Towards Detection of Usability Issues by Measuring Emotions

Elena Stefancova, Robert Moro[0000−0002−3052−8290], and Maria Bielikova

Slovak University of Technology in Bratislava,
Faculty of Informatics and Information Technologies,
Ilkovicova 2, 842 16 Bratislava, Slovak Republic
{xstefancovae, robert.moro, maria.bielikova}@stuba.sk

Abstract. User Experience is one of the most important criteria when designing and testing user interfaces with emotions as its essential element. To assess, how emotions could be used for automatic detection of usability issues, we carried out a user study with a website which included intentionally inserted usability issues. We classified valence of emotions, i.e., negative vs. positive ones based on data from electroencephalography (EEG) and facial expressions recognition. The study results confirmed that usability issues cause negative emotional response of the user and that presence of a negative emotion is a good predictor of a usability issue presence. When detecting negative and positive emotional states from the acquired dataset, we achieved the accuracy of 94% for samples with seconds granularity and 70% for the task granularity.

Keywords: usability, emotions, EEG, facial expressions, data analysis

1 Introduction and Related Work

User experience (UX) according to ISO-9241-210 is "a person's perceptions and responses that result from the use or anticipated use of a product, system or service". One of its factors is usability [9] which is "extent to which a product can be used by specific users to achieve specific goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO/TS 20282-2:2013). Usability of a website can be measured by how easily and effectively the user can browse and perform specific tasks on the website.

One of the components of UX are *emotions*, which can be measured by questionnaires or by various sensors [7, 9]. They are commonly represented by Ekman's model of six basic emotions (joy, fear, sadness, disgust, anger, and surprise) [3] or a dimensional approach [11], which distinguishes two main dimensions, namely *valence* (how positive the emotion is) and *arousal* (the strength of an emotion). When using sensors, emotions can be detected from many physiological responses, such as changes in blood pressure, skin conductance (GSR), or the character of brain waves (EEG) [5]. An interesting approach is to recognize emotions from facial expressions, since it does not require a special equipment

(a normal webcam is usually sufficient). The accuracy of facial expression recognition can near 90% [13], but the participant usually has to stay still during the measurement, otherwise the recording can be damaged [4]. In [7], Matlovic et al. showed that the accuracy for seven classes of emotions measured by EEG using EMOTIV Epoc device was 58%, while it was only 19% in case of facial expressions recognition software Noldus FaceReader. The problem with FaceReader probably was that they took into account only the dominant emotion.

The researchers have not found a connection between significantly positive emotions and a level of usability yet, but it seems that a bad usability level can result into negative ones [9], especially when the user performs a task with a specific goal. This field of study is still to be explored, because the majority of papers on detection of emotions are concerned with affective corpora, such as watching videos or other multimedia content [7], instead of the impact of a user interface. The interface gives to a user more freedom in interaction, thus resulting into more challenges in emotions detection.

One work that did focus on emotion detection in context of usability, was the work of Aggarwal et al. They used a combination of EEG data and facial expression recognition [1] and designed two websites with the same functionality, but the first one with good and the other one with poor level of usability, established according to Shneiderman's eight golden rules of interface design [12]. The results of their exploratory study showed that user experience indeed influences emotions; in this case, it was excitement and frustration. The frustration occurred mostly during interaction with the worse website.

In our paper, we aim to verify these findings on a larger user sample. Moreover, our primary hypothesis is that a low level of usability not only causes emotional response, but that negative emotional response of a user can help to automatically expose usability issues (no automatic classification was performed in the aforementioned work [1]). This is significant during development and testing of user interfaces. A reliable method of usability issues detection would be useful mainly in such UX testing setups, where it is possible to work with more participants at once and, thus, save significant amount of (recording as well as data analysis) time or for remote testing setups.

We proposed a method of binary emotions classification using a combination of features extracted from EEG and facial expressions analysis. For the evaluation of our proposed method, we used data from sensors and self-report surveys collected during a user study, in which participants had to fulfill tasks on a website, the interface of which was modified by adding usability issues.

2 Study Methodology

Our main aim was to find out, how usability can influence emotional state of a user and if emotional response is strong enough to detect issues of usability. We used EEG and facial expressions as data sources for emotions classification. The

study was conducted in the User Experience and Interaction Research Centre at the Slovak University of Technology in Bratislava[1].

For gaining EEG data we used Emotiv Epoc[2]. It is a wearable device, commonly used for interaction with applications. For analysis of facial expressions, we employed Noldus FaceReader[3], which provides basic emotion features (valence, arousal) for every frame and can recognize neutral, contempt, happy, sad, angry, surprised, scared, and disgusted facial expressions. It uses a common web camera as its input. Tobii Studio[4] was used for orchestration of our study.

