AMUSED: A Multi-Modal Dataset for Usability Smell Identification

Flávia de S. Santos, Marcos Treviso, Kamila R. H. Rodrigues, Renata P. M. Fortes, and Sandra P. Gama

***Abstract*—Understanding how users interact with systems and experience usability issues is vital in Human-Computer Interaction (HCI). Existing approaches often focus on isolated data types, such as interaction logs or subjective reports, providing only a partial view of the user experience. This work introduces AMUSED, a multimodal dataset that integrates user interaction logs, physiological signals, facial emotion features, and self-assessment reports. The dataset includes expert annotations of eleven types of usability smells, creating a rich resource for investigating relationships between usability problems, user behavior, and emotional states. We conducted experiments with 70 participants interacting with three social networks containing usability issues. We then evaluated various machine learning models to assess the feasibility of automatically detecting these issues. The dataset comprises 24 h of user recordings, with over 20,000 user events, such as clicks, scrolls, and input changes. Our analysis reveals that (i) usability smells frequently co-occur, (ii) negative emotions predominate when severe usability issues arise, and (iii) Gradient Boosting models achieve up to 92% accuracy in detecting usability smells, demonstrating the potential for computational methods in automated usability evaluation. Our findings emphasize the value of emotional metrics in HCI research and highlight promising uses of machine learning to automatically detect usability issues.**

***Index Terms*—Dataset, usability evaluation, user experience, usability smells, EEG, BVP, facial recognition, emotion recognition, multimodal data.**

I. Introduction

In the field of Human-Computer Interaction (HCI), capturing the full scope of user interaction is crucial for evaluating and improving system usability. Traditional approaches often focus on a single type of data, such as interaction logs or self-assessment reports, providing a limited view of the overall user experience. As digital platforms, particularly social media, continue to shape our daily lives, it becomes essential to gather comprehensive data that reflect not only how users interact with systems, but also how these interactions can highlight usability issues. For example, detecting and recognizing users' emotions during interactions with computational systems can be a key aspect of understanding the full user experience and is increasingly important in HCI research [1]. Emotional data allows not only the evaluation but also the improvement of various aspects of the interaction, interface design, and overall user experience. As such, emotional design plays a crucial role in enriching user engagement, fostering emotional resonance [2], and inspiring greater user confidence [3].

Recent advances in neurological signal processing have increasingly enabled Electroencephalography (EEG) applications in user interface and usability evaluation [4], [5], while the automation of usability evaluation with machine learning techniques has emerged as a prominent research direction [6], [7]. Building on the framework proposed by de Souza Santos et al. [6], we present a new dataset[1] that integrates multiple forms of interaction data in order to offer a holistic view of user-system interactions. Concretely, our datasets include interaction logs to capture the sequence of user actions, EEG and Blood Volume Pulse (BVP) signals along with facial features to capture insights into the user's cognitive and emotional state during interactions, and self-assessment reports to capture subjective user feedback. To support a structured analysis of usability issues, experts have annotated each interaction with usability smells [8], which indicate inefficient or problematic interaction patterns, such as high cognitive load or unclear feedback mechanisms. By annotating these smells across 70 user sessions, our dataset enables researchers to explore the relationship between user actions and potential usability issues in depth, offering a more comprehensive understanding of how interaction patterns impact user experience.

The following sections provide a structured overview of our study. We begin by reviewing related work (Section II), followed by a description of the dataset structure and the data collection process, including setup details (Section III). Next, we outline the data annotation process (Section IV) and the methodology used for dataset construction (Section V), before presenting an

Received 12 March 2025; revised 8 November 2025; accepted 11 November 2025. Date of publication 14 November 2025; date of current version 10 March 2026. This work was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) Finance Code 001, and by Conselho Nacional de Desenvolvimento Científico e Tecnolgico (CNPq). Recommended for acceptance by S. Zhao. *(Corresponding author: Flávia de S. Santos.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Ethics Committee of the University of São Paulo (Comitê de Ética em Pesquisa da Universidade de São Paulo . USP) under protocol No. 46999521.0.0000.5390 (opinion no. 4,785,203, June 2021), and by the Ethics Committee of Instituto Superior Técnico – University of Lisbon (Comitê de Ética do Instituto Superior Técnico) under Application No. 29/2024, and performed in line with the 1964 Declaration of Helsinki and its later amendments.

Flávia de S. Santos, Kamila R. H. Rodrigues, and Renata P. M. Fortes are with the University of São Paulo, São Carlos 13566-590, Brazil (e-mail: flaviasantos@usp.br; kamila.rios@icmc.usp.br; renata@icmc.usp.br).

Marcos Treviso is with Instituto de Telecomunicações & Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisboa, Portugal (e-mail: marcos.treviso@tecnico.ulisboa.pt).

Sandra P. Gama is with the INESC-ID & Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisboa, Portugal (e-mail: sandra.gama@tecnico.ulisboa.pt).

Digital Object Identifier 10.1109/TAFFC.2025.3632675

[1]Dataset available at: https://doi.org/10.5281/zenodo.15870704

in-depth analysis of the dataset (Section VI). We then report results on automatic detection of usability smells on our dataset by using different machine learning models (Section VII). Finally, we discuss our conclusions and outline the study's limitations (Section VIII).

II. Related Work

Traditional usability evaluation relies on expert inspection methods that suffer from subjective interpretation variations and evaluator expertise dependencies. While automated usability smell detection provides behavioral indicators for potential problems [9], [10], [11], such detection has evolved from rule-based pattern matching to sophisticated machine learning approaches that leverage multimodal data [6].

Recent advances in emotion recognition have driven the development of multimodal emotional expression datasets that include physiological signals such as EEG and BVP. In this section, we review key datasets labeled with emotions that have contributed significantly to the field by collecting different data sources, such as EEG/BVP signals, facial expressions, self-assessments, among others.

DEAP [12], one of the most widely used datasets, contains EEG, BVP, electromyogram (EMG), electrooculogram (EOG), skin temperature, and galvanic skin response (GSR) recordings from 32 participants watching music videos, with 22 of them also having recorded facial videos. Participants rated their responses using the Self-Assessment Manikin (SAM) [18] in terms of valence-arousal-dominance (VAD), along with a "liking" scale (thumbs-up, neutral, thumbs-down). MAHNOB HCI [13] contains EEG, eye movements, audio, and facial expressions from 27 participants who watched emotional video clips, labeled on valence and arousal scales to the target emotions. The eNTERFACE dataset [14] consists of two main subsets. The first includes functional near-infrared spectroscopy (fNIRS), peripheral signals, and EEG data from five participants who viewed images from the International Affective Picture System (IAPS) [19]. The second subset contains facial expression and fNIRS recordings of 16 participants exposed to a series of images and video sequences. Like other datasets, eNTERFACE is labeled based on valence and arousal scales. DREAMER [15] focuses on EEG, and electrocardiogram (ECG) recordings, alongside heart rate variability (HRV) derived from ECG signals, with participants reporting their emotional responses on the SAM scale. Unlike previous datasets, the DECAF dataset includes brain signals from 30 participants exposed to music and videos, acquired through magnetoencephalography (MEG), near-infrared facial videos, horizontal electrooculogram (hEOG), ECG, and trapezius electromyogram (tEMG). Similar to the other datasets, participants rated their emotional responses on the VAD scale. While these multimodal datasets support emotion recognition research, datasets integrating user interaction data remain limited. The DUX dataset [17] addresses this gap by combining keyboard and mouse activity with facial emotion analysis from 50 participants using a travel expense report application with embedded emotional triggers (intentional interface flaws violating usability principles to induce negative emotions—referred to as "bad UX triggers" in our work) to explore the relationship between user interactions and affective states. However, DUX does not include physiological signals such as EEG and BVP and lacks data specifically focused on usability-related contexts.

Although these datasets provide valuable insights, most do not incorporate user interaction data or usability-related contexts, which are crucial for understanding emotions in HCI scenarios. To address this gap, we introduce the Affective Metrics & Usability Smell Evaluation Dataset (AMUSED), a comprehensive multimodal resource that incorporates facial expressions, self-reported emotions, EEG and BVP signals, detailed user interaction data, and usability smell annotations. Collected from 70 participants interacting with interfaces containing usability issues, AMUSED uniquely enables the exploration of the relationship between user behavior, usability challenges, and emotional responses. This combination fosters new opportunities for analyzing and improving user experience through the lens of emotion recognition and usability research. Table I summarizes the discussed datasets, detailing the number of participants, EEG channels, along with the presence of heart rate data, facial emotion analysis, self-assessment methods, and interaction data. We also detail the types of emotional stimuli used, and the target emotions captured, showcasing the diversity of multimodal emotion recognition resources. Finally, AMUSED does not contain data from other sources such as fNIRS and tEMG, which could further improve the dataset.

