An Experience of Textual Evaluation Using the MALTU Methodology

Marilia S. Mendes[1](✉) and Elizabeth Furtado[2]

[1] Federal University of Ceará (UFC), Russas, CE, Brazil
marilia.mendes@ufc.br
[2] University of Fortaleza (UNIFOR), Fortaleza, CE, Brazil
elizabet@unifor.br

Abstract. This paper presents an experience of textual evaluation in Usability and User eXperience in an academic system. The methodology used for the textual evaluation was MALTU. In this study, we analyzed 650 postings from an academic system and these posts have gone through a process of textual evaluation whose results will be presented in this paper.

Keywords: Textual evaluation · Usability · User eXperience (UX)
Postings related to the use (PRUs)

1 Introduction

Methods of collecting user opinion about a system, such as: interviews, questionnaires, schedules and Experience Sampling Method (ESM) [18] have been used to obtain User eXperience (UX) data from a product. Although such methods provide valuable data, they do not provide rich UX descriptions of users' daily life, primarily because they are applied at predefined times by researchers (for example developers and evaluators) of systems [9].

In [13–15], the authors of this paper investigated post messages of the users of Social Systems (SS): Facebook and Twitter. Postings that revealed reports of users' experiences will be called herein as Postings Related to the Use (PRUs). Unlike the other textual evaluation works [5, 9, 13, 21, 26], in which users are asked to write about their experience, these posts are spontaneous and report the user's perceptions about the system during its use. A PRU is a post in which the user refers to the system in use, for example: "*I can't change the Twitter profile photo*". A non-PRU is any post that does not refer to the use of the system, such as: "*Let's go to the show on Friday?*". The capture of spontaneous posts is obtained because we collect posts exchanged by the users in the system itself, when it has a forum or space to exchange messages.

In [12] we proposed the Maltu methodology and since then we have been experimenting with textual evaluation in different systems [4, 25]. The purpose of this paper is to present a detailed textual evaluation and discuss interesting points of this new form of systems evaluation. In this work, we analyzed 650 postings of an academic system with social characteristics (e.g., communities, forums, chats, etc.).

G. Meiselwitz (Ed.): SCSM 2018, LNCS 10914, pp. 236–246, 2018.
https://doi.org/10.1007/978-3-319-91485-5_18

This paper is organized as follows: in the next section, we present a background on textual evaluation of systems and the Maltu Methodology. In Sect. 3, we present some researches related to ours. In Sect. 4, we describe the Textual evaluation with Maltu Methodology, followed by results, conclusion and future works.

2 Background

2.1 Textual Evaluation of Systems

The textual evaluation of systems consists of using user narratives in order to evaluate or obtain some perception about the system to be evaluated [12]. It is possible to evaluate one or more criteria of quality of use with textual evaluation, such as usability, UX and/or its facets (satisfaction, memorability, learnability, efficiency, effectiveness, comfort, support, etc.) [6, 9, 12, 13, 21]. Other criteria can be evaluated, such as privacy [11], credibility [3, 11] and security [23]. Evaluation forms vary from identifying the context of use to identifying the facets of Usability or UX. Some papers have analyzed specifically the most satisfactory and unsatisfactory user experiences with interactive systems [5, 20, 21, 26].

The textual evaluation can be manual, through questionnaires with questions about the use of the system or experience reports, in which the users are requested to describe their perceptions or sentiments about the use of the system. The other way is automatic: evaluators can collect product evaluations on rating sites [6] or extract PRUs from Social Systems (SS) [10, 12–15, 17, 19]. The automatic form allows more spontaneous reports, including doubts when using the system, but, on the other hand, may also contain many texts that are not related to the use of the system, and these must be discarded.

Textual evaluation has its advantages and disadvantages, similar to other types of HCI assessment, such as user testing, heuristic evaluation, among others. The main advantage is to consider users' spontaneous opinions about the system, including their doubts. The main disadvantage is the long time of texts analysis. However, there are few initiatives of automatic textual evaluations [16], since it is an new evaluation type.

