Interplay between AI and HCI for UX evaluation: the SERENE case study

Giuseppe Desolda[1], Andrea Esposito[1], Rosa Lanzilotti[1], Maria Francesca Costabile[1]

[1] *University of Bari Aldo Moro, via Orabona, 4, Bari, Italy*

Abstract

User eXperience (UX) is an important quality of software products. However, its evaluation is often neglected, mainly because developers think that it is resource-demanding and because of the scarce automation of the evaluation process. This paper presents an ongoing research work on SERENE, a platform that exploits Artificial Intelligence (AI) for semi-automatic UX evaluation. Specifically, this platform helps understand users' emotions by analyzing the log data of the users' interactions with websites. SERENE has been developed following the Human-Centered Artificial Intelligence approach since user control of the AI model is supported through the customization of some model features; in addition, the explanation of the AI model is provided through heatmap visualizations, thus supporting augmentation rather than automation of the UX evaluation activity.

Keywords

Human-Centered Artificial Intelligence; UX evaluation; UX smells

1. Introduction

Artificial Intelligence (AI) typically aims at defining prediction models that, while improving performances over the existing ones, have the final goal of replacing humans in the decision-making process or in any other activity that requires human-level reasoning [12]. However, recent failures of AI-based technologies highlight the limitations of the approaches to designing and evaluating the interaction with AI systems [14]. Most of these failures are caused by a lack of explainability of and user control of the current state-of-the-art models [6]. To overcome these limitations, the Human-Centered AI (HCAI) is an emerging research area that studies methods to design and evaluate AI systems aiming to amplify, augment, and enhance human performances, in ways that make AI systems reliable, safe, and trustworthy [14]. HCAI pursues two main goals. The first one relates the definition of AI algorithms that result in transparent models to understand and trust the system decisions by explaining the black-box models and their outcomes [6; 10]. The second goal regards user control on AI systems by allowing users to interactively manipulate and control model parameters while observing the effect of such changes in real-time [11; 13].

In this paper, we present the last advances on SERENE ("uSer ExpeRiENce dEtector"), a Web platform for UX evaluation that, starting from the visitors' interaction log of a given website, predicts their emotions. SERENE supports both *explainability* and *control*. The former is provided by heatmap visualizations that show the intensity of emotions on the web pages. The latter is enabled by allowing UX experts to control some AI model parameters to tune the model accuracy and precision.

Proceedings of CoPDA2022 - Sixth International Workshop on Cultures of Participation in the Digital Age: AI for Humans or Humans for AI? June 7, 2022, Frascati (RM), Italy
EMAIL: giuseppe.desolda@uniba.it (G. Desolda); andrea.esposito@uniba.it (A. Esposito); rosa.lanzilotti@uniba.it (R. Lanzilotti); maria.costabile@uniba.it (M. F. Costabile)
ORCID: 0000-0001-9894-2116 (G. Desolda); 0000-0002-9536-3087 (A. Esposito); 0000-0002-2039-8162 (R. Lanzilotti); 0000-0001-8554-0273 (M. F. Costabile)


While the design of the AI model of SERENE to predict emotions has been already described in [4], in this paper, we present the Web platform that provides UX experts with a simple tool to *i*) track the visitors' interaction logs on a given website, and *ii*) visualize the AI model predictions through interactive heatmaps that help identify UX smells. We present and discuss the interaction between the users of SERENE and the AI model that powers the system, highlighting all the choices that were made to ensure full user control and to prefer augmentation rather than automation, in line with the HCAI view.

2. SERENE Architecture and Design

SERENE provides emotion predictions by supporting UX-expert control rather than attempting to replace them. Before detailing the design of the interaction and cooperation between the UX experts and the AI model of SERENE, we introduce the overall system to allow an easier comprehension of the following sections.

UX experts use SERENE to detect UX smells throughout the various pages of a website. In this context, "UX smell" [2] can be identified when visitors tend to lose engagement in the web page or when a "negative emotion" is recognized: in fact, UX is defined as the set of a person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service [7].

The proposed platform allows UX experts to manage their evaluation "project", each one related to the UX analysis of a given website. A project is created by providing the website URL and the project name. For each project, a data-collector component is generated by SERENE. This component consists of a JavaScript asynchronous script that must be included in all the website pages, as it happens for similar platforms like Google Analytics. Once the component is installed it starts collecting visitors' interaction logs (i.e., mouse movements, idle time, and the frequency of key presses, without collecting sensitive data) in real-time and sends them to the SERENE web server, which stores these logs into a MySQL DBMS. Notice that SERENE does not collect a video recording of either the user's facial expression or their interaction with the screen: it only uses non-sensitive information regarding the standard input methods.

