we present the background on affect modelling and emotion classification using biofeedback (Sect. 2.2) and voice analysis (Sect. 2.3). Finally, we discuss relevant related work in the broader area of emotions in RE (Sect. 2.4) and use of biometrics in software engineering (Sect. 2.5), and discuss our contribution to the field.

2.1 Engagement and Emotions

Affective states, ranging from personality traits, which are stable features of an individua, to emotions, that are, conversely, dynamic, episodic and rapidly changing events depending on individual and contextual factors (Cowie et al., 2011). Psychologists have investigated the nature and triggers of emotions for decades. As a consequence, a plethora of theories of emotions emerged in the last decades. Cognitive models describe emotions as reactions to cognition. For example, the OCC model (Ortony et al., 1988) defines a taxonomy of emotions and identifies them as *valenced* reactions (either positive or negative) to the cognitive processes involved in evaluating objects, events, and agents. Analogously, 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: when a situation is congruent with the person's goals positive emotions arise; otherwise, negative emotions are triggered when one's goal are threatened (Lazarus, 1991).

In line with these theories and consistently with the operationalization adopted in our previous study Girardi et al. (2020a), we use emotions as a

proxy for users' *engagement* during interviews. Our choice is further supported by previous empirical findings demonstrating how emotions can be leveraged for detecting engagement in speech-based analysis of conversations (Yu et al., 2004) or to detect students' motivation (Barhenke et al., 2011). 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 (Müller and Fritz, 2015; Graziotin et al., 2015; Girardi et al., 2020b), we adopt a dimensional representation of developers' emotions. In particular, we refer to the Circumplex Model of Affect by Russel (Russell, 1991), which models emotions along two continuous dimensions, namely *valence*, that is the pleasantness of the emotion stimulus, ranging from pleasant to unpleasant, and *arousal*, that is the level of emotional activation, ranging 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, instead, the level of emotional activation. Some emotions are associated with the person being inactive, thus experiencing *low* arousal, as in *sadness* or *relaxation*. Conversely, high level of arousal are associated to high emotional activation, as in *anger* or *excitement*.

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).

2.2 Biofeedback-based Classification of Emotions

The use of physiological signals for emotion recognition has been largely investigated by affective computing research (Canento et al., 2011; Kim and André, 2008; Koelstra et al., 2012; Soleymani et al., 2016; Girardi et al., 2017). Previous work studied the relationship between emotions and biometrics such as the electrical activity of the brain—e.g., using electroencephalogram (EEG) (Kramer, 1990; Reuderink et al., 2013; Soleymani et al., 2016; Li and Lu, 2009), the electrical activity of the skin, or elecrodermal activity (EDA) (Burleson and Picard, 2004; Kapoor et al., 2007), the electrical activity of contracting muscles measured using electromyogram (EMG) (Koelstra et al., 2012; Nogueira et al., 2013; Girardi et al., 2017), and the blood volume

pulse (BVP) from which heart rate (HR) and its variability (HRV) are derived (Canento et al., 2011; Scheirer et al., 2002b). In this study, we leverage metrics based on the electrodermal activity (EDA), hearth rate

Electroencephalogram (EEG) captures the electrical activity of the brain through electrodes placed on the surface of the scalp. Changes in the EEG spectrum correlate with increased or decreased overall levels of arousal or alertness (Kramer, 1990) as well as with the valence of the emotion experienced (Reuderink et al., 2013; Soleymani et al., 2016).

The electrodermal activity (EDA) measures the electrical conductance of the skin due to the sweat glands activity. EDA correlates with the arousal dimension (Lang and Bradley, 2007) 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 (Burleson and Picard, 2004; Kapoor et al., 2007).

Heart-related metrics have been successfully employed for emotion detection (Canento et al., 2011; Scheirer et al., 2002b). In particular, blood volume pressure (BVP) measures 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 presence of increased cognitive and mental load (Kushki et al., 2011). Increases in HR occur when the body needs a higher blood supply, for example in presence of mental or physical stressors (Greene et al., 2016).

Finally, several studies demonstrated the high predictive power of facial EMG for emotion recognition (Koelstra et al., 2012; Nogueira et al., 2013). However, it leads to poor results when the sensors are placed on body parts other than the face (i.e., the arms (Girardi et al., 2017)).

In a recent study, Girardi et al. (Girardi et al., 2020b) identify a minimum set of sensors including EDA, BVP, and HR for valence and arousal classification. To collect such physiological signals, 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 3.2), thus increasing the ecological validity of our study. Furthermore, we combine biofeedback and with *voice* features, which were not considered in previous works.

2.3 Voice Analysis and Classification of Emotions

Classification of emotions based on the analysis of voice features is a well-developed research field, normally referred as *speech emotion recognition* (SER). Different surveys have been recently published on the topic (Akçay and Oğuz, 2020; Schuller, 2018; Sailunaz et al., 2018), which highlight the maturity of research, but also point out the limits in terms of real-world applications, mainly

due to limited gold standard datasets available for SER systems' training and assessment.

Speech is composed by a diverse set of acoustic features, and its information content is usually accompanied by other so-called supporting modalities, including linguistic features (i.e., the textual content equivalent to a verbal utterance), visual features, and physiological signals such as those discussed in the previous section.

Acoustic features used in SER are normally classified into *prosodic* (e.g., pitch, tone), *spectral* (i.e., frequency-based representation of the sound produced), *voice quality* (e.g., measuring the stability of the voice) and Teager energy operator *(TEO)-based features*, specifically developed to detect stress from the voice signal. Prosodic and spectral features are the most commonly used in the literature (Akçay and Oğuz, 2020; Sailunaz et al., 2018). In particular, most commonly used features are Mel-scaled spectrogram and Mel-frequency cepstral coefficients (MFCCs), which are spectral features that mimic the reception pattern of sound frequency intrinsic to a human (Issa et al., 2020). Issa et al. (2020) uses also Chromagrams—typically used for music representation—since the other features were recognised to be poor in distinguishing pitch classes and harmony (Beigi, 2011).

Research in SER initially focused on identifying relevant features and combination thereof to optimise the performance of traditional classification algorithms (Lee and Narayanan, 2005; Ververidis and Kotropoulos, 2006), leading to good recognition rates especially with Support Vector Machines (Chen et al., 2012; Schuller et al., 2004). With the advent of deep neural networks, and the possibility of overcoming the feature engineering problem altogether, the focus shifts to the selection of appropriate network architectures, and promising results are achieved through Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) models (Zhao et al., 2019; Trigeorgis et al., 2016). To address the problem of data scarcity, the recent development of transfer learning methods have been also experimented by Ottl et al. (2020). Another avenue of research in SER, still under development, is the combination of pure audio features with other contextual cues, including videos (Chao et al., 2015; Ren et al., 2019; Ottl et al., 2020).

Applications of the SER techniques are mostly in the field of human-computer interaction (HCI). Ramakrishnan and El Emary (2013) lists a set of ten different possible applications, including lie detection, treatment of language disorders, driving support systems (Schuller et al., 2004), surveillance (Nanda et al., 2017), and smart assistants. These latter have become commercially available in recent years—e.g., Siri, Cortana—and represent one of the natural area of exploitation of SER research. Another traditional application, closely related to our context, is the recognition of customer's emotions during conversations with call center operators (Han et al., 2020; Li et al., 2019; Morrison et al., 2007; Batliner et al., 2003; Petrushin, 1999). These works are oriented to identify critical phases of the dialogue between a customer and an artificial operator. This can be generally useful to understand when the

artificial agent is irritating the customer, thus to deciding to transfer the call to a human operator.

With respect to works in SER, we are among the first ones that bring these techniques to the software engineering domain, and to the RE area in particular, therefore identifying a novel application field.

2.4 Sentiment and Emotions in Requirements Engineering

Researchers acknowledge the role and importance of users' emotions in RE activities (Sutcliffe, 2011). As far as written communication is concerned, data such as stakeholders' conversational traces and users' feedback (e.g., tweets and app reviews) can be 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 (NLP) to textual artefacts to mine the stakeholders' emotions and opinions about a given product or feature. For example, Guzman et al. (2017) use sentiment analysis to analyze 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 and satisfaction about existing features. Martens and Maalej (2019) apply sentiment analysis to 7M app reviews over 245 free and paid apps. They showed that specific categories are characterized by either positive or negative sentiment. Furthermore, they observed a correlation between the users' sentiment polarity and the rating (e.g., 1–5 stars). Researchers also leveraged emotion mining from reviews in app stores reviews to evaluate single app features (Guzman and Maalej, 2014; Johann et al., 2017; Shah et al., 2019).

In particular, several studies propose supervised machine learning approaches that leverage sentiment information extracted from a textual source to support RE tasks. For example, Maalej and Nabil (2015) propose a method that uses sentiment scores to classify app reviews into bug reports or feature requests, in order to support the stakeholder involved in a software artifact maintenance and evolution to fruitfully deal with large amount of feedback. Kurtanović and Maalej (2017) 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 (Werner et al., 2019).

Finally, emotions have been also considered in early-stage RE activities, such as requirements elicitation and modelling. For example, Colomo-Palacios et al. (2011) asked users to rank requirements according to Russel's Valence-Arousal theory, which is the one that we adopt in the present study (see Section 2.1) They use this information to enhance effective resolution of conflicting requirements. Other researchers leverage information regarding users' emotions gathered through psychometrics (e.g., surveys) to augment tradi-

tional requirements goal modelling approaches (Taveter et al., 2019; Miller et al., 2015) and artefacts, such as user stories (Kamthan and Shahmir, 2017).

With respect to previous work on emotions and RE, we focus on the users' feedback phase. Differently from existing works that leverage app reviews and NLP techniques, thereby considering purely textual or structured feedback (e.g., star ratings), we investigate emotions in oral interviews.

2.5 Biofeedback and Voice Analysis in Software Engineering and RE

Biometric sensors have been leveraged in several software engineering studies for recognizing cognitive and affective states of software developers.

In one of the early studies in this field, Parnin (2011) envisions an approach to infer developers' cognitive states based on the analysis of sub-vocal signals, that is the electrical signal the brain transfers to the mouth and vocal cords while performing complex cognitive activities. While presenting preliminary findings, this study demonstrate that it is possible to use EMG to capture the subvocalization associated to programming and leverage this information to distinguish between easy and hard development tasks. Fucci et al. (2019) use multiple biometrics, including EEG, EDA, HR, and BVP, to distinguish between code and prose comprehension tasks during software development. Fritz et al. (2014) employ EEG, BVP, and eye tracker to measure the difficulty of programming tasks, thus preventing developers from introducing bugs. The authors use the same set of sensors in a follow-up work aimed at classifying the emotional valence of developers during programming tasks (Müller and Fritz, 2015). Girardi et al. (2020b) replicate previous findings by Müller and Fritz (2015) regarding the use of non-invasive sensors for valence classification during software development tasks. Furthermore, Girardi et al. also address the classification of arousal. Other studies also propose approaches for predicting developers' interruptibility (Züger et al., 2018) and for identifying code quality concerns (Müller and Fritz, 2016) by leveraging EDA, HR, and HRV.

