Journal on Interactive Systems, 2026, 17:1
ISSN: 2763-7719 • doi: 10.5753/jis.2026.7625

Research Paper

User Experience Evaluation in Non-Immersive 3D Digital Environments Using Facial Emotion Recognition

Érico de Souza Veriscimo [Sao Paulo Federal Institute | *veriscimo.erico@ifsp.edu.br*]
João Luiz Bernardes Júnior [University of Sao Paulo | *jlbernardes@usp.br*]
Luciano Antonio Digiampietri [University of Sao Paulo | *digiampietri@usp.br*]

Institute of Computing, Federal Institute of São Paulo – São Miguel Paulista Campus, Rua Ten. Miguel Delia, 105, São Miguel Paulista, São Paulo, SP, 08021-090, Brazil.

Abstract. User experience (UX) is fundamental for the acceptance and use of information systems. Although there are well-known and widely used UX evaluation techniques for traditional interfaces, a literature review revealed several gaps regarding UX evaluation in non-immersive 3D interaction. One significant gap is the predominant use of pragmatic criteria in assessments, while another is the lack of an approach that evaluates hedonic aspects using facial emotion recognition. This work proposes an approach for automatically evaluating user experience in non-immersive three-dimensional environments, focusing on its hedonic aspects based on facial emotion recognition. An experimental protocol was developed and approved by the Ethics Committee. The experiment was conducted with 52 participants. Throughout the testing period (before, during, and after the interaction), participants' faces were recorded using a low-cost camera. The experiment involved participants playing a game and answering questionnaires, including categorization and mood profile instruments, as well as the UEQ-S and the PLEX Framework. The Face-api.js library was used for facial emotion recognition. The hypothesis that automatic facial emotion recognition can support user experience evaluation was confirmed. This method enabled the estimation of UEQ-S and PLEX questionnaire responses with an average error of approximately ±1 point using only emotion extraction through an artificial intelligence model. Given that UX evaluation is crucial for the acceptance of new software or functionality, this work contributes to improving system quality and acceptance.

Keywords: Emotion Recognition, User Experience, 3D Interaction, Facial Expression Analysis

Edited by: Emanuel Felipe Duarte | **Received:** 16 February 2026 • **Accepted:** 21 May 2026 • **Published:** 30 May 2026

1 Introduction

User Experience (UX) is one of the main factors used to design, describe, or improve the way users interact with a system and how they feel during this interaction [Rajeshkumar *et al.*, 2013]. This is particularly relevant when user feedback can influence others, such as in an app store [Mennig *et al.*, 2019]. Therefore, UX is crucial for the acceptance, engagement, and competitive advantage of technological products, systems, or services [Veriscimo *et al.*, 2020; Martinelli *et al.*, 2022].

ISO 9241 [International Organization for Standardization, 2010] defines UX as people's perceptions and responses resulting from the use of a product, system, or service, including all emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviors, and realizations that occur before, during, and after use. With so many variables, choosing instruments to evaluate UX is not a simple task.

There are several ways to evaluate UX, which can be used in combination or alone, such as observation, event recording, and self-assessment through questionnaires, among others. However, most evaluations still focus mainly on user performance criteria without giving due importance to criteria related to user emotions and enjoyment [Veriscimo *et al.*, 2020]. This could be a flaw, since, according to ISO 9241 [International Organization for Standardization, 2010], emotion is fundamental to UX.

Among all available assessment instruments, questionnaires are still the most used [Veriscimo *et al.*, 2020]. Despite the existence of well-established techniques for evaluating UX in traditional interfaces [Bevan, 2009], a review of the literature revealed several gaps in the evaluation of UX in 3D interaction. One such gap is the predominant reliance on primarily pragmatic criteria, often to the exclusion of most or all hedonic criteria. Pragmatic criteria refer to user task performance in the system, while hedonic criteria are related to user emotions and pleasure [Veriscimo *et al.*, 2020].

Another literature review highlighted gaps in the assessment of user experience, or some aspect of the experience, through facial emotion recognition [Veriscimo *et al.*, 2021]. Among these gaps is the absence of an approach that considers more aspects of UX and in combination, rather than isolated aspects, using facial emotion recognition.

The main objective of this work is to compare, during the use of a 3D application, the recognition of emotions through the user's facial expressions with their self-assessment through two standardized *UX* questionnaires and, based on the analysis from the results of this comparison, propose and develop an approach for the automatic evaluation of the user experience in this context, especially its hedonic aspects, based on facial emotion recognition.

This new approach focuses on the hedonic aspects of the user experience, which are often neglected by traditional evaluations that tend to emphasize pragmatic criteria.

As a secondary contribution, we explored the use of user responses from one UX questionnaire to assist in inferring the responses to another. This problem differs from the primary focus of the present work, which aims to infer the experience of users who did not respond to any UX questionnaire. How-

ever, it can serve as a valuable tool for comparative analyses between studies that employed different questionnaires.

Inferring user experience or predicting potential responses to a UX questionnaire is an extremely relevant problem, as it automates a time-consuming and often unpopular task for users—filling out lengthy questionnaires. Similarly, the second problem addressed in this article, which involves using responses from one questionnaire to infer responses to another, is equally significant, as it enables meaningful comparisons between studies that employ different questionnaires.

The remainder of this paper is organized as follows. Section 2 presents the related work, highlighting existing approaches and research gaps in UX evaluation and emotion recognition. Section 3 describes the proposed methodology, including the experimental design, data collection, and processing steps. Section 4 presents and discusses the results obtained from the predictive models. Finally, Section 5 summarizes the main findings, discusses limitations, and outlines directions for future work.

2 Related Work

A systematic review on user experience assessment [Veriscimo *et al*., 2020] identified that self-assessment via questionnaire is the most common method in UX evaluation. Regarding the evaluation criteria, hedonic aspects received little consideration in the reviewed works. Thus, it is clear that there is a greater concern with efficiency than with the user's mood, emotions, and pleasure.

It is possible to argue that in several areas of application (such as education, entertainment, and even assistance in decision-making), hedonic aspects are more decisive for the adoption and use of technology than pragmatic ones.

The most commonly used standardized questionnaire in UX assessment, discussed by Veriscimo *et al*. [2020], is the User Experience Questionnaire (UEQ). The UEQ aims to evaluate UX with 26 items distributed across six categories: attractiveness, perspicuity, efficiency, reliability, stimulation, and novelty [Laugwitz *et al*., 2008]. There is also a short version of this questionnaire, called short version of the User Experience Questionnaire (UEQ-S), with only eight questions. Both the short and traditional versions use a seven-point semantic differential scale between antonyms, such as "difficult" and "easy".

In another systematic review on the evaluation of user experience or some aspect of the user experience through facial emotion recognition [Veriscimo *et al*., 2021], it was noted that there is an absence of work evaluating UX as a whole, and not just some aspects, such as accessibility or usability.

In the same review [Veriscimo *et al*., 2021], most studies relied on facial images as the primary data type for emotion recognition. Additionally, facial point tracking and facial muscle activity analysis were used, with the face serving as the main data source across the reviewed studies.

In this context, this work proposes the development of an approach for the automatic evaluation of user experience, using as a basis the gaps found: the prevalence of pragmatic aspects in evaluations [Veriscimo *et al*., 2020] and, when there is the use of facial emotion recognition in UX evaluation, the lack of an approach that evaluates UX as a whole [Veriscimo *et al*., 2021].

Considering user experience evaluation using machine learning and facial expressions, Santos and Digiampietri [2024] performed a systematic literature review. The authors highlighted a range of studies aimed at evaluating and improving UX, though few focused on predicting UX by considering factors such as a user's emotional state before interaction or their declared experience with the system. A significant challenge identified was the lack of a consistent definition of UX, which varies greatly across articles, complicating the standardization of methodologies and results. Many studies emphasized the development of recommendation algorithms for content like music and news, leveraging emotional data to personalize suggestions and enhance user satisfaction and engagement.

Recent advances in emotion recognition have been largely driven by deep learning techniques, particularly convolutional neural networks (CNNs) and transformer-based architectures, which have significantly improved performance in tasks such as facial expression analysis, speech-based emotion detection, and multimodal affect recognition. Contemporary studies increasingly emphasize multimodal approaches, combining facial expressions, voice, physiological signals, and contextual data to enhance robustness and reliability. Despite these advances, the field still faces important challenges, including limited generalization across datasets, sensitivity to environmental conditions (e.g., lighting, occlusion), and potential demographic and cultural biases embedded in training data. In the context of Human-Computer Interaction, there is a growing body of research exploring the integration of automatic emotion recognition with user experience (UX) evaluation; however, scholars highlight that emotional responses represent only one dimension of UX and should be interpreted with caution. As a result, current trends advocate hybrid approaches that combine automated emotion recognition with traditional UX methods, aiming to balance scalability, interpretability and methodological rigor [Rehman *et al*., 2025].

These findings reinforce the observation that, although there has been significant progress in the use of machine learning and facial emotion recognition for UX-related applications, the focus of most existing approaches remains on supporting or improving user experience rather than explicitly modeling it as a predictive outcome. Moreover, as highlighted by Santos and Digiampietri [2024], the reliance on facial data as the primary source for emotion detection, combined with the absence of standardized UX definitions and evaluation frameworks, poses additional challenges for developing generalizable and comparable models. In this sense, the literature still lacks approaches capable of integrating emotional signals with established UX constructs in a unified and predictive manner, particularly in interactive and dynamic contexts such as three-dimensional environments.

3 Method

Figure 1 presents a flowchart of the main tasks developed throughout this work.

Based on information obtained in the related literature,

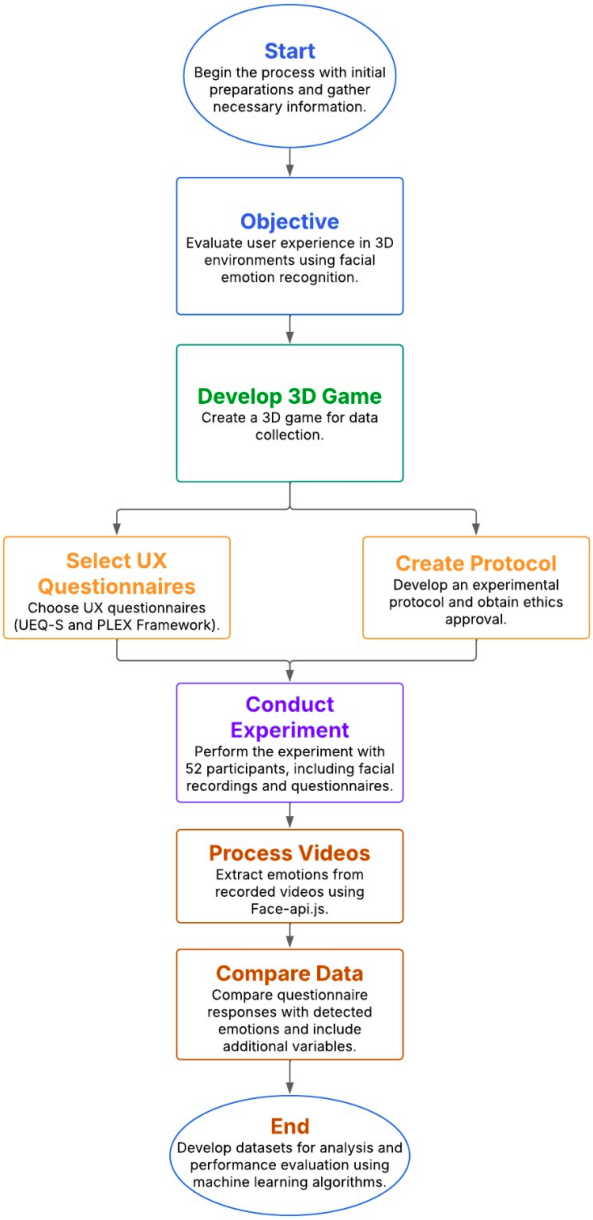

Figure 1. Flowchart of the main tasks developed throughout this work.

such as the importance of user emotions in UX evaluation, evaluation methodologies and criteria and strategies for emotion recognition, we designed an innovative UX evaluation methodology based on facial emotion recognition. We compare user emotion recognition (through their facial expressions) during the use of a 3D application with their self-assessment through two standardized UX questionnaires after the experience and, based on the analysis of the results of this comparison, develop and propose an approach for the automatic evaluation of user experience in this context, especially its hedonic aspects, based on facial emotion recognition.

