# Textual evaluation vs. User testing: a comparative analysis

**Thiago Hellen O. da Silva**[1], **Lavínia Matoso Freitas,**
**Marília Soares Mendes**[2], **Elizabeth Sucupira Furtado**[1]

[1]PPGIA – Universidade de Fortaleza (UNIFOR) – Fortaleza, CE – Brasil

[2]Universidade Federal do Ceará (UFC) – Russas, CE – Brasil

thiagoliveira@unifor.br, laviniamatosof@gmail.com,

marilia.mendes@ufc.br, elizabet@unifor.br

***Abstract.*** *One of the ways to evaluate a system is from Textual Evaluation. This type of evaluation consider the textual user's opinions to infer aspects of the interaction with the system. Although this method covers many texts and considers spontaneous narratives of the users, it takes a lot of time and effort. Some authors have reported on the need to compare evaluations techniques in order to investigate their effectiveness in revealing issue or to supplement the results of a systems assessment. This study presents a comparative analysis between the textual evaluation and user testing. A case study was performed evaluating the usability and user experience of a health app. As a result, the techniques were analyzed based on aspects that involved describing the results, resources needed and description of problems and users.*

## 1. Introduction

A system can be evaluated through investigation, observation of use or inspection [Barbosa and Silva 2010]. The investigative methods, such as interviews, questionnaires and user reports, collect user opinions to evaluate a system. A method for investigation that has been gaining attention lately is the Textual Evaluation. It consists in gathering and analyzing the opinion of the users expressed in the form of text in order to obtain a perception of the system. This type of evaluation takes user opinions - expressed as texts - to evaluate a system [Mendes 2015]. The textual evaluation can be performed by gathering the opinion of users through reviews collected from rating sites (e.g. AppStore, Play Store) or extract the Postings Related to the Use (PRUs) from social systems, such as Twitter, Facebook, forums, etc. [Mendes 2015].

The textual evaluation method is: (a) frequently applied [Fetter and Gross 2014, Hedegaard and Simonsen 2013, Mendes 2015, da Silva et al. 2017, Freitas et al. 2016]; (b) covers many user texts; and (c) considers spontaneous reports. Some studies have suggested that this method may: (a) require a lot of time or effort [Freitas et al. 2016, da Silva et al. 2017] if no methods are used to analyze the texts automatically; (b) does not precisely describe the use problems because it is limited to what has been reported by users about the system [Hedegaard and Simonsen 2013]; and, (c) may have influence of very satisfied or very dissatisfied users (who are more prone to express themselves) [Hedegaard and Simonsen 2014].

In [da Silva et al. 2017, Bach and Scapin 2010, Gray and Salzman 1998] the authors presented the need to compare the results obtained in this type of evaluation with other techniques. Each evaluation method has itself peculiarities, such as: objectives, available resources and capacity to reveal (or not) the problems in the system being evaluated [Barbosa and Silva 2010]. In the particular case of usability testing, one of the pioneering tests in the area of IHC, there are several variability for its application, depending on the environment to be controlled or not, the types of posttest forms applied, which can affect further effort of the tester. Many articles found in the literature perform an analysis between evaluation techniques, but since textual evaluation is a recent technique, it was not addressed in these articles. The main goal of this paper was to make a comparative analysis of the Textual Evaluation method with user testing.

The work methodology took the following steps: (a) to exemplify the use of a textual evaluation methodology by observing the resources used; (b) to analyze the Textual Evaluation with user testing; and (c) to investigate if the user testing can be useful to complement the results of the Textual Evaluation. The analysis involved evaluating the quantity of problems, the number of users involved, the effort and the time taken to apply each technique. We made the decision to carry out these steps with the case study to evaluate the usability and User eXperience (UX) with the MyFitnessPal application. One of the reasons that lead to choose MyFitnessPal as the evaluated application was its category. People use health apps continuously to insert datas of the day-to-day. So that makes easy to collect reports from different perspectives and user experiences.

This paper is organized as follows: first, we have the related work section, followed by an explanation of the textual evaluation technique, the case study, a brief discussion about the results and the conclusion section.

## 2. Related Work

Textual evaluation is a recent technique and some studies have analyzed its effectiveness compared to other techniques. Textual evaluations provided by software and video game users were used in [Hedegaard and Simonsen 2013] to investigate Usability and UX (UUX). The authors focused on extracting information of individual evaluations rather than comprehensive evaluations in order to explore the content of the reports with more precision. With the results, was suggest that it is possible to extract more information about UUX, but they stress the limitations of textual evaluations compared to traditional methods, such as the low level of detail on the use context that the reports provide.