2.2 The Maltu Methodology

The MALTU methodology [12] for the Usability and UX (UUX) textual evaluation, mentioned in the introduction, consists in using user-generated narratives (postings) done in the own system, usually a SS, where spontaneous comments about the system are reported by users while using it; or from the extraction of postings on product/service evaluation websites [4, 25]. A user's posting can have more than one sentence, which in turn has multiple terms (words, symbols, scores), and those can help investigate what motivated (the cause of the problem) the user to write their posting, as well as what their reaction (behavior) was to the system in use, for example.

The methodology uses five steps for evaluation: (1) definition of the evaluation context; (2) extraction of PRUs; (3) classification of PRUs; (4) results and (5) report of results. In step 1, we define the system under evaluation; the users whose opinion matters to the evaluators; and the purpose of the evaluation. In step 2, the extraction of PRUs can be carried out either manually or automatically, by using the patterns of extraction proposed by the methodology described in [12]. When the extraction is manually done, the evaluators should use the search fields of the system under evaluation by informing the extraction patterns for the recovery of PRUs. When extraction is done automatically, the evaluators should use a posting extraction tool [16]. In step 3, we apply a process of classification of PRUs. This step can also be performed either manually or automatically (by using a tool [16]). When this step is performed manually, the sentences are analyzed by specialists for classification. The methodology proposes the minimum number of two specialists for classification. In addition to the previously mentioned criteria (classification by UUX facets, type of posting: complaint, doubt, praise), it is possible to analyze the user's feelings and intentions regarding the system in use and identify the functionality that may be the cause of the problem. In step 4, we interpret the results, and in step 5 we report them. In the next section, these steps will be more detailed in the evaluation of the academic system.

3 Related Works

Some studies that have focused on user narratives in order to study or evaluate usability or UX. In [5], the authors, focusing on studying UX from positive experiences of users, collected 500 texts written by users of interactive products (cell phones, computers etc.) and presented studies about positive experiences with interactive products. In [9], the authors collected 116 reports of users' experiences about their personal products (smartphones and MP3 players) in order to evaluate the UX of these products. Users had to report their personal feelings, values and interests related to the moment at which they used those. In [20], the authors collected 90 written reports of beginners in mobile applications of augmented reality. The focus was also evaluating the UX of these products, and the analysis consisted in determining the subject of each text and classifying them, by focusing attention on the most satisfactory and most unsatisfactory experiences. Following this line, in [26], the authors studied 691 narratives generated by users with positive and negative experiences in technologies in order to study the UX from them.

In the four studies mentioned above, the information was manually extracted from texts generated by users. The users were specifically asked to write texts or answer a questionnaire, unlike the spontaneous gathering of what they post on the system.

In [6], the authors extracted reviews of products from a reviews website and did a study in order to find relevant information regarding UUX in texts classified by specialists. However, they did not investigate SS, but other products used by users. In this case, the texts were written by products reviewers. It is believed that the posture of users in a product review website is different from that when they are using a system and face a problem, then deciding to report this problem just to unburden or even to suggest a solution. In addition, in none of these studies was a methodology used to present system evaluation results. In this work, we focused on considering the opinions of users about the system in use from their postings on the system being evaluated. We intend thereby to capture the user spontaneously at the moment they are using the system and evaluate the system.

4 Textual Evaluation Using the MALTU Methodology

The evaluation will be described, following the steps of the Maltu methodology.

(1) **Definition of the evaluation context**

The investigations were carried out in PRUs written in Brazilian Portuguese, collected from the database of an academic system with social characteristics (communities, discussion forums, chats, etc.) called SIGAA [24], which is the academic control system of the Federal Universities in Brazil. In this system, students can have access to several functionalities, such as: proof of enrollment, academic report, enrollment process, etc. The system allows the exchange of messages from a discussion forum. Its users are students and employees from the university. The system can be accessed by a browser on computers and mobile phones.

(2) **Extraction of PRUs**

For this work, 650 PRUs were selected from a part of the database coming from a previous work [12]. In this previous work, from a total of 295,797 posts, this sample of posts was collected by IHC specialists. The selection criteria was to collect postings in which users were talking about the system. An example of a PRUs collected was: "*I cannot stand this SIGAA anymore!*". Postings from students asking questions about their graduation courses, grades, location, etc. were not selected, for example: "*Will the coordination work during vacation?*" and "*Professors did not post the grades yet*".