2.1 Test Scenario

The participants performed eight tasks on a groupon-like website Zlava Dna[5], half of which contained an artificially added issue. The order of the tasks was counterbalanced (using *Wiliams Design*[6]). The issues were inserted by a web browser extension Greasemonkey[7]. It allowed us to customize display and behaviour of a web page by using JavaScript scripts, which could be de/actvivated during the usage of a website. We designed the issues as a violation of the commonly used heuristics of usability by Nielsen [8] and Schneiderman [12]. Some of the tasks were essentially the same, e.g., *product search*, but they differed in the product that was searched for and also in the presence or absence of an issue. There were four tasks without usability issues; the participants were asked to:

- *Find the cheapest ticket to a water park.*
- *Find, what the number of offers for a specific meal is.*
- *Find three offers for a sauna.*
- *Find a contact email on the website.*

There were four tasks with usability issues; the participants were asked to:

- *Find a specific product* – in this task, the search button was disabled unless the user clicked on it three times in a row.
- *Find five specific products on the map* – we added an issue causing problems with loading the map.
- *Find a specific product and buy three pieces of it* – the button for increasing the number of products in the basket was disabled, so the participants had to write the number manually.
- *Register* – we modified the registration so that it was necessary to enter password longer than 20 characters, but there was no error message notifying the participants of this constraint.

[1] http://uxi.sk
[2] https://www.emotiv.com/epoc
[3] http://www.noldus.com/human-behavior-research/products/facereader
[4] http://www.tobiipro.com/product-listing/tobii-pro-studio
[5] https://www.zlavadna.sk
[6] http://statpages.info/latinsq.html
[7] https://addons.mozilla.org/firefox/addon/greasemonkey

After every task, the participants had to answer three questions: (i) How *intensive* were the emotions that you felt?, (ii) How *positive* were the emotions that you felt?, and (iii) What was the *strongest* emotion you felt? They answered on a 5-point Likert scale for the first two questions (options ranging from non-intensive to intensive for the first one and from negative to positive for the second one) and selected one of the seven options (joy, fear, sadness, disgust, anger, surprise, neutral) for the last one.

2.2 Collected Dataset

We collected data from 21 participants (18 men and 4 women with ages ranging from 18 to 30 years, all being students). We excluded recordings of insufficient quality, i.e., when FaceReader was able to analyze less than 70% of the video (this happened, e.g., when participants obstructed the view of their face with their hands). In the rest of the videos, we filled the missing data by linear interpolation. We also excluded tasks where participants gave contradictory answers, i.e., when:

- participants labeled their emotion on a negative scale in the answer to the second post-study question, but then selected a positive emotion, such as joy as their strongest emotion or vice versa,
- the participants' answer to one of the questions was "neutral", since we were interested in binary classification of emotions (positive vs. negative); however, they were left for the first part of the analysis (see Fig. 1).

Tasks were labeled *positive* or *negative* based on the emotion a participant felt during solving of the task. The total number of labeled solved tasks was 147 for EEG and 35 for both EEG and facial expressions, out of which 47% were labeled negative and 53% as positive. We worked with the collected data at seconds precision; we labeled each second as positive or negative based on the recording label. The final dataset consisted of 13 336 seconds of EEG (70.33% labeled negative) and 3 564 seconds of both EEG and facial expressions (74.76% labeled negative). Each second in the dataset was described by raw EEG features, participant's emotional state self-reported at the end of a task, and valence and arousal from analysis of facial expressions.

3 Method of Emotion Detection

Since we hypothesize that usability issues cause negative emotions and these in turn indicate a usability issue, we formulate the task of emotion detection as a binary classification problem. We extract features from raw EEG data and from facial expressions; we aim to evaluate these two data sources individually as well as their combination. We process the EEG data as follows:

1. The patterns captured by EEG are divided by frequency [6]. For this purpose, we employ Discrete wavelet transform (DWT) similarly as in [7]. To measure emotions, the most important are the *alpha* (8-13 Hz) and *beta* (14-30 Hz) waves and we must also take into account the intensity of these waves, given their localization in areas of the brain [2].

2. For arousal, high values of beta waves in the frontal lobe are typical [10]. We compute the arousal using following formula [2]:

$$Arousal = \frac{\alpha(AF3 + AF4 + F3 + F4)}{\beta(AF3 + AF4 + F3 + F4)} \quad (1)$$

where α and β are the strength of the waves and a letter with a number is a label of the electrode, on which the waves are measured. The layout and naming of electrodes for measuring EEG is standardized.