Biofeedback has been used also in RE, mainly to capture users' emotions *while* using an app. For example, Scherr et al. (2019b) and Mennig et al. (2019) uses the mobile phone cameras to recognize facial muscle movements and associate them to the users' emotions when using different features of an app. This methodology was recently applied to enable user validation of new requirements (Scherr et al., 2019a) and to for identifying usability issues (Johanssen et al., 2019a) with minimal privacy concerns (Stade et al., 2019). Part of the authors of the current paper previously proposed using biometrics in requirements elicitation interviews (Spoletini et al., 2016). Our previous work focused on ambiguity, and remained at the research preview stage, as it evolved in the current work after pilot experiments.

With respect to using biofeedback and voice analysis 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 studies (e.g., Scherr et al. (2019b); Mennig et al. (2019)) focus on detecting the user's

engagement experienced *while* using the software features. In our case, we aim to detect users' engagement about certain features when users reflect on the features and speak about them. This captures 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 1), which is not possible when performing observations without interacting with users. Finally, to our knowledge, our work is among the first ones that use voice features to predict the emotion of a speaker involved in a software engineering activity.

3 Research Design

Our study is *exploratory* in nature, aiming to investigate a certain area of interest—i.e., *engagement* in interviews—and identify possible avenues of research. We adopt a quantitative experimental approach involving human subjects, and oriented to compare software-based artifacts (i.e., machine learning algorithms and feature configurations). The study was approved by the Kennesaw State University review board (study 16-068).

In the following, we illustrate research questions (Sect. 3), study participants (Sect. 3.1), adopted artifacts (Sect. 3.2, 3.3, and 3.4), and experimental protocol (Sect. 3.5, 3.6, and 3.7).

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

- **RQ1:** *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 with biofeedback. More specifically, we want to assess to what extent we can recognize emotional valence and arousal—i.e., the two dimensions we use for the operationalization of engagement. To collect trainig and testing data, we first 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. During the interviews with users, we acquire their raw biofeedback signals. We use features extracted from the signals, and consider intervals of reported engagement as classes to be predicted. Based on these data, we evaluate and compare different supervised machine learning classifiers.
- **RQ2:** *To what extent can we predict users' engagement using voice analysis and supervised classifiers?* This question aims to understand whether we can recognize engagement with automatic voice analysis. To this extent, we record the audio of the interviews, and we extract voice features from the audio signals. We then use the voice features to train and compare the previously used supervised classifiers.

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

– **RQ3:** *To what extent can we predict users' engagement by combining voice and biofeedback features?* This questions aims to use voice and biofeedback features in conjunction. By training and comparing the classifiers, we check if and in which way the combination of features allows improving the performance of the approach.

3.1 Study Participants

We recruited 31 participants among the students of Kennesaw State University with an opportunistic sampling. The 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 (Bent et al., 2020), 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. We collected information about the ethnicity of participants because the reserach demonstrated that heart-rate optical sensors might give more/less reliable readings based on the skin tones. Having a diverse pool of participants in terms of ethnicity strengthens the validity of our empirical findings. Participants received a monetary incentive of $25 for up of one hour of their time.

3.2 Biofeedback Device and Signals

The device we use to acquire the biofeedback is the Empatica E4[3] wristband. We selected it as it is used in several studies in affective computing (Greene et al., 2016) as well as in the field of software engineering (Müller and Fritz, 2015; Fucci et al., 2019)). 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

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

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 (Critchley and Nagai, 2013).
- **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 (Allen, 2007; Sinex, 1999).
- **Heart Rate**: HR is measured by Empatica E4 based on elaboration of the BVP signal with a proprietary algorithm.

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 (Girardi et al., 2020b).

3.3 Audio Device and Signals

The interviews' audios were captured using the default audio recorder of a Mac OS laptop, and the files were stored in the classical Waveform Audio File Format (`.wav`), which is an uncompressed representation of the raw signal. Audio is a complex, information-rich signal, and a largely variable set of classical features are used to characterise its salient aspects (Schuller and Batliner, 2013). Among these features, we consider the following ones:

- **Mel Spectrogram**: it represents the acoustic time-frequency representation of sound.
- **Mel-Frequency Cepstral Coefficients (MFCC)**: MFCC are the representation of the short-term power spectrum of sound. More in details, Cepstral features contain information about the rate changes in the different spectrum bands and they have the ability to separate the impact of the vocal cords and the vocal tract in a signal. In the MFCC, these features are extracted at the frequency more audible by human ears.
- **Chromagram**: the Chromagram is a 12-element feature vector indicating how much energy of each pitch class is presented in the signal. This is typically used to model harmonic and melodic characteristics of music, and it is recognised as useful also to model the emotional aspect of voice (Schuller and Batliner, 2013).

We choose these features as they are amongst the most common in speech emotion recognition (Issa et al., 2020; Schuller and Batliner, 2013; Issa et al., 2020).

3.4 Supervised Learning Algorithms

We address the problem of predicting user engagement (RQ2, RQ3, RQ4) using machine learning. In line with previous research on biometrics (Fucci et al.,

2019; Müller and Fritz, 2015; Koelstra et al., 2012; Girardi et al., 2020b), we chose popular algorithms—i.e., Naive Bayes (NB), C4.5-like decision trees (DTree), Support Vector Machines (SVM), Multi-layer Perceptron for neural network (MLP), and Random Forest (RF).

3.5 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, checking that audio recording is operational, 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.

Device calibration and emotion triggering In line with previous research (Müller and Fritz, 2015; Girardi et al., 2020b) 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 biofeedback signal by letting the wristband record the raw signals for all sensors under the experimenter scrutiny. Second, the collected data will be needed to adjust the scores obtained during the self-assessment questionnaire (see Sect. 3.6). 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 (Dan-Glauser and Scherer, 2011) previously used in software engineering studies by Müller and Fritz (2015). 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 (Müller and Fritz, 2015), 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.

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

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 1. For each topic, we include multiple questions, to allow users sufficient time to get immersed in the topic, and have more stable biometric parameters in relation to the topic. Questions related to topics we expect to raise more engagement, (i.e., privacy, relationship, money, and 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, the audio recorder acquires the voice of the speakers, 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.

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 3.6).

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

3.6 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 (Müller and Fritz, 2015; Girardi et al., 2020b). First, we adjust the valence and arousal scores based on the mean values reported while watching the emotion-triggering pictures (see Section 3.5). This step is necessary to take into account fluctuations due to 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 biofeedback and voice signals with the self-assessment, we (1) save the timestamp corresponding to the interviewer

[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. 4.

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

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 biofeedback signal samples (for EDA, BVP and HR) within the time interval associated to Q_i;
- a voice signal sample in the form of a `.wav` file—the segment of the `.wav` file of the whole interview for the time interval associated to Q_i;
- two labels, one representing arousal ($q_A(Q_i)$) and the other representing 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 (Müller and Fritz, 2015). 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 (Canento et al., 2011). 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) (Braithwaite et al., 2015). We extract the two components using the cvxEDA algorithm (Greco et al., 2016). For the audio signal, we use the Python package Librosa (McFee et al., 2015)[6] for audio and music analysis to process the files, and we extract the different features (Mel Spectrogram, MFCC, Chromagram).

Table 2: Machine learning features grouped by physiological and voice 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)
Voice	- mean Mel Spectrogram
	- mean MFCC (mean of the first 20 MFCC features)
	- mean Chromagram (mean of the 12 Chroma features)

After signals pre-processing, we extracted the features presented in Table 2, which we use to train our classifiers. We select biofeedback features based on previous studies using the same signals (Müller and Fritz, 2015; Fucci et al., 2019; Girardi et al., 2020b) and we choose audio features according to recommendations from the specialised literature (Schuller and Batliner, 2013; Issa et al., 2020).

[6] https://librosa.org

3.7 Analysis Procedure

The analysis procedure aims at answering the three RQs, as detailed in the following. We first collect descriptive data and provide qualitative considerations. To this end, we 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. Then, for each user, we use the biometrics gathered in the user's interview phase as input features for the different classifiers listed in Sect. 3.4. To answer the different questions, we first consider solely biofeedback features (RQ1), then voice features (RQ2) and finally their combination (RQ3).

In line with previous research (Müller and Fritz, 2015; Girardi et al., 2020b), 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[7].

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, which allows having a balanced set of instances from the different classes in both sets. For each algorithm, we search for the optimal hyper-parameters (Tantithamthavorn et al., 2016, 2019) using leave-one-out cross validation—i.e., the recommended approach for small training sets (Raschka, 2018) such as ours. The resulting model is then evaluated on the test set to assess its performance on unseen data and avoid overfitting. We repeat this process 10 times with different splits of the train and test sets to further increase the validity of the results. The performance is then evaluated by computing the mean for precision, recall, F1-measure, and accuracy over the different runs. This setting is directly comparable to the one implemented by Müller and Fritz (Müller and Fritz, 2015) and by Girardi et al. (Girardi et al., 2020b), which includes data from the same subject in both training and test sets.

The process outlined above is repeated with a maximum of 8 different settings, based on the three following data preparation options oriented to improve the performance of the machine learning algorithms without losing validity of the results:

- **Standard Scaling:** the features in the training set are standardised so that their distribution will have a mean value 0 and standard deviation of 1. The standardisation parameters from the training set are then applied to scale the test set. This way, information from the test set (i.e., its standard deviation) is not passed to the training set, which could bias the learning process. Standard scaling is essential for machine learning algorithms that

[7] 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 3: Label distribution in the gold standard for biofeedback feature vectors and for experiments using imputation.

calculate distances between data, in our case SVM and MLP. If not scaled, the feature with a higher value range starts dominating when calculating distances. Scaling should not affect rule-based algorithms that consider each feature separately, and are not affected by monotonic transformations of the variables, such as standard scaling. Standardisaiton is performed by means of the `StandardScaler` of `Scikitlearn`.