(3) **Classification of PRUs**

The PRUs contained between one and six sentences each. That is why many times the post starts praising the system and ends up criticizing it, for example: "*I think this new system has a lot to improve*" (Negative Feeling)…"*However, it is already much better than the previous one*" (Positive Feeling). In this way, we divided the PRUs into sentences. After this division, we performed another analysis in order to verify the related and unrelated sentences to the use of the system, because there were sentences such as: "*Good morning*", "*Thank you*", "*Sincerely…*", which were not related to the use of the system. In this way, we discarded such sentences.

The rating process consists of categorizing a post into an evaluation category. There are seven types of classification categories for evaluation: (i) type of message to be investigated; (ii) intention of the user; (iii) polarity of Sentiment; (iv) intensity os sentiment; (v) quality in use criterion; (vi) functionality; and (vii) platform.

(i) **Type of message:** this type of classification refers to investigating what type of message the user is sending over the system in use, which can be: **(a) critical**: containing complaint, error, problem or negative comment regarding to the system; **(b) praise** or positive comment about the system; **(c) help** (giving of) to carry out an activity in the system; **(d) doubt** or question about the system or its functionalities; **(e) comparison** with another system; and **(f) suggestion** about a change in the system;

(ii) **Intention of the user:** the intention classification aims to classify the PRUs according to the user's intention with the system. In [17], a classification of PRUs was made in the categories: visceral, behavioral and reflexive. The definitions that emerged from the PRU were as follows:

(a) **Visceral PRU:** has greater intensity of user's sentiment, usually to criticize or praise the system. It is mainly related to attraction and first impressions. It does not contain details of use or system features. These are two examples: "*I'm grateful to SIGAA which has errors all the time: (*" and "*This System does not work!!! < bad language > !!*";

(b) **Behavioral PRU:** has lower intensity of user's sentiment and is also characterized by objective sentences, which contain details of use, actions performed, functionalities, etc. Two examples are the following: "*I would like to know how you can add disciplines to SIGAA*"; and "*It's so cool to be able to enter here*";

(c) **Reflective PRU:** is characterized by being subjective, presenting affection or a situation of reflection on the system. One example: "*The system looks much better now than it did last semester, when it was installed*".

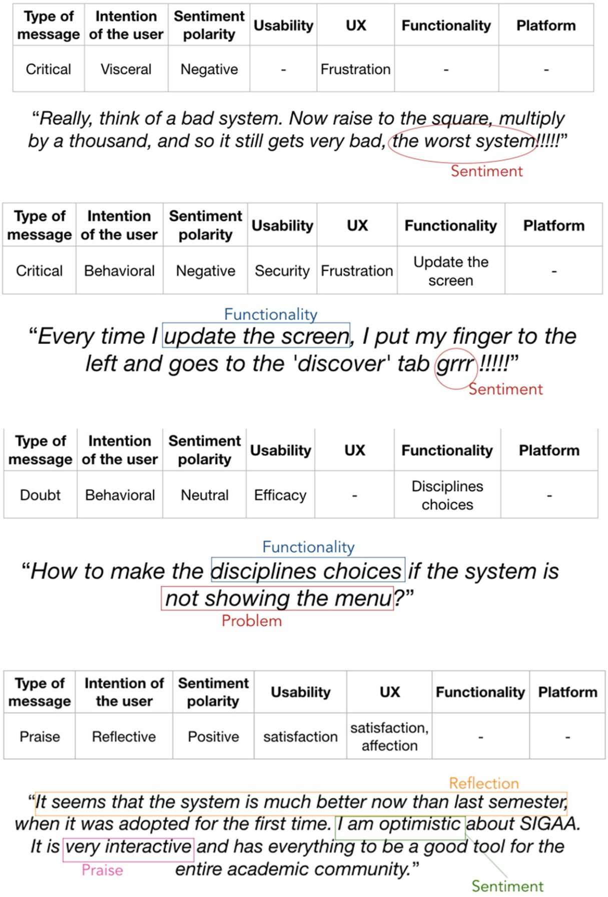
Information of Sentiment: in this category, two forms of classification are presented to analyze the sentiment in the PRUs: **(iii) polarity:** a PRU can demonstrate positive sentiment, neutral sentiment and negative sentiment; and **(iv) intensity:** allows us to classify how much of sentiment (positive or negative) is expressed in a PRU. In the examples: "*I like this system…*" and "*I really love using this system*". The positive sentiment observed is more intense in the second PRU. This type of classification is only performed automatically [12].