3. The most important electrodes for valence are $F3$ and $F4$ [2], which are placed also near the frontal lobe [10]. Activity of the lobe indicates positive emotions. The valence is computed as follows [2]:

$$Valence = \frac{\alpha(F4)}{\beta(F4)} - \frac{\alpha(F3)}{\beta(F3)} \quad (2)$$

Similarly, we extract valence and arousal using facial expressions analysis on videos of participants; this is provided by the Noldus FaceReader. The resulting features are normalized using *z-score normalization*[8]. From normalized data we can derive the rest of the features. To smooth the normalized valence and arousal values, we use the so-called rolling time windows, which take into account 5 seconds before and 5 seconds after the current one (the shortest tasks duration are about 10 seconds). After applying the smoothing, we get a new set of features including window mean, maximum, minimum, maximum deviation from the mean and difference of the current second from the mean. All of these together with the normalized and raw data are used as input to the emotion classifier. We also use features unrelated with valence and arousal, namely order of the current second to the whole duration of the task, and how the previous task was labeled (i.e., the initial emotional state of a participant).

For classification, we use a decision tree, which is commonly used to detect emotions. We divide data into training set (90%) and test set (10%) following a standard division and aiming to maximize the amount of data used for training. For feature selection, we use logistic regression with regularization L1 (LASSO). In order to find the optimal hyperparameters of the used machine learning algorithm, we perform 5-fold cross-validation on the training set; the results reported in the paper are from the application of the trained model on the test set.

4 Results

Firstly, we analyzed whether usability issues do in fact cause significantly negative emotions as suggests the distribution of participants' answers for different tasks with or without usability issues (see Fig. 1). To determine whether the observed difference is significant, we used a *paired t-test* comparing the mean self-reported values of valence for tasks with added usability issues and

[8] http://www.statisticshowto.com/probability-and-statistics/z-score

without them for each participant. We found the difference to be significant ($T(83) = -9.77, p < 0.0001$), thus confirming our hypothesis. In addition, if we considered solely the reported negative emotion as indicative of a usability issue, such a classifier would have 80.95% accuracy (82.5% precision, 78.57% recall). These results are very promising and suggest the potential of using emotions for automatic usability issues detection.

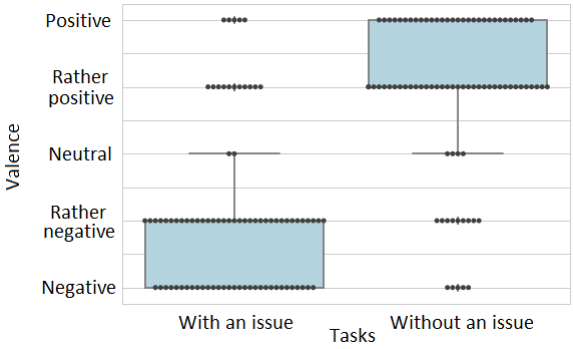

Fig. 1. Distribution of valence values for tasks self-reported by participants.

However, we first need to be able to reliably distinguish positive emotions from negative ones. Therefore, we explored the usefulness of features described in the previous section for this task; we performed automatic feature selection using logistic regression with L1 regularization on the training set. The best feature turned out to be mean value of arousal for the 10s window as measured by EEG, followed by features derived from normalized value of valence (based both on EEG and facial expressions). We explain the usefulness of arousal feature by the fact that if participants felt positive emotions, these were usually mild, while with the negative ones (caused by a usability issue) they were more aroused.

We classified emotions using our proposed method on the collected dataset (using every second of tasks as samples). We report the results for each data source (EEG and facial expressions) as well as their combination in Table 1. The combination of EEG and facial expressions outperformed the individual classifiers. These results were achieved by optimized versions of the decision tree ($criterion =$ "$gini$", $max\_depth = 9$, $min\_samples\_in\_leaf = 1$). We used grid search with 5-fold cross-validation on the training set to find the optimal values the classifier hyperparameters.

Next, we wanted to evaluate the capability of our approach to generalize for unseen task or user. We created a new training set, where the ratio of the negative second samples was the same as the positive ones, and all three sets of extracted features (EEG, facial expression, and their combination) were trained on this training set. Leave-one-out cross-validation was used, i.e., we used seconds of

Table 1. The results of emotion classification using EEG, facial expressions, and their combination.

	Accuracy	**Precision**	**Recall**	**F1**
EEG	85.6%	85.0%	76.4%	79.3%
Facial expressions	89.6%	89.4%	83.0%	85.4%
Combination	94.4%	94.6%	90.6%	92.4%

one task of one user, which were not included in the training, as validation set in each iteration. This was done for every solved task and the results were averaged. All three classifiers achieved much lower accuracy around 55%, which is above random, but suggests problems with over-fitting.

To overcome this problem, we tried to apply our proposed method on the data averaged for the whole tasks. Since we had 147 observation of task solving for EEG data, but only 35 for combination of EEG and facial expressions, we carried out this part of evaluation only for EEG data. The trained decision tree classifier had accuracy of about 70% (on approximately balanced data). The best features in this case were maximal and minimal values of arousal and maximal deviation of valence from its mean (computed based on the whole task duration).