In the next section, we describe the structure of AMUSED and detail the data collection process.

III. Data Collection

In our study, we collected data from three social networks (SNs) to investigate usability issues and their impacts on user interactions. We adopt the concept of *usability smells*, where "bad smells" are indicators of inadequate application design, with the potential to compromise usability [9].

Prior to the commencement of the study, approval was obtained from the Research Ethics Committee of the University of São Paulo (USP) under protocol number 46999521.0.0000.5390, with substantiated opinion no. 4.785.203, in June 2021, and from the Ethics Committee of Instituto Superior Técnico - Universidade de Lisboa (approval no. 29/2024).

A. Interfaces Information

The social networks used in this work include a custom-designed website and two prototype websites with publicly available source code, none of which were commercially available to users. To highlight the presence of usability smells during user interactions, usability issues were deliberately introduced in Perspective (SN1).[2] This platform offers a variety of standard social networking features, including searching and adding friends, viewing profiles, sending private messages, adding photos, creating HTML-rich posts on personal profiles or

[2]SN1: https://github.com/flassantos/perspective

TABLE I
OVERVIEW OF MULTIMODAL EMOTION DATASETS

Dataset	N. of Participants	EEG chs.	HR	Facial Emotion	Self-assess.	Emotion stimulation	Target Emotions	Int. data
DEAP [12]	32	32	Yes	Yes	Yes	Music videos	Valence, arousal, dominance, liking, and familiarity	No
MAHNOB HCI [13]	27	32	Yes	Yes	Yes	Movies and pictures fragments	Valence and arousal	No
eNTERFACE [14]	5	54	Yes	No	Yes	IAPS images	Calm, positive exciting, negative exciting	No
	16	-	-	Yes	No	Images, Videos	Happy, disgust, neutral	No
DREAMER [15]	23	14	Yes	No	Yes	Film clips	Valence, arousal, and dominance	No
DECAF [16]	30	-	Yes	Yes	Yes	Music and video clips	Valence, arousal, and dominance	No
DUX [17]	50	-	-	Yes	Yes	Bad UX triggers	Anger, confusion, contempt, disgust, engagement, fear, joy, neutral, sadness, sentimentality, surprise, and judgement	**Yes**
AMUSED (this work)	70	2	Yes	Yes	Yes	Bad UX triggers	**Facial:** anger, disgust, fear, enjoyment, contempt, sadness, and surprise. **SAM:** valence and arousal	**Yes**

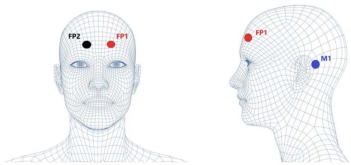

Fig. 1. Electrode placement on the forehead and behind the ear for EEG acquisition in 10-20 system configuration [22].

friends' timelines, commenting on and liking posts, and changing account information. Love Social (SN2)[3] and Social-Network (SN3),[4] are also simple social networks similar to SN1, allowing users to carry out typical tasks. We identified usability issues in both of these platforms as well.

B. EEG and BVP Setup

We use BITalino (r)evolution Plugged Kit BT[5] with the addition of a Photoplethysmography (PPG) sensor,[6] and the software application OpenSignals[7] to record the electroencephalography (EEG) and Blood Volume Pulse (BVP) signals during user interactions. The BITalino's EEG sensor has a bipolar configuration, comprising three electrodes. In its bipolar setup, two of these electrodes measure the electrical potentials in a targeted area of the scalp, while a reference electrode is positioned in a region with minimal muscular activity. Electrodes I and II were placed at FP1 and FP2 (forehead), respectively, according to the 10-20 system [20], [21], and electrode III was placed at M1 (mastoid process, behind the left ear), as illustrated in Fig. 1. The sensor connection cable was plugged on BITalino's channel 1. The two-channel EEG configuration was chosen for practical constraints, such as minimal user interaction interference and reduced cost.

To record the BVP, we use the PPG sensor (also known as PulseSensor), placed in the ear lobe (positioned with the help of an ear clip), to record the heart-rate. The sensor connection cable was plugged on BITalino's channel 2. BITalino was attached to the back collar of the participants' shirts using its protective cover. We connected the hardware device and the software application via Bluetooth. Once connected, we configured channel 1 for EEG and channel 2 for BVP, as they were connected to the hardware. Additionally, we set the sampling rate to 1,000 Hz.

C. Experiment Setup

Participants were recruited using convenience methods of recruitment. Eligible participants included anyone interested in taking part, provided they were at least 18 years old. The participants were informed about the experiment and their rights in a verbal introduction and through a consent form. They were briefed on the objectives of the experiment, the types of data that would be collected, and the methods of collection. After addressing any questions or concerns, they were invited to complete the consent form, as well as provide demographic information before starting the session. There were no potential risks and no expected benefits for the individual participants. Each session lasted approximately 1 h. We used the Wildfire browser extension[8] to collect the user interaction logs, and a screen recorder extension to capture the screen.

We invited participants to attend sessions in calm, quiet, and comfortable environments. To ensure that all usability measurements reflected authentic first-time user interactions, all participants were confirmed to have no previous experience with the social network platforms used in the study, as these were prototypes rather than commercially available applications. For participants where only interaction logs were collected, tasks were performed using Google Chrome on an available computer, and for participants whose EEG, BVP, and other physiological data were collected, interactions were conducted on a notebook running the Ubuntu 18.04 operating system, with a Core i7 (eighth generation). EEG and BVP sensors positioned according to the predefined setup and connected to OpenSignals, alongside screen recording and interaction log capture.

[3]SN2: https://github.com/flassantos/react-social-network
[4]SN3: https://github.com/flassantos/social-network
[5]https://www.pluxbiosignals.com/collections/bitalino
[6]https://www.pluxbiosignals.com/products/photoplethysmography-ppg-sensor
[7]https://support.pluxbiosignals.com/article-categories/opensignals
[8]https://wildfire.ai/

TABLE II
DATA STATISTICS OF THE AMUSED DATASET

	SN1	SN2	SN3	Total
Outline				
Users	32	30	8	70
Num. Tasks	17	11	16	44
Full Duration	13h56m	07h20m	03h28m	24h46m
Features				
Num. EEG & BVP	23	30	3	56
Num. Face	22	28	0	50
Num. SAM	1	28	0	29

At the start of data collection, participants were shown a slideshow of six selected images for 5 min to induce a calm state. The slideshow consisted of a repeated set of images from the International Affective Picture System (IAPS) [19], specifically images categorized as neutral to positive. Participants were then instructed to complete the tasks outlined for the website under evaluation, with the option to move on to the next task at any time. Once the tasks were completed, the data recordings were stopped, and the sensors were removed.

In a subset of tests, facial images were captured using the notebook's built-in camera. Additionally, in a small subset, participants were also asked to complete a shortened SAM questionnaire [18] after each task, reporting valence and arousal levels, where valence ranges from positive (happiness) to negative (sadness), and arousal from calm to excitement [23]. Before these assessments, participants were introduced to the concepts of arousal and valence.

D. Participants

A total of 70 participants agreed to participate in the study after the recruitment process, with an equal distribution of 35 women and 35 men, with ages ranging from 18 to 61 years (median age of 27 years). Among them, 35 subjects have completed or are currently pursuing postgraduate studies, 23 have completed undergraduate degrees, seven are currently enrolled in undergraduate programs, and five have completed high school. Regarding computer usage, 64 participants reported frequent use, while only two indicated they rarely use the device. As for smartphone usage, 56 participants use it multiple times a day, twelve use it moderately throughout the day, and only two reported using it sparingly during the day. Consequently, 55 participants classified themselves as experienced users, 14 as moderate users, and one as a beginner.