A three-dimensional game was developed to test the user experience and, through its use, collect users' facial images. This game may be considered a non-immersive or monitor-based virtual reality application [Milgram *et al.*, 1995]. While we considered conducting the experiment in Augmented Reality or immersive Virtual Reality settings, such setups would require devices that occlude users' faces, limiting reliable facial emotion capture without a substantially more complex experimental design. Therefore, we adopted a non-immersive approach in this study. Nevertheless, the proposed method is not restricted to non-immersive environments and can be extended to general 3D contexts. Future work should investigate alternative acquisition strategies—such as external camera setups or sensor fusion—to enable the application of this approach in immersive environments without compromising facial expression analysis. Two types of input devices were used, the keyboard and *Leap Motion*, allowing the comparison between a traditional and three-dimensional interaction. The option for a game, rather than another system, was made especially because it was believed that it could be a quick way of eliciting emotions in users, including different emotions (such as satisfaction for achieving an objective or frustration due to difficulty) when presenting levels with different degrees of difficulty. The game is presented in Section 3.1.

It was necessary to choose one or more standardized UX questionnaires to be used as labels (templates) in the training and evaluation of the methodology proposed in this work. It led to the selection of UEQ-S and PLEX Framework and will be discussed in Section 3.2.

An experimental protocol was developed for user testing and submitted to the Research Ethics Committee of EACH-USP. The project was approved under number 38540620.0.0000.5390 [1] . The primary objective of this experiment is to assess the feasibility of evaluating user experience through the automatic detection of facial expressions.

Section 3.3 details the experiment. Section 3.4 presents the strategy used for facial emotion recognition, as well as the creation of the datasets to compare the answers to the questionnaire and the emotions detected in Section 3.5. Finally, this comparison is made in section 3.7. It also presents the procedures for this comparison. Finally, Section 4 presents the results of this work.

3.1 3D game development

Using the *Unity 3D* game engine and *C#* programming language, we developed a three-dimensional game with which

[1]Download link to the original document in Portuguese `https://drive.google.com/file/d/126OPREsnoyOHbZYgduV21GsN7VyOVa_W/view?usp=sharing`. Accessed on 26 May 2026.

users interact in our experiments. The goal of the game is to use a rolling ball to "capture" (i.e., collide with) three points spread throughout the scenario. As in classic games such as the *Super Monkey Ball* series [Amusement Vision, 2001], instead of controlling the ball directly, the player tilts the entire scenario, changing the terrain inclination, and the ball rolls according to the simulated effect of gravity. By controlling terrain slope in this way, the user makes the ball change direction, accelerate or slow down. There are three levels with varying difficulties: the initial level is the easiest, followed by an intermediate level, with the third and final level being the most complex. Figures 2, 3, 4 and 5 illustrate different aspects of the game, including the initial level configuration, movement dynamics, spatial layout, and interaction with in-game objectives.

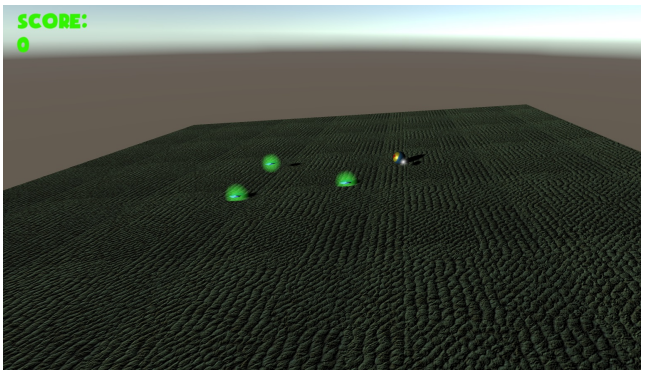

Figure 2. Screenshot of the initial scene from the first level of the game. This level has no obstacles, allowing players to familiarize themselves with the controls and mechanics.

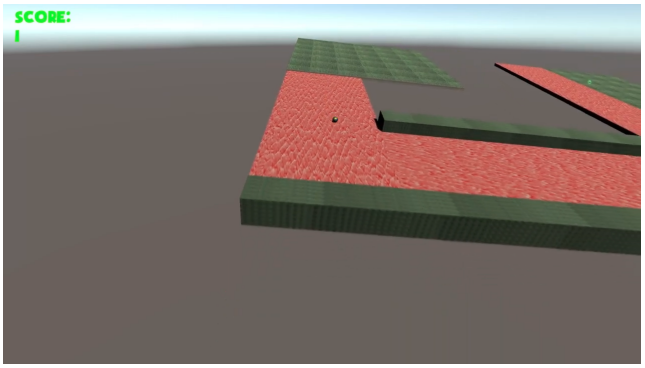

Figure 3. Screenshot of a gameplay moment showing the sphere navigating through the level. This scene highlights movement dynamics and the influence of control on the player's trajectory.

The scenario can be tilted around two axes, changing its pitch (rotation around our Y axis) and roll (rotation around X), in order to make the ball move according to a physics simulation of gravity. In this way, even controlling only two dimensions (pitch and roll), the player can make the ball move in a complex manner in 3D space and the game requires good 3D visualization and manipulation skills. In similar classic games, terrain pitch and roll are usually controlled using a gamepad stick (often the analog stick), with the back and forward movement of the stick mapping to changes in pitch and left and right movements controlling roll. We adapted this control scheme to be used with a computer keyboard in a way used very often in gaming, using the WASD keys to

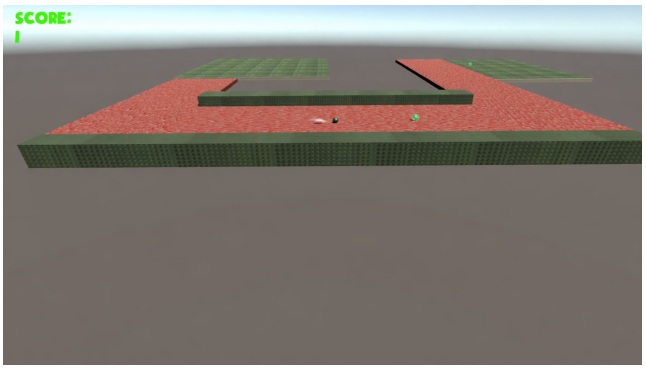

Figure 4. Screenshot providing a general view of the second level of the game. This scene illustrates the spatial layout and navigation paths, offering an overview of the interaction environment.

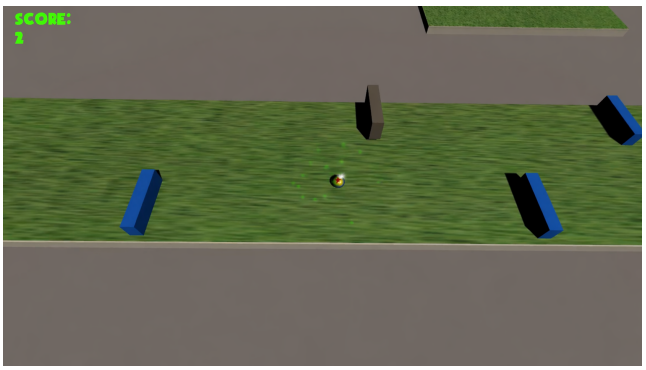

Figure 5. Screenshot of a gameplay moment in which the player captures a collectible item. The particle effects surrounding the sphere indicate successful interaction and highlight the presence of in-game objectives.

map directions, with "W" representing the stick's forward movement, "S" representing pulling it back, and "A" and "D" mapping to left and right movements, respectively. In this way, "W" and "S" control terrain pitch, with "W" tilting it forward and "S" tilting it back, while "A" and "D" control its roll to the left and right, respectively.

In addition to the keyboard control scheme, we also implemented a free-hand scheme, in which the user freely moves a single hand in space to tilt the scenario, as exemplified in Figures 6 and 7. In this scheme, the goal is to map variations in the hand's pitch and roll directly to the scenario, in what might be a more natural interaction. The free-hand tracking is performed by the *Leap Motion* sensor, which provides three-dimensional positioning (x, y, and z) of the participant's fingers and wrist.

In this second version, roll is controlled by the relative positions of the user's thumb and little finger, while pitch is controlled by the relative positions of the middle finger and wrist. Both control mappings incorporate a variable named "weight", which is used to filter sensor noise by disregarding values below a defined threshold.

A *log* is recorded every time the user loses the game, captures any point, or wins the game.

A demo video of the developed game can be seen at the following link: `https://drive.google.com/file/d/18drXYn88lwLnOAGLwn4Kguu5qSpohivG/view?usp=sharing`. Accessed on 26 May 2026.

3.2 Questionnaire selection

We selected two questionnaires for the experiment. The first is the short version of the User Experience Questionnaire

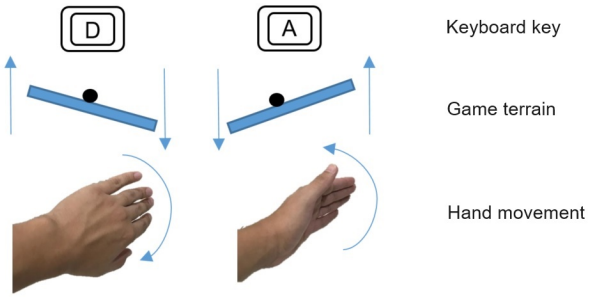

Figure 6. Controlling terrain roll. Example of the free-hand control scheme, where the user moves their hand in space to tilt the scenario. Hand movements are tracked by the Leap Motion sensor, mapping pitch variations directly to the scenario.

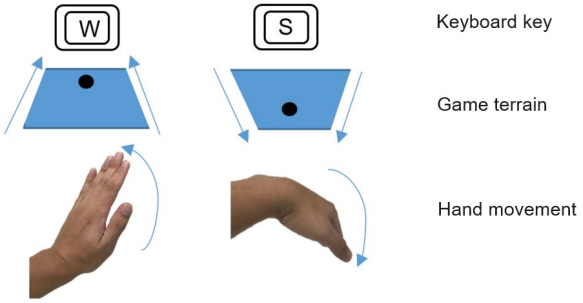

Figure 7. Controlling terrain pitch. Example of the free-hand control scheme, illustrating how hand roll movements are mapped to tilt the scenario. The Leap Motion sensor captures the three-dimensional positioning of the fingers and wrist to enable natural interaction.

(UEQ-S), chosen because the UEQ is one of the most widely used standardized tools for UX evaluation, as discussed by Veriscimo *et al*. [2020]. The second is the PLEX Framework, which offers an alternative approach to UX evaluation, focusing specifically on the playful aspects of the experience. This framework consists of 22 items, each tailored to fit the context of applicability [Lucero *et al*., 2013; Arrasvuori *et al*., 2011]. Although playfulness can manifest in diverse ways, humans are inherently playful by nature, and this quality is closely linked to humor.

3.3 Experiment

An experimental protocol for user testing was developed with the title *Evaluation of user experience in 3D interaction through facial expressions*. It was submitted to the Research Ethics Committee of EACH-USP and obtained its approval under the number 38540620.0.0000.5390. The objective of the test is to find out whether it is possible to evaluate the user experience through automatic facial expression recognition, particularly when interacting in 3D.

We invited students of technical and higher education courses in computing at the following institutions: University of São Paulo (EACH); Federal Institute of São Paulo (São Miguel Paulista); State Technical School (ETEC Guaianases); Faculty of Technology of the State of São Paulo (FATEC), of these, 52 agreed to participate in the experiment.

The majority of participants were aged between 16 and 18 due to technical course students having the greatest participation in the experiment. Table 1 shows the age of the participants separated into four groups: 16 to 18 years old, 19 to 21 years old, 22 to 24 years old, and 25 years or older. The male-to-female ratio is 56% (29) men and 44% (23) women. Another important piece of information is that most participants have already had some contact with three-dimensional games (87%).

Table 1. Number of participants according with their age

Age of participants	Number of participants
16 to 18 years old	29
19 to 21 years old	13
22 to 24 years old	5
25 years or older	5

Before the test began, each user underwent an interview to explain the free and informed consent form (the Download link to the original document in Portuguese `https://drive.google.com/file/d/1_NrUzG84R0K3ms_jvF2PK9TPkW425mAA/view?usp=sharing`, accessed on 26 May 2026). The interview included questions categorizing users by gender, age, education/profession, previous experience with this type of interaction, and a mood profile questionnaire adapted from Viana *et al*. [2001] to assess the user's mood before the experiment. The mood profile questionnaire is presented in Table 2. The same questionnaire was applied at the end of the experiment to identify the mood after the test. Figures 8 and 9 present a boxplot of the answer values provided by users for the mood state profile questionnaires, respectively before and after the test.