While these data are automatically collected, they are also 'translated' into emotion by the AI model adopted by SERENE, which was built as a random forest, since an analysis comparing different models showed that it outperformed others like decision trees, AdaBoost, and Multi-Layered Perceptron [4]. As better explained in the following sections, interactive heatmaps are then used to depict the visitors' emotions. As reported in **Figure 1**, the heatmap shows the emotions felt by the visitors on a given webpage. According to the EMFACS (Emotional Facial Action Coding System) model used to train the SERENE random forest model, seven emotions are considered, i.e., Anger, Contempt, Disgust, Joy, Fear, Sadness, and Surprise. The emotion to be shown in the heatmap can be selected in the menu on the right side. The heatmap colors range from red to blue, with red that means high values of a given emotion, while blue meaning low values of a given emotion. **Figure 1** represents a heatmap generated by the system. In this example, containing fake and generated data created just for the purpose of this example, we can see that the UX expert has selected the *contempt* emotion for the visualization and high values of contempt are concentrated on the left-bottom, where a product in the e-commerce website is placed. This concentration of negative emotion must be then analyzed by the UX expert for, for example, investigating if the product description is too vague or the interaction with that product is complex due to bad interactive system behavior.

Figure 1. SERENE dashboard: the heatmap that overlays the analyzed webpage shows the (averaged) emotions felt by the visitors on that webpage; red color means high values of a given emotion, while blue color means low values of a given emotion (the colors in-between represent half-values obtained as the average). A white color highlights the lack of data for a specific section of the web page. Different emotions can be selected on the right-side menu.

3. How SERENE provides user control in UX analysis

A crucial step in the development of SERENE was the design of the interaction between the UX experts and the SERENE AI model. To allow the UX expert to control the model, we first identified the aspects of the model to be controlled by the UX expert to empower the UX analysis. To this aim, we conducted a semi-structured interview with 5 UX experts. They were presented with a prototype of SERENE and a description of the technical details of the model. Each interview lasted around 15 minutes and two main requirements emerged.

The first requirement was the *possibility to select only one emotion at a time* to be shown in the heatmap. Indeed, since 7 emotions are predicted by the AI model starting from the visitors' logs, cognitive overload can occur if all data on all the emotions are shown at the same time in the heatmap. In addition, UX smells are typically found by identifying the concentration of negative emotions, thus positive emotions are less useful in this analysis, and it could be useful to filter out them.

The second requirement was the *granularity of the model prediction* that must be controlled by the UX expert. Indeed, the emotion values predicted by the SERENE AI model range from 0 to 100, where 100 indicates maximum emotion intensity. To improve the performance of SERENE, we decided to stratify the emotion prediction in n classes C_i, with n chosen between either 3, 5, or 7. For example, if $n = 3$, class C_1 contains all the predictions in the range 0–33, class C_2 contains all the predictions in the range 34–66, and class C_3 contains all the predictions in the range 67–100. This decision was made to reduce the codomain of the output function, thus mapping a greater real range to a smaller discrete set: this improves performances as it increases the probability of guessing the right output value (supposing a random classifier) from 10^{-2} to n^{-1}. It is clear that aggregating the predictions in a low number of classes results in a higher accuracy but also in fewer details, while aggregating the predictions in a high number of classes results in low accuracy but higher details.

4. Augmentation rather than Automation through explanation

A typical AI-based system that aims at the recognition of UX and usability issues in a web page, would typically attempt to indicate the sections on the page where there is a loss of engagement or an increase in the perception of negative emotions. For example, [8] shows how a peak in negative emotions can be related to a usability issue, thus it can be assumed that a usability issue is modifying the user's emotions. As another example, [5] presents the Usability Smell Finder system that is able to automatically recognize usability smells and provide to the expert a report describing all of them, alongside a possible solution. However, such a solution is the typical approach that emphasizes "automation" rather than "augmentation": in fact, this would be done without requiring any kind of user input: once the system has been set up, the user would receive a report of the found issues when he/she requests it.

On the contrary, our goal is to provide a tool that empowers UX experts rather than replaces them: in other words, our aim is toward augmentation rather than automation. To accomplish such a goal, we designed the system in order to allow its users to decide by themselves where a UX smell might be: in fact, the SERENE provides them with a heatmap of the recognized emotions. Heatmaps are a solution often adopted to visualize model outcomes, in particular in the case of Deep Neural Networks for image classification [1]. This falls in scope with the "augmentation" goal: given that the data is collected on all the available user base, the UX experts no longer must approximate their decision by only evaluating a sample of the user (that may introduce biases, especially if the sample is not chosen correctly [9]).