- **Balancing:** Synthetic Minority Oversampling Technique (SMOTE) is a traditional data augmentation technique applied to train machine learning models in case of class imbalance (Chawla et al., 2002). Indeed, in case of class imbalance, machine learning algorithms tend to perform poorly on the minority class, as they do not have a sufficient amount of example items to learn from and build a fair classification model. To overcome this issue, SMOTE creates synthetic examples of the minority class, in our case based on the k-nearest neighbour algorithm. To prevent data leakage, SMOTE is applied solely to the training set, therefore the test set does not contain synthetic data. SMOTE is performed through the `SMOTE` class from the `imblearn` Python package.
- **Imputation:** data imputation is normally adopted when some features have missing data. In our case, we miss voice feature data for 66 arousal vectors and 60 valence vectors. We can however infer (*impute*) the data by using the corresponding biofeedback features. Imputation is performed by means of k-nearest neighbors imputation, using the `KNNImputer` from `Scikitlearn`.

4 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 scat-

Arousal			*Valence*		
High	**Low**	Neut.	**Positive**	**Negative**	Neut.
159 (43%)	211 (57%)	-	300 (80%)	74 (20%)	-

Table 4: Label distribution in the gold standard for voice feature vectors and combined feature vectors without imputation.

ter 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. 3.7, 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*.

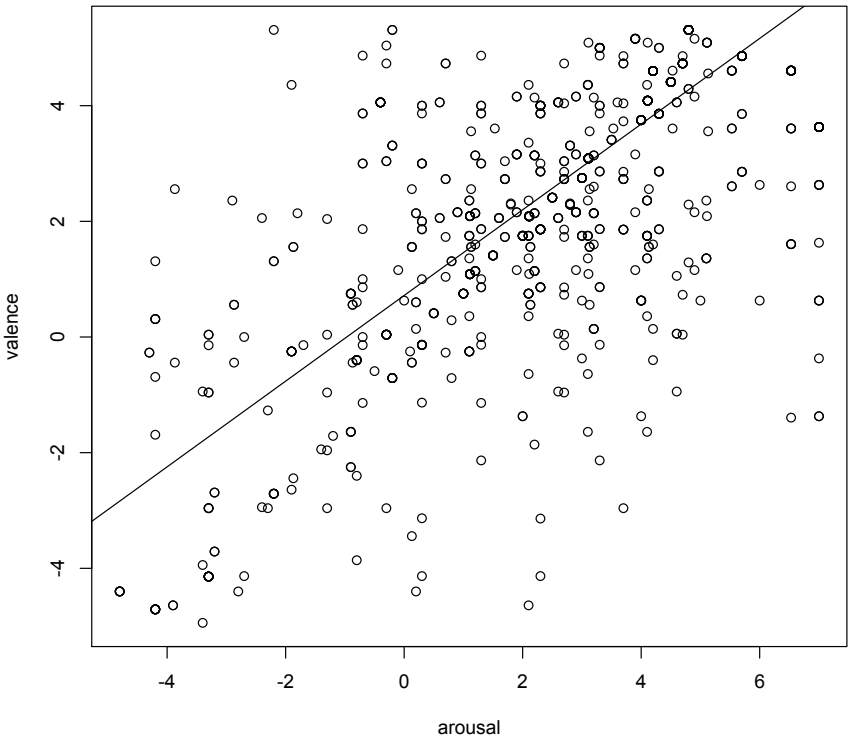

Fig. 1: Scatter plots of normalised valence and arousal according to the self-assessment questionnaire.

As our goal for the RQs 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. Table 3 reports the gold standard dataset with valence and arousal distribution, when considering biofeedback features (for RQ1). Voice feature vectors corresponding to each biofeedback vector could not be identified for part of th gold standard items, as the audio recording was not reliable for some subjects. Therefore, the gold standard dataset for audio only (RQ2) and for combined features (RQ3) without imputation is a subset of the original gold standard, and is reported in Table 4.

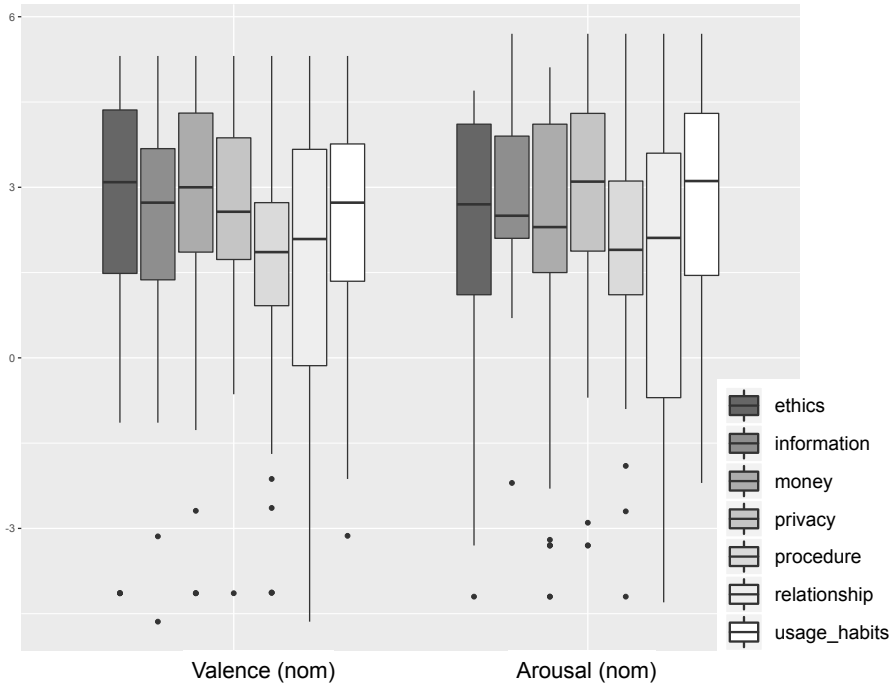

Fig. 2: Box plots of valence and arousal for each group of questions, according to the self-assessment questionnaire.

4.1 Descriptive Statistics

In the following, we report some descriptive statistics on the data.

Table 5 reports the ranges of valence and arousal, according to the self-assessment questionnaire. We report both original values and normalised ones

	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 5: Descriptive statistics of the reported engagement.

("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 5) 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[8]. 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 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*.

4.2 RQ1: To what extent can we predict users' reported engagement using biofeedback measurements and supervised classifiers?

In Figure 3 we report the performance of the different classifiers in terms of their precision, recall, F1-measure and accuracy, considering the different configurations. 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 (Sebastiani, 2001). 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 training instances on specific classes.

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

Green cells indicate cases with good performance, while red cells indicate poor performance, while different shades indicate intermediate values.

We see that the best performance (in **bold**) for valence are achieved by the Multi-layer Perceptron (MLP), while for arousal the best algorithm is Random Forest (RF), when applying both balancing and standard scaling. The worst performing algorithm is Naive Bayes (NB), regardless of the configuration. In general, we see that balancing the dataset (**Bal.** set to Y in the table) leads to the most relevant improvement for all the algorithms—the top-2 lines are green—except NB, allowing to pass from F1-measures in the range of $0.5-0.6$ to values in the range of $0.8-1.0$. This indicates that compensating for class imbalance in the training set substantially boosts the performance of the algorithms in this setting, characterised by a limited number of data points, especially for negative valence (see Table 3). Given that high performance are observed by most algorithms, it suggests that the biofeedback features considered are effective in discriminating between positive (high) and negative (low) valence (arousal).

As expected, standard scaling (**Scale** set to Y) has a particularly positive influence on SVM (F1 passes from 0.648 to 0.923 for valence, and from 0.684 to 0.890 for arousal), since this algorithm is strongly influenced by the scale of the features. Scaling does not have a strong influence on rule-based algorithms that consider feature values individually, such as Decision Trees (DT) and Ranform Forest (RF). Actually, for DT, we even see a decrease in performance for arousal (F1 decreases from 0.934 to 0.823) when applying standard scaling, and similarly for NB.

	Precision	**Recall**	**F1**	**Accuracy**
Valence				
Multi-layer Perceptron	0.98	0.99	0.98	0.99
Baseline	0.40	0.50	0.45	0.79
Improvement	0.58 (145%)	0.50 (98%)	0.53 (117%)	0.20 (25%)
Arousal				
Random Forest	0.97	0.97	0.97	0.97
Baseline	0.28	0.50	0.36	0.56
Improvement	0.69 (246%)	0.47 (94%)	0.61 (170%)	0.41 (73%)

Table 6: Performance of the best classifiers based on F1, using EDA, BVP, and HR features with respect to majority class baseline classifier. Improvement over the baseline is also shown.

In Table 6, we report the result of the two best algorithms with the best configurations, and we compare them with a baseline. Following previous research on sensor-based emotion recognition in software development (Girardi et al., 2020b), 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.

	Options		Valence				Arousal			
Algorithm	Bal.	Scale	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc
SVM	Y	Y	0.902	0.954	0.923	0.946	0.889	0.895	0.890	0.891
		N	0.652	0.723	0.648	0.700	0.699	0.697	0.684	0.684
	N	Y	0.581	0.579	0.576	0.743	0.596	0.595	0.592	0.600
		N	0.537	0.503	0.462	0.783	0.549	0.548	0.547	0.556
MLP	Y	Y	**0.977**	**0.990**	**0.983**	**0.989**	0.965	0.963	0.963	0.964
		N	0.868	0.906	0.880	0.915	0.871	0.870	0.866	0.868
	N	Y	0.590	0.575	0.578	0.746	0.577	0.576	0.576	0.584
		N	0.614	0.581	0.582	0.773	0.552	0.550	0.548	0.565
DTree	Y	Y	0.954	0.984	0.966	0.976	0.824	0.825	0.823	0.824
		N	0.894	0.948	0.915	0.939	0.938	0.939	0.934	0.934
	N	Y	0.576	0.569	0.569	0.731	0.622	0.622	0.617	0.622
		N	0.591	0.563	0.563	0.726	0.635	0.634	0.631	0.636
NB	Y	Y	0.520	0.529	0.428	0.444	0.541	0.513	0.372	0.456
		N	0.520	0.529	0.424	0.439	0.539	0.537	0.536	0.552
	N	Y	0.550	0.525	0.497	0.773	0.493	0.499	0.384	0.466
		N	0.568	0.511	0.475	0.787	0.514	0.510	0.468	0.544
RF	Y	Y	0.959	0.950	0.954	0.970	**0.969**	**0.969**	**0.969**	**0.969**
		N	0.959	0.962	0.960	0.974	0.954	0.957	0.955	0.956
	N	Y	0.757	0.538	0.519	0.804	0.640	0.634	0.635	0.647
		N	0.650	0.534	0.512	0.800	0.651	0.646	0.646	0.656

Fig. 3: Performance of the different algorithms with the different configurations in terms of data balancing with SMOTE (Bal.) and in terms of standard scaling (Scale) when using **biofeedback** features. Y = Yes, the configuration option is applied; N = No, the configuration option is not applied.