(v) **Quality in use criterion:** this category involves determining the criterion of quality in use. The Maltu uses the following criteria: **(a) usability** and/or **(b) UX**. This category involves relating a facet of each criterion to a PRU. Maltu uses the following facets for Usability: efficacy [7], efficiency [7], satisfaction [7], security [22], usefulness [22], memorability [22] and learning [22]. For UX, the facets used are: satisfaction [7], affection [1], confidence [1], aesthetics [1], frustration [1], motivation [2], support [8], impact [8], anticipation [8] and enchantment [8];

(vi) **Functionality:** there are PRUs that detail the use of the system, making it possible to classify the functionality of the system and is referred to by the user or the cause of the problem to which the user refers. In the exemplo: "*I can not exclude disciplines. Can someone help me?*", the functionality is "exclude disciplines"; and

(vii) **Platform:** this category consists of identifying the operating system and device that the user was using at the time of the relative posting. There are systems, like Twitter and Facebook, for example, where the PRUs extracted from the system can come from different devices. On SIGAA, as access is by browser, it can also be accessed from different devices.

We illustrate (Fig. 1) the following some examples of classification of postings.



Type of message	Intention of the user	Sentiment polarity	Usability	UX	Functionality	Platform
Critical	Visceral	Negative	-	Frustration	-	-

"Really, think of a bad system. Now raise to the square, multiply by a thousand, and so it still gets very bad, the worst system!!!!!"

Type of message	Intention of the user	Sentiment polarity	Usability	UX	Functionality	Platform
Critical	Behavioral	Negative	Security	Frustration	Update the screen	-

"Every time I update the screen, I put my finger to the left and goes to the 'discover' tab grrr !!!!!"

Type of message	Intention of the user	Sentiment polarity	Usability	UX	Functionality	Platform
Doubt	Behavioral	Neutral	Efficacy	-	Disciplines choices	-

"How to make the disciplines choices if the system is not showing the menu?"

Type of message	Intention of the user	Sentiment polarity	Usability	UX	Functionality	Platform
Praise	Reflective	Positive	satisfaction	satisfaction, affection	-	-

"It seems that the system is much better now than last semester, when it was adopted for the first time. I am optimistic about SIGAA. It is very interactive and has everything to be a good tool for the entire academic community."

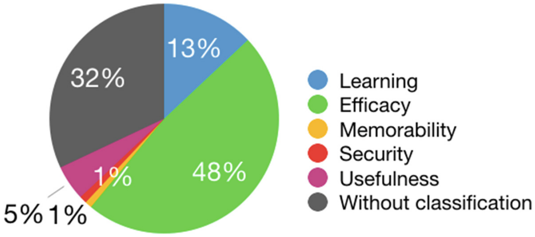
Fig. 1. Examples of classification of postings

According to the examples presented, it is not always possible to categorize a post in all proposed classification forms. The classification form took place as follows: 500 PRUs were classified by 10 undergraduate students and 150 by IHC specialists, totaling 650 PRUs, corrected by two IHC specialists.

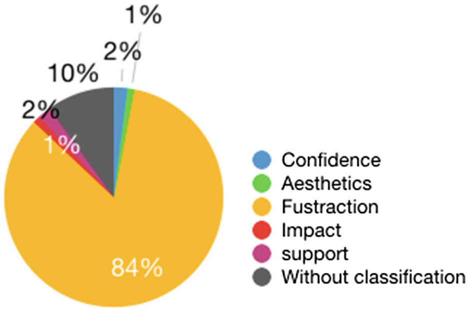
(4) Results and (5) Report results

The graphs and tables presented below, in this section, present the relationship between the classifications obtained, providing an overview of the evaluated system. Graph 1 illustrates the percentages obtained in each usability facets related to PRUs of the critical type. The efficacy facet, for example, obtained a higher percentage (48%). Graph 2 shows the percentages obtained in each UX facet related to PRUs of the critical type. The frustration facet, for example, obtained higher percentage (84%).