Table 2. Mood State Profile Questionnaire adapted from Viana *et al*. [2001]

Tired	○ ○ ○ ○ ○ ○ ○	Rested
Irritated	○ ○ ○ ○ ○ ○ ○	Calm
Sad	○ ○ ○ ○ ○ ○ ○	Cheerful
Bad-Tempered	○ ○ ○ ○ ○ ○ ○	Good-Humored
Unmotivated	○ ○ ○ ○ ○ ○ ○	Excited
Impatient	○ ○ ○ ○ ○ ○ ○	Patient
Anxious	○ ○ ○ ○ ○ ○ ○	Calm

Throughout the test period (before, during, and after the interaction), the user had their face recorded by a low-cost video camera (webcam). Participants who had to wear glasses or other accessories that cover part of the face were preferably not selected, to simplify facial emotion recognition. In addition to the captured face video for later processing, certain game events and task execution times were logged.

Users were introduced to the game developed in Section 3.1, with its three levels. After completing each level, each participant answered the UEQ-S (*User Experience Questionnaire*) [Schrepp *et al*., 2017]. Table 3 presents this questionnaire.

The questionnaire consists of eight questions, each requiring the user to select one option out of seven. Each question presents a pair of opposing words, for example, "uninteresting" and "interesting". Choosing the leftmost option indicates that the user perceives the system as completely uninteresting while selecting the rightmost option (seventh) indicates the system is absolutely interesting. If the participant chooses an option other than the first or seventh, the result is expressed as a percentage. For instance, selecting the fifth option in-

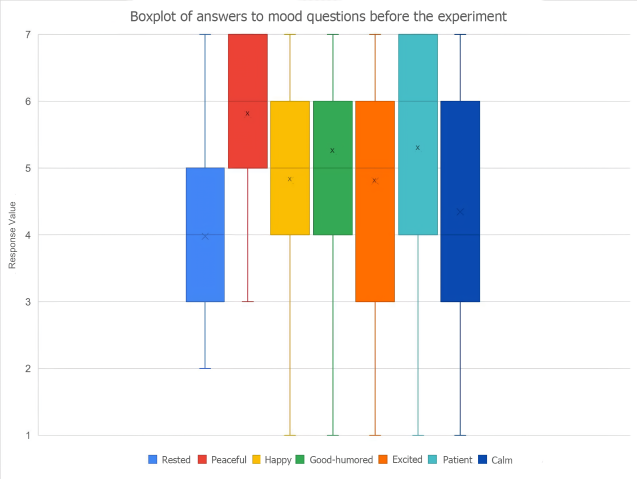

Figure 8. Boxplot of user responses to the mood state profile questionnaire administered before the experiment.

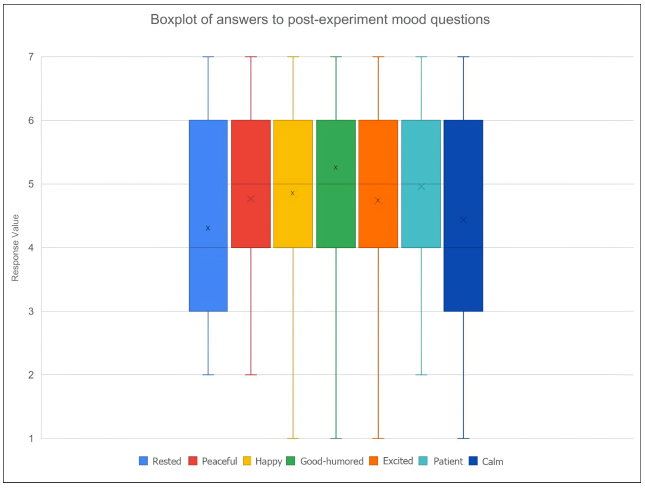

Figure 9. Boxplot of user responses to the mood state profile questionnaire administered after the experiment.

Table 3. UEQ-S Questionnaire

Obstructive	○ ○ ○ ○ ○ ○ ○	Supportive
Complicated	○ ○ ○ ○ ○ ○ ○	Easy
Inefficient	○ ○ ○ ○ ○ ○ ○	Efficient
Confusing	○ ○ ○ ○ ○ ○ ○	Clear
Boring	○ ○ ○ ○ ○ ○ ○	Exciting
Not interesting	○ ○ ○ ○ ○ ○ ○	Interesting
Conventional	○ ○ ○ ○ ○ ○ ○	Inventive
Usual	○ ○ ○ ○ ○ ○ ○	Leading edge

dicates the user finds the system 65% interesting and 35% uninteresting - more interesting than uninteresting.

The terms used in the questionnaire are explained below:

- Impeditive/Facilitating – refers to how much the user felt that the system helped or hindered him in using it;
- Complicated/Easy – is related to the level of difficulty in using the application;
- Inefficient/Efficient - refers to the system's functionalities and performance;
- Confusing/Clear – when using the application, a lot of explanations and training are necessary, whether it is confusing or not;
- Boring/Exciting – refers to the user's emotion, what they felt when using the application between something that makes them bored or sad and something that makes them attracted or happy and what the level is.
- Uninteresting/Interesting - is related to the participant's level of interest;
- Conventional/Original – refers to the system's level of originality;
- Common/Innovative – is related to how innovative the application is.

Each participant had up to three attempts for each level with the traditional input device (keyboard). On the other hand, with the three-dimensional input device (*Leap Motion*), the user had three attempts to go as far as possible in the game. If he lost three times in the first level, the user would not play in the second level and would end the experiment. The user always initiated the interaction using the standard input device, as it was verified in preliminary tests that some users (due to the shape of the hand) could have difficulty recognizing their hand in *Leap Motion*, which could cause some damage to the experience. No maximum or minimum time was set to complete the task. As mentioned earlier, the user filled the UEQ-S user experience questionnaire after each game. This approach was chosen to prevent forgetfulness or confusion between tasks, which could lead to inaccurate responses if the questionnaire were only filled out once at the end of all games.

Figure 10 presents a *boxplot* of the UEQ-S answer values provided by users. As expected, different users generally had different experiences using the same system.

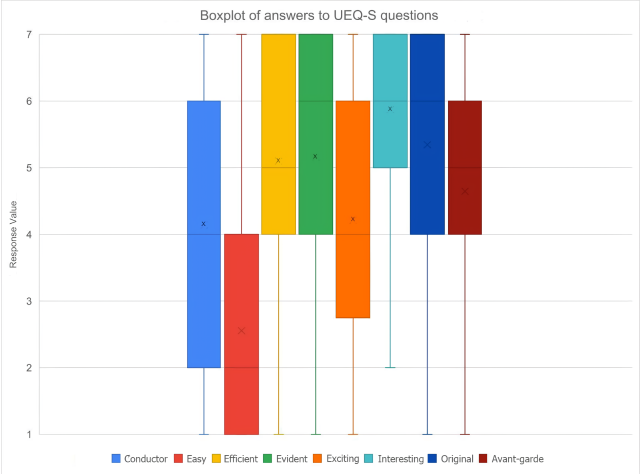

Figure 10. Boxplot illustrating the distribution of UEQ-S answer values provided by users. The variability in responses reflects the diverse user experiences with the same system, as expected.

After completing each level, participants also answered the PLEX *Framework* [Arrasvuori *et al.*, 2011] questionnaire. Table 4 presents the questionnaire.

This questionnaire consists of fourteen questions and in each answer the user must fill in the intensity level from one to five, that is, one option out of five such that one is the lowest intensity and five the highest. Each question is made up of a word or phrase, for example: Captivating, if the user fills in the first option, intensity one, it means that for the user the system is not captivating, whereas if he fills in the closest option, intensity five, represents that for the user the system is absolutely captivating.

Table 4. PLEX questionnaire adapted from Arrasvuori *et al.* [2011]

	1	2	3	4	5
Captivating	○	○	○	○	○
Challenging	○	○	○	○	○
Competitive	○	○	○	○	○
Fun	○	○	○	○	○
Frustrating	○	○	○	○	○
Felt in control	○	○	○	○	○
Felt like completing an important task	○	○	○	○	○
Found something new or unknown	○	○	○	○	○
Expressive	○	○	○	○	○
Relaxing	○	○	○	○	○
Simulates something from real life	○	○	○	○	○
Adrenaline (derived from risk and/or danger)	○	○	○	○	○
Part of a bigger structure	○	○	○	○	○
Experience that needs imagination	○	○	○	○	○

The terms used in the questionnaire are explained below:

- Captivating – refers to how captivating the user felt the system was;
- Challenging – is related to the challenge level;
- Competitive - refers to the level of competitiveness;
- Fun – refers to how much fun the user felt the system was;
- Frustrating – is related to the level of frustration;
- Felt in control - is related to how much the user felt in control of the system;
- If you felt like you were finishing an important task – it refers to the emotion of accomplishing something important for the user, that is, when you finished the task, you had a emotion of satisfaction;
- Found something new or unknown – related to games, refers to the user browsing a scenario/map and finding something new or unknown;
- Expressive – refers to how much the user felt that the system had an expression, in other words, whether the system managed to express some emotion in the user;
- Relaxing – refers to how relaxing the user felt the system was;
- Simulates something from real life – is related to the level of simulation of something from real life;
- Adrenaline - refers to the level of adrenaline that the user felt, derived from risk or danger;
- Is part of a larger structure – refers to the extent to which the user realized that their task is part of something larger;
- Experience that requires imagination – is related to the level of imagination needed to carry out the task.

Figure 11 presents a *boxplot* of PLEX answer values provided by users. As expected, as with UEQ-S, users generally had different experiences using the same system.

3.4 Emotion recognition

The detection of facial emotions began after the end of the experiment, using videos of the participants' faces. As facial emotion detection itself is not part of the contribution of this work, an already trained model was used (a ready-made *API*).

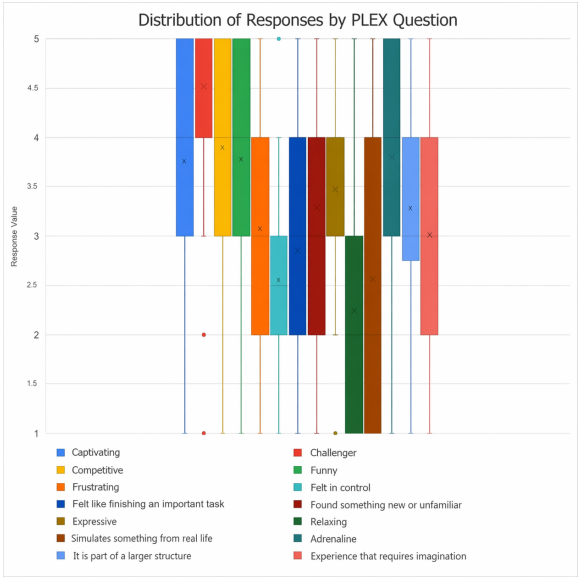

Figure 11. Boxplot illustrating the distribution of PLEX answer values provided by users. Similar to the UEQ-S results, the responses indicate that users generally had varied experiences while interacting with the same system.

According to Deshmukh and Jagtap [2017], the libraries with the highest accuracy in detecting emotions are: *FaceReader* [Den Uyl and Van Kuilenburg, 2005], *Affectiva* [Magdin and Prikler, 2018], and *OpenFace* [Baltrusaitis *et al.*, 2018]. All of them achieve at least 90% accuracy. However, they are proprietary and come with prohibitively high costs, rendering their use in this project unfeasible. An alternative approach is to use open-source libraries. In the Aranha *et al.* [2020]'s study, a comparison is made between some of these libraries, with the best result being *Face-api.js* [Mühler, 2020], which achieves 64% accuracy. This library is capable of detecting the following emotions: happiness, sadness, anger, surprise, fear, disgust, and neutrality. Therefore, the library chosen was *Face-api.js*.

Before the extraction of emotions, it was necessary to edit the videos of the experiments as, in each experiment, one recording was made from start to finish. Thus, we cut the videos into levels, creating a video for each level. The program *Sony Vegas* [Magix, 2023] was used to edit the videos. Each video had a resolution of 1280 *pixels* wide and 720 *pixels* high with the *frames* approximately 30 per second.

The emotion extraction process started by isolating each frame from the experiment recordings. Since the participant's screen and face were recorded, a Python algorithm was developed to extract only the user's face from the video, generating an image for each frame. Figures 12 and 13 present two *frames* from the experiment in which one is winning and the other is losing the match. The participants' faces were blurred to preserve their identity. Just below the user's face, in the right corner of the *frame*, a rounded object with a green or red color can be seen in the image, representing the gain of a point or victory of the match and the defeat of the level, respectively. This device is a visual *log* to facilitate synchronization in extracting emotion with the knowledge that the user won, lost, or scored in the match.