E. AMUSED Dataset

As previously mentioned, we collected interactions from a total of 70 participants: 32 from SN1, 30 from SN2, and eight from SN3, as outlined in Table II. The assignment of participants to each SN was performed randomly. However, at the beginning of the experiments, SN2 temporarily faced access issues, which were later solved. Additionally, the SN3 was later removed from the testing platform, leaving only its source code accessible. Although the code for SN3 remained available, we were unable to execute it locally, which hindered the continuation of sessions on this network. This limitation impacted the volume and diversity of data collected for SN3, as reflected in the table.

AMUSED dataset captures diverse interactions across the three social networks, with tasks designed according to the typical user activities and platform affordances. A total of 44 tasks were defined,[9] distributed as 17 for SN1, eleven for SN2, and 16 for SN3. The total duration time was approximately 24h46 min, with SN1 accounting for the largest portion (13h56 min). In addition to interaction data, we collected EEG and BVP data, facial images, and SAM questionnaire responses. EEG and BVP were recorded in 56 sessions, facial images in 50 and SAM questionnaires were completed in 29. The heterogeneous data distribution also reflects progressive refinement of data collection protocols, with SAM questionnaires introduced after initial SN1 sessions were completed.

IV. DATA ANNOTATION

To develop a machine learning model capable of identifying usability issues, it is essential to annotate the collected interaction data. In HCI, this stage is often referred to as the evaluation phase, in which experts analyze user interactions to identify usability problems. This process ensures that the dataset captures meaningful usability insights, which can then be leveraged to train models for automated detection.

For the data annotation process, we considered eleven subjective usability smells, as described by de Souza Santos et al. [6], categorized into two levels of specificity: task-level and action-level. **Task-level** smells refer to broader usability issues that impact the *overall completion of a task*. These smells are typically related to the general structure and flow of tasks within the user interface and include Laborious Task, Cyclic Task, Too Many Layers, Missing Task Feedback, High Interaction Distance, Repetition in Text Fields, and Late Validation. **Action-Level** focus on specific user actions and include Undescriptive Element, Missing Action Feedback, Unnecessary Action, and Misleading Action.[10]

The evaluation was conducted by 20 experts with varying levels of experience in usability assessment. The most experienced group consisted of four experts with over seven years of experience in usability evaluation. Additionally, five experts had more than four years of experience. One expert had three years of experience, and another expert had two years of experience. Four experts had one year of experience. Finally, five experts participated without prior practical experience but with theoretical knowledge of usability evaluation.

Each expert received an annotation guide and a link to the online annotation tool for each assigned annotation. This tool allowed them to review screen recordings of interactions and annotate perceived usability smells. Each interaction was evaluated by three experts, with different experts assigned to distinct users. To maximize the identification of usability issues, we rotated experts across different evaluation trios rather than using

[9]The full list of tasks and the full dataset structure are available in https://github.com/flassantos/amused/tree/main/supplementary.

[10]A full description of all usability smells is available at: https://github.com/flassantos/annotation-usability-smells/wiki/Usability-Smells.

TABLE III
SMELLS STATISTICS

	SN1	SN2	SN3	Total	
Task-level Smells					
Laborious task	326	145	44	515	(0.27)
Cyclic task	114	25	7	146	(0.08)
Too many layers	103	51	31	185	(0.10)
Missing task feedback	371	96	31	498	(0.27)
High interaction distance	58	69	22	149	(0.08)
Repetition in text fields	71	5	1	77	(0.04)
Late validation	234	48	21	303	(0.16)
Overall	1277	439	157	1873	(1.00)
	(0.68)	(0.23)	(0.08)		
Action-level Smells					
Undescriptive element	212	192	45	449	(0.13)
Missing action feedback	633	458	78	1169	(0.35)
Unnecessary action	65	75	14	154	(0.05)
Misleading action	987	526	51	1564	(0.47)
Overall	1897	1251	188	3336	(1.00)
	(0.57)	(0.37)	(0.06)		

fixed teams. To ensure a high level of expertise in the annotation process, each evaluation trio included at least one expert with seven years or more of experience in usability evaluation.[11] For task-level smells, when identified in a task, the expert should select the corresponding checkbox. For action-level smells, when identified, the experts should record the time (in minutes and seconds) when the action occurred, according to the interaction video's timestamp. Multiple occurrences of the same action smell can be identified within a single task.

We provide a comprehensive overview of the usability annotation in Table III, where we analyze the occurrence of both task and action-level usability smells across the three social networks. The distribution of usability smells at task and action-level aligns with the overall event distribution across social networks. Notably, SN1, which logged the majority of events (58%), also contributed the highest number of task-level (68%) and action-level smells (57%), reflecting the prominence of user effort and feedback issues in this network. At the task level, the most frequent issues are Laborious Task and Missing Task Feedback, each accounting for 27% of the total task-level smells. These are followed by Late Validation (16%) and Too Many Layers (10%). Less frequent task-related smells include Repetition in Text Fields (4%), Cyclic Task (8%), and High Interaction Distance (8%). At the action level, Misleading Action was the most frequently encountered smell, with a total of 1564 occurrences, making up 47% of all action-level smells, followed by Missing Action Feedback (35%), Unnecessary Action (5%), and Undescriptive Element (13%).

A. Annotation Agreement

We start by computing the Fleiss' kappa metric [24] for measuring the annotator agreement of binarized labels, which represent whether an event was marked with a smell or not. However, since our annotation consists of assigning a set of smells for an event (e.g., for action-level, an event can be considered both as "unnecessary" and "misleading"), we also compute the annotation agreement using two standard metrics: the weighted agreement percentage (intersect-over-union, IOU), and the inter-rater agreement with Krippendorff's alpha [25]. The results are illustrated in Table IV. For task-level smells, we only consider URL change events, while for action-level we consider all other events. Therefore, most action-level events receive a "null" label, indicating that none of the annotators detected a smell for that specific interaction.

[11]The annotation tool is available at: https://github.com/flassantos/annotation-usability-smells/.

TABLE IV
AGREEMENT BETWEEN ANNOTATORS IN TERMS OF FLEISS' KAPPA (F'κ), INTERSECT-OVER-UNION (IOU), AND KRIPPENDORFF'S ALPHA (K'α)

	Task-level Smells			*Action-level Smells*		
	F'κ	**IOU**	**K'α**	**F'κ**	**IOU**	**K'α**
SN1	0.40 ± .19	0.40 ± .12	0.32 ± .14	0.28 ± .13	0.88 ± .04	0.26 ± .12
SN2	0.33 ± .16	0.52 ± .14	0.21 ± .11	0.18 ± .09	0.78 ± .05	0.14 ± .07
SN3	0.30 ± .19	0.59 ± .16	0.19 ± .14	0.15 ± .08	0.94 ± .02	0.11 ± .07
All	0.36 ± .19	0.47 ± .15	0.26 ± .14	0.22 ± .12	0.84 ± .07	0.19 ± .12

Following Landis [26], our κ values indicate a fair agreement between annotators for the binary case. For the multilabel case, the IOU values suggest a moderate level of agreement, typically falling between 0.4 and 0.6. Moreover, as expected, Krippendorff's alphas are usually lower, as this metric is more punitive. However, they remain positive, indicating a modest level of reliability. We note that the agreement is higher for task-level smells compared to action-level ones, underscoring the inherent subjectivity of the action-level events, which are not only less frequent in the dataset but also more susceptible to interpretation challenges. Moreover, for action-level events, annotators must identify a specific set of smells and pinpoint the exact timestamp of the "bad" event, which demands a high level of precision and increases the complexity of the annotation process.

To integrate the annotated usability smells into the dataset, we structured them alongside user interaction logs and physiological data. We detail this preprocessing in the next section.

V. DATASET CONSTRUCTION

The AMUSED dataset integrates multiple data sources, including user interaction logs, physiological signals (EEG and BVP), facial emotion features, and self-reported emotions. These data streams were synchronized and preprocessed to ensure consistency, enabling a comprehensive analysis of user interactions and emotional states during tasks on social networks. In Fig. 2 we provide an overview of the data alignment process, illustrating how different data sources—interaction logs, physiological signals, facial expressions, and self-reported emotions—are structured and synchronized. Each component (A–D) represents a key aspect of dataset construction, which will be further detailed in the following subsections.