In [Hedegaard and Simonsen 2014] the user evaluations of three devices were extracted and automatically classified in categories: (a) usability, for sentences specifying usability issues; (b) vague, for requests of resources or hypothetical problems; and (c) general, when the reported problem was not caused by the interface. In this same study, 10 participants performed a Think Aloud and 9 professionals performed the heuristic evaluation. The authors highlight that the results of the user reviews vary greatly, just as the results obtained through traditional methods vary from evaluator to evaluator. Although the automated methods didn't extract all the usability problems, the traditional methods didn't identify several problems that the textual assessment identified. The authors also underline that the textual evaluations may be more expensive than traditional methods.

An evaluation of Spotify was performed in [Freitas et al. 2016] and reinforces the need to use other techniques to support textual evaluations. In this evaluation, data was obtained on the UUX of the application based on 100 posts from users of the application.

A heuristic evaluation was also carried out that identified problems that were not found in the texts of the users. The authors point out that the main problems reported could only be identified through textual evaluation. The results indicate that the application receives many criticisms in several functionalities; most of the problems were related to support and effectiveness and this has caused the frustration of users of the application.

The need to analyze HCI evaluation techniques is also revealed in [Bim et al. 2016], where three inspection methods are compared: heuristic evaluation, cognitive walkthrough and semiotic inspection. The study shows the role of each technique and how they differ in the support they offer to the HCI professional.

## 3. Textual Evaluation Technique

Textual Evaluation is a technique that uses user texts to infer aspects about their interaction with a system [Mendes 2015]. There is a textual evaluation methodology called MALTU. This methodology seeks to evaluate the UUX of systems through a set of PRUs. The methodology has five evaluation steps: (i) definition of the evaluation context; (ii) extraction of PRUs; (iii) classification of PRUs; (iv) interpretation of results; and (v) reporting of the results. In step 1, the system to be evaluated, its users and the evaluation objectives are defined. In step 2, the PRUs are extracted, either manually or using automatic extraction tools. In step 3, the PRUs are classified by at least two specialists and a validator in up to six different categories: (a) type; (b) intention; (c) emotion analysis; (d) functionality; (e) quality-in-use criteria; and (f) artifact. The results are then interpreted (step 4) and, finally, reported (step 5).

The classification by functionality seeks to identify the functionality or the cause that motivated the user to publish the report. The classification by type seeks to identify if the post is a Praise, Criticism, Doubt, Comparison, Suggestion and/or Help. The classification by emotion analysis seeks to identify the polarity of the PRU (Positive, Negative or Neutral). The classification by intention seeks to identify the intention of the user regarding the system, which may be Visceral, Behavioral or Reflexive. This type of classification is based on Norman's emotional design [Norman 2004] and proposed by [Mendes 2015] in posts. The classification by quality-in-use criteria involves identifying the criterion (Usability or UX) and its facets.

The MALTU methodology has been applied in the evaluation of different types of systems, such as: an academic system [Mendes 2015], social networks [Freitas et al. 2016], internet-of-things systems [de Souza Filho et al. 2018] and mobile applications, such as Waze, Google Maps, Spotify [da Silva et al. 2017]. However, although in [Freitas et al. 2016, de Souza Filho et al. 2018] more than one type of evaluation has been used besides the textual evaluation to assess the system, this requires a more accurate comparison with other techniques in order to discuss its advantages/disadvantages and the context of its application.

## 4. Case Study - Evaluation of MyFitnessPal

Next, we'll show you the processes for selecting and evaluating the app. All techniques were applied by undergraduate students with three years of participation in HCI projects under the supervision of an advising professor.

### 4.1. The Selected App

The app used in the case study was MyFitnessPal. Some selection criteria were used to assist the decision: (a) availability in the official Android app store (Google Play); (b) having free functionalities; and (c) having the largest number of recent evaluations.

MyFitnessPal is an information system that works as a calorie calculator. The application allows its users to register their meals and exercises to obtain a daily ingested and expended calorie count [MyFitnessPal 2019]. The MyFitnessPal encourages its users to control their nutrition and physical exercises based on pre-established goals [MyFitnessPal 2019]. The application allows users to: record water consumption and bodily measures, identify the nutritional composition of the reported foods, add reminders and connect with other applications to synchronize physical exercise data [MyFitnessPal 2019]. The users can also add new foods on database of application. The users can access graphs to monitor their progress.