Then, *Face-api.js* is used in each *frame*, which provides recognition of emotions in seven emotions, namely: anger, disgust, fear, happiness, sadness, surprise, and neutral. *Face-*

api.js provides the percentage of each sentiment for each *frame*, thus generating a *JSON* file with the result of all *frames* in a single file . With the percentages of emotions, we include other attributes in the *JSON* file: the name of the *frame* ("image"), the chronological order ("seq"), whether it corresponds to a punctuation moment ("scored"), if it corresponds to a moment of victory ("won") or if it corresponds to a moment of defeat ("lost"), in Figure 14 the result of a *frame*.

Figure 12. Screenshot from the experiment showing the participant winning the match. The green rounded object in the bottom-right corner indicates the gain of a point or victory in the match. The participant's face has been blurred to protect their identity.

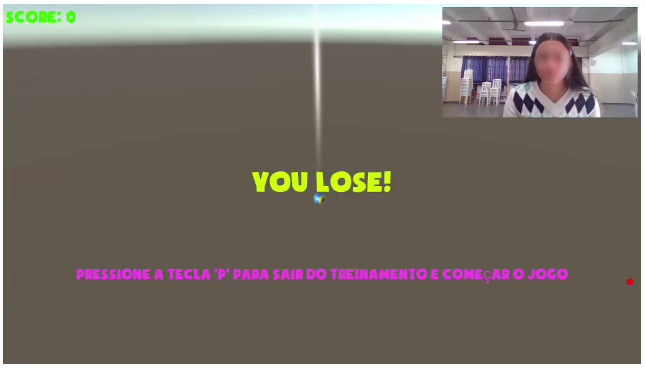

Figure 13. Screenshot from the experiment showing the participant losing the match. The red rounded object in the bottom-right corner signifies the defeat in the level. The participant's face has been blurred to preserve their privacy.

3.5 Dataset

The sequence of emotions was converted as a string. The most intense emotion of each frame was mapped to a letter (excluding the neutral emotion because preliminary tests including it had worse results), for example, the most intense emotion in Figure 14 is happiness and was mapped to the letter 'D'. The mapping of all emotions into letters is represented by: anger (A); disgust (B); fear (C); happiness (D); neutral (E); sadness (F); surprise (G).

All emotions extracted from a given level are represented by a vector of characters respecting chronological order. For example, fear, fear, sadness, and surprise are represented as CCFG. This character vector is one of the attributes of the dataset called emotions.

In addition to emotions as a string of letters, the dataset has the overall average of each of the seven emotions generating seven numeric attributes with a range from 0 to 1.0. The other attributes that make up the dataset are:


```
{
   "angry":  1.8174779015112108e-8,
   "disgusted":  2.7608818342628183e-8,
   "fearful":  5.7865485736385835e-9,
   "happy":  0.9998815059661865,
   "neutral":  0.000047970352170523256,
   "sad":  0.00006902758468640968,
   "surprised":  0.0000014412647715289,
   "seq":  "101",
   "image":  "101-frame.jpg",
   "lost":  false,
   "won":  false,
   "scored":  true
}
```


Figure 14. JSON with an example of emotion detection from a frame.

- **User identification (PersonId)** - is a unique identifier per user formed by a letter corresponding to the location where the test was performed (USP (U), Fatec (F), IFSP (I), and ETEC (E)) concatenated with a sequential number. For example, U1, F3, and E14.
- **Input device (Type)** – represents the device used in the interaction [Keyboard or Leap]
- **Game level (Level)** – corresponds to the game level number, it is represented by an integer from 1 to 3.
- **Final Status (FinalStatus)** – is the status that represents the final result of the match, with number one meaning victory and number zero meaning defeat.
- **Plex Framework Answers (PP1 to PP14)** - Plex Framework answers are represented numerically in a range from 1 to 5, each question represents an attribute of the dataset, adding 14 new characteristics, such as: Captivating (PP1); Challenging (PP2); Competitive (PP3); Fun (PP4); Frustrating (PP5); Felt in control (PP6); Felt like finishing an important task (PP7); Found something new or unknown (PP8); Expressive (PP9); Relaxing (PP10); Simulates something from real life (PP11); Adrenaline (PP12); It is part of a larger structure (PP13); Experience that requires imagination (PP14).
- **Questionnaire Answers *UEQ-S* (PUEQ1 to PUEQ8)** - these are the questionnaire answers *UEQ-S* represented numerically in a range from 1 to 7, each question represents an attribute of the dataset, adding eight new characteristics, such as: Impediment or facilitator (PUEQ1); Complicated or easy (PUEQ2); Inefficient or efficient (PUEQ3); Confused or clear (PUEQ4); Boring or exciting (PUEQ5); Uninteresting or Interesting (PUEQ6); Conventional or original (PUEQ7); Common or innovative (PUEQ8).
- **Answers to the Questions of the Mood State Profile Questionnaire (PHA1 to PHA7 and PHD1 to PHD7)** - these are the answers to the mood state profile questionnaire, applied before and after the experiment, so we will have two answers for this questionnaire, the first before the experiment with the prefix PHA and another after the experiment with the prefix PHD, the answers are represented numerically in a range from 1 to 7, each question represents an attribute in the data set, adding 14 new features, such as: Tired or rested (PHA1 and PHD1); Irritated or calm (PHA2 and PHD2); Sad or happy (PHA3

and PHD3); Bad-tempered or good-natured (PHA4 and PHD4); Discouraged or excited (PHA5 and PHD5); Impatient or patient (PHA6 and PHD6); Anxious or calm (PHA7 and PHD7).

This dataset described with 48 attributes is called in this project the base dataset, as variations will be made from it to form other data sets, including the class (label) that will be included according to the classification object. If it is the UEQ-S as a classification object, the dataset will use the answers to the UEQ-S questions as a label, creating a variation for each answer in the questionnaire, generating eight datasets, such that each set has as its class the answer to a specific question in the UEQ-S. In all these sets, the other responses to UEQ-S questions will be removed (we will not use the answer to some UEQ-S questions to infer the answer to other of its questions). In the same way, when the classification object is *Plex Framework*. Table 5 presents an instance of the base dataset.

The base dataset has 218 instances. 62 instances are noted for *Leap Motion* and 156 instances for Keyboard. Regarding the level of the game executed, there are 104 instances of Level 1, 59 instances of Level 2, and 55 instances of Level 3. Regarding the result of the match, there are 161 instances of defeat and 57 instances of victory. The values of the other characteristics are covered in the previous sections. The original document, in Portuguese, may be downloaded at `https://drive.google.com/file/d/1EP5a1sRHTL75fucTgzL-LZRyQcj2yIy2/view?usp=sharing`. Accessed on 26 May 2026.

3.6 Preprocessing

Before pre-processing, variations of the base data set were created, such as, with only the attributes related to the mood profile questionnaire applied before the experiment (MoodBefore), with only the attributes related to the mood profile questionnaire applied after the experiment (MoodAfter); with attributes related to the mood profile questionnaire applied before and after the experiment (MoodBeforeANDAfter); with the emotions attribute considering only the initial 33.3% (*Start*); with the emotions attribute considering only the final 33.3% (*End*); with the emotions attribute considering only the central 33.3% (*Middle*); with the attributes of emotions completely; with the general average sentiment attributes (*Average*); and only the attributes related to the *UEQ-S* or *Plex Framework* questionnaire depending on which classification object is (*Plex* or *UEQ*). With these new sets generated, other sets were created by combining the previous sets.

The emotions attribute, which correspond to a sequence of letters, was represented using *n-grams* by counting and *TF-IDF*, the maximum size of the n-gram to be considered was seven (heptagram). Two groups of datasets were generated, one by counting and the other using TF-IDF.

In each group, three options related to dimensionality reduction were used. *PCA* - Principal Component Analysis, which corresponds to a projection of the attributes into a lower dimensionality, *KBEST*, which is a selector of the "most important" attributes, and without reduction. In *PCA*, the number of ten dimensions used by the dimensionality reducer was defined. For *KBEST*, thirty was the number of defined attributes to be selected by the *KBEST* attribute selector.

3.7 Automatic assessment and performance analysis

Due to the novelty of the approach proposed in this study, no indications were found in the literature regarding the best artificial intelligence algorithms for inferring user experience based on facial expressions. Therefore, this work investigates the use of various algorithms of different types.

After preprocessing and generating dataset variations, we used the following classifiers/regressions for each variation in order to predict each answer of the *UEQ* and *PLEX* questionnaires:

- *Dummy* – standard for comparison (uses the response mode)
- RF - *Random Forest*
- *SVM - Support Vector Machine - kernel* linear
- *SVM Poly - Support Vector Machine - kernel* polynomial
- *SVR - Support Vector Regression - kernel* linear
- *SVR Poly - Support Vector Regression - kernel* polynomial
- GBR – *Gradient Boosting Regressor*
- EN – *Elastic Net*
- *KernelR – Kernel Ridge*
- *RegLog* - Logistic Regression
- *RegLin* - Linear Regression
- *MLP - Multilayer Perceptron*
- *KNN - K-nearest neighbors*
- DT - *Decision Tree*

We defined a maximum 10,000 iterations for the classifiers and considered 3 neighbors for *KNN* algorithm.

The models were evaluated with cross-validation, using ten subsets (*folds*) for validation (10-fold cross-validation). The chosen evaluation metric was mean absolute error. *t-tests* were also run to check whether different forms of representation (different pre-processing/features) produced statistically different results.

These configurations were applied to solve both problems: inferring users' experience from facial expressions and inferring users' experience by combining facial expressions with responses from another UX questionnaire.

4 Results

This section presents the results, organized into four subsections. The first and second subsections present the predictions for the *UEQ-S* and *Plex Framework* questionnaires, focusing on emotion recognition 4.1 and based on responses from another questionnaire 4.2. A summary of these results is presented in Tables 6 and 7. The third subsection discusses the correlations between the questionnaires. Finally, the fourth subsection includes tables showing the p-values of all predictions considering the attributes used for predicting the values of both questionnaires.

4.1 Predictions focusing on emotion recognition

Although we had logged data about game level, match status and input device, these were not used in the predictive model, which attempted to infer User Experience by focusing mostly on the emotions extracted during the interaction.

Table 5. An instance of the base dataset

PersonId	E10
Type	Keyboard
Level	1
FinalStatus	1
PP1 to PP14	[3, 5, 2, 3, 5, 2, 2, 1, 2, 4, 2, 3, 4, 1]
PUEQ1 to PUEQ8	[6, 2, 5, 7, 4, 5, 3, 2]
PHA1 to PHA7	[4, 3, 2, 3, 3, 6, 1]
PHD1 to PHD7	[2, 2, 2, 4, 3, 5, 3]
Anger	2.7334392376e-06
Disgust	1.36935855207062e-09
Fear	9.55444740683634e-09
Happiness	4.01130631914284e-07
Neutral	0.99980726316975
Sadness	3.23366928950876e-06
Surprise	0.000387848338800073
Emotions	AFGAAGFGAAGAAAAAAFAA...GGGFGGGGFFFGGG

In all predictions for *UEQ-S*, the following attributes were used : questionnaire answers *UEQ-S* that correspond to the attributes to be predicted, whether or not the following attributes are included (generating several variations): Emotions, EmotionsBeginning (emotions extracted in the first third of each interaction), EmotionsMiddle (emotions extracted during the second third of each interaction), EmotionsEnd (emotions extracted from the third third of each interaction), average of emotions (average of the scores assigned to each sentiment by the emotion detection algorithms), the responses to the mood questionnaire before, the responses to the mood questionnaire after. The concatenation of the attributes emotions, average emotions, the answers to the mood questionnaire before and the responses to the mood questionnaire after will be described by the word "all".

In all predictions for *PLEX*, the following attributes were used: *PLEX* responses that correspond to the attributes to be predicted, whether or not the following attributes were included (generating several variations): Emotions, EmotionsBeggining, EmotionsMiddle, EmotionsEnd, the average of emotions, the mood quiz answers before, mood quiz answers after. The concatenation of emotions, average emotions, the responses to the mood questionnaire before and the responses to the mood questionnaire after will be described by the word "all".

The emotions, converted as arrays of letters, were used by the classifiers or regressors as character n-grams. To represent which values of n were used, the number corresponding to n will be presented, together with the respective test set. When a range has been used, two numbers will be displayed. For example, 1-2 indicates the combined use of unigrams and bigrams for the entire *string* of emotions.