A. User Interaction Logs

We filtered the log, which is in JSON format, to consider only five event types: *URL change*, *click*, *input change*, *keypress*, and *scroll*. Each event is composed of a timestamp, current URL,

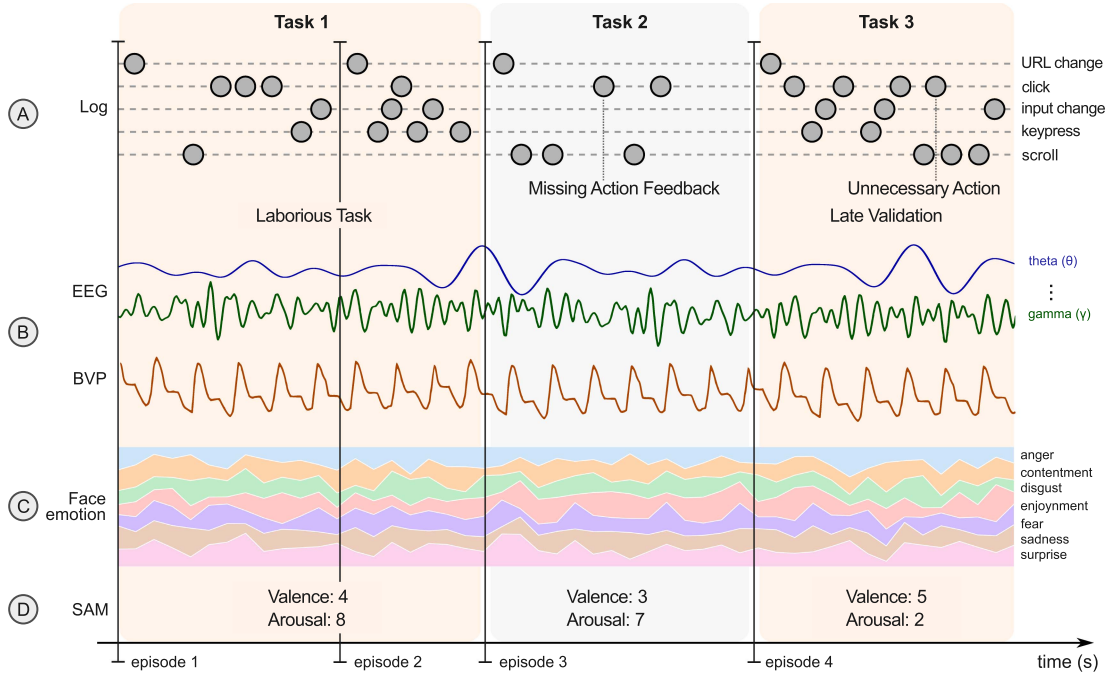

Fig. 2. Illustration of the data alignment process along the interaction timeline. Part (A) shows how the interaction log is segmented into tasks and episodes, with usability smells marked at both task (across episodes) and action levels (single events inside an episode). Part (B) aligns EEG and BVP signals to episodes, capturing users' physiological responses over time. Part (C) depicts facial emotion recognition data synced to the same timeline, and Part (D) shows the corresponding SAM (valence and arousal) reports for each task.

DOM information, and additional event-specific metadata, such as the x and y coordinates for *clicks*, duration for *scrolls*, and the text typed by the user for *input changes*. We further structured the log into **episodes**, which consists of a sequence of events tied to a particular URL, representing distinct user interactions on a specific webpage. An episode starts with a URL change and may include various other types of events. Therefore, the full user interaction is a collection of episodes, and a task is composed of one or more episodes, as illustrated in Fig. 2 (Part A). Importantly, **all episodes** in a task are labeled with **task-level smells**, while **a single event** in an episode is labeled with an **action-level smell**. Finally, to simplify the structure of our dataset, we collapsed successive *keypress* and *scroll* events into a single event, registering the total number of collapsed events as metadata. Note that the content typed by the user in input elements is still registered via the *input change* event.

B. EEG and BVP Signals

Despite having forehead sensors, BITalino outputs a single array describing the EEG signal and another single array describing the BVP signal, both with a sampling rate of 1,000 Hz (see Section III-B for more details). The EEG and BVP signals were normalized between -1 and 1 to standardize the data and reduce potential variability across different sessions or subjects. Given the structured log, we delimited both signals per episode and stored them in isolated arrays. Next, for each episode, we extracted specific frequency bands (theta, alpha low, alpha high, beta, and gamma) using the BioSPPY[12] library for EEG signals, as these frequency bands are known to relate to different cognitive and emotional states [21], making them relevant for subsequent analyses. Fig. 2 (Part B) illustrates how we structured EEG and BVP signals by aligning them with the logs.

C. Facial Expressions

Facial emotion recognition involves detecting and interpreting emotions expressed through facial expressions. For our study, we focused on using techniques that analyze the spatial features of the face to classify these emotions effectively. In order to identify faces within each frame, we employed a pre-trained model based on the widely used Viola-Jones method [27] with a rate of 30 frames per second, allowing us to segment the video into individual facial images for further analysis. Each detected face was pre-processed by converting it to grayscale, resizing it to a standard 48x48 pixel format, and adjusting the contrast to improve facial feature visibility.

Then, we apply VGG16 [28], a pre-trained Convolutional Neural Network (CNN) model, which enjoys high accuracy for image recognition tasks [29]. After finetuning the model on the undersampled Facial Expression Recognition 2013 (FER-2013) dataset [30], it achieved a validation accuracy of 84.6%, indicating a strong performance in classifying facial emotions into one of the seven Ekman categories [31]: joy, sadness, anger, surprise, fear, disgust, or contempt. As before, we aligned the recognized emotions for each frame according to the logs, such

[12]https://biosppy.readthedocs.io/en/latest/biosppy.html

that all events between frames A and B were marked with the emotions detected in frame A. We illustrate this synchronization in Fig. 2 (Part C).

D. Self-Assessment Manikin (SAM)

For a subset of the experiments, we asked participants to complete a shortened version of the SAM questionnaire [18], where they provided self-reports on valence and arousal experienced during task execution. This procedure occurred at the end of each task, capturing participants' emotional states at specific moments. Since self-reports were collected at the task level (e.g., Task 1), we applied the same valence and arousal values to all constituent episodes of that task (e.g., episodes 1 and 2). This approach enabled emotional data integration at the episode level despite the task-level granularity of the original assessments. Fig. 2 (Part D) illustrates the temporal alignment of SAM data with the task logs.

E. Ground Truth Labels

As discussed in Section IV-A, each expert marked events with a set of possible usability smells. Since annotators may agree on a subset of those and disagree on others, we take the union set from all three annotators to create the final ground truth label. This creates a ground truth in a "multilabel" format. For example, if expert 1 marked `{A, B}`, expert 2 `{A}`, and expert 3 `{B, C}`, the final label will be `{A, B, C}`. Therefore, after the individual evaluations, the results from all the experts were consolidated into a single list of issues, undergoing steps of grouping and duplicate removal. Furthermore, to simplify the process of identifying usability smells, we also generate a binary ground truth label, by assigning zero (`0`) if the multilabel ground truth is the empty set, and one (`1`) otherwise.

VI. Dataset Analysis

In this section, we provide an analysis of the dataset, focusing on relationships between user interactions, emotional responses, and usability smells. Our analysis is divided into key areas, each exploring different aspects of the dataset.

A. Interaction Data

Our analysis starts with the interaction logs, which detail user behavior across different tasks and social networks. We focus on identifying and interpreting interaction patterns, which reveal unique behavioral insights. We detail the distribution of event types—click, URL change, scroll, input change, and keypress—across all three social networks (SN1, SN2, SN3) and the overall dataset in Table V.

A total of 20,050 events were captured across the social networks, with SN1 contributing for the majority (58%), followed by SN2 (28%) and SN3 (14%). Click events were the most frequent, accounting for 52% of all recorded interactions, representing user interactions with buttons, links, or other clickable elements. The second most frequent event was URL change, indicating user navigation within the social network.[13] These results highlight a strong click-based interaction pattern on these platforms, supplemented by navigation actions (e.g., URL changes and scrolling) and limited form-based inputs. The low frequency of keypress events suggests that textual inputs are infrequent, hinting at a user behavior focused more on browsing and selection rather than text-based interactions. This pattern underscores potential areas for interface optimization, such as improving click-based functionalities and streamlining navigation to align with predominant user behaviors.