### 4.2. Textual Evaluation of MyFitnessPal

*1) Definition of the evaluation context.* The goals for the evaluation were: (a) to identify problems in the interaction and in the interface; and (b) to investigate user satisfaction.

*2) Extraction of the PRUs.* We use the comments posted by the users to evaluate MyFitnessPal on the app's page [MyFitnessPal 2019] in Google Play. The extraction was performed automatically using a crawler available in the Apify tool, to extract 3,000 posts written in Brazilian Portuguese. The PRUs used on this evaluation were posted in the period from November 3rd, 2017, to April 10th, 2018. A PRU of a system can contain one or more sentences that are not related to its use [Mendes 2015]. As such, the MALTU methodology recommends that the posts are transformed into sentences before the classification step [Mendes 2015]. The period sign "." was used to delimit and segment the posts, which yielded 3,963 sentences. The obtained sentences were analyzed to see if they were PRUs or non-PRUs, which resulted in 382 sentences being discarded because they were not related to the system, for example: "*Thank you for the dedication*". In this way, 3,581 sentences were used on this evaluation. This amount represents 0.19% of the total app ratings in the Google Play (on day of extraction) [MyFitnessPal 2019].

*3) Classification of the PRUs.* The PRUs were classified in the following categories proposed by the methodology [Mendes 2015]: (a) functionality; (b) type; (c) sentiment analysis; and (d) the quality-in-use criteria in HCI. For this last type of classification, the following Usability facets were used [Rogers et al. 2013]: Effectiveness, Efficiency, Security, Usefulness, Memorability and/or Learnability. For UX facets were used [Bargas-Avila and Hornbæk 2011]: Affection, Trust, Aesthetics, Frustration, Motivation, Support, Engagement, Impact, Enchantment and/or Fun. The Satisfaction facet is characteristic of both criteria. In the classification by quality-in-use criteria, a post can be classified in more than one facet, such as in: "*It is quick to insert data and it is beautiful.*" This sentence is classified in the "efficiency" facet of usability and in the "Aesthetics" facet of the UX.

Using a spreadsheet with posts, three evaluators classified 3,581 sentences for 20 hours. All sentences were classified by two evaluators. If there was any disagreement about a classification, the sentence was analyzed by a third evaluator. The classification was made by evaluators because there is no tool that classifies the posts automatically yet, considering all the criteria worked by the methodology; and this classification is a subjective analysis requiring more than one evaluator to obtain an evaluation consensus.

All the evaluators have extensive experience in analysis and classification of texts using the methodology. It is not part of the scope to analyze the number of posts that were classified as equal or different between the evaluators. However, in general, most of the posts obtained the same ranking among the evaluators. Only in a few postings did the reviewers have questions about how to rank them, the third reviewer had to give her opinion on the classification. Classification doubts may arise in posts that have specific information about the application domain being evaluated. Thus, evaluators usually flag such posts so that a survey and discussion between the evaluators be done and concluded. An example posting that the reviewers had doubts about how to classify was: "*I lost 10 kg with his help and I am now gaining muscle in the right measure with the correct division of macro nutrients and calculating the right calories to avoid gaining fat again.*". In this post, the user reports that with the application he was able to lose weight. This fact is related to his satisfaction over the use of the system; however, the user only reported a result about the service of application.

*4) Interpretation of the Results.* We counted the number of PRUs that cite an application's functionality (n = 880). From these PRUs, it was identified that 113 different features were cited by users in their posts. The most frequently cited functionalities were: calorie count (132 posts), progress (120 posts) and registration (115 posts). Some PRUs didn't mention functionalities, such as: "*Today I uninstalled the app and the problem continues*". In the classification by type of PRU, 77.9% were of the type "Praise" and 21.9% to "Criticism". In the classification by sentiment analysis, 73.9% of the PRUs were positive, 15.7% were negative and 10.5% neutral.With the classification according to quality-in-use criteria the facets most identified in the classification were: "Satisfaction" (70.7%); and "Efficacy" (17.6%). The total percentage of the result of the classification by type and by Usability and UX facet exceeds 100%, given that some sentences were classified in more than one type or more than one facet. Look at this post, for example: "*I am enjoying it, but it would be better if it counted strengthening exercises.*". In the classification by type, this post has been classified as "Praise"and "Suggestion".