For valence, the MLP classifier distinguishes between negative and positive emotions with an F1 of 0.98, thus obtaining an increment of 117% with respect to the baseline. Furthermore, we observe an improvement in precision of 145% (from 0.40 of the baseline to 0.98 of MLP) and 98% in recall (from 0.50 to 0.99). These results indicate that the classifiers' behavior is substantially better than the baseline classifier that always predicts the positive class.

As for arousal, we observe a comparable performance. The RF classifier distinguishes between high and low activation with an F1 of 0.97, representing an improvement of 170% over the baseline (0.36). Again, the classifier substantially outperforms the baseline with an improvement of 246% for precision (from 0.28 to 0.97) and 94% for recall (from 0.50 to 0.97). The improvement with respect to the baseline is particularly high for arousal in terms of precision since arousal data (cf. Table 3) are more balanced. Therefore, the baseline that always predicts the positive class is inherently less effective, with a precision of 0.28.

	Options			Valence				Arousal			
Algorithm	Bal.	Scale	Imp.	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc
			Y	0.942	0.978	0.958	0.971	0.979	0.975	0.977	0.977
		Y	N	**1.000**	**1.000**	**1.000**	**1.000**	0.941	0.942	0.941	0.942
			Y	0.714	0.790	0.730	0.788	0.756	0.758	0.756	0.759
	Y	N	N	0.712	0.805	0.726	0.782	0.761	0.764	0.761	0.763
			Y	0.630	0.603	0.604	0.782	0.690	0.681	0.681	0.693
		Y	N	0.703	0.680	0.687	0.810	0.697	0.690	0.692	0.701
			Y	0.645	0.550	0.544	0.782	0.691	0.683	0.684	0.695
SVM	N	N	N	0.712	0.605	0.622	0.813	0.687	0.680	0.680	0.691
			Y	0.964	0.976	0.969	0.979	**0.980**	**0.979**	**0.979**	**0.979**
		Y	N	**1.000**	**1.000**	**1.000**	**1.000**	0.947	0.950	0.947	0.948
			Y	0.849	0.893	0.858	0.894	0.854	0.855	0.847	0.848
	Y	N	N	0.897	0.946	0.917	0.943	0.834	0.818	0.814	0.820
			Y	0.653	0.642	0.643	0.767	0.688	0.682	0.681	0.691
		Y	N	0.688	0.680	0.680	0.802	0.689	0.687	0.687	0.693
			Y	0.585	0.544	0.532	0.776	0.652	0.645	0.645	0.656
MLP	N	N	N	0.582	0.584	0.578	0.771	0.646	0.640	0.634	0.643
			Y	0.909	0.956	0.929	0.950	0.948	0.949	0.948	0.949
		Y	N	0.992	0.994	0.993	0.996	0.870	0.869	0.865	0.867
			Y	0.871	0.928	0.893	0.923	0.933	0.937	0.934	0.935
	Y	N	N	0.939	0.980	0.956	0.970	0.926	0.930	0.927	0.927
			Y	0.636	0.596	0.597	0.756	0.642	0.631	0.631	0.646
		Y	N	0.666	0.632	0.638	0.798	0.630	0.627	0.625	0.634
			Y	0.614	0.575	0.580	0.756	0.630	0.626	0.625	0.634
DTree	N	N	N	0.613	0.603	0.602	0.773	0.633	0.626	0.626	0.638
			Y	0.589	0.627	0.584	0.658	0.611	0.612	0.607	0.608
		Y	N	0.596	0.635	0.598	0.688	0.612	0.613	0.606	0.608
			Y	0.578	0.608	0.573	0.653	0.626	0.628	0.625	0.628
	Y	N	N	0.580	0.614	0.577	0.669	0.667	0.669	0.663	0.664
			Y	0.582	0.601	0.584	0.689	0.619	0.621	0.618	0.621
		Y	N	0.611	0.652	0.612	0.698	0.630	0.631	0.625	0.626
			Y	0.561	0.569	0.558	0.695	0.621	0.620	0.619	0.624
NB	N	N	N	0.595	0.604	0.588	0.709	0.624	0.624	0.620	0.625
			Y	0.953	0.946	0.950	0.967	0.945	0.944	0.944	0.945
		Y	N	0.964	0.957	0.960	0.975	**0.967**	**0.966**	**0.966**	**0.967**
			Y	0.948	0.943	0.945	0.964	0.945	0.944	0.944	0.945
	Y	N	N	0.966	0.953	0.959	0.974	0.962	0.961	0.961	0.962
			Y	0.724	0.556	0.553	0.802	0.670	0.656	0.656	0.673
		Y	N	0.731	0.559	0.556	0.820	0.676	0.664	0.664	0.678
			Y	0.708	0.533	0.514	0.796	0.675	0.665	0.665	0.678
RF	N	N	N	0.712	0.546	0.540	0.808	0.666	0.657	0.658	0.671

Fig. 4: Performance of the different algorithms with the different configurations in terms of data balancing with SMOTE (Bal.), standard scaling (Scale) and Imputation (Imp.) when using **voice** features. Y = Yes, the configuration option is applied; N = No, the configuration option is not applied.

4.3 RQ2: To what extent can we predict users' engagement using voice analysis and supervised classifiers?

Figure 4 reports the comparison of the performance between the different machine learning algorithms, evaluated in terms of precision, recall, F1-measure and accuracy (macro-average), for the different configurations.

Differently from the biofeedback case, here we apply also data imputation (**Imp.** column) as configuration option, by synthesising data for vectors with missing voice data, based on the corresponding biofeedback vectors. Therefore, the gold standard considered when Imp. is set to Y (Yes) is analogous to the one used for biofeedback and reported in Table 3. When Imp. is set to N (No), the gold standard is the one in Table 4.

This is an important aspect to notice for at least two reasons: (1) the two gold standards have slightly different distributions, and thus the default baselines for comparison will differ. In particular, when using imputation the baselines are the same as the one used in Table 6. When imputation is not used, the baselines need to be recomputed, and in particular the majority class baseline for arousal will always predict the negative class, as this is the most frequent in Table 4; (2) the usage of imputation in a real-world context assumes that, although solely voice data are used for classification, biofeedback data are collected anyway, so the practical advantage, both economic and logistic, is limited.

As it happened for the biofeedback case, the actual boost in performance for all the algorithms, except Naive Bayes (NB), is provided by the application of SMOTE (**Bal.** set to Y)—green lines are the top 4 of each algorithm.

The best performance (in **bold**) for valence and arousal is achieved by the Multi-layer Perceptron (MLP) and by Support Vector Machines (SVM), while NB again remains far behind the other algorithms, regardless of the applied configuration options.

For valence, both MLP and SVM lead to perfect classification (all measures equal 1.000) when balancing and scaling are applied, and imputation is *not* applied. Therefore, voice features alone appear to be particularly effective in discriminating the quality of the engagement (positive or negative valence), thereby confirming that our set-up is effective in capturing the so called *emotional prosody* (Buchanan et al., 2000), which reveals the sentiment of the speaker.

Lower performance is achieved with imputation (F1=0.969 for MLP and 0.958 for SVM). These values are lower also with respect to the best performance obtained using biofeedback features (F1=0.983 for MLP, cf. Fig. 3). The decrease in performance when using imputation is common to all the algorithms, when considering valence—F1 for valence is always lower in the top line of each algorithm in Figure 4.

For arousal, the best performance (**bold**) is obtained by MLP, when using *all* the configuration options, including imputation. The performance is comparable to the best one obtained with biofeedback features (F1=0.979 vs F1=0.969 for Random Forest (RF), cf. Fig. 3). When imputation is not used,

the best performance is obtained by RF, and remains comparable with the best case for biofeedback (F1=0.966 vs F1=0.969). Differently from the valence case, here imputation appears to improve the performance for all the algorithms—cf., e.g., top lines of each algorithm for arousal in Fig. 4—except RF and NB.

Overall, our findings suggest that voice features represents a valid alternative to biofeedback for the emotion recognition during requirements elicitation. In fact, uur classifiers' performance demonstrate that in absence of biofeedback information, both valence and arousal can still be successfully predicted with voice-only features.

	Precision	**Recall**	**F1**	**Accuracy**
Valence				
Multi-layer Perceptron & SVM	1.00	1.00	1.00	1.00
Baseline	0.40	0.50	0.45	0.80
Improvement	0.60 (150%)	0.50 (100%)	0.55 (122%)	0.20 (25%)
Arousal				
Random Forest	0.97	0.97	0.97	0.97
Baseline	0.29	0.50	0.37	0.57
Improvement	0.68 (235%)	0.47 (94%)	0.60 (162%)	0.40 (70%)

Table 7: Performance of the best classifiers, according to F1, using voice features, and without imputation, with respect to majority class baseline classifier. Improvement over the baseline is also shown.

To have additional insights, Table 7 compares the result of the best algorithms, with respect to the majority class baselines. As it is more interesting, given the practical advantages given by the usage of voice features only, we consider the case in which imputation is not applied. For valence, the perfect classification achieved leads to an increase of 150% in terms of precision, 100% for recall and 122% for F1. For arousal, the increase in performance is again higher, with 235% for precision, 94% for recall and 162% for F1. These numbers are basically equivalent to those obtained with biofeedback, confirming that using voice-only features is sufficient to predict engagement in interviews.

4.4 RQ3: To what extent can we predict users' engagement by combining voice and biofeedback features?

Figure 5 reports the comparison of the performance of the different algorithms with the different configurations when using voice and biofeedback features combined.

General trends are analogous to those observed when features are treated in separation, with increased performance for all the algorithms except Naive Bayes thanks to the usage of SMOTE, and to the application of standard scaling for Support Vector Machines (SVM) and Multi-layer Perceptron (MLP).