TABLE V
Event Type Statistics in the AMUSED Dataset

Event Types	SN1		SN2		SN3		Total	
URL change	2106	(0.18)	1522	(0.27)	422	(0.15)	4050	(0.20)
Input change	1877	(0.16)	416	(0.07)	281	(0.10)	2574	(0.13)
Click	5450	(0.47)	3027	(0.54)	1897	(0.67)	10374	(0.52)
Scroll	2077	(0.17)	603	(0.11)	184	(0.06)	2864	(0.14)
Keypress	105	(0.01)	23	(0.00)	60	(0.02)	188	(0.01)
Overall	11615	(0.58)	5591	(0.28)	2844	(0.14)	20050	(1.00)

Building on this analysis of event types, Fig. 3 (a) shifts focus to the duration of these interactions by examining the distribution of episode durations across the same three social networks. As illustrated in the figure, user engagement tends to be short-lived, with episode frequency peaking within the first few seconds of interaction and declining after around 15 s across all platforms. SN1 exhibits the longest sustained engagement, with a peak of nearly 1,000 around the 10 s mark, followed by SN2, which peaks slightly above 800 at around 6 s. SN3, in contrast, shows a much lower peak frequency, at approximately 100 around 12 s, then gradually decreasing to almost zero by 100 seconds. The distribution suggests that users on these platforms may quickly browse or scan through content, with limited sustained interactions. This could reflect design elements that encourage quick scanning rather than prolonged engagement in each episode.

Fig. 3 (b) delves deeper into event duration by examining how long each event type persists across networks. Click events remain consistently brief and frequent across all networks, typically lasting less than ten seconds, aligning with their role as quick, repeated actions. Both URL change and scroll events display short durations, concentrated at the lower end of the time scale, with URL changes in SN1 lasting less than one second. Notably, while input change events have a lower frequency than scroll events in all networks, they have a higher duration, suggesting a consistent trend where users engage in form-based tasks for longer than they scroll. This duration trend of input changes, especially on SN1, indicate that input tasks demand more user engagement time. Moreover, the consistent pattern of input changes outpacing scroll events highlights an area for potential UI improvements. Reducing the need for frequent scrolling could streamline the user experience while optimizing form-based interactions for efficiency could help minimize time

[13] Scroll and keypress events are reported as "collapsed" occurrences rather than their total count during user interactions, as described in Section V.

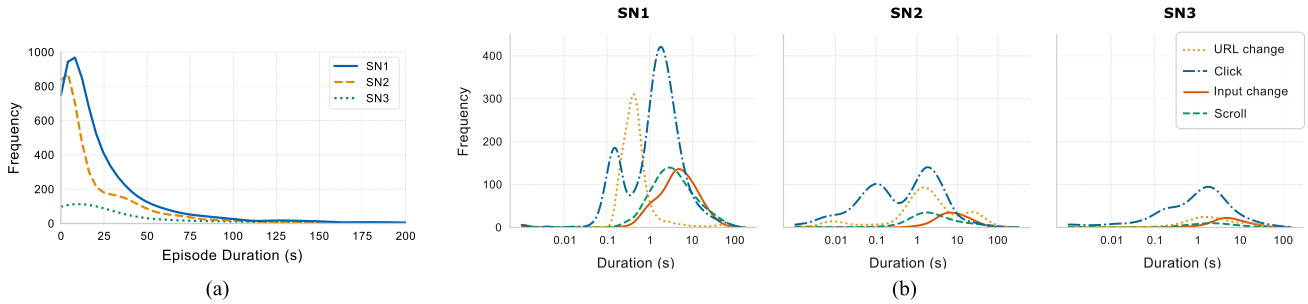

Fig. 3. Duration distribution of episode (a), and event types (b), per Social Network.

spent on input tasks. Such adjustments can improve user experience by supporting more intuitive and efficient interactions across these platforms.

B. Usability Smells Data

Table III summarizes the distribution of usability smells at both task and action levels across the three social networks. SN1 exhibits the highest share of usability smells, with 68% of task-level smells and 57% of action-level smells, aligning with its overall dominance in event activity (58%). SN2 follows with 23% of task-level smells and 37% of action-level smells, while SN3 accounts for the smallest share, with 8% of task-level smells and 6% of action-level smells. Analyzing the distribution of usability smells across individual networks reveals that certain issues are prevalent in all of them. Notably, Laborious Task consistently emerges as the most frequent smell, especially in SN2, where it indicates significant user effort. Missing Task Feedback follows closely, highlighting problems related to the lack of guidance or confirmation during tasks. This pattern suggests that users often face incomplete feedback loops and high-effort tasks, potentially leading to frustration and inefficiency. In contrast, Repetition in Text Fields is the least frequent smell in SN2, SN3, and the overall distribution. These issues provide valuable design insights, emphasizing the need to improve task feedback mechanisms and minimize task complexity.

At the action level, although the number of identified smells varies across the three social networks, Misleading Action and Missing Action Feedback are consistently prevalent in all of them. In contrast, Unnecessary Action remains consistently infrequent across the networks. The high frequency of Misleading Action suggests that users are often confused or misled by the interface, indicating that improving the clarity and intent of actions could significantly improve usability. Similarly, the prevalence of Missing Action Feedback underscores the need for more immediate and informative responses to user actions, ensuring that users are aware of the outcomes and can interact with the system more effectively.

Following our analysis of individual usability smells, we now turn our attention to the co-occurrence of these smells at both task and action levels. Fig. 4 provide insights into how frequently different usability smells appear together, offering a deeper understanding of potential compounding effects that could exacerbate user frustration or confusion.

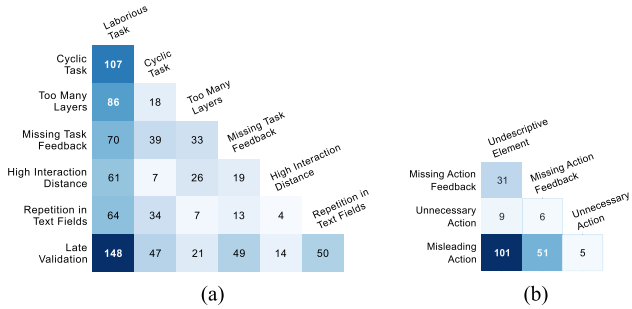

Fig. 4. Task-level (a) and action-level smells (b) co-occurrence heatmap.

In Fig. 4(a), we observe the pairwise co-occurrence of task-level usability smells across the social networks. The smell with the highest co-occurrence is Laborious Task, which frequently appears in combination with other smells. The most prominent pairing is Laborious Task and Late Validation, co-occurring 148 times, reflecting the compounding effect of user effort and delayed system responses. Similarly, Laborious Task and Cyclic Task appear together 107 times, suggesting that tasks requiring repetitive steps are often associated with excessive laboriousness. The co-occurrence of less frequent smells, such as Repetition in Text Fields and High Interaction Distance, although lower, highlights specific interaction problems that, when combined, could significantly hinder task completion. The presence of these patterns can inform where improvements in user experience design could alleviate multiple usability problems at once.

In Fig. 4(b), we focus on the co-occurrence of action-level usability smells. The most frequent pair is Misleading Action and Undescriptive Element, which appear together 101 times. This pairing suggests that users often encounter unclear or misleading interface elements, leading them to perform actions that result in unintended outcomes. Such a scenario creates a significant barrier to effective interaction, as users may struggle to navigate or complete tasks due to confusion. On the other hand, the co-occurrence of Misleading Action and Unnecessary Action is much lower. This indicates that while users may be misled by certain actions, they are less likely to perform entirely redundant actions during these interactions.

By examining these pairwise relationships, we gain deeper insights into how specific interaction patterns amplify usability issues. Understanding these combinations allows for more precise and targeted design interventions that can address multiple

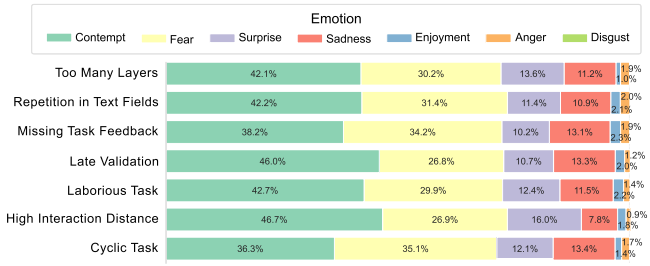

Fig. 5. Facial emotion distribution across task-level usability smells, highlighting the emotional responses triggered by each usability issue.

problems simultaneously, ultimately improving the overall user experience by reducing confusion and inefficiency.