*5) Reporting of the Results.* For the reporting of the results, the methodology suggests establishing a relationship between the evaluation categories, as shown below. A) Functionality x Criticism Type: 67 of the 113 functionalities found had at least one PRU classified as Criticism. The most criticized functionalities were: Progress (115), User Registration (111), Login (22) and Connecting with other apps (22). B) Functionality x Doubt Type: 8 functionalities related to the Doubt type were identified. Namely: User Registration (7), Progress (3), Posting photos (2), Updating status (1), Tracking exercises (1), Connecting with other apps (1) and Calorie count (1). C) Functionality x UUX Facets x Sentiment Analysis: the functionalities identified in UUX facets were also classified according to their polarity. As such, it was possible to identify if the author of the post referred to the functionality facet in a positive or negative way. No functionalities with a positive polarity were found related to the facets of: Memorability, Learnability, Trust, Aesthetics, Frustration and Support. No functionalities with a negative polarity were related to the facets of: Satisfaction, Affection, Motivation and Impact. No functionalities were found related to the facets Enchantment and Fun.

### 4.3. User testing of MyFitnessPal

This method was applied in 2 days. In the first day, the planning was made (it lasted about 1 hour and 30 minutes) and in the following day the test was applied, with an average duration of 12 minutes for each session. After the application of all tests with users, the results were analyzed (this lasted about 1 hour and 30 minutes). Were recruited 6 people voluntarily and were differentiated based on their experience in using health applications. The participants P1, P2 e P3 never used the application and the participants P4, P5 e P6 already used some type app for at least 1 month. No participant had experience with MyFitnessPal until the day the test was performed.

One behavior was observed in the textual evaluation: the functionality "Registration" gave rise to many criticisms related to the effectiveness of the system. Some posts describe a possible cause: the minimum age allowed. Such as in the following post: "*Just because I'm 15 you won't accept me.*" To test the effectiveness of this functionality, participant (P6) was a person recruited with under 18 years of age. The use scenario defined for the test was: "*You just got to know the MyFitnessPal app and start using it for the first time.*" The functionalities tested were: registration; meal record, exercise and current weight; and progress monitoring. The device used in all test sessions was a Xiaomi Mi 6 with the Android 8.0.0 operating system.

Each test session was performed involving three people: one participant, one facilitator and one observer. The facilitator was responsible for conducting the test, and the observer for making relevant notes and capturing the audio and video of the sessions. An authorization to record the audio and video of the participants was requested with the signing of a consent form. Each participant was then invited to perform actions in the MyFitnessPal application. These actions were: a) T1 - register in the application; b) T2 - add a meal; c) T3 - add an exercise; d) T4 - add the current weight; and, e) T5 - monitor progress. The test sessions were conducted in an environment where the participant chose (uncontrolled), that is, where he felt more comfortable or where s/he would use the application daily. The results of each tested task are described below.

**T1 - Register in the Application:** participants P2, P3 and P5 had difficulties informing their date of birth, as they did not find a faster way to enter the year of their birth. The zip code field does not have a validation mask and two users (P1 and P3) had to remember its formatting. Participant P3 made two attempts before he understood that the height had to be reported in centimeters. During use, participant P1 reported that the "save" button was wrong, since it said "configure". Participant P2 wondered if he should accept the terms of use of the application, as he worried that sensitive data from the app could be shared with an unknown company. Participant P6 was unable to complete this task and had to abandon the test because the application restricted the use in the third step of the registration when the user informed his date of birth. However, it's important to note that Google Play's classification says "Free" and that nowhere on the page it says that you must be at least 18 years old to use it.

**T3 - add an exercise:** When reporting an exercise, the application did not report the approximate number of calories and participant P3 could not estimate one. Three participants (P2, P3, P5) were in doubt about having to choose between the equal options listed when searching an exercise.

**T4 - report current weight:** one participant (P5) did not complete this task correctly. Instead of reporting his weight in the application, participant P5 edited the previously entered weight in the application settings. Two participants (P1 and P3) were slow to find the weight reporting functionality.

**T5 - monitor progress:** participants were expected to identify their progress chart after updating the weight (T4), in the screen following the one for reporting current weight. Except for participant P4, however, all participants had difficulty understanding what their progress would be. The participants P1 and P5 were unable to complete the task correctly, identifying another progress presented in the application.

## 5. Textual evaluation vs. Users test

In this section some aspects of comparison between the two methodologies are shown.

**Number of problems found:** from 3,581 sentences, 70 problems (sum of criticism, doubts and comparison) were identified in the Textual Evaluation. For user testing, 7 problems were found from the test performed by 6 participants.