	Options			Valence				Arousal			
Algorithm	Bal.	Scale	Imp.	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc
			Y	0.991	0.998	0.994	0.996	0.951	0.950	0.950	0.951
		Y	N	0.982	0.985	0.983	0.989	**1.000**	**1.000**	**1.000**	**1.000**
			Y	0.756	0.836	0.778	0.828	0.812	0.811	0.811	0.814
	Y	N	N	0.756	0.836	0.778	0.828	0.812	0.811	0.811	0.814
			Y	0.653	0.650	0.651	0.775	0.711	0.701	0.700	0.712
		Y	N	0.653	0.650	0.651	0.775	0.711	0.701	0.700	0.712
			Y	0.661	0.588	0.599	0.789	0.687	0.677	0.678	0.690
SVM	N	N	N	0.661	0.588	0.599	0.789	0.687	0.677	0.678	0.690
			Y	**1.000**	**1.000**	**1.000**	**1.000**	0.996	0.996	0.996	0.996
		Y	N	**1.000**	**1.000**	**1.000**	**1.000**	0.998	0.998	0.998	0.998
			Y	0.879	0.910	0.882	0.911	0.881	0.872	0.873	0.878
	Y	N	N	0.879	0.910	0.882	0.911	0.881	0.872	0.873	0.878
			Y	0.671	0.638	0.650	0.789	0.678	0.676	0.676	0.682
		Y	N	0.671	0.638	0.650	0.789	0.678	0.676	0.676	0.682
			Y	0.685	0.609	0.615	0.791	0.646	0.643	0.642	0.650
MLP	N	N	N	0.685	0.609	0.615	0.791	0.646	0.643	0.642	0.650
			Y	0.996	0.997	0.997	0.998	0.966	0.967	0.966	0.966
		Y	N	0.951	0.983	0.962	0.973	0.969	0.973	0.970	0.970
			Y	0.995	0.999	0.997	0.998	0.988	0.989	0.988	0.989
	Y	N	N	0.995	0.999	0.997	0.998	0.988	0.989	0.988	0.989
			Y	0.606	0.586	0.589	0.750	0.657	0.657	0.655	0.659
		Y	N	0.606	0.586	0.589	0.750	0.657	0.657	0.655	0.659
			Y	0.615	0.585	0.591	0.762	0.633	0.631	0.631	0.638
DTree	N	N	N	0.615	0.585	0.591	0.762	0.633	0.631	0.631	0.638
			Y	0.643	0.687	0.650	0.730	0.626	0.625	0.613	0.614
		Y	N	0.673	0.707	0.684	0.781	0.620	0.621	0.614	0.615
			Y	0.629	0.667	0.636	0.721	0.622	0.623	0.620	0.622
	Y	N	N	0.629	0.667	0.636	0.721	0.622	0.623	0.620	0.622
			Y	0.621	0.617	0.611	0.734	0.614	0.613	0.604	0.607
		Y	N	0.621	0.617	0.611	0.734	0.614	0.613	0.604	0.607
			Y	0.594	0.604	0.595	0.717	0.608	0.609	0.606	0.610
NB	N	N	N	0.594	0.604	0.595	0.717	0.608	0.609	0.606	0.610
			Y	0.987	0.967	0.976	0.985	0.970	0.972	0.971	0.971
		Y	N	0.984	0.979	0.981	0.988	0.964	0.968	0.965	0.966
			Y	0.977	0.972	0.974	0.983	0.970	0.973	0.971	0.972
	Y	N	N	0.977	0.972	0.974	0.983	0.970	0.973	0.971	0.972
			Y	0.800	0.540	0.521	0.807	0.688	0.676	0.677	0.691
		Y	N	0.800	0.540	0.521	0.807	0.688	0.676	0.677	0.691
			Y	0.678	0.532	0.509	0.799	0.682	0.675	0.675	0.685
RF	N	N	N	0.678	0.532	0.509	0.799	0.682	0.675	0.675	0.685

Fig. 5: Performance of the different classifiers with the different configurations when using voice features.

The best performance for valence (in **bold**) are achieved by MLP, when applying SMOTE and scaling, and regardless of the application of imputation.

Alg.	Feature	*Valence*				*Arousal*			
		Prec	Rec	F1	Acc	Prec	Rec	F1	Acc
SVM	biofeedback	0.902	0.954	0.923	0.946	0.889	0.895	0.890	0.891
	voice	**1.000**	**1.000**	**1.000**	**1.000**	0.979	0.975	0.977	0.977
	combined	0.991	0.998	0.994	0.996	**1.000**	**1.000**	**1.000**	**1.000**
MLP	biofeedback	0.977	0.990	0.983	0.989	0.965	0.963	0.963	0.964
	voice	**1.000**	**1.000**	**1.000**	**1.000**	0.980	0.979	0.979	0.979
	combined	**1.000**	**1.000**	**1.000**	**1.000**	**0.998**	**0.998**	**0.998**	**0.998**
DTree	biofeedback	0.954	0.984	0.966	0.976	0.938	0.939	0.934	0.934
	voice	0.992	0.994	0.993	0.996	0.948	0.949	0.948	0.949
	combined	**0.996**	**0.997**	**0.997**	**0.998**	**0.988**	**0.989**	**0.988**	**0.989**
NB	biofeedback	0.520	0.529	0.428	0.444	0.539	0.537	0.536	0.552
	voice	0.596	0.635	0.598	0.688	**0.667**	**0.669**	**0.663**	**0.664**
	combined	**0.673**	**0.707**	**0.684**	**0.781**	0.622	0.623	0.620	0.622
RF	biofeedback	0.959	0.962	0.960	0.974	0.969	0.969	0.969	0.969
	voice	0.964	0.957	0.960	0.975	0.967	0.966	0.966	0.967
	combined	**0.984**	**0.979**	**0.981**	**0.988**	**0.970**	**0.973**	**0.971**	**0.972**

Table 8: Comparison of the performance for all the algorithms, considering their best configurations for the different feature combinations (Feature column). In **bold**, we report the best performance for each algorithm considering F1.

Instead, the best performance for arousal are obtained by the SVM algorithm, when considering data without imputation. For these cases, perfect classification is achieved, with all measures equal to 1.000. These results suggest that voice and biofeedback features can play complementary roles in engagement recognition, and can lead to the best classification results compared to those obtained when using only one source of data, when specific algorithms are selected. These observations are confirmed by Table 8, where the best performance of each algorithm are compared, considering different feature combinations. The table shows that, except for some cases where voice features lead to the best F-1 values, best performance is generally achieved when combined features are used.

	Precision	**Recall**	**F1**	**Accuracy**
Valence				
Multi-layer Perceptron	1.00	1.00	1.00	1.00
Baseline	0.40	0.50	0.45	0.79
Improvement	0.60 (150%)	0.50 (100%)	0.65 (144%)	0.21 (27%)
Arousal				
Support Vector Machines	1.00	1.00	1.00	1.00
Baseline	0.29	0.50	0.37	0.57
Improvement	0.71 (245%)	0.50 (100%)	0.63 (170%)	0.47 (82%)

Table 9: Performance of the best classifiers, according to F1, using combined features with respect to majority class baseline classifiers. Improvement over the baseline is also shown.

As for the other cases, in Table 9 we compare the performance for the best algorithms with the majority class baselines. In all cases, the improvement for precision, recall and F1 is always greater or equal to 100%, thus confirming that the combination of voice and biofeedback features allow obtaining fine-grained distinction of classes for both valence and arousal, which cannot be achieved without combining the features.

5 Discussion

The main take-away messages of this study are:

- users' interviews are activities that can trigger positive engagement in the involved users;
- 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;
- by combining biofeedback features into vectors and by training the Multilayer Perceptron (MLP) and Random Forest (RF) algorithms, it is possible to predict the engagement in a way that outperforms a majority-class baseline, with F1-measure of 98% for valence when using MLP, and 97% for arousal with RF;
- using voice features only when training MLP, Support Vector Machines and RF, performance are increased. Engagement can be predicted through voice features alone with F1-measure 100% (valence, SVM or MLP) and 97% (arousal, RF)
- the combination of biofeedback and voice features maximises the performance, with F-1 measure 100% (valence) and 100% (arousal). In this case, the best performance is achieved with SVM for arousal, and with MLP for valence.
- The major boost in performance is generally obtained by all the algorithms when applying data augmentation by means of SMOTE. Regardless of the types of features, its usage increases the performance by allowing to pass from the range of 0.5-0.6 in terms of F1 to the range of 0.8-1.0.

In the following sections, we discuss our results in relation to existing literature and outline possible applications and timely avenues of research that are enabled by the current study.

5.1 Engagement and Topics

Our descriptive statistics 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 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. (2019)) 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)[9]. 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.

5.2 Performance Comparison with Related Studies

According to the theoretical model of affect described in Sect. 2, in this study we use emotions as a proxy for engagement. Specifically, we operationalize emotions along the valence and arousal dimensions of the Circumplex Model of affect (Russell, 1991), which we recognize using biometrics and voice. Using machine learning, we are able to classify emotions of users engaged in requirements elicitation interviews by distinguishing between positive and negative valence and high and low arousal. We experimented with different experimental setting, i.e. with/without data balancing using SMOTE, data scaling, and data imputation (for voice data only).

As for biometrics, we observe that the performance significantly increases when SMOTE is applied for balancing our training data, achieving and F1-measures up to 0.98 and 0.97 for valence and arousal, respectively (see Table 6), thus outperforming our previous classifier (Girardi et al., 2020a). A direct comparison is possible also possible with the performance achieved in the empirical study by Girardi et al. (2020b), as we use the same device (i.e., Empatica E4 wristband) and include the same metrics for EDA, BVP, and HR. Our classifier performance for arousal (F1 = 0.97 accuracy = .99) and valence (F1 = 0.98 and accuracy = 0.97) outperforms the one they obtain using Empatica—i.e., 0.55 for arousal and 0.59 for valence. They report a slightly better performance, though still lower than ours, when including also the EEG helmet (F1 = 0.59 for arousal and F1 = 0.60 for valence). Müller and Fritz (2015) 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. Overall, tasks are different from ours, as neither voice nor active expression of emotions were

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

triggered in these related works. Our considerably better performance may be also linked to the specific task of interviewing and the actual use of voice, not only as a feature for emotion prediction, but as a mean for emotion expression Laukka (2017); Scherer (2003). Indeed, the simple act of vocalizing can be regarded as an explicit, although not necessarily voluntary, expression of emotion that have an effect on biometric aspects, thus improving the performance of our classifiers also when using biofeedback only as a predictor.

Previous studies in affective computing report comparable performance—e.g., accuracy of 0.97 for arousal (Soleymani et al., 2015; Chen et al., 2015; García et al., 2016) and 0.91 for valence (Nogueira et al., 2013). However, it is worth highlighting that such studies rely on high-definition EEG helmets (Soleymani et al., 2015; Chen et al., 2015; García et al., 2016) and facial electrodes for EMG (Nogueira et al., 2013) which are not comfortable to wear and, thus, could not be used outside a laboratory settings—e.g., during real interviews with users or in remote interviews.