C. Emotional Data

Our emotional data analysis starts with the distribution of recognized facial emotions in response to different usability task smells. This emotional data offers insight into user responses when encountering specific interface issues, allowing us to gauge which emotional reactions are most frequently associated with different types of task-level usability smells. The data presented in Fig. 5, summarized in a series of horizontally stacked bar charts, reveals the percentage distribution of seven primary facial emotions (anger, contempt, disgust, enjoyment, fear, sadness, and surprise) across task-level smells.

Contempt is the most frequently observed emotion across all task-level smells, particularly for those requiring high interaction distances, where it peaks at 46.7%. This suggests that users feel frustrated or undervalued when the interface requires excessive navigation or effort, indicating dissatisfaction with a lack of efficiency or ease in task completion. Fear is the second most prominent emotion across all task smells. The presence of fear could be attributed to uncertainty or confusion when interacting with complex or unintuitive interfaces, highlighting areas where clarity and user guidance are lacking. Surprise and sadness are present, but not dominant in the emotional profiles. Surprise may occur due to unexpected interface behavior or outcomes, while sadness could reflect disappointment with the user experience. Their lower percentages indicate these emotions are less commonly associated with the usability smells analyzed. Both enjoyment and anger show relatively low presence across task-level smells, with enjoyment barely reaching notable percentages. The low levels of enjoyment suggest that these usability issues rarely lead to positive user experiences. Similarly, the minimal presence of anger may indicate that while users feel frustrated (as reflected by contempt and fear), they do not necessarily experience intense irritation or hostility towards the interface. Disgust is notably absent or has a negligible presence in the emotional data for task-level smells. This absence may suggest that, although users are frustrated or anxious about usability issues, they do not feel a strong sense of repulsion or aversion toward the interface.

The emotional data analysis reveals that negative emotions like Contempt and Fear are predominant in response to task-level usability smells. The consistent presence of these emotions underscores the significant impact that usability issues have on user experience, inducing feelings of frustration.

Building upon the insights derived from facial emotion recognition, we further explore user emotional responses through the Self-Assessment Manikin (SAM) questionnaire. This additional data allows us to interpret users' self-reported emotional states along two core dimensions: arousal (activation level) and valence (positive or negative emotional tone). We specifically selected these two dimensions from the SAM framework due to their established correlations with physiological signals, including EEG and BVP, which were collected as part of our multimodal dataset. Although the complete SAM framework includes a dominance dimension, our focus on valence and arousal is motivated by the potential to map these emotional dimensions directly to neural oscillations and cardiovascular patterns observed in the physiological data [32]. The SAM questionnaire results can be plotted on a Cartesian plane, displaying data points representing user responses to the tasks plotted along the valence (x) and arousal (y) axes. Each point's position on the graph helps characterize the emotional tone associated with specific tasks, categorized as follows [33], [34]:

- *High valence, high arousal:* corresponds to excitement. Emotions here suggest that users found these tasks stimulating and enjoyable;
- *High valence, low arousal:* indicates calmness or satisfaction. Emotions here suggest contentment or a relaxed state, implying that users felt positively engaged without being overstimulated;
- *Low valence, high arousal:* represents distress or anxiety. Emotions here indicate that users experienced a high level of activation with a negative emotional tone, potentially indicating frustration or discomfort with the task;
- *Low valence, low arousal:* corresponds to boredom or sadness. Emotions here reflect a lack of interest or negative feelings with low energy, suggesting user disengagement or dissatisfaction.

VII. Automatic Detection of Usability Smells

Next, we describe our approach to automatically detect usability smells in user interactions using Machine Learning models, along with reporting and discussing our results.

A. Experimental Setup

We formulate the detection of usability smells as a supervised Machine Learning problem. Two types of classification scenarios were addressed: binary classification (presence or absence of usability smells) and multilabel classification (simultaneous detection of multiple usability smells). Additionally, we engineered new features and performed normalization to improve the model's ability to identify usability smells.

1) Features and Feature Engineering: Our dataset contains event-specific features (e.g., event durations, scroll counts), structural HTML features (e.g., XPath depth), and aggregated measures at the episode level. Additional features include physiological (EEG and BVP) and emotional signals (facial expressions and SAM). We carried out feature engineering in order to generate global aggregation features (e.g., max, min, mean)

and local contextual features (e.g., values for preceding and succeeding episodes). To ensure consistency across sessions, we applied Z-score and min-max normalization to numerical features.[14]

2) Classification Tasks: We address the detection of usability smells at two levels, **task-level classification** for predicting usability smells for entire episodes, and **action-level classification** for predicting usability smells for individual actions. We perform binary classification and multilabel classification at both levels.

3) Classifiers: Ideally, our classifiers should consider the entire structure of the input, meaning that the model would consider the full timeline illustrated in Fig. 2. However, in this study, we relax this assumption and choose simpler yet fast and interpretable classifiers that can use global features from the entire episode or from neighboring events/episodes. Concretely, we perform feature engineering to consider neighboring events/episodes and experiment with Logistic Regression, K-Nearest Neighbors (KNN), Linear SVM, Multi-Layer Perceptron (MLP), Random Forest, and Gradient Boosting classifiers, which were implemented using the scikit-learn library [35]. We carry out hyper-parameter tuning with Grid Search using the macro F1-score as the optimization metric. For multilabel classification, we use wrappers, such as `MultiOutputClassifier` and `OneVsRestClassifier` to adapt classifiers to predict a set of labels. Finally, we also include a chance baseline that always predicts the majority label for the binary case and a classifier that generates random predictions uniformly for the multilabel setting.

4) Evaluation Metrics: We evaluate the classifiers using 5-fold cross-validation to ensure a robust performance assessment. For both binary and multilabel classification tasks, we use multiple evaluation metrics to capture various performance aspects. These include accuracy, precision, recall, and the F1-score, the latter reported as macro average to better account for class imbalance. Additionally, for the multilabel setting, we also report Jaccard and Hamming scores [36].

B. Results and Discussion

1) Task-Level Classification: We present the results in Table VI. Overall, the Gradient Boosting classifier consistently outperforms other classifiers across all metrics in both binary and multilabel settings. Specifically, in the binary case, Gradient Boosting achieves the highest accuracy (92.1%), precision (90.3%), recall (82.5%), and F1-score (86.2%). These results highlight its ability to detect usability smells with very high accuracy while balancing false positives and false negatives. Simpler models like Logistic Regression, KNN, and LinearSVM perform better than the baseline but show limitations, particularly in precision and recall. For the multilabel classification setting, Gradient Boosting achieves a strong F1-score (91.3%) and Jaccard score (83.9%). As before, simpler models struggle. Overall, the gap between simple and advanced classifiers highlights the need for more sophisticated techniques that can not only detect the presence of usability smell, but also distinguish their specific types.

TABLE VI
TASK-LEVEL RESULTS FOR DIFFERENT CLASSIFIERS IN TERMS OF ACCURACY (ACC.), PRECISION (P), RECALL (R), F1 SCORE, EXACT MATCH (EM), JACCARD SCORE (JAC.), AND HAMMING SCORE (HAM.). FOR ALL METRICS, THE HIGHER THE BETTER. **BOLD** REPRESENTS TOP RESULTS; UNDERLINE REPRESENTS SECOND-BEST.

Classifier	*Binary*			
	Acc.	**P**	**R**	**F1**
Baseline	0.812	0.406	0.500	0.448
K-Nearest Neighbor	0.838	0.756	0.628	0.686
Logistic Regression	0.743	0.654	0.723	0.687
Linear SVM	0.820	0.777	0.522	0.624
Multi-Layer Perceptron	0.859	0.774	0.737	0.755
Random Forest	0.876	0.885	0.688	0.774
Gradient Boosting	**0.921**	**0.903**	**0.825**	**0.862**

Classifier	*Multilabel*					
	EM	**P**	**R**	**F1**	**Jac.**	**Ham.**
Baseline	0.009	0.285	0.503	0.364	0.211	0.502
K-Nearest Neighbor	0.400	0.676	0.564	0.615	0.441	0.809
Logistic Regression	0.122	0.462	0.696	0.555	0.377	0.710
Linear SVM	0.130	0.469	0.686	0.557	0.381	0.718
Multi-Layer Perceptron	0.362	0.679	0.655	0.667	0.504	0.833
Random Forest	0.567	0.881	0.722	0.794	0.655	0.901
Gradient Boosting	**0.776**	**0.952**	**0.877**	**0.913**	**0.839**	**0.957**

TABLE VII
ACTION-LEVEL RESULTS FOR DIFFERENT CLASSIFIERS IN TERMS OF ACCURACY (ACC.), PRECISION (P), RECALL (R), F1 SCORE, EXACT MATCH (EM), JACCARD SCORE (JAC.), AND HAMMING SCORE (HAM.). FOR ALL METRICS, THE HIGHER THE BETTER. **BOLD** REPRESENTS TOP RESULTS; UNDERLINE REPRESENTS SECOND-BEST.