The prediction results are divided into two subsections: one presenting the results of predictions for *UEQ-S*, and another the results for the predictions for the *PLEX framework*. Additionally, a summary of the best results is provided in subsection 4.1.3.

4.1.1 Predictions for the UEQ-S questionnaire

Predictions for the Impediment/Facilitator question (P1): The best predictive result occurred using the EmotionsMiddle with pentagrams (n=5), obtaining an error reduction of 31.14% compared to *dummy*, which corresponds to an approximate average error of 1.67, being close to the best overall result. *KBEST* was used for attribute reduction, counting to assign weight to n-grams, and the regressor used was *Gradient Boosting Regressor* (GBR).

Predictions for the Complicated/Easy question (P2): The best predictive result achieved used the average of emotions with humor before and after, obtaining an error reduction of 15.47% compared to the *dummy*, which corresponds to an average error of 1.31. In this case, no attribute reduction strategy was used, and the model with the best performance was *Random Forest*.

Using only emotions as an attribute, the best predictive result achieved used the average of emotions, achieving an error reduction of 13.09% compared to the *dummy*, which corresponds to an approximate average error of 1.35. No attribute reduction strategy was used, and the model with the best performance was *Support Vector Regression* - with polynomial *kernel* (SVR – Poly).

Predictions for the Inefficient/Efficient question (P3): The best predictive result achieved used the set of attributes "all" with the combined use of unigrams and pentagrams, obtaining an error reduction of 32.39% in relation to *dummy*, which corresponds to the approximate average error of 1.29. Among the results obtained without using *PLEX* responses, it was the best error reduction compared to *baseline (dummy)*. It is worth noting that the user's perception of the system's efficiency or inefficiency can, to a certain degree, be inferred from their facial expressions. For this result, the *KBEST* strategy was used for attribute reduction, counting to assign weight to n-grams, and the model with the best performance was GBR.

Using only emotions as an attribute, the best predictive result achieved used the set of EmotionsEnd attributes with the use of unigrams (n=1), achieving an error reduction of 26.56% in relation to *dummy*, which corresponds to an average error of 1.40. No feature reduction strategy was used, TFIDF was used to assign weight to the n-grams, and the best-performing model was *Support Vector Regression* (SVR).

Predictions for the Confused/Clear question (P4): The best predictive result achieved used the set of EmotionsMiddle attributes with the use of bigrams (n=2), obtaining an error reduction of 15.82% in relation to *dummy*, which corresponds to the approximate average error of 1.55, being close to the best overall result. The *PCA* strategy was used for attribute reduction, TFIDF was used to assign weight to n-grams, and the model with the best performance was SVR.

Predictions for the question Boring/Exciting (P5): The best predictive result achieved used the set of attributes "all" with the use of unigrams (n=1), obtaining an error reduction of 15.56% in relation to *dummy*, which corresponds to the approximate average error of 1.43. The *KBEST* strategy was used for attribute reduction, TFIDF was used to assign weight to n-grams, and the model with the best performance was GBR.

Using only emotions as attributes, the best predictive result achieved using the set of attributes EmotionsBeginning with the use of bigrams (n=2), obtaining an error reduction of 3.36% in relation to *dummy* which corresponds to an average error of 1.64. This small reduction indicates that, for the experiments carried out, the extracted emotions were not able to significantly help in predicting this response. The *PCA* strategy was used for attribute reduction, TFIDF was used to assign weight to n-grams, and the model with the best performance was GBR.

Predictions for the Uninteresting/Interesting question (P6): The best predictive result achieved used the set of MoodAfter attributes, obtaining an error reduction of 28.04% compared to the *dummy*, which corresponds to the approximate average error of 0.81. No attribute reduction strategy was used, i.e., all attributes were used as input to the regressor, and the model with the best performance was SVR.

Using only emotions as an attribute, the best predictive result achieved used the average of emotions, achieving an error reduction of 21.92% compared to the *dummy*, which corresponds to an approximate average error of 0.88. No feature reduction strategy was used, and the model with the best performance was SVR – Poly.

Predictions for the Conventional/Original question (P7): The best predictive result achieved used the set of MoodAfter attributes, obtaining an error reduction of 28.74% compared to the *dummy*, which corresponds to the approximate average error of 1.18. No attribute reduction strategy was used, i.e., all attributes were passed to the regressor, and the model with the best performance was SVR.

The best predictive result achieved using only emotions as attributes considered the set of Emotions attributes with the use of heptagrams (n=7), obtaining an error reduction of 24.88% compared to the *dummy*, which corresponds to an average error of 1.24. In this case, the strategy *KBEST* was used to reduce attributes, TFIDF was used to assign weight to n-grams, and the model with the best performance was SVR Poly.

Predictions for the Common/Innovative question (P8): The best predictive result achieved used the set of attributes "all" with the use of unigrams (n=1), obtaining an error reduction of 20.82% compared to the *dummy*, which corresponds to the approximate average error of 1.30. No attribute reduction strategy was used, counting was employed to assign weight to the n-grams, and the model with the best performance was the *Support Vector Machine* (SVM).

The best predictive result considering only emotions as an attribute was achieved using the set of emotions attributes using heptagrams (n=7), achieving an error reduction of 15.76% in relation to *dummy*, which corresponds to an average error of 1.38. The *KBEST* strategy was used to reduce attributes, TFIDF was employed to assign weight to n-grams, and the model with the best performance was *Support Vector Machine - kernel* polynomial (SVM – Poly).

4.1.2 Predictions for the Plex Framework

Predictions for the Captivating question (P1): The best predictive result was achieved using the set of attributes "all" with the use of unigrams (n=1), obtaining an error reduction of 8.17 % compared to the *dummy* which corresponds to the approximate average error of 0.72, being close to the best overall result. *KBEST* was used to reduce attributes, TFIDF was used to assign weight to n-grams, and the model used was *Random Forest*.

The best result considering only emotions-related attributes was achieved using the *dummy* model. I.e., the use of emotions was not able to improve the prediction for this question.

Predictions for the Challenging question (P2): The best predictive result achieved used the set of EmotionsEnd attributes with the use of heptagrams (n=7), obtaining an error reduction of 2.07% compared to *dummy*, which corresponds to an approximate average error of 0.47. Among the results obtained without using the *UEQ-S* questionnaire responses as attributes, this was the lowest error value. No attribute reduction strategy was used, TFIDF was employed to assign weight to the n-grams, and the model used was *Support Vector Machine - Polynomial Kernel*. It is noteworthy that the *baseline* error value was already quite low.

Predictions for the Competitive question (P3): The best predictive result achieved used the set of attributes EmotionsMiddle with the use of trigrams (n=3), obtaining an error reduction of 13.26% compared to the *dummy*, which corresponds to the approximate average error of 0.96. *KBEST* was used to reduce attributes, TFIDF was used to assign weight to n-grams, and the model used was *Random Forest*.

Predictions for the question Fun (P4): The best predictive result achieved using the set of attributes "all" with the combined use of unigrams and tetragrams (n varying from 1 to 4), obtaining a reduction of error by 0.66% in relation to *dummy* which corresponds to an approximate average error of 0.75. No attribute reduction strategy was used, counting was employed to assign weight to the n-grams, and the model used was *Support Vector Machine – with polynomial kernel*.

The best result achieved considering only emotions attributes used the *dummy* model. It means the analysis of emotions extracted from facial expressions did not contribute to inferring the user's response to this question.

Predictions for the Frustrating question (P5): The best predictive result achieved used the set of average sentiment attributes, obtaining an error reduction of 44.38% compared to *dummy*, which corresponds to the approximate average error of 1.23. This was among the results obtained without using the *UEQ-S* questionnaire responses as attributes the

greatest error reduction compared to *dummy*. No attribute reduction strategy was used, i.e., all attributes were used, and the model used was *Support Vector Machine*.

Predictions for the question Did you feel in control (P6): The best predictive result achieved employed the set of attributes "all" using the concatenation from unigrams to heptagrams (n=1-7), obtaining a reduction in error by 13.82% compared to *dummy*, which corresponds to an average error of approximately 1.00. *PCA* was used to reduce attributes, counting to assign weight to n-grams, and the model used was *Elastic Net*.

Using only emotions as an attribute, the best predictive result achieved used the set of attributes EmotionsBeginning with the use of bigrams (n=2), achieving an error reduction of 13.82% compared to *dummy*, which corresponds with an average error of approximately 1.00. The *PCA* strategy was used to reduce attributes, counting to assign weight to n-grams, and the model with the best performance was *Elastic Net*.

Predictions for the question Did you feel like you were finishing an important task (P7): The best predictive result used the set of attributes "all" with the use of unigrams, obtaining an error reduction of 41.91% compared to *dummy*, which corresponds to an approximate average error of 1.09. *KBEST* was used to reduce attributes, TFIDF was used to assign weight to n-grams, and the model used was *Gradient Boosting Regressor*.

Using only emotions as an attribute, the best predictive result achieved used the set of attributes EmotionsBeginning with the use of bigrams (n=2), achieving an error reduction of 38.24% compared to *dummy*, which corresponds with an approximate average error of 1.16. The *KBEST* strategy was used for attribute reduction, counting to assign weight to n-grams, and the best-performing model was GBR.

Predictions for the question Found something new or unknown (P8): The best predictive result achieved used the set of attributes mood before, obtaining an error reduction of 5.81% compared to *dummy*, which corresponds with an approximate average error of 1.05. In this case, no technique was used to reduce attributes, and the model used was *Gradient Boosting Regressor*.

Using only emotions as attributes, the best predictive result utilized the set of EmotionsEnd attributes with the use of heptagrams (n=7), obtaining an error reduction of 4.20% compared to *dummy*, which corresponds with an approximate average error of 1.07. In this case, the *PCA* strategy was used to reduce attributes, counting to assign weight to n-grams, and the model with the best performance was SVR.

Predictions for the Expressive question (P9): The best predictive result used the set of mood attributes and the average of emotions, obtaining an error reduction of 18.16% compared to *dummy* , which corresponds to an approximate average error of 0.83. No attribute reduction strategy was used, and the model used was *Multilayer Perceptron*.

The best predictive result considering only emotions as attributes used the average of emotions, achieving an error reduction of 8.30% compared to *dummy*, which corresponds to an average error of 0.94. In this case, no attribute reduction strategy was used, and the model with the best performance was SVR Poly.

Predictions for the Relaxing question (P10): The best predictive result used the attributes mood before, mood after, and the average of emotions, obtaining an error reduction of 21.20% compared to *dummy*, which corresponds to an approximate average error of 0.98. No attribute reduction strategy was used, and the model used was GBR.

The best predictive result considering only emotions as attributes used the set of attributes EmotionsBeggining with the use of bigrams (n=2), achieving an error reduction of 16.95% compared to *dummy*, which corresponds to an average error of 1.03. The *PCA* strategy was used to reduce attributes, counting to assign weight to n-grams, and the model with the best performance was *Elastic Net*.

Predictions for the question Simulates something from real life (P11): The best predictive result used the "all" attribute with the combined use of unigrams and hexagrams (n=6), obtaining an error reduction of 25.72% compared to *baseline (dummy)*, which corresponds to the approximate average error of 1.16. In this case, the *PCA* strategy was used to reduce attributes, TFIDF was used to assign weight to n-grams, and the model with the best performance was GBR.

The best predictive result considering only the use of emotions as attributes used the set of EmotionsEnd attributes with the use of tetragrams (n=4), achieving an error reduction of 24.50% compared to *dummy*, which corresponds to the approximate average error of 1.18. The *PCA* strategy was used to reduce attributes, counting to assign weight to n-grams, and the model with the best performance was Linear Regression.

Predictions for the question Adrenaline (derived from risk and/or danger) (P12): The best predictive result used EmotionsEnd as an attribute with the use of pentagrams (n=5), obtaining an error reduction of 5.80% compared to *baseline (dummy)*, which corresponds to an approximate average error of 0 .86, that is, the error is less than 1 for more or less of the answer the user responded. In this case, the strategy *KBEST* was used to reduce attributes, TFIDF was used to assign weight to n-grams and the model with the best performance was SVR Poly.

Without using only the EmotionsEnd attribute, the best predictive result used the average of emotions as an attribute, obtaining an error reduction of 4.27% compared to *dummy*, which corresponds to an approximate average error of 0.87. No attribute reduction strategy was used, and the model with the best performance was SVR.

Predictions for the question Is it part of a larger structure (P13): The best predictive result used the attributes mood before, mood after, and the average of emotions, obtaining an error reduction of 1.37% compared to *dummy*, which corresponds to an approximate average error of 0.93. No attribute reduction strategy was used, and the model used was *Support Vector Machine*(SVM).