Our approach also achieves comparable performance when using voice features only for both arousal (F1 = 0.97, accuracy = 0.97) and valence (F1 = 1.00 accuracy = 1.00) recognition. Furthermore, the model relying on voice-based features paves the way to future replications for *in vivo* studies that do not require the use of wearable sensors. The voice-based classifier we trained and tested in the scope of this study outperfors most state of the art approaches on speech emotion recognition (Akçay and Oğuz, 2020). However, further research is needed with a larger pool of participants, to further assess the classifier performance on new, unseen speakers.

Previous work also tried to recognize *discrete* emotions instead of valence and arousal. Lin and Wei (2005) used HMM and SVM to classify five emotions, namely anger, happiness, sadness, surprise, and a absence of emotion (i.e, the neutral condition), achieving a performance up to to 99.5% accuracy.

As far as the combination of biofeedback and voice is concerned, our classifier outperforms the approach recently proposed by Aledhari et al. (2020), based on deep learning. As in our study, they use the Empatica E4 wristband for collecting biofeedback and report an accuracy of 85% on test set and 79% in the validation set for recognition of emotional valence.

5.3 Implications for Research and Practice

This is an exploratory study, which is not specifically oriented to direct applications, but rather to have a first understanding of engagement in user interviews, and on the potential usage of biofeedback devices and voice analysis in this context. However, we argue that our results, once consolidated, can have multiple applications and can open new avenues of research.

Applications in User Feedback In user interviews similar to those staged in our experiment, biometric information in the form of biofeedback and voice features can be exploited to better investigate possible discrepancies between

user engagement and the reported relevance of features, to facilitate requirements prioritization tasks, similarly to sentiment analysis applied to textual user feedback Sutcliffe (2011). Furthermore, the usage of these technologies can be extended to identify the engagement of the user *on-the-fly*—i.e., during the interview—to guide analysts steering the flow of the dialogue. These applications, which support human analysts in their activity, become particularly important when *artificial agents* are used to elicit feedback or provide customer support, as shown by related research on voice analysis for call centers (Han et al., 2020; Li et al., 2019). In these works, detection of negative emotions is used to understand when a human operator needs to replace an artificial one, because the latter is irritating the customer. Therefore, our work also opens to further applications on emotion-aware, voice-based chatbots for user interviews.

The Role of Voice The introduction of voice features is particularly decisive in this sense. Biofeedback needs to be locally acquired with specialised devices such as Empatica E4, which: (i) costs about $1,690.00 at the time of writing; (ii) needs to locally register the different signals; (iii) does not remotely send the signal in an automated manner; (iv) can raise privacy reserves in users who are not accustomed to this type of devices. Therefore, their usage is realistic only during face-to-face interviews, in which a certain level of mutual trust can be achieved and all data can be acquired *in loco*. Instead, the analysis of voice is particularly appropriate in remote communication scenarios—involving either human or artificial agents—which are increasingly common due to the COVID-19 pandemic. Voice is voluntarily produced and transmitted by users, and can be remotely recorded and processed without the need to resort to specialised devices, with evident savings in terms of costs. The cost reduction extends the applicability of the idea to large-scale scenarios. With voice analysis, automated user feedback campaigns become feasible, and companies can improve automated A/B testing of web apps or pages. Specifically, they can ask multiple users to interact with different versions of an interface, and speak-up their reflections on the experience. The recording and the analysis of the engagement can be used to facilitate the identification of preferred versions, appreciated features, or relevant interaction problems.

Applications in RE and Software Engineering In the case of more classical requirements elicitation interviews (Davis et al., 2006; Zowghi and Coulin, 2005), the usage of biometrics can support these activities to improve the analyst's ability to create a trustworthy relationship wit the customer, and improve the quality of the interview and the collected data. In this context, it is relevant to extend the work to identify the customer's frustration, which often corresponds to the first step to create mistrust in the analyst (Distanont et al., 2012). Frustration can be detected using biofeedback by analyzing the changes in the heart-rate, temperature, and other vitals (Haag et al., 2004; Wagner et al., 2005; Mandryk et al., 2006; Scheirer et al., 2002a) and used to warn the analyst. Furthermore, frustration is strictly related with stress, which

can be detected in voice signals through Teager energy operator (TEO)-based features (Zhou et al., 2001; Bandela and Kumar, 2017).

Overall, we argue that the analysis of voice, with its relative cost-effectiveness, can be broadly applied not only to RE, but to all software engineering scenarios in which conversations are central (e.g., SCRUM stand-up meetings, information exchange between developers, etc.) to investigate the emotional side of these human-intensive activities that have a relevant impact on the development, but are currently ephemeral in terms of data permanence.

Tacit Knowledge It is worth noting that the improved performance obtained with voice features, and the lower cost of the approach, do not rule-out biofeedback. We have shown that the best performance are actually obtained with a combination of both types of features. In addition, biofeedback captures involuntary body signals that the speaker cannot fully control, while voice tone can, to a certain extent, be manipulated to deceive (Kim and André, 2006). Biofeedback somehow reveals a more faithful representation of emotions, and one can compare discrepancies between emotion prediction with biofeedback and with voice to identify situations in which what the voice appear to tell is different from what the speaker feels. This can happen in requirements elicitation interviews, which can involve controversial political aspects (Milne and Maiden, 2012), or domain experts who need to be interviewed to gather process-related information, but may be reluctant to share their knowledge (Gervasi et al., 2013). Therefore, our research contributes to further scratch the surface of the open problem of tacit knowledge in requirements engineering (Gervasi et al., 2013; Ferrari et al., 2016b; Sutcliffe et al., 2020).

Recently, SE reserachers proposed to identify software usability problem by relying on user emotions derived from facial expression analysis (Johanssen et al., 2019b). In follow-up studies we plan to investigate the combination of audio and visual signals, which have been already proven to be complementary to each other. Specifically, we envisage an approach where facial expression analysis is combined with voice-based emotion detection, thus implementing multimodal arousal and valence classification, in line with previous research on affective computing (Sebe et al., 2006; Pantic and Rothkrantz, 2003; Busso et al., 2004; Tzirakis et al., 2017).

6 Threats to Validity

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

External validity. The generalizability of our results is limited by the amount of subjects (and associated data points) who took part in the study. Although with some imbalance, our sample includes multiple ethnic groups and genders to account for physiological differentiation (Bent et al., 2020). Further replications with a confirmatory design should engage more participants, and consider balance between ethnicity, culture, age, and gender to account for the

differences in emotional reactions due to these aspects. As for the topic of the interviews, we selected features from a commonly-used social media app for which no particular expertise is needed.

To support generalizability, we share the materials and procedures described in this paper, and encourage researchers to adapt them to other domains (e.g., gaming apps) and populations (e.g., children). Moreover, we make this study reproducible and extensible to an new set of data by sharing the scripts necessary to run our analysis[10].

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 (Tantithamthavorn et al., 2018). Following consolidated guidelines for machine learning, we split our data into train-test subsets. The training is performed using cross-validation and the final model performance is assessed on a hold-out test set. The entire process is repeated ten times for each algorithm, to account for random variations in the data. Moreover, our classifiers configuration included scaling and data balancing techniques. To increase the validity of our study, we report all the results related to such configurations.

Construct validity. This threat refers to the reliability of the operationalization of the study constructs. Our study may suffer by threats to construct validity in capturing emotions using self-reports. To address this issue, we performed data quality assurance and excluded participants who did not show engagment with the task (e.g., who provided always the same score or scores with overall low standard deviation). We believe that the designed interview script is sufficiently representative of typical users' interviews in terms of triggered engagement.

Internal validity. Threats to internal validity deal with 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 (Adair, 1984). 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, settings, and the presence of the experimenter. Furthermore, self-assessment questionnaires were filled *immediately after* the interview. This choice was driven by 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) nor their personality traits . We acknowledge that

[10] https://github.com/alessioferrari/VoiceBiofeedEmo

an emotionally-charged event in the life of a participant, either sad or happy, before the interview took place can impact the results.

7 Conclusion and Future Work

This paper presents the first study about engagement prediction in user interviews. In particular, we show that it is possible to predict the positive or negative engagement of a user during an interview about a product. This can be achieved through the usage of *biofeedback* measurements acquired through a wristband, the analysis of emotional prosody (Buchanan et al., 2000) through speech processing, and the application of supervised machine learning algorithms. Furthermore, in budget-constrained development contexts, the usage of voice analysis alone can lead to sufficiently good results. 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, or resorting to voice features only. Furthermore, the approach can be particularly promising to equip artificial agents with some form of emotional sensitivity, so to upgrade relational abilities of voice-based chatbots for gathering product feedback, as well as automated interviewers for requirements elicitation.

The study is exploratory in nature, and application of our results requires 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 (Soleymani et al., 2016) and electroencephalographic (EEG) activity data (Girardi et al., 2020b; Müller and Fritz, 2015); (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.

References

Adair JG (1984) The hawthorne effect: a reconsideration of the methodological artifact. Journal of applied psychology 69(2):334

Akçay MB, Oğuz K (2020) Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication 116:56–76

Aledhari M, Razzak R, Parizi RM, Srivastava G (2020) Deep neural networks for detecting real emotions using biofeedback and voice. In: Bimbo AD, Cucchiara R, Sclaroff S, Farinella GM, Mei T, Bertini M, Escalante HJ, Vezzani R (eds) Pattern Recognition. ICPR International Workshops and Challenges - Virtual Event, January 10-15, 2021, Proceedings, Part IV, Springer, Lecture Notes in Computer Science, vol 12664, pp 302–309, DOI 10.1007/978-3-030-68799-1_21, URL https://doi.org/10.1007/978-3-030-68799-1_21

Allen J (2007) Photoplethysmography and its application in clinical physiological measurement. Physiological measurement 28(3)

Aranda AM, Dieste O, Juristo N (2015) Effect of domain knowledge on elicitation effectiveness: an internally replicated controlled experiment. IEEE Transactions on Software Engineering 42(5):427–451

Bakalova Z, Daneva M (2011) 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, pp 74–80

Bandela SR, Kumar TK (2017) Stressed speech emotion recognition using feature fusion of teager energy operator and mfcc. In: 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, pp 1–5

Bano M, Zowghi D (2015) A systematic review on the relationship between user involvement and system success. Information and Software Technology 58:148–169

Bano M, Zowghi D, Ferrari A, Spoletini P, Donati B (2019) Teaching requirements elicitation interviews: an empirical study of learning from mistakes. Requirements Engineering 24(3):259–289