Classifier	*Binary*			
	Acc.	**P**	**R**	**F1**
Baseline	0.856	0.428	0.500	0.461
K-Nearest Neighbor	0.842	0.664	0.628	0.642
Logistic Regression	0.861	0.716	0.581	0.641
Linear SVM	0.856	0.428	0.500	0.461
Multi-Layer Perceptron	0.841	0.672	**0.660**	0.665
Random Forest	0.873	**0.771**	0.614	0.684
Gradient Boosting	**0.876**	0.766	0.651	**0.704**

Classifier	*Multilabel*					
	EM	**P**	**R**	**F1**	**Jac.**	**Ham.**
Baseline	0.062	0.042	**0.491**	0.077	0.039	0.502
K-Nearest Neighbor	0.836	0.238	0.114	0.154	0.085	0.953
Logistic Regression	0.856	0.302	0.046	0.080	0.042	0.958
Linear SVM	0.832	0.150	0.064	0.090	0.046	0.957
Multi-Layer Perceptron	0.815	0.234	0.197	0.214	0.125	0.946
Random Forest	**0.862**	0.375	0.078	0.130	0.071	**0.961**
Gradient Boosting	0.860	**0.390**	0.161	**0.228**	**0.135**	0.960

2) Action-Level Classification: Table VII illustrates the results for action-level usability smell detection. In the binary classification setting, the Gradient Boosting classifier once again delivers the best overall performance, with an accuracy of 87.6% and an F1-score of 70.4%. However, an MLP gets a higher recall (66.0%), suggesting it is better at capturing more instances of usability smells, even if it introduces more false positives. Linear models lag behind, indicating that finer-grained detection requires deeper methods. Finally, we observe that multilabel predictions at the action level pose a greater challenge, as reflected in the consistently lower F1 scores across all models. We attribute this to three key factors. First, the action-level dataset

[14] A full description of features and preprocessing steps is provided at: https://github.com/flassantos/amused/.

TABLE VIII
TOP 10 GRADIENT BOOSTING FEATURES FOR TASK AND ACTION-LEVEL SMELLS

Task-level	
Binary	**Multilabel**
Relative Task Distance	Relative Task Distance
Input Change Text Length	Click Text Length
Relative Episode Distance	Input Change Text Length
Finished Task Indicator	Num. of URL Changes
Num. of Tags	Finished Task Indicator
URL Change Duration	Relative Episode Distance
Average HTML Distance	SAM: Arousal
Facial Emotion: Surprise	Facial Emotion: Enjoyment
SAM: Valence	Facial Emotion: Disgust
Episode Duration	SAM: Valence
Action-level	
Binary	**Multilabel**
DOM Object	DOM Object
Num. of Input Changes	Click Text Length
Relative Episode Distance	Click XY Coordinates
Event Duration	Event Duration
Click Text Length	Relative Task Distance
Num. of Unique XPath Tags	Relative Episode Distance
Click XY Coordinates	Num. of Tags
Click Text	Num. of Unique XPath Tags
Most Frequent Facial Emotion	Relative Event Distance
Num. of Tags	Average HTML Distance

is highly imbalanced, with non-empty action smells comprising only 16% of the dataset (3,336 out of 20,050 instances). Second, as shown in Table IV, annotator agreement is lower at the action level ($K'\alpha = 0.19$) compared to the task level ($K'\alpha = 0.26$), indicating an inherently more difficult classification problem. Finally, accurately predicting action-level smells requires capturing contextual information preceding the event in which the smell occurs. This necessitates more sophisticated models that are able to account for the sequential structure of the input, such as Recurrent Neural Networks [37]. Given the strong performance of our current classifiers in the binary setting, we defer the analysis of sequence models for multilabel prediction to future work. Next, we examine the most important features and error cases of the Gradient Boosting classifier, our best-performing model.

3) Feature and Error Analysis: We computed feature importance scores from the Gradient Boosting classifier by averaging the importance values across all individual estimators of our 5-fold cross-validation. The top ten features, ranked by their normalized importance, are shown in Table VIII. At the task level, contextual features like Relative Task Distance and Relative Episode Distance consistently rank high, highlighting the role of spatial and temporal dynamics in user interactions. The length of text in clicks and input changes also receive high importance scores, which reflect the complexity and cognitive effort required for interactions. Additionally, emotional signals such as SAM (valence and arousal) and facial emotions highlight the value of incorporating user emotions into usability evaluation. For action-level smells, interface-specific features such as the DOM object, click-related information, and XPath tags dominate. Interestingly, these features capture critical information about how users engage with specific elements, such as buttons and forms, where action-level usability issues are likely to occur. In turn, this reflects the importance of structural and semantic properties of UI elements to minimize confusion and ensure harmonious interactions.

TABLE IX
GRADIENT BOOSTING'S PERFORMANCE PER USABILITY SMELL ALONG WITH THE TOP 2 CO-OCCURRING LABELS

Smell	**F1**	**Top 2 Co-occurring Smells**
Task-level		
Laborious Task	0.954	Missing Task Feedback, Late Validation
Cyclic Task	0.922	Laborious Task, Missing Task Feedback
Too Many Layers	0.888	Laborious Task, Missing Task Feedback
High Inter. Distance	0.879	Laborious Task, Too Many Layers
Repet. in Text Fields	0.904	Laborious Task, Late Validation
Missing Task Feedback	0.915	Laborious Task, Late Validation
Late Validation	0.927	Laborious Task, Missing Task Feedback
Action-level		
Unnecessary Action	0.063	Misleading Action, Undescriptive Element
Misleading Action	0.320	Missing Act. Feedback, Unnec. Element
Missing Act. Feedback	0.443	Misleading Action, Undescriptive Element
Undescriptive Element	0.061	Missing Act. Feedback, Misleading Action

Building on the previous insights, Table IX presents a detailed analysis of Gradient Boosting's classification performance for each usability smell, highlighting also the most frequent co-occurring smells. Task-level smells such as Laborious Task and Missing Task Feedback achieve high precision and recall, indicating the model's robust ability to identify these prominent issues. However, we note a slightly lower recall for smells like Too Many Layers, suggesting more difficulties in capturing subtler patterns. Co-occurrence analysis reveals significant overlaps between smells, such as Laborious Task frequently pairing with Missing Task Feedback and Late Validation, which underscores the compounding nature of usability problems. Interestingly, the top 2 co-occurring smells are closely aligned with the ground-truth co-occurrence ranking illustrated in Fig. 4, suggesting that our classifier was able to capture well the relation between usability smells. At the action level, the model identifies Misleading Action and Missing Action Feedback effectively but struggles with more rare smells like Unnecessary Action and Undescriptive Element (*c.f.* Table III), which is expected, since action-level smells present greater challenges due to their granular nature and reliance on contextual data.

VIII. FINAL REMARKS

We introduced AMUSED, a comprehensive multimodal HCI dataset annotated with usability smells within social networks. By integrating interaction logs, physiological signals, facial emotion features, and self-assessment reports, we provided a rich resource for studying the interplay between user behavior, emotional responses, and usability issues. Our findings demonstrated that usability smells often co-occur and can impact user experience, with emotions like contempt and fear frequently associated with problematic interactions. Additionally, our proposed Machine Learning approach achieved strong performance in detecting usability smells, particularly at the task level, highlighting the potential of computational methods in usability evaluation. However, challenges remain, particularly in action-level detection and the integration of methods that

can leverage the sequential structure of the data. Future work should focus on expanding the dataset's cultural and application scope and leveraging more sophisticated sequence-based models to further improve the automatic detection of usability smells.