Predictions for the question Experience that requires imagination (P14): The best predictive result used the set of attributes "all" with the combined use of unigrams and bigrams (n=2), obtaining an error reduction of 22 .45% compared to *dummy*, which corresponds to an approximate average error of 1.16, close to the best overall result. The *KBEST* strategy was used for attribute reduction, counting to assign weight to n-grams, and the best-performing model was the GBR.

The best predictive result considering only the use of

emotions as attributes used the set of attributes EmotionsBeginning with the use of bigrams (n=2), achieving an error reduction of 22.45% compared to *dummy*, which corresponds to the approximate average error of 1.16. The *KBEST* strategy was used for attribute reduction, counting to assign weight to n-grams, and the best-performing model was the GBR.

4.1.3 Synthesis

In summary, using only emotions as an attribute for predictions for *UEQ-S*, we obtained an approximate average error of 1.39. This value is considered satisfactory since it is common for users to have doubts when choosing between two answer options while answering the question, as empirically observed during the application of the questionnaires in our experiments. In the predictions for the *Plex Framework*, using only emotions as attributes, we obtained an approximate average error of 0.97, that is, the error is, on average, less than 1 for more or less of the users' responses. This value is considered satisfactory, as users often face uncertainty when choosing between two response alternatives when answering the question.

The only question in which using the *UEQ-S* questionnaire responses as an attribute did not result in the best result was Plex12, in this case, the best result obtained was using emotions (the EmotionsEnd attribute with the use of pentagrams) as an attribute.

Tables 6 and 8 present a summary of the best results. In these tables, there are columns defining which attributes were used, namely, emotions only, emotions and mood (mood state questionnaire) and better global (which can be one of the previous columns or a combination of two or more columns as attributes).

4.2 Predictions based on responses from another questionnaire

In the same way as for predictions using emotion recognition the logged data about game level, match status and input device, these were not used in the predictive model, which attempted to infer User Experience by focusing mostly on the emotions extracted during the interaction.

In all predictions for *UEQ-S*, the following attributes were used: questionnaire answers *UEQ-S* that correspond to the attributes to be predicted, whether or not the following attributes are included (generating several variations): Emotions, EmotionsBeggining, EmotionsMiddle, EmotionsEnd, average of emotions (average of the scores assigned to each sentiment by the emotion detection algorithms), the responses to the mood questionnaire before, the responses to the mood questionnaire after and the *PLEX framework* responses. The concatenation of the attributes emotions, average emotions, the answers to the mood questionnaire before and the responses to the mood questionnaire after will be described by the word "all". It is worth highlighting we are also using the responses from one questionnaire to predict another.

In all predictions for *PLEX*, the following attributes were used: *PLEX* responses that correspond to the attributes to be predicted, whether or not the following attributes were included (generating several variations): Emotions, EmotionsBeggining, EmotionsMiddle, EmotionsEnd, the average of emotions, the mood quiz answers before, mood quiz answers after and *UEQ-S* questionnaire answers. The concatenation of emotions, average emotions, the responses to the mood questionnaire before and the responses to the mood questionnaire after will be described by the word "all".

The emotions, converted as arrays of letters, were used by the classifiers or regressors as character n-grams. To represent which values of n were used, the number corresponding to n will be presented, together with the respective test set. When a range has been used, two numbers will be displayed. For example, 1-2 indicates the combined use of unigrams and bigrams for the entire *string* of emotions.

The prediction results are divided into two subsections: one presenting the results the predictions for *UEQ-S*, and another the results the predictions for the *PLEX framework*. Additionally, a summary of the best results is provided in subsection 4.2.3.

4.2.1 Predictions for the UEQ-S questionnaire

Predictions for the Impediment/Facilitator question (P1): The best predictive result achieved used the *PLEX framework* responses as attributes, obtaining an error reduction of 37.30% compared to the *baseline (dummy)* which corresponds to an approximate average error of 1.5, that is, the error is between 1.5 and more or less of the answer the user responded. This value is considered satisfactory, as it is common for the user to have doubts between two answer options when answering the question. In this case, no attribute reduction strategy was used, meaning all attributes were passed to the regressor, and the model with the best performance was Linear Regression (RegLin).

Predictions for the Complicated/Easy question (P2): The best predictive result for the second question used as attributes the answers from the *PLEX framework* added the set of attributes "everything" with the combined use of unigrams and bigrams, obtaining an error reduction of 29.65% compared to the *baseline (dummy)* which corresponds to the approximate average error of 1.10, that is, the error is very close to 1 for more or less the answer the user responded. In this case, the *KBEST* strategy was used to reduce attributes, counting to assign weight to n-grams, and the model with the best performance was *Random Forest*.

Predictions for the Inefficient/Efficient question (P3): The best predictive result achieved used the answers from the *PLEX framework* as attributes, adding the set of attributes "everything" with the use of unigrams, obtaining an error reduction of 37.69% in relation to *baseline (dummy)* which corresponds to the approximate average error of 1.19, that is, the error is close to 1 plus or minus the answer the user responded. Among the results obtained, this was the greatest reduction in error compared to *baseline (dummy)*. In this case, the *KBEST* strategy was used for attribute reduction, counting to assign weight to n-grams, and the model with the best performance was GBR.

Predictions for the Confused/Clear question (P4): The best predictive result for the fourth question used the answers from the *PLEX framework* as attributes, obtaining an error reduction of 18.64% in relation to *baseline (dummy)*, which corresponds to an approximate average error of 1 .5, that is, the error is between 1.5 and more or less of the answer the user responded. It was among the results obtained with the

Table 6. Summary of best results for UEQ-S prediction

Question	Just Emotions	Emotions and Humor	Best Overall
Obstructive/Supportive	1.665	1.674	1.516
Complicated/Easy	1.354	1.317	1.096
Inefficient/Efficient	1.399	1.288	1.187
Confusing/Clear	1.549	1.567	1.497
Boring/Exciting	1.640	1.433	1.215
Not interesting/Interesting	0.880	0.824	0.717
Conventional/Inventive	1.244	1.249	1.115
Usual/Leading edge	1.384	1.301	1.284
Overall Average	**1.389**	**1.332**	**1.203**

Table 7. Summary of best results for PLEX prediction

Question	Just Emotions	Emotions and Humor	Best Overall
Captivating	0.783	0.719	0.659
Challenging	0.473	0.483	0.410
Competitive	0.962	0.988	0.875
Fun	0.760	0.755	0.685
Frustrating	1.228	1.249	1.116
Felt in control	1.004	1.004	0.825
Felt like completing an important task	1.161	1.092	1.046
Found something new or unknown	1.071	1.116	0.957
Expressive	0.939	0.838	0.783
Relaxing	1.034	0.981	0.895
Simulates something from real life	1.180	1.161	1.161
Adrenaline	0.861	0.899	0.861
Part of a bigger structure	0.942	0.933	0.874
Experience that needs imagination	1.164	1.164	1.094
Overall Average	**0.969**	**0.956**	**0.874**

smallest error reduction compared to *baseline (dummy)* using the *PLEX framework* responses as attributes. This result was achieved with no attribute reduction strategy employed, i.e., all attributes were used by the regressor, and the model with the best performance was Linear Regression (RegLin).

Predictions for the question Boring/Exciting (P5): The best predictive result for the fifth question used as attributes the answers from the *PLEX framework* added the set of attributes "everything" with the combined use of unigrams and tetragrams(n=4), obtaining a reduction in error in 28.40% in relation to *baseline (dummy)* which corresponds to the approximate average error of 1.21, that is, the error is very close to 1 for more or less the answer the user responded. In this case, the *PCA* strategy was used to reduce attributes, counting to assign weight to n-grams, and the model with the best performance was *Support Vector Machine*(SVM).

Predictions for the Uninteresting/Interesting question (P6): The best predictive result achieved used the answers from the *PLEX framework* as attributes, obtaining an error reduction of 36.38% compared to the *baseline (dummy)* which corresponds to an approximate average error of 0.71, that is, the error is, on average, less than 1 for more or less of the answer the user responded. This result was the lowest error value among the results obtained. In this case, no attribute reduction strategy was used, i.e., all attributes were passed to the regressor, and the model with the best performance was GBR.

Predictions for the Conventional/Original question (P7): The best predictive result for the seventh question used as attributes the answers from the *PLEX framework* added the set of attributes "everything" with the combined use of unigrams to tetragrams (n varying from 1 to 4), obtaining a reduction of the error by 32.67% compared to the *baseline (dummy)* which corresponds to the approximate average error of 1.11, that is, the error is very close to 1 for more or less the answer the user answered. No feature reduction strategy was used, counting was used to assign weight to the n-grams, and the model with the best performance was SVR.

Predictions for the Common/Innovative question (P8): The best predictive result achieved used as attributes the answers from the *PLEX framework* added the set of attributes "everything" with the combined use of unigrams and tetragrams(n=4), obtaining an error reduction of 21.85 % in relation to *baseline (dummy)* which corresponds to the approximate average error of 1.28. No attribute reduction strategy was used, counting was employed to assign weight to the n-grams, and the model with the best performance was SVR Poly.

4.2.2 Predictions for the Plex Framework

Predictions for the Captivating question (P1): The best predictive result achieved used as attributes the answers to the *UEQ-S* questionnaire along with the average emotions and mood before and after, obtaining a reduction in error by 15.84% compared to the *baseline (dummy)* which corresponds to the approximate average error of 0.66, that is, the error is

less than 1 for more or less of the answer the user responded. This results is considered satisfactory, as it is common for the user to have doubts between two answer options when answering the question. In this case, no attribute reduction strategy was used, i.e., all attributes were used, and the model with the best performance was *Random Forest*.

Predictions for the Challenging question (P2): The best predictive result for the second question used the *UEQ-S* questionnaire responses as attributes, obtaining an error reduction of 15.11% compared to the *baseline (dummy)* which corresponds to the approximate average error of 0.41, that is, the error is less than 0.5 for more or less of the answer the user responded. This result was the lowest error value among the results obtained. In this case, no attribute reduction strategy was used, i.e., all attributes were passed to the regressor, and the model with the best performance was Logistic Regression (RegLog).

Predictions for the Competitive question (P3): The best predictive result achieved used as attributes the answers to the questionnaire *UEQ-S* with humor before and after, obtaining an error reduction of 21.10% compared to the *baseline (dummy)* which corresponds to the error approximate average of 0.87, that is, the error is less than 1 for more or less of the answer the user responded. In this case, no attribute reduction strategy was used, i.e., all attributes were used, and the model with the best performance was *Random Forest*.

Predictions for the question Fun (P4): The best predictive result for the fourth question used the *UEQ-S* questionnaire responses as attributes, obtaining an error reduction of 9.87% in relation to *baseline (dummy)* which corresponds to the approximate average error of 0.68, that is, the error is less than 1 for more or less of the answer the user responded. This result is considered satisfactory. In this case, no attribute reduction strategy was used, i.e., all attributes were used, and the model with the best performance was *Kernel Ridge*.

Predictions for the Frustrating question (P5): The best predictive result achieved used as attributes the answers to the questionnaire *UEQ-S* together with the attribute "all" which combined the use of unigrams and heptagrams (n=7), obtaining an error reduction of 49.46% compared to *baseline (dummy)*, which corresponds to the approximate average error of 1.17, i.e., the error is, on average, 1.17 plus or minus the answer the user responded. This result was, among the obtained, the greatest error reduction compared to *dummy*. *PCA* was used for attribute reduction, counting to assign weight to n-grams, and the model used was *Support Vector Machine – polynomial kernel*.

Predictions for the question Did you feel in control (P6): The best predictive result for the sixth question used as attributes the answers to the questionnaire *UEQ-S* together with the attribute "everything" with combined use of unigrams and heptagrams (n=7), obtaining an error reduction of 29 .18% compared to *baseline (dummy)*, which corresponds to the approximate average error of 0.82, that is, the error is 1 for more or less of the answer the user responded. *PCA* was used for attribute reduction, counting to assign weight to n-grams, and the model used was *Support Vector Machine – polynomial kernel*.

Predictions for the question Did you feel like you were finishing an important task (P7): The best predictive result achieved used as attributes the answers to the *UEQ-S* questionnaire along with the attribute "everything" using unigrams, obtaining an error reduction of 44.36% compared to *baseline (dummy)*, which corresponds to the approximate average error of 1.05. In this case, no reducing attributes technique was used, TFIDF was employed to assign weight to the n-grams, and the model used was GBR.