Barhenke A, Miller AL, Brown E, Seifer R, Dickstein S (2011) Observed emotional and behavioral indicators of motivation predict school readiness in head start graduates. Early Childhood Research Quarterly 26(4):430–441

Batliner A, Fischer K, Huber R, Spilker J, Nöth E (2003) How to find trouble in communication. Speech communication 40(1-2):117–143

Beigi H (2011) Speaker recognition. In: Fundamentals of Speaker Recognition, Springer, pp 543–559

Bent B, Goldstein BA, Kibbe WA, Dunn JP (2020) Investigating sources of inaccuracy in wearable optical heart rate sensors. NPJ digital medicine 3(1):1–9

Braithwaite JJ, Watson DG, Jones R, Rowe M (2015) A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments. Tech. rep., University of Birmingham, UK, University of Birmingham, UK

Buchanan TW, Lutz K, Mirzazade S, Specht K, Shah NJ, Zilles K, Jäncke L (2000) Recognition of emotional prosody and verbal components of spoken language: an fmri study. Cognitive Brain Research 9(3):227–238

Burleson W, Picard RW (2004) Affective agents: Sustaining motivation to learn through failure and state of "stuck". In: Social and Emotional Intelli-

gence in Learning Environments Workshop.
Busso C, Deng Z, Yildirim S, Bulut M, Lee C, Kazemzadeh A, Lee S, Neumann U, Narayanan S (2004) Analysis of emotion recognition using facial expressions, speech and multimodal information. pp 205–211, DOI 10.1145/1027933.1027968
Canento F, Fred A, Silva H, Gamboa H, Lourenço A (2011) Multimodal biosignal sensor data handling for emotion recognition. In: SENSORS, IEEE, pp 647–650, DOI 10.1109/ICSENS.2011.6127029
Chao L, Tao J, Yang M, Li Y, Wen Z (2015) Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, pp 65–72
Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16:321–357
Chen L, Mao X, Xue Y, Cheng LL (2012) Speech emotion recognition: Features and classification models. Digital signal processing 22(6):1154–1160
Chen M, Han J, Guo L, Wang J, Patras I (2015) Identifying valence and arousal levels via connectivity between eeg channels. In: Proc. of the 2015 Int. Conf. on Affective Computing and Intelligent Interaction (ACII), IEEE Computer Society, USA, ACII '15, pp 63–69, DOI 10.1109/ACII.2015.7344552
Colomo-Palacios R, Casado-Lumbreras C, Soto-Acosta P, García-Crespo Á (2011) Using the affect grid to measure emotions in software requirements engineering
Cowie R, Sussman N, Ben-Ze'ev A (2011) Emotion: Concepts and definitions. In: Emotion-oriented systems, Springer, pp 9–30
Critchley H, Nagai Y (2013) Electrodermal Activity (EDA), Springer New York, New York, NY, pp 666–669. DOI 10.1007/978-1-4419-1005-9_13
Dan-Glauser ES, Scherer KR (2011) The geneva affective picture database (gaped): a new 730-picture database focusing on valence and normative significance. Behavior research methods 43(2):468
Davis A, Dieste O, Hickey A, Juristo N, Moreno AM (2006) Effectiveness of requirements elicitation techniques: Empirical results derived from a systematic review. In: 14th IEEE Int. Requirements Engineering Conf. (RE'06), IEEE, pp 179–188
Distanont A, Haapasalo H, Vaananen M, Lehto J (2012) The engagement between knowledge transfer and requirements engineering. IJKL 1(2):131–156
Fernández DM, Wagner S, Kalinowski M, Felderer M, Mafra P, Vetrò A, Conte T, Christiansson MT, Greer D, Lassenius C, et al. (2017) Naming the pain in requirements engineering. Empirical software engineering 22(5):2298–2338
Ferrari A, Spoletini P, Gnesi S (2016a) Ambiguity and tacit knowledge in requirements elicitation interviews. Requirements Engineering 21(3):333–355
Ferrari A, Spoletini P, Gnesi S (2016b) Ambiguity cues in requirements elicitation interviews. In: 2016 IEEE 24th International Requirements Engineering

Conference (RE), IEEE, pp 56–65

Fritz T, Begel A, Müller SC, Yigit-Elliott S, Züger M (2014) 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, pp 402–413, DOI 10.1145/2568225.2568266, URL https://doi.org/10.1145/2568225.2568266

Fucci D, Girardi D, Novielli N, Quaranta L, Lanubile F (2019) 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, pp 311–322, URL https://dl.acm.org/citation.cfm?id=3339126

García HF, Álvarez MA, Orozco ÁÁ (2016) Gaussian process dynamical models for multimodal affect recognition. In: 38th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society, EMBC 2016, pp 850–853, DOI 10.1109/EMBC.2016.7590834

Gervasi V, Gacitua R, Rouncefield M, Sawyer P, Kof L, Ma L, Piwek P, De Roeck A, Willis A, Yang H, et al. (2013) Unpacking tacit knowledge for requirements engineering. In: Managing requirements knowledge, Springer, pp 23–47

Girardi D, Lanubile F, Novielli N (2017) 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, pp 125–130, DOI 10.1109/ACII.2017.8273589

Girardi D, Ferrari A, Novielli N, Spoletini P, Fucci D, Huichapa T (2020a) The way it makes you feel predicting users' engagement during interviews with biofeedback and supervised learning. In: Breaux TD, Zisman A, Fricker S, Glinz M (eds) 28th IEEE International Requirements Engineering Conference, RE 2020, Zurich, Switzerland, August 31 - September 4, 2020, IEEE, pp 32–43, DOI 10.1109/RE48521.2020.00016, URL https://doi.org/10.1109/RE48521.2020.00016

Girardi D, Novielli N, Fucci D, Lanubile F (2020b) Recognizing Developers' Emotions while Programming. In: 42nd Int. Conf. on Software Engineering (ICSE '20), May 23–29, 2020, Seoul, Republic of Korea, DOI 10.1145/3377811.3380374, URL https://doi.org/10.1145/3377811.3380374

Graziotin D, Wang X, Abrahamsson P (2015) Do feelings matter? on the correlation of affects and the self-assessed productivity in software engineering. Journal of Software: Evolution and Process 27(7):467–487, DOI 10.1002/smr.1673, URL https://doi.org/10.1002/smr.1673

Greco A, Valenza G, Lanata A, Scilingo EP, Citi L (2016) cvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing. IEEE Transactions on Biomedical Engineering 63(4):797–804, DOI 10.1109/TBME.2015.2474131

Greene S, Thapliyal H, Caban-Holt A (2016) A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health. IEEE Consumer Electronics Magazine 5(4):44–56

Groen EC, Seyff N, Ali R, Dalpiaz F, Doerr J, Guzman E, Hosseini M, Marco J, Oriol M, Perini A, et al. (2017) The crowd in requirements engineering: The landscape and challenges. IEEE software 34(2):44–52

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

Guzman E, Alkadhi R, Seyff N (2017) An exploratory study of twitter messages about software applications. Requirements Engineering 22(3):387–412

Haag A, Goronzy S, Schaich P, Williams J (2004) Emotion recognition using bio-sensors: First steps towards an automatic system. In: ADS'04, Springer, pp 36–48

Hadar I, Soffer P, Kenzi K (2014) The role of domain knowledge in requirements elicitation via interviews: an exploratory study. Requirements Engineering 19(2):143–159

Han W, Jiang T, Li Y, Schuller B, Ruan H (2020) Ordinal learning for emotion recognition in customer service calls. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 6494–6498

Hanssen GK, Fægri TE (2006) Agile customer engagement: a longitudinal qualitative case study. In: Proc. of the 2006 ACM/IEEE Int. symposium on Empirical software engineering, pp 164–173

Heiskari J, Lehtola L (2009) Investigating the state of user involvement in practice. In: 2009 16th Asia-Pacific Software Engineering Conf., IEEE, pp 433–440

Issa D, Demirci MF, Yazici A (2020) Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control 59:101894

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

Johanssen JO, Bernius JP, Bruegge B (2019a) 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, pp 1–7

Johanssen JO, Bernius JP, Bruegge B (2019b) Toward usability problem identification based on user emotions derived from facial expressions. In: Proceedings of the 4th International Workshop on Emotion Awareness in Software Engineering, IEEE Press, SEmotion '19, p 1–7, DOI 10.1109/SEmotion.2019.00008, URL https://doi.org/10.1109/SEmotion.2019.00008

Kamthan P, Shahmir N (2017) Effective user stories are affective. In: Int. Conf. on Ubiquitous Computing and Ambient Intelligence, Springer, pp 605–611

Kapoor A, Burleson W, Picard RW (2007) Automatic prediction of frustration. Int Journal Human-Computer Studies 65(8):724–736, DOI 10.1016/j.ijhcs.2007.02.003

Kim J, André E (2008) Emotion recognition based on physiological changes in music listening. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(12):2067–2083, DOI 10.1109/TPAMI.2008.26, URL https://doi.org/10.1109/TPAMI.2008.26

Kim J, André E (2006) Emotion recognition using physiological and speech signal in short-term observation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 4021 LNAI:53–64, DOI 10.1007/11768029_6

Koelstra S, Mühl C, Soleymani M, Lee J, Yazdani A, Ebrahimi T, Pun T, Nijholt A, Patras I (2012) DEAP: A database for emotion analysis using physiological signals. IEEE Transaction on Affective Computing 3(1):18–31, DOI 10.1109/T-AFFC.2011.15, URL https://doi.org/10.1109/T-AFFC.2011.15

Kramer AE (1990) Physiological Metrics of Mental Workload: A Review of Recent Progress. DOI https://doi.org/10.21236/ada223701

Kurtanović Z, Maalej W (2017) Mining user rationale from software reviews. In: 2017 IEEE 25th Int. Requirements Engineering Conf. (RE), IEEE, pp 61–70

Kurtanović Z, Maalej W (2018) On user rationale in software engineering. Requir Eng 23(3):357–379, DOI 10.1007/s00766-018-0293-2, URL https://doi.org/10.1007/s00766-018-0293-2

Kushki A, Fairley J, Merja S, King G, Chau T (2011) Comparison of blood volume pulse and skin conductance responses to mental and affective stimuli at different anatomical sites. Physiological measurement 32(10):1529

Lang PJ, Bradley M (2007) The int. affective picture system (iaps) in the study of emotion and attention. In: Coan JA, Allen JJB (eds) Handbook of Emotion Elicitation and Attention, Oxford University Press, chap 2, pp 29–46

Laukka P (2017) Vocal Communication of Emotion, Springer International Publishing, Cham, pp 1–6. DOI 10.1007/978-3-319-28099-8_562-1, URL https://doi.org/10.1007/978-3-319-28099-8_562-1