5 Conclusions and Future Work

We demonstrated the potential of using emotions to automatically detect usability issues during usability testing. The results of our study suggest that usability has a strong connection to emotions—mainly negative ones—and the detected negative emotion is a good indicator of a usability issue. We also proposed a method of emotions detection or more precisely, their valence that combines features from EEG device and facial expressions recognized with a basic webcam. The achieved classification results are promising.

Nevertheless, the presented evaluation had several limitations. First, the usability issues were added into the tested webpage intentionally, which might have lowered their ecological validity. An experiment with natural usability issues is needed in the future. Further experiments are also needed to determine, how the severity of the usability issues affects the emotions, i.e., what level of severity is a borderline to illicit an emotion and thus for issues to be recognized by our proposed method. Additionally, the participant sample was quite homogeneous in our study. It remains an open question, how other factors (age, gender, computer literacy, etc.) can impact the emotional effects of the usability issues.

Second, although we worked with the data at seconds precision, they were labeled for the whole task; a more fine-grained labelling of changing emotional states might lead to better results. Lastly, we did not test, how the trained model generalizes for a different tested interface, i.e., whether it would be possible to train the model during testing of one interface and apply it (with sufficient accuracy) to a new one. This remains a future work as well.

Acknowledgement. This work was partially supported by grants No. APVV-15-0508, VG 1/0646/15 and VG 1/0667/18 and it was created with the support of the Ministry of Education, Science, Research and Sport of the Slovak Republic within the Research and Development Operational Programme for the project "University Science Park of STU Bratislava", ITMS 26240220084, co-funded by the ERDF.

References

1. Aggarwal, A., Niezen, G., Thimbleby, H.: User experience evaluation through the brain's electrical activity. In: Proceedings of the 8th Nordic Conference on Human-Computer Interaction Fun, Fast, Foundational - NordiCHI '14. pp. 491–500. ACM Press, New York, USA (2014). https://doi.org/10.1145/2639189.2639236
2. Blaiech, H., Neji, M., Wali, A., Alimi, A.M.: Emotion recognition by analysis of EEG signals. In: 13th International Conference on Hybrid Intelligent Systems (HIS 2013). pp. 312–318. IEEE (2013). https://doi.org/10.1109/HIS.2013.6920451
3. Ekman, P.E., Davidson, R.J.: The nature of emotion: Fundamental questions. Oxford University Press (1994)
4. Landowska, A.: Towards Emotion Acquisition in IT Usability Evaluation Context. Proceedings of the Mulitimedia, Interaction, Design and Innnovation pp. 1–9 (2015). https://doi.org/10.1145/2814464.2814470
5. Levenson, R.W.: Autonomic Nervous System Differences Among Emotions. Psychological Science **3**(1), 23–27 (1992)
6. Lin, Y.P.L.Y.P., Wang, C.H.W.C.H., Wu, T.L.W.T.L., Jeng, S.K.J.S.K., Chen, J.H.C.J.H.: EEG-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 489–492 (2009). https://doi.org/10.1109/ICASSP.2009.4959627
7. Matlovic, T., Gaspar, P., Moro, R., Simko, J., Bielikova, M.: Emotions detection using facial expressions recognition and eeg. In: Proc. of 11th Int. Workshop on Semantic and Social Media Adaptation and Personalization (SMAP'16). pp. 18–23. IEEE (2016). https://doi.org/10.1109/SMAP.2016.7753378
8. Nielsen, J.: Enhancing the explanatory power of usability heuristics. In: Proc. of the SIGCHI Conf. on Human Factors in Computing Systems. pp. 152–158. CHI '94, ACM, New York, NY, USA (1994). https://doi.org/10.1145/191666.191729
9. Raita, E., Oulasvirta, A.: Mixed feelings?: The relationship between perceived usability and user experience in the wild. In: Proc. of the 8th Nordic Conf. on Human-Computer Interaction: Fun, Fast, Foundational. pp. 1–10. NordiCHI '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2639189.2639207
10. Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology **39**, 1161–1178 (1980)
11. Scherer, K.R.: What are emotions? And how can they be measured? Social Science Information **44**(4), 695–729 (2005). https://doi.org/10.1177/0539018405058216
12. Shneiderman, B.: Designing the User Interface: Strategies for Effective Human-computer Interaction. Addison-Wesley Longman Publ. Co., Inc., Boston, MA, USA (1986)
13. Takahashi, K.: Remarks on emotion recognition from multi-modal bio-potential signals. 2004 IEEE International Conference on Industrial Technology, 2004. IEEE ICIT '04. **3**, 186–191 (2004). https://doi.org/10.1109/ICIT.2004.1490720