Predictions for the question Found something new or unknown (P8): The best predictive result for the eighth question used as attributes the answers to the questionnaire *UEQ-S* together with the attribute "everything" with combined use of unigrams and heptagrams (n=7), obtaining an error reduction of 14 .40% compared to *baseline (dummy)*, which corresponds to the approximate average error of 0.96, that is, the error is less than 1 for more or less of the answer the user responded. *KBEST* was used for attribute reduction, counting to assign weight to n-grams, and the model used was GBR.

Predictions for the Expressive question (P9): The best predictive result used the *UEQ-S* questionnaire responses as attributes, obtaining an error reduction of 23.54% compared to *baseline (dummy)*, which corresponds to an average error of approximately 0 .78, that is, the error is less than 1 for more or less of the answer the user responded. In this case, no attribute reduction strategy was used, i.e., all attributes were used, and the model with the best performance was *Kernel Ridge*.

Predictions for the Relaxing question (P10): The best predictive result for the tenth question used as attributes the answers to the questionnaire *UEQ-S* together with the attribute "everything" with the use of unigrams (n=1), obtaining an error reduction of 28.11 % compared to *baseline (dummy)*, which corresponds to the approximate average error of 0.89, that is, the error is less than 1 for more or less of the answer the user responded. In this case, no attribute reduction strategy was used, counting was employed to assign weight to the n-grams, and the model with the best performance was SVR Poly.

Predictions for the question Simulates something from real life (P11): The best predictive result used as attributes the answers to the questionnaire *UEQ-S* together with the attribute "all" with the combined use of unigrams(n=1) and heptagrams (n=7), obtaining an error reduction of 25.08% compared to *dummy*, which corresponds to an approximate average error of 1.17. The *PCA* strategy was used for attribute reduction, TFIDF was used to assign weight to n-grams, and the model with the best performance was GBR.

Predictions for the question Adrenaline (derived from risk and/or danger) (P12): Using the *UEQ-S* questionnaire responses as attributes, the best predictive result was adding the attribute "everything" with the combined use of unigrams and heptagrams (n=7), obtaining a reduction in error by 1, 20% compared to *dummy*, which corresponds to an approximate average error of 0.90. In this case, the *PCA* strategy was used to reduce attributes, counting to assign weight to n-grams and the model with the best performance was SVR.

Predictions for the question Is it part of a larger structure (P13): The best predictive result used the *UEQ-S* questionnaire responses as attributes, obtaining an error reduction of 7.61% compared to *baseline (dummy)*, which corresponds to an average error of approximately 0.87, that is, the error is

less than 1 for more or less of the answer the user responded. This result was, among the ones obtained using the *UEQ-S* questionnaire responses as attributes, the smallest error reduction compared to *dummy*. In this case, no attribute reduction strategy was used, counting was employed to assign weight to the n-grams, and the model with the best performance was SVR.

Predictions for the question Experience that requires imagination (P14): The best predictive result used, as attributes, the answers to the *UEQ-S* questionnaire, along with the attribute "everything" with the use of unigrams(n=1), obtaining an error reduction of 27.12% compared to o *baseline (dummy)*, which corresponds to the approximate average error of 1.09, that is, the error is 1 for more or less of the answer the user responded. This result is considered satisfactory, as it is common for the user to have doubts between two answer options when answering the question. In this case, no feature reduction strategy was used, and the best-performing model was GBR.

4.2.3 Synthesis

In summary, the results using *Plex Framework* responses as attributes for predictions for *UEQ-S* had better results compared to the others, with an average error of approximately 1.24. It is important evidence as it allows for the mapping of answers from one questionnaire to another. In the results using the *UEQ-S* questionnaire responses as an attribute for the predictions for the *Plex Framework* had a better average result compared to the others with an average error of approximately 0.88, even though it is not part of the main objective of this paper, it constitutes relevant evidence, as it makes it possible to map responses from one questionnaire to another.

Tables 8 and 9 present a summary of the best results. In these tables, there are columns defining which attributes were used, namely, only *Plex* (using only the answers from *Plex Framework*) or only *UEQ-S* (using only the responses from the *UEQ-S* questionnaire), and better global.

4.3 T-test (p-value) of predictive results

The normality of the responses was verified using the Kolmogorov-Smirnov test. The t-test with two-tailed distribution was performed for all predictive results from the *UEQ-S* questionnaire and all predictive results from *Plex Framework*, Tables 10 and 11 present the results of each questionnaire, respectively. The purpose of these tables is to summarize the p-values for each pair of feature sets used. Lighter shades of green indicate values close to 1, while lighter shades of red indicate values close to zero.

In both, the use of only emotions as attributes (regardless of the variation, the value of *n* in the *n-gram* and weight of the *n-gram* used) are statistically not considerably different, i.e., there is no significantly different contribution in using one variation or another.

On the other hand, when incorporating data from additional questionnaires, several cases present p-values below 0.05 (5%). This indicates statistically significant differences between the results, demonstrating that the inclusion of these additional questionnaires meaningfully contributes to improving the models' performance.

5 Discussion and Conclusion

The present work contributes to user experience analysis through facial emotion recognition.

In Section 2, three systematic reviews of the literature were presented, showing a gap in the use of hedonic criteria in UX assessment and that the most used tool to carry out UX evaluation was self-assessment via questionnaire. The second review confirmed the existence of a gap in the evaluation of UX through automatic emotion recognition.

In this context, the first contribution of this work is related to verifying the relationship between questionnaire self-assessment and facial emotion recognition, making it possible to complete a questionnaire validated by an artificial intelligence model using the automatic extraction of facial emotions as parameters. This work makes a research contribution by developing an innovative method to achieve these objectives using a low-cost device.

Although the best-predicted results for *UEQ-S* questionnaire were achieved using *Plex* responses as attributes (average error 1.24), the results using only emotions extracted automatically had a satisfactory result (average error 1.39). The difference in the average error between the results using *Plex* responses and just emotions was 0.15, which corresponds to 12.10%. When placing this result in the context of completing a self-assessment questionnaire, the difference between these results becomes minimal since, in the *UEQ-S* questionnaire, it is not possible to choose an option with decimal values and only whole numbers in the range from 1 to 7. The result takes on an even more rewarding dimension when we consider that users often express hesitation when faced with the choice between two alternative answers.

When considering the predictions for *Plex*, in the same way as those from the *UEQ-S* questionnaire, the best predicted results achieved used the responses from the other questionnaire as attributes, in this case, the *UEQ-S* questionnaire (average error 0.90), the results using only emotions extracted automatically had a great result (average error 0.97). The difference in the average error between the results using *Plex* responses and just emotions was 0.09, which corresponds to around 7.77%. When placing this result in the context of filling out a self-assessment questionnaire, the difference between these results becomes minimal, since in *Plex*, it is also not possible to choose an option with decimal values.

One of the possible reasons for *Plex* predictions to have a lower average error than *UEQ-S* questionnaire predictions is the range of options that the user can choose when answering a question, *Plex* has a range of 1 to 5, on the other hand, *UEQ-S* has a range of 1 to 7.

When analyzing *Plex* predictions, the only question for which using *UEQ-S* questionnaire responses as attributes did not result in better performance compared to using just emotions was Plex12 (*Adrenalina*). In this case, the best overall result obtained was using the EmotionsEnd attribute with the combination of unigrams to pentagrams. This can be justified by the nature of the question being closely linked to emotions and because the Plex12 question does not have any average, strong, or perfect correlation with any question in the *UEQ-S* questionnaire.

To verify whether there is a considerable statistical dif-

Table 8. Summary of best results for UEQ-S prediction

Question	Just Plex Framework	Best Overall
Obstructive/Supportive	1.516	1.516
Complicated/Easy	1.099	1.096
Inefficient/Efficient	1.344	1.187
Confusing/Clear	1.497	1.497
Boring/Exciting	1.281	1.215
Not interesting/Interesting	0.717	0.717
Conventional/Inventive	1.162	1.115
Usual/Leading edge	1.297	1.284
Overall Average	**1.239**	**1.203**

Table 9. Summary of best results for PLEX prediction

Question	Just UEQ-S	Best Overall
Captivating	0.665	0.659
Challenging	0.410	0.410
Competitive	0.926	0.875
Fun	0.685	0.685
Frustrating	1.149	1.116
Felt in control	0.829	0.825
Felt like completing an important task	1.083	1.046
Found something new or unknown	0.972	0.957
Expressive	0.783	0.783
Relaxing	0.946	0.895
Simulates something from real life	1.181	1.161
Adrenaline	0.914	0.861
Part of a bigger structure	0.874	0.874
Experience that needs imagination	1.150	1.094
Overall Average	**0.898**	**0.874**

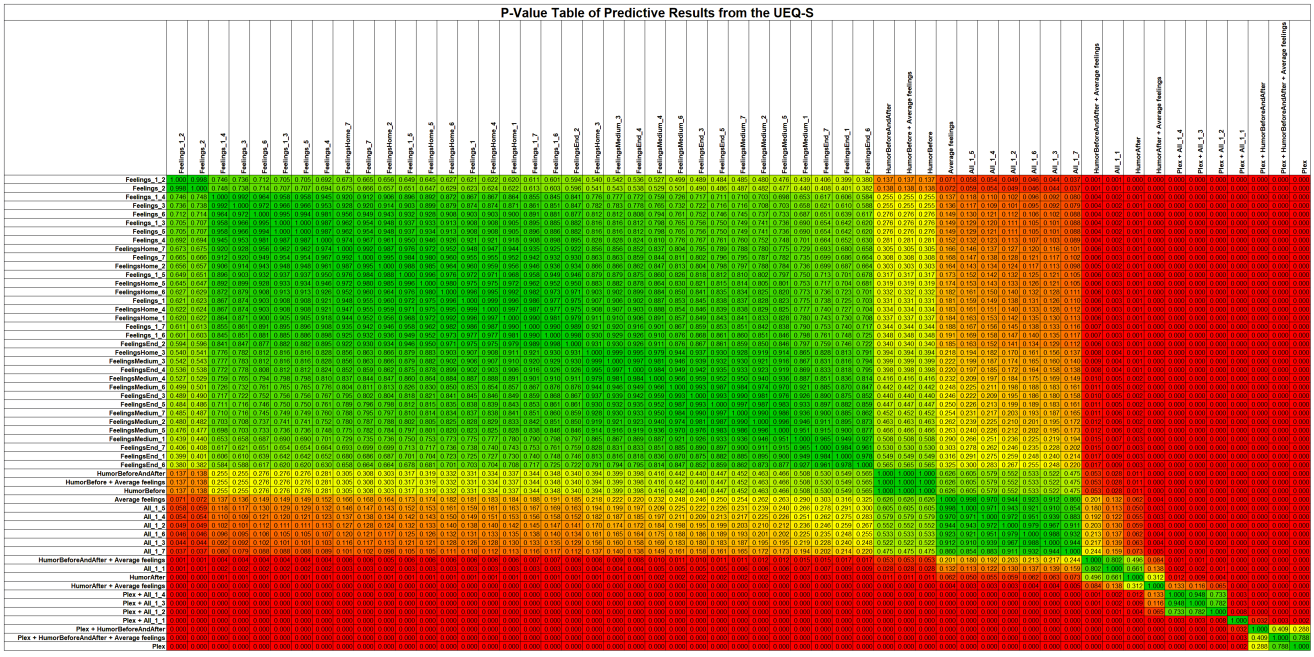

Table 10. T-test of all predictive results for the UEQ-S. Lighter shades of green indicate values close to 1, while lighter shades of red indicate values close to zero.

ference in the use of attributes, Section 4.3 presents the t-test with a two-tailed distribution for all predictive results. It was found that the different representations of emotions, do not achieve significant differences. However, when other questionnaires are added, a notable distinction emerges, contributing significantly to improved performance.