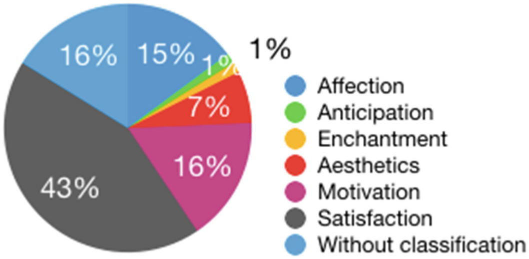
Graph 3 illustrates the percentages obtained in each UX facet related to praise PRUs. The satisfaction facet, for example, obtained a higher percentage (43%).



Graph 1. Quality of use criteria = usability x type of PRU = critical



Graph 2. Quality of use criteria = UX x type of PRU = critical



Graph 3. Quality of use criteria = UX x type of PRU = praise

Table 1 presents the functionalities collected from the critical-type PRUs in each usability facet. In the memorization facet, the criticisms were referring to: "*a lot of information*", "*how to register*", "*visual*". Table 2 presents the percentages and functionalities collected from PRUs of praise type in each usability facet. The highest percentage, satisfaction facet, indicates that users are satisfied with SIGAA for the following reasons: "communication", "interaction", "beauty", "new features", "practicality" and "sociable".

Table 3 presents the functionalities collected from the critical-type PRUs in each UX facet. The frustration facet, for example, presents a greater number of causes cited in the PRUs. The others have few functionalities, because, through the analysis

Table 1. Quality of use criteria = usability x type of PRU = criticism x cause.

Facets of usability	Functionalities
Learning	View, download or insert file; view or error in the disciplines; edit information; view or history error; error in calculating the media; perform, display or error in the registration; view notes, classes, frequency or faults; lock registration
Efficacy	View, download, open or insert file; credits less; view or history error; Perform, display or error in registration; view notes or times; error in calculating the media; blocking in the system, system in general
Security	Perform, display or error in registration; view or error in the disciplines
Usefulness	Browser; room location
Memorability	A lot of information; how to registration; visual

Table 2. Quality of use criteria = usability x type of PRU = praise x cause.

Percentage	Usability facet	Functionalities
2%	Learning	General system
5%	Efficacy	Make registration; General system
63%	Satisfaction	Communication; interaction; Beauty; new features; practical; sociable
15%	Usefulness	Warnings; Communication; interactivity; Discussion forums; system in general
15%	Without classification	–

Table 3. Quality of use criteria = UX x type of PRU = critical x cause

UX Facet	Functionalities
Fustraction	Menu unavailable; View, download or insert file; Calendar; accounting for claims; view or history error; Make registration; view or understand the disciplines' schedules; access only by the Firefox browser; error in calculating the media; view groups
Support	Make registration
Impact	Previous system
Confidence	Grades; Registration
Esthetics	Visual
Without classification	–

performed PRUs – UX classifications, the users did not present details of the system. Table 4 presents the main functionalities that the users had doubts and Table 5 presents suggestions of functionalities for the system.

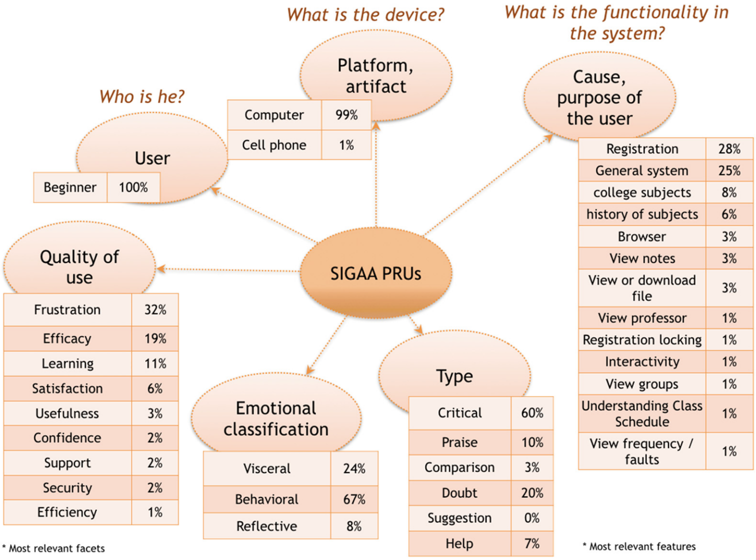
Table 4. Main features that users had doubts