Lazarus RS (1991) Emotion and Adaptation. Oxford University Press

Lee CM, Narayanan SS (2005) Toward detecting emotions in spoken dialogs. IEEE transactions on speech and audio processing 13(2):293–303

Li B, Dimitriadis D, Stolcke A (2019) Acoustic and lexical sentiment analysis for customer service calls. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 5876–5880

Li M, Lu BL (2009) Emotion classification based on gamma-band eeg. In: 2009 Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society, pp 1223–1226, DOI 10.1109/IEMBS.2009.5334139

Lin YL, Wei G (2005) Speech emotion recognition based on hmm and svm. pp 4898–4901, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-28444440262&partnerID=40&md5=7e4f407dc2b41b65a5164f25938c2ef2, cited By 132

Maalej W, Nabil H (2015) Bug report, feature request, or simply praise? on automatically classifying app reviews. In: 2015 IEEE 23rd Int. requirements engineering Conf. (RE), IEEE, pp 116–125

Maalej W, Nayebi M, Johann T, Ruhe G (2015) Toward data-driven requirements engineering. IEEE Software 33(1):48–54

Mandryk R, Inkpen K, Calvert T (2006) Using psychophysiological techniques to measure user experience with entertainment technologies 25(2):141–158

Martens D, Maalej W (2019) Release early, release often, and watch your users' emotions: Lessons from emotional patterns. IEEE Software 36(5):32–37

McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, Nieto O (2015) librosa: Audio and music signal analysis in python. Citeseer

Mennig P, Scherr SA, Elberzhager F (2019) Supporting rapid product changes through emotional tracking. In: 2019 IEEE/ACM 4th Int. Workshop on Emotion Awareness in Software Engineering (SEmotion), IEEE, pp 8–12

Miller T, Pedell S, Lopez-Lorca AA, Mendoza A, Sterling L, Keirnan A (2015) Emotion-led modelling for people-oriented requirements engineering: the case study of emergency systems. Journal of Systems and Software 105:54–71

Milne A, Maiden N (2012) Power and politics in requirements engineering: embracing the dark side? Requirements Engineering 17(2):83–98

Morrison D, Wang R, De Silva LC (2007) Ensemble methods for spoken emotion recognition in call-centres. Speech communication 49(2):98–112

Müller SC, Fritz T (2015) 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, pp 688–699, DOI 10.1109/ICSE.2015.334

Müller SC, Fritz T (2016) 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, pp 452–463, DOI 10.1145/2884781.2884803, URL https://doi.org/10.1145/2884781.2884803

Murukannaiah PK, Ajmeri N, Singh MP (2016) 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, pp 176–185

Nanda A, Sa PK, Choudhury SK, Bakshi S, Majhi B (2017) A neuromorphic person re-identification framework for video surveillance. IEEE Access 5:6471–6482

Nogueira PA, Rodrigues RA, Oliveira EC, Nacke LE (2013) 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, pp 253–260, DOI 10.1109/WI-IAT.2013.117

Ortony A, Clore G, Collins A (1988) The Cognitive Structure of Emotion, vol 18. DOI 10.2307/2074241

Ottl S, Amiriparian S, Gerczuk M, Karas V, Schuller B (2020) Group-level speech emotion recognition utilising deep spectrum features. In: Proceedings

of the 2020 International Conference on Multimodal Interaction, pp 821–826
Pantic M, Rothkrantz L (2003) Toward an affect-sensitive multimodal human-computer interaction. Proceedings of the IEEE 91(9):1370–1390, DOI 10.1109/JPROC.2003.817122
Parnin C (2011) Subvocalization - toward hearing the inner thoughts of developers. pp 197 – 200, DOI 10.1109/ICPC.2011.49
Petrushin V (1999) Emotion in speech: Recognition and application to call centers. In: Proceedings of artificial neural networks in engineering, vol 710, p 22
Ramakrishnan S, El Emary IM (2013) Speech emotion recognition approaches in human computer interaction. Telecommunication Systems 52(3):1467–1478
Raschka S (2018) Model evaluation, model selection, and algorithm selection in machine learning. CoRR abs/1811.12808, URL http://arxiv.org/abs/1811.12808, 1811.12808
Ren M, Nie W, Liu A, Su Y (2019) Multi-modal correlated network for emotion recognition in speech. Visual Informatics 3(3):150–155
Reuderink B, Mühl C, Poel M (2013) Valence, arousal and dominance in the eeg during game play. Int Journal of Autonomous and Adaptive Communications Systems 6(1):45–62, DOI 10.1504/IJAACS.2013.050691, URL http://dx.doi.org/10.1504/IJAACS.2013.050691
Russell J (1991) Culture and the categorization of emotions. Psychological Bulletin 110 (3):426–450, DOI 10.1037/0033-2909.110.3.426
Sailunaz K, Dhaliwal M, Rokne J, Alhajj R (2018) Emotion detection from text and speech: a survey. Social Network Analysis and Mining 8(1):28
Scheirer J, Fernandez R, Klein J, Picard R (2002a) Frustrating the user on purpose: a step toward building an affective computer. Interacting with Computers
Scheirer J, Fernandez R, Klein J, Picard RW (2002b) Frustrating the user on purpose: a step toward building an affective computer. Interacting with Computers 14:93–118, DOI 10.1016/S0953-5438(01)00059-5, URL https://ieeexplore.ieee.org/document/8160759
Scherer KR (2003) Vocal communication of emotion: A review of research paradigms. Speech Communication 40(1):227–256, DOI https://doi.org/10.1016/S0167-6393(02)00084-5, URL https://www.sciencedirect.com/science/article/pii/S0167639302000845
Scherr SA, Kammler C, Elberzhager F (2019a) 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, pp 169–173
Scherr SA, Mennig P, Kammler C, Elberzhager F (2019b) On the road to enriching the app improvement process with emotions. In: 2019 IEEE 27th Int. Requirements Engineering Conf. Workshops (REW), IEEE, pp 84–91
Schuller B, Batliner A (2013) Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. Chichester: John Wiley Sons.

Schuller B, Rigoll G, Lang M (2004) Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE, vol 1, pp I–577

Schuller BW (2018) Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Communications of the ACM 61(5):90–99

Sebastiani F (2001) Machine learning in automated text categorization. ACM Computing Surveys 34:1–47, DOI 10.1145/505282.505283

Sebe N, Cohen I, Gevers T, Huang T (2006) Emotion recognition based on joint visual and audio cues. vol 1, pp 1136–1139, DOI 10.1109/ICPR.2006.489

Shah FA, Sirts K, Pfahl D (2019) Using app reviews for competitive analysis: tool support. In: Proc. of the 3rd ACM SIGSOFT Int. Workshop on App Market Analytics, pp 40–46

Sinex JE (1999) Pulse oximetry: principles and limitations. The American journal of emergency medicine 17(1):59–66

Soleymani M, Pantic M, Pun T (2015) 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, pp 491–497, DOI 10.1109/ACII.2015.7344615

Soleymani M, Asghari-Esfeden S, Fu Y, Pantic M (2016) Analysis of EEG signals and facial expressions for continuous emotion detection. IEEE Transaction on Affective Computing 7(1):17–28, DOI 10.1109/TAFFC.2015.2436926, URL https://doi.org/10.1109/TAFFC.2015.2436926

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

Stade M, Scherr SA, Mennig P, Elberzhager F, Seyff N (2019) Don't worry, be happy–exploring users' emotions during app usage for requirements engineering. In: 2019 IEEE 27th Int. Requirements Engineering Conf. (RE), IEEE, pp 375–380

Sutcliffe A (2011) Emotional requirements engineering. In: 2011 IEEE 19th Int. Requirements Engineering Conf., IEEE, pp 321–322

Sutcliffe A, Sawyer P, Stringer G, Couth S, Brown LJ, Gledson A, Bull C, Rayson P, Keane J, Zeng Xj, et al. (2020) Known and unknown requirements in healthcare. Requirements engineering 25(1):1–20

Tantithamthavorn C, McIntosh S, Hassan AE, Matsumoto K (2016) 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, pp 321–332, DOI 10.1145/2884781.2884857

Tantithamthavorn C, McIntosh S, Hassan AE, Matsumoto K (2018) The impact of automated parameter optimization on defect prediction models. IEEE Transactions on Software Engineering 45(7):683–711

Tantithamthavorn C, McIntosh S, Hassan AE, Matsumoto K (2019) The impact of automated parameter optimization on defect prediction models. IEEE Transactions on Software Engineering 45(7):683–711, DOI 10.1109/

TSE.2018.2794977
Taveter K, Sterling L, Pedell S, Burrows R, Taveter EM (2019) 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, pp 100–105
Trice M, Potts L, Small R (2019) Values versus rules in social media communities: How platforms generate amorality on reddit and facebook. In: Digital Ethics, Routledge, pp 33–50
Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou MA, Schuller B, Zafeiriou S (2016) Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 5200–5204
Tzirakis P, Trigeorgis G, Nicolaou M, Schuller B, Zafeiriou S (2017) End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal on Selected Topics in Signal Processing 11(8):1301–1309, DOI 10.1109/JSTSP.2017.2764438
Ververidis D, Kotropoulos C (2006) Emotional speech recognition: Resources, features, and methods. Speech communication 48(9):1162–1181
Wagner J, Kim J, Andre E (2005) From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification. In: ICME 2005, IEEE
Werner C, Li ZS, Damian D (2019) Can a machine learn through customer sentiment?: A cost-aware approach to predict support ticket escalations. IEEE Software 36(5):38–45
Williams G, Mahmoud A (2017) Mining twitter feeds for software user requirements. In: 2017 IEEE 25th Int. Requirements Engineering Conf. (RE), IEEE, pp 1–10
Yu C, Aoki PM, Woodruff A (2004) 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, URL http://www.isca-speech.org/archive/interspeech_2004/i04_1329.html
Zhao Z, Bao Z, Zhao Y, Zhang Z, Cummins N, Ren Z, Schuller B (2019) Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition. IEEE Access 7:97515–97525
Zhou G, Hansen JH, Kaiser JF (2001) Nonlinear feature based classification of speech under stress. IEEE Transactions on speech and audio processing 9(3):201–216
Zowghi D, Coulin C (2005) Requirements elicitation: A survey of techniques, approaches, and tools. In: Engineering and managing software requirements, Springer, pp 19–46
Züger M, Müller SC, Meyer AN, Fritz T (2018) 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), p 591, DOI 10.1145/3173574.3174165, URL https://doi.org/10.1145/3173574.3174165