5. Conclusions

In this position paper, we presented and discussed the case study of SERENE, as an example of the design of an AI-based system that puts its focus on the cooperation between its users (i.e., UX experts) and the AI system by having the goal of augmentation rather than automation. The ongoing research presented here highlights the need to design human-centered AI systems that focus on the users' needs besides on the AI model and performances. This requires a change of view on the design of AI-oriented systems; they should not always automate and replace the human, but they should be designed to create a valuable and effective collaboration among AI systems and their end-users, to take the best of the abilities of algorithms (e.g., computational power, knowledge inferred from large data sets) and humans (creativity, emotionality, multidisciplinary knowledge). Such an idea is already being implemented in other fields: for example, [3] presents an AI model for Alzheimer's prediction that provided the probability of being ill rather than a crisp decision, in the attempt of leaving the control to the physician while still providing a suggestion. Although these are some steps in the right direction, still a lot of effort needs to be put into shifting the current AI-design perspective to a more human-centered point of view.

6. References

[1] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R. and Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7), e0130140. DOI: 10.1371/journal.pone.0130140.

[2] Buono, P., Caivano, D., Costabile, M.F., Desolda, G. and Lanzilotti, R. (2020). Towards the Detection of UX Smells: The Support of Visualizations. *IEEE Access*, 8, 6901-6914. DOI: 10.1109/access.2019.2961768.

[3] Castellano, G., Esposito, A., Mirizio, M., Montanaro, G. and Vessio, G. (2021). Detection of Dementia through 3d Convolutional Neural Networks Based on Amyloid Pet. In *Proc. of the 2021 IEEE Symposium Series on Computational Intelligence (Ssci 2021)*. IEEE. DOI: 10.1109/SSCI50451.2021.9660102.

[4] Desolda, G., Esposito, A., Lanzilotti, R. and Costabile, M.F. (2021). Detecting Emotions through Machine Learning for Automatic UX Evaluation. In *Proc. of the 18th IFIP TC 13 International Conference on Human-Computer Interaction (INTERACT 2021)*. Springer, 12934, 270–279. DOI: 10.1007/978-3-030-85613-7_19.

[5] Grigera, J., Garrido, A., Rivero, J.M. and Rossi, G. (2017). Automatic Detection of Usability Smells in Web Applications. *International Journal of Human-Computer Studies*, 97, 129-148. DOI: https://doi.org/10.1016/j.ijhcs.2016.09.009.

[6] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5), Article 93. DOI: 10.1145/3236009.

[7] ISO/IEC (2018). 9241-11 Ergonomics of Human-System Interaction — Part 11: Usability: Definitions and Concepts. *ISO/IEC 9241-11:2018*.

[8] Johanssen, J.O., Bernius, J.P. and Bruegge, B. (2019). Toward Usability Problem Identification Based on User Emotions Derived from Facial Expressions. In *Proc. of the 2019 IEEE/ACM 4th International Workshop on Emotion Awareness in Software Engineering (SEmotion) (Toward Usability Problem Identification Based on User Emotions Derived from Facial Expressions)*, 1-7. DOI: 10.1109/SEmotion.2019.00008.

[9] Lewis, J.R. (2014). Usability: Lessons Learned … and yet to Be Learned. *International Journal of Human–Computer Interaction*, 30(9), 663-684. DOI: 10.1080/10447318.2014.930311.

[10] Lundberg, S.M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In *Proc. of the Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, California, USA, Curran Associates Inc.,4768–4777.

[11] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. and Hassabis, D. (2015). Human-Level Control through Deep Reinforcement Learning. *Nature*, 518(7540), 529-533. DOI: 10.1038/nature14236.

[12] Russell, S.J. and Norvig, P. 2016. *Artificial Intelligence: A Modern Approach*. Pearson Education. ISBN: 9781292153964.

[13] Schmidt, A. (2020). Interactive Human Centered Artificial Intelligence: A Definition and Research Challenges. *Proceedings of the International Conference on Advanced Visual Interfaces.* Association for Computing Machinery, Article 3. DOI: 10.1145/3399715.3400873.

[14] Shneiderman, B. 2022. *Human-Centered Ai*. Oxford University Press. ISBN: 9780192845290.