Our dataset was exclusively gathered from Portuguese speakers, which may limit the generalizability of the findings across different cultural and linguistic contexts. For example, cultural factors may influence how users express emotions and perceive usability issues, limiting cross-cultural applicability. Additionally, the dataset is heterogeneous, as not all participants contributed the same types of data; some provided only interaction logs, while others included EEG signals, facial images, and SAM reports. While this diversity enriched the dataset, it introduced challenges in maintaining uniformity and consistency during analysis. Furthermore, the scope of this work was restricted to social networks with usability issues, which limits the applicability of the findings to other systems with higher usability standards.

Challenges also arose during the annotation process, particularly at the action level. Task-level annotations benefited from well-defined tasks within the evaluated Websites, leading to higher consistency among annotators. However, action-level annotations—such as identifying misleading clicks—required experts to pinpoint exact timestamps, often differing by only a few seconds. This margin for alignment error introduces variability in the dataset and highlights the inherent difficulty of annotating granular usability issues.

Lastly, the automatic detection methods did not fully leverage the sequential nature of the data. While our approach achieved high scores for task-level classification and binary action-level detection, incorporating sequential models could improve performance, especially for multilabel action-level classification. Sequential analysis could also better capture temporal patterns in EEG and BVP signals, providing a more comprehensive understanding of action-level usability smells. Thus, future work includes experimenting with LLM-based approaches to more comprehensively validate our dataset.

REFERENCES

[1] A. Alslaity and R. Orji, "Machine learning techniques for emotion detection and sentiment analysis: Current state, challenges, and future directions," *Behav. Inf. Technol.*, vol. 43, no. 1, pp. 139–164, 2024.

[2] X. Chen, R. Huang, X. Li, L. Xiao, M. Zhou, and L. Zhang, "A novel user emotional interaction design model using long and short-term memory networks and deep learning," *Front. Psychol.*, vol. 12, 2021, Art. no. 674853.

[3] R. A. C. Xavier and V. P. de Almeida Neris, "Measuring users' emotions with a component-based approach," in *Proc. 14th Int. Conf.*, 2013, pp. 393–409.

[4] A. A. Syahidi and K. Kiyokawa, "Evaluation of user interface, user experience, and usability in software through electroencephalography (EEG) signal detection: A mapping review," in *Proc. IEEE Int. Workshop Artif. Intell. Image Process.*, 2023, pp. 133–138.

[5] C. Bellos, K. Stefanou, A. Tzallas, G. Stergios, and M. Tsipouras, "Methods and approaches for user engagement and user experience analysis based on electroencephalography recordings: A systematic review," *Electronics*, vol. 14, no. 2, 2025, Art. no. 251.

[6] F. de Souza Santos, M. V. Treviso, S. P. Gama, and R. P. de Mattos Fortes, "A framework to semi-automated usability evaluations processing considering users' emotional aspects," in *Proc. Int. Conf. Hum.- Comput. Interact.*, 2022, pp. 419–438.

[7] R. Torres-Molina and M. Seyam, "The intersection of usability evaluation and machine learning in software systems," in *Proc. IEEE 5th Int. Conf. Cogn. Mach. Intell.*, 2023, pp. 122–127.

[8] A. Garrido, G. Rossi, and D. Distante, "Refactoring for usability in Web applications," *IEEE Softw.*, vol. 28, no. 3, pp. 60–67, May/Jun. 2011.

[9] F. Paternò, A. G. Schiavone, and A. Conti, "Customizable automatic detection of bad usability smells in mobile accessed web applications," in *Proc. 19th Int. Conf. Hum.-Comput. Interact. Mobile Devices Serv.*, 2017, pp. 1–11.

[10] J. Grigera, A. Garrido, J. M. Rivero, and G. Rossi, "Automatic detection of usability smells in web applications," *Int. J. Hum.- Comput. Stud.*, vol. 97, pp. 129–148, 2017.

[11] V. Lelli, A. Blouin, B. Baudry, F. Coulon, and O. Beaudoux, "Automatic detection of GUI design smells: The case of blob listener," in *Proc. Symp. Eng. Interactive Comput. Syst.*, 2016, pp. 263–274.

[12] S. Koelstra et al., "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan.-Mar. 2011.

[13] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan.-Mar. 2012.

[14] A. Savran et al., "Emotion detection in the loop from brain signals and facial images," in *Proc. eINTERFACE'06-SIMILAR NoE Summer Workshop Multimodal Interfaces*, 2006, pp. 69–80.

[15] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 1, pp. 98–107, Jan. 2018.

[16] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-based multimodal database for decoding affect. physiological responses," *IEEE Trans. Affective Comput.*, vol. 6, no. 3, pp. 209–222, Jul.-Sep. 2015.

[17] D. Leppich, C. Bieber, K. Proschek, P. Harms, and U. Schubert, "Dux: A dataset of user interactions and user emotions," *I-COM*, vol. 22, no. 2, pp. 101–123, 2023.

[18] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Ther. Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[19] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual," University of Florida, Gainesville, Tech. Rep. A-8, 2008.

[20] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems," *Neuroimage*, vol. 34, no. 4, pp. 1600–1611, 2007.

[21] S. M. Alarcão and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 374–393, Jul.-Sep. 2019.

[22] BiosignalsPlux, "Electroencephalography (EEG) sensor user manual," 2021. Accessed: Jul. 07, 2025. [Online]. Available: https://support.pluxbiosignals.com/wp-content/uploads/2021/11/Electroencephalography_EEG_User_Manual.pdf

[23] E. P. Torres, E. A. Torres, M. Hernández-Álvarez, and S. G. Yoo, "Eeg-based BCI emotion recognition: A survey," *Sensors*, vol. 20, no. 18, 2020, Art. no. 5083.

[24] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, 1971, Art. no. 378.

[25] K. Krippendorff, "Computing krippendorff's alpha-reliability," Dept. Papers (ASC), Annenberg School Commun., Univ. Pennsylvania, Philadelphia, PA, USA, 2011. [Online]. Available: https://repository.upenn.edu/handle/20.500.14332/2089

[26] J. Landis, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, Mar. 1977.

[27] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2001, pp. I–I.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[29] A. A. Rodrigues, F. de Souza Santos, and S. P. Gama, "The emotionality tool: Evaluating usability with facial emotions analysis," *Int. J. Hum.–Comput. Interact.*, vol. 41, pp. 1–16, 2025.

[30] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in *Proc. 20th Int. Conf. Neural Inf. Process.*, 2013, pp. 117–124.

[31] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

[32] F. Galvão, S. M. Alarcão, and M. J. Fonseca, "Predicting exact valence and arousal values from EEG," *Sensors*, vol. 21, no. 10, 2021, Art. no. 3414.
[33] J. A. Russell, "A circumplex model of affect," *J. Pers. Social Psychol.*, vol. 39, no. 6, 1980, Art. no. 1161.
[34] M. Yik, J. A. Russell, and J. H. Steiger, "A 12-point circumplex structure of core affect," *Emotion*, vol. 11, no. 4, 2011, Art. no. 705.
[35] F. Pedregosa et al., "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[36] F. Herrera, F. Charte, A. J. Rivera, and M. J. D. Jesus, *Multilabel Classification*. Berlin, Germany: Springer, 2016.
[37] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

Flávia de S. Santos received the PhD degree from the University of São Paulo (USP), Brazil. Her research specializes in Human-Computer Interaction, with a particular emphasis on developing methods for usability and accessibility evaluation. She is affiliated with the Intermidia Lab (USP, Brazil) and the Human Factors in Interaction (HUMAN) Lab (INESC-ID, Portugal).

Marcos V. Treviso is an assistant professor at IST (University of Lisbon) and a researcher at IT, where his work centers on developing efficient and interpretable natural language processing (NLP) models.

Kamila R. H. Rodrigues is an assistant professor with ICMC/USP. She conducts research in the areas of Human-Computer Interaction, Interactive Multimedia Systems, and Serious Digital Games. She is head of the Intermidia research laboratory.

Renata P. M. Fortes is a senior professor with ICMC/USP, Brazil. Her research interests include Human-Computer Interaction, Usability Engineering, and Interactive Multimedia Systems, with a particular focus on methods to improve user interfaces.

Sandra P. Gama is an assistant professor with IST (University of Lisbon) and a senior researcher, INESC-ID. Sandra's research interests fall within the scope of Human-Computer Interaction, Information Visualization, and Digital Gamification. She is head of the Human Factors in Interaction (HUMAN) Lab at the Graphics and Interaction (GI) group at INESC-ID.

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - ROR identifier: 00x0ma614