Type of PRU = doubt x functionalities
Edit information; View, download or insert file; Lock in system; View or error in the disciplines; how to hide board registration numbers; View or error in history; To visualize or to understand the class schedule, notes, amendment of the disciplines; media calculation error; How to make a lock

Table 5. Main features suggestions

Type of PRU = suggestion x features
option to "enjoy", to do tests at home; Location map of the room allied to the disciplines; improvement of the system, explanation of the time code;

The Fig. 2 illustrates the system usage context obtained from the evaluation of a set of PRUs.



User	
Beginner	100%

Platform, artifact	
Computer	99%
Cell phone	1%

Cause, purpose of the user	
Registration	28%
General system	25%
college subjects	8%
history of subjects	6%
Browser	3%
View notes	3%
View or download file	3%
View professor	1%
Registration locking	1%
Interactivity	1%
View groups	1%
Understanding Class Schedule	1%
View frequency / faults	1%

Quality of use	
Frustration	32%
Efficacy	19%
Learning	11%
Satisfaction	6%
Usefulness	3%
Confidence	2%
Support	2%
Security	2%
Efficiency	1%

Emotional classification	
Visceral	24%
Behavioral	67%
Reflective	8%

Type	
Critical	60%
Praise	10%
Comparison	3%
Doubt	20%
Suggestion	0%
Help	7%

Fig. 2. Context of use of the SIGAA system evaluation

5 Final Considerations and Future Work

The results obtained using the methodology pointed to UUX problems, the main functionalities in which the users have doubts, criticisms and suggestions about SIGAA. As for the evaluation experience using Maltu, the classification stage was sometimes not simple, since the extracted PRUs were characterized by an average of 3 lines each, being at least 1 and at most 10 lines. In this way, the classification has become, at times, a slow and tiring process for the evaluators.

This paper reported a textual evaluation experience of UUX of SIGAA. The results have shown that the application receives many criticisms from various causes, mainly being support and efficacy problems that cause frustration to users of the application. Maltu is a recent methodology. Its use in this work consisted in the validation of the methodology by the application in different contexts. Future work will seek new ways to improve the classification process of PRUs with Maltu, in order to simplify and automate the extraction, classification and interpretation of results. Other suggested forms of classification will also be used. Another activity to be carried out is the expansion of the database, since only a specific source of complaints was used.

Acknowledgment. This project is funded by the Fundação Cearense de Apoio a Desenvolvimento Científico e Tecnológico (FUNCAP) and supported by LINCE Laboratories (Laboratory of Computer Science and Software Engineering) and LUQS (Laboratory of user studies and Quality of use of the Systems).