Considering these results, we believe our main goal was achieved, confirming the hypothesis that it is possible, to

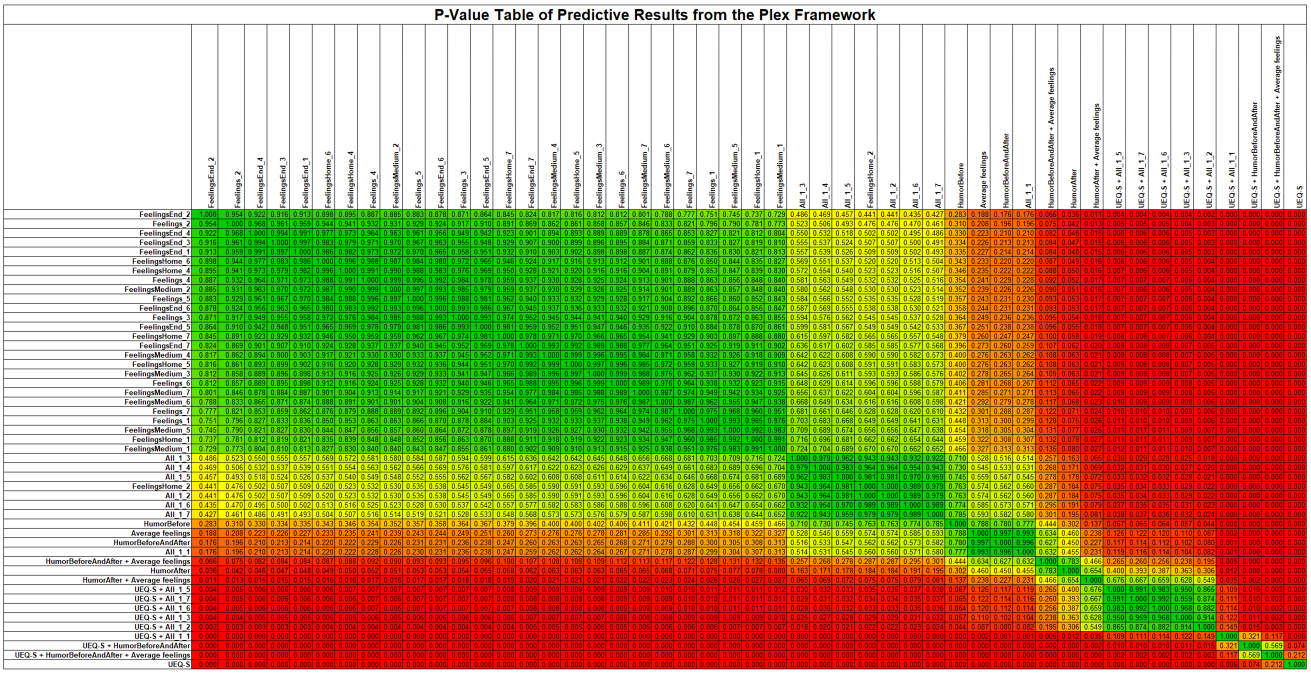

Table 11. T-test all predictive results for Plex Framework. Lighter shades of green indicate values close to 1, while lighter shades of red indicate values close to zero.

a certain extent, to evaluate the user experience using only sentiment extraction through an artificial intelligence model, making it possible to estimate the answers to UX questionnaires *UEQ- S* or *Plex* with an approximate average error of plus or minus one for their questions.

Another intriguing contribution is the importance of the user's mood in the experience and its evaluation, since the addition of responses from the mood profile questionnaire improved the predictions for both *UEQ-S* and *Plex* questionnaires. Even though using mood as a predictor was not part of our scope, it is an interesting future work, since answering a mood profile questionnaire is a less complex task than answering the *UEQ-S* or *Plex* questionnaires and, besides, some automated mechanisms could be explored to extract mood information.

The predictive identification of one questionnaire using the questions of another as attributes is a significant discovery. It is possible to infer *Plex* with an average error of 0.898 using only the *UEQ-S* responses. To infer *UEQ-S* using only *Plex* responses, the average error is 1.239. By adding facial expressions as attributes, the accuracy of inferences improves minimally, reducing the average error to 0.887 (a reduction of 0.011) in the case of *Plex* and to 1.214 (a decrement of 0.025) in the case of *UEQ -S*. In addition to allowing the matching of responses between different questionnaires, it also enables the use of a less complex questionnaire to predict the results of a more elaborate one.

The dataset produced for this article is a considerable contribution, given the scarcity of similar datasets, and could serve as a basis for other future projects, such as the use of *transfer learning* for use with other applications or questionnaires.

Section 3.2 presents a new method for evaluating questionnaires using seven dimensions or UX factors as evaluation criteria. The selection of these dimensions/factors constitutes a contribution to future research aiming to compare two or more UX questionnaires. The presented evaluation method generates a final score for each evaluated questionnaire, identifying their strengths and weaknesses. This facilitates a more informed choice regarding which questionnaire to apply in a given context, thereby improving the final evaluation of the user experience.

Once the possibility of carrying out, even partially, the evaluation of UX through automatic recognition of facial emotions has been demonstrated, a possible future work is to improve the accuracy rate in emotion recognition. In this work, an open source library was used called *Face-api.js*, which has an accuracy of 64%. Another possibility is the application in a context other than a three-dimensional game, such as a commercial application or using other standardized questionnaires.

Another interesting future research is the comparison of the user experience in using the three-dimensional input device with the traditional one in this context, providing more data for the hypothesis that, in the three-dimensional scenario, it is more attractive and intuitive to use devices with more degrees of freedom than traditional devices (2D). However, because users are potentially not used to this type of device, it is possible to obtain varied results that must be carefully analyzed.

As UX evaluation is fundamental for the acceptance of new software or functionality, it is understood that this work contributes by increasing the quality and acceptance of new systems.

Despite the promising results, some limitations of this work must be acknowledged. One important limitation concerns the absence of a manually annotated ground truth for the detected emotions. Since emotional states were obtained using an automated emotion recognition model (*Face-api.js*), there is an inherent level of uncertainty regarding the correct-

ness of the extracted emotional signals.

It is important to highlight, however, that the main objective of this study is not to assess the accuracy of emotion recognition itself, but rather to investigate whether automatically extracted emotional signals, even when subject to noise, can serve as predictors of user experience questionnaire responses.

In this context, the proposed approach reflects a realistic and practical scenario, in which low-cost and non-intrusive emotion recognition tools are employed in real-world applications. The results obtained, with an average prediction error close to ± 1 point, indicate that even imperfect emotion detection can provide meaningful information for UX evaluation.

Nevertheless, the absence of a ground truth for emotions limits the ability to isolate the impact of emotion recognition accuracy on the predictive performance of the models.

A potential direction for future work is the construction of a manually annotated emotional dataset, enabling a controlled comparison between models trained on annotated data and those based on automatically extracted emotions. Such an analysis would provide deeper insights into how improvements in emotion recognition accuracy affect UX prediction performance.

Furthermore, it is important to note that even manually annotated emotions are subject to inter-annotator variability, reflecting the inherently subjective nature of emotional expression. This suggests that, although improvements in emotion recognition may enhance predictive performance, such gains may not necessarily be linear or unlimited, constituting an open research question.

Declarations

Funding

This research was partially funded by the Coordination for the Improvement of Higher Education Personnel (CAPES).

Authors' Contributions

LAD, JLBJ and ESV worked together to the conception of this study. ESV performed the experiments. ESV is the main writer of this manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The constructed dataset is public available in GitHub: `https://github.com/digiampietri/basesDeDados/blob/main/UEQ_PLEX_Sentiments.csv`. Accessed on 26 May 2026.

Further relevant information

This research was submitted to the Research Ethics Committee of EACH-USP and obtained its approval under the number 38540620.0.0000.5390.

Part of the references used in this paper were selected from three systematic literature reviews, based on their relevance to the present study [Veriscimo *et al*., 2020, 2021; Santos and Digiampietri, 2024].

References

Amusement Vision (2001). Super monkey ball. Video game, Nintendo GameCube.

Aranha, R. V., Casaes, A. B., and Nunes, F. L. S. (2020). Influence of environmental conditions in the performance of open-source software for facial expression recognition. In *Proceedings of the Brazilian Symposium on Human Factors in Computing Systems*, pages 1–10. DOI: https://doi.org/10.1145/3424953.3426630.

Arrasvuori, J. *et al*. (2011). Applying the plex framework in designing for playfulness. In *Proceedings of the Conference on Designing Pleasurable Products and Interfaces*, pages 1–8. ACM. DOI: https://doi.org/10.1145/2347504.2347531.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 59–66. DOI: https://doi.org/10.1109/FG.2018.00019.

Bevan, N. (2009). What is the difference between the purpose of usability and user experience evaluation methods. In *Proceedings of the Workshop on UX Evaluation Methods*, pages 1–4.

Den Uyl, M. J. and Van Kuilenburg, H. (2005). The facereader: Online facial expression recognition. In *Proceedings of Measuring Behavior*, pages 589–590.

Deshmukh, R. S. and Jagtap, V. (2017). A survey: Software API and database for emotion recognition. In *Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 284–289. DOI: https://doi.org/10.1109/ICCONS.2017.8250727.

International Organization for Standardization (2010). ISO 9241-210:2010 ergonomics of human-system interaction – human-centred design for interactive systems. `https://www.iso.org/standard/52075.html`. Accessed on 26 May 2026.

Laugwitz, B., Held, T., and Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In *HCI and Usability for Education and Work*, pages 63–76. Springer. DOI: https://doi.org/10.1007/978-3-540-89350-9_6.

Lucero, A., Holopainen, J., Ollila, E., Suomela, R., and Karapanos, E. (2013). The playful experiences (PLEX) framework as a guide for expert evaluation. In *Proceedings of the International Conference on Designing Pleasurable Products and Interfaces*, pages 221–230. ACM. DOI: https://doi.org/10.1145/2513506.2513530.

Magdin, M. and Prikler, F. (2018). Real-time facial expression recognition using webcam and sdk affectiva. *International Journal of Interactive Multimedia and Artificial Intelligence*. DOI: https://doi.org/10.9781/ijimai.2017.11.002.

Magix (2023). Sony vegas. `https://www.vegascreativesoftware.com/br/vegas-pro/`. Version 21.0.0. Accessed on 26 May 2026.

Martinelli, S., Lopes, L., and Zaina, L. (2022). Ux research in the software industry: An investigation of long-term ux practices. In *Proceedings of the Brazilian Symposium on Human Factors in Computing Systems*. SBC. DOI: https://doi.org/10.1145/3554364.3559126.

Mennig, P., Scherr, S. A., and Elberzhager, F. (2019). Supporting rapid product changes through emotional tracking. In *Proceedings of the IEEE/ACM 4th*

International Workshop on Emotion Awareness in Software Engineering (SEmotion), pages 8–12. DOI: https://doi.org/10.1109/SEmotion.2019.00009.

Milgram, P., Takemura, H., Utsumi, A., and Kishino, F. (1995). Augmented reality: A class of displays on the reality-virtuality continuum. *Telemanipulator and Telepresence Technologies*, 2351:282–292. DOI: https://doi.org/10.1117/12.197321.

Mühler, V. (2020). face-api.js – JavaScript API for face recognition in the browser with TensorFlow.js. https://itnext.io/face-api-js-javascript-api-for-face-recognition-in-the-browser-with-tensorflow-js-bcc2a6c4cf07. Accessed on 26 May 2026.

Rajeshkumar, S., Omar, R., and Mahmud, M. (2013). Taxonomies of user experience (ux) evaluation methods. In *Proceedings of the International Conference on Research and Innovation in Information Systems (ICRIIS)*, pages 533–538. DOI: https://doi.org/10.1109/ICRIIS.2013.6716765.

Rehman, A., Mujahid, M., Elyassih, A., Alghofaily, B., and Saeed, A. (2025). Comprehensive review and analysis on facial emotion recognition: Performance insights into deep and traditional learning with current updates and challenges. *Computers, Materials, & Continua*, 82(1):41–72. DOI: https://doi.org/10.32604/cmc.2024.058036.

Santos, B. and Digiampietri, L. (2024). User experience evaluation using machine learning and facial expressions: A systematic review. In *Anais do XXI Encontro Nacional de Inteligência Artificial e Computacional*, pages 930–941, Porto Alegre, RS, Brasil. SBC. DOI: https://doi.org/10.5753/eniac.2024.245150.

Schrepp, M., Hinderks, A., and Thomaschewski, J. (2017). Design and evaluation of a short version of the user experience questionnaire (ueq-s). *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(6):103–108. DOI: https://doi.org/10.9781/ijimai.2017.09.001.

Veriscimo, E. S., Bernardes Junior, J. L., and Digiampietri, L. A. (2020). Evaluating user experience in 3D interaction: A systematic review. In *Proceedings of the Brazilian Symposium on Information Systems*, pages 1–8. DOI: https://doi.org/10.1145/3411564.3411640.

Veriscimo, E. S., Bernardes Junior, J. L., and Digiampietri, L. A. (2021). Facial emotion recognition in ux evaluation: A systematic review. In *International Conference on Universal Access in Human-Computer Interaction*. Springer. DOI: https://doi.org/10.1007/978-3-030-78462-1_40.

Viana, M. F., Almeida, P., and Santos, R. C. (2001). Adaptação portuguesa da versão reduzida do perfil de estados de humor – POMS. *Análise Psicológica*, 19(1):77–92. DOI: https://doi.org/10.14417/ap.345.