References

1. Bargas-avila, J.A., Hornbæk, K.: Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In: CHI 2011, Vancouver, pp. 2689–2698 (2011)
2. Bevan, N.: What is the difference between the purpose of usability and user experience evaluation methods? In: UXEM 2009, INTERACT 2009, Uppsala, Sweden, pp. 24–28 (2009)
3. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on Twitter. In: Proceedings of the 20th international Conference on World Wide Web, Hyderabad, India, pp. 675–684 (2011)
4. Freitas, L., Silva, T., Mendes, M.: Avaliação do Spotify – uma experiência de avaliação textual utilizando a metodologia MALTU. In: IHC 2016, São Paulo, Brazil (2016)
5. Hassenzahl, M., Diefenbach, S., Goritz, A.: Needs, affect, and interactive products. J. Interact. Comput. **22**(5), 353–362 (2010)
6. Hedegaard, S., Simonsen, J. G.. Extracting usability and user experience information from online user reviews. In: Proceedings of CHI 2013, Paris, France, pp. 2089–2098 (2013)
7. ISO DIS 9241 - 210:2008: Ergonomics of human system interaction - Part 210: Human - centred design for interactive systems (formerly known as 13407)
8. Ketola, P., Roto, V.: Exploring user experience measurement needs. In: 5th COST294-MAUSE Open Workshop on Valid Useful User Experience Measurement (2008)
9. Korhonen, H., Arrasvuori, J., Väänänen-vainio-mattila, K.: Let users tell the story. In: Proceedings of CHI 2010, pp. 4051–4056 (2010)
10. Lima, A., Silva, P., Cruz, L., Mendes, M.: Investigating the polarity of user postings in a social system. In: 19th International Conference on Human-Computer Interaction, HCII 2017 (2017)
11. Mao, H., Shuai, X., Kapadia, A.: Loose tweets: an analysis of privacy leaks on Twitter. In: Proceedings of WPES 2011, Chicago, IL, USA, pp. 1–12 (2011)

12. Mendes, M.S.: MALTU - model for evaluation of interaction in social systems from the users textual language, 200 f. Thesis (Ph.D. in Computer Science) – Federal University of Ceará (UFC), Fortaleza, CE – Brazil (2015)
13. Mendes, M.S., Furtado, E.S., Furtado, V., Castro, M.F.: Investigating usability and user experience from the user postings in social systems. In: HCI International (2015)
14. Mendes, M.S., Furtado, E.S., Militao, G., Castro, M.F.: Hey, I have a problem in the system. Who can help me? An investigation of Facebook users interaction when facing privacy problems. In: HCI International, pp. 391–403 (2015)
15. Mendes, M.S., Furtado, E., Castro, M.F.: Do users write about the system in use? An investigation from messages in natural language on Twitter. In: 7th Euro American Association on Telematics and Information Systems, Valparaiso, Chile (2014)
16. Mendes, M., Furtado, E.: UUX-Posts: a tool for extracting and classifying postings related to the use of a system. In: VIII Latin American Conference on Human-Computer Interaction, CLIHC 2017 (2017)
17. Mendes, M., Furtado, E., Furtado, V., Castro, M.: How do users express their emotions regarding the social system in use? A classification of their postings by using the emotional analysis of Norman. HCI Int. **2014**, 229–241 (2014)
18. Obrist, M., Roto, V., Väänänen-vainio-mattila, K.: User experience evaluation – do you know which method to use? In: CHI 2009, pp. 2763–2766 (2009)
19. Oliveira, D., Furtado, E., Mendes, M.: Do users express values during use of social systems? A classification of their postings in personal, social and technical values. In: HCI International, Los Angeles, CA, USA (2016)
20. Olsson, T., Salo, M.: Narratives of satisfying and unsatisfying experiences of current mobile augmented reality applications. In: CHI 2012, pp. 2779–2788 (2012)
21. Partala, T., Kallinen, A.: Understanding the most satisfying and unsatisfying user experiences: emotions, psychological needs, and context. Proc. Interact. Comput. **24**(1), 25–34 (2012)
22. Preece, J., Rogers, Y., Sharp, H.: Interaction Design: Beyond Human-Computer Interaction. Wiley, New York (2002)
23. Reynolds, B., Venkatanathan, J., Gonçalves, J., Kostakos, V.: Sharing ephemeral information in online social networks: privacy perceptions and behaviours. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) INTERACT 2011. LNCS, vol. 6948, pp. 204–215. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23765-2_14
24. SIGAA: Integrated Management System for Academic Activities. https://si3.ufc.br/sigaa/verTelaLogin.do. Accessed 10 Mar 2018
25. Silva, T., Freitas, L., Mendes, M.: Beyond traditional evaluations - users view in app stores. In: IHC 2017, Joinville, Brazil (2017)
26. Tuch, A.N., Trusell, R.N., Hornbæk, K.: Analyzing users' narratives to understand experience with interactive products. In: Proceedings of CHI 2013, pp. 2079–2088 (2013)