**Application time of the technique:** the textual evaluation without using automated analysis can be very long as it depends on the number of posts to be analyzed, the number of evaluators and their experience. Although the user testing is faster to apply, the time can vary according to the evaluation scope and availability of participants, whether these are evaluators or users. In this study the evaluators of Textual Evaluation took 20 hours (non-consecutive) to evaluate 3,581 posts, devoting 1 hour per day, on average. To apply the Test method to users, 7 hours and 12 minutes were used. This time was distributed between planning, execution and analysis of the results.

**Number of users involved:** in textual evaluation users only participate publishing PRUs. It is not possible to estimate how many users are participating in each textual evaluation, because it depends on the source of extraction, the number of posts extracted, and a user can publish one or more PRUs. In this case, the extraction source does not have a unique user id that can be extracted and analyzed. For user testing, the author of [Nielsen 2012] states that it takes 5 users to find most usability issues. When the evaluator increases the number of users, the same problems start to emerge repeatedly.

**Information about users:** with the textual evaluation, we could see that that the posts contained information about the users, such as: time of use and age. In some posts, users reported: "*...I've been using it for 3 weeks...*", "*...I'm 15 years old...*". However, this type of information was identified in a few posts. An example is the Twitter which can provide information such as: age, location, language, account creation date and others. In this study, the source used did not provide information about users. In the user testing, this information could be identified because some questions were asked in order to categorize the participants, asking your age and health application experience.

## 6. Discussion

Next are presented and discussed some important points of the article and factors that may have influenced the results.

**Web Social:** users' opinions on web about use of products and/or services take advantage of collaboration between individuals on the web. Before purchasing a product or service, users try to analyze the opinions (experiences) of other users. Those who had their experiences try to be more specific in their assessments, indicating usage details that can help (or influence) future users or encourage their manufacturers to improve them. This type of assessment, individual (by each user - future or experienced), is increasingly present in the context of the social web. Textual Evaluation leverages this concept of collaboration to provide insight into the evaluation of a product/service.

**Spontaneity of the users:** by the spontaneity of the users, we mean that they express themselves without the influence of someone. This is important to get the most accurate information about usage. In this study, the posts presented evidences of spontaneity of the users, because they were extracted from a social system where the users publish their reviews to other users. The presence of an evaluator could influence the user experience during use [Korhonen et al. 2010, Fetter and Gross 2014], so we can't say if there was spontaneity in user testing.

**System evaluated:** the selection of the system may influence the characteristics of the extracted posts. In previous evaluations, applying the same methodology, in an evaluation on Twitter [Mendes 2015] resulted in more criticism; an academic system [Mendes 2015] resulted in more doubts; and mobile applications [Freitas et al. 2016, da Silva et al. 2017] resulted in more praise. It is also possible that even the means of extracting the post influences usage reports. In the case of the academic system, the reports were collected from a discussion forum by students and teachers, which may have resulted in more descriptive reports.

**Time of application x Evaluators' experience:** the application time of the techniques depends on the amount of posts extracted and the level of experience of the evaluators. In an initial evaluation [Freitas et al. 2016], the classification time was high even with 100 posts. The time decreased as the classification experience rose. For example, 1,018 PRUs were classified in 10 days in a previous study [da Silva et al. 2017], and 3,851 posts in 20 hours, in this study. Studies involving machine learning have been conducted in our research group to automate the process of classification of those posts. As soon as we finish these algorithms, the classification in some categories of evaluation can be done automatically. This will make the classification more efficient and facilitate the application of the method by anyone, since the tool currently functions as a shared worksheet in which evaluators classify the posts. The tool assisted in the identification of posts with different classifications and in the accounting of the results and generation of charts to be interpreted by the evaluators.

**Characteristics of the methods used to complement each other:** as a way of complimenting the results of the Textual Evaluation, we can use the technique of Users testing. An example of this would be to conduct the evaluation in the following steps: (a) do the Textual Evaluation and observe which features are most critical, doubts and comparison; (b) conduct a user test to understand how the use of these features is done in a real-world environment. Thus, getting a better description of the problems.

**Usefulness of the results of a Textual Evaluation versus Other Techniques:** one of the uses of the results of textual evaluation is that one can be more assertive in identifying: (a) what goes well or wrong in the system; (b) what in the system needs to be modified; (c) obtain evolutionary requirements for the system. In comparison with the other techniques used, only problems can be identified. Textual Evaluation is limited to systems that already exist; however, would this technique be able to obtain requirements for a new system based on users' opinions?

**HCI and other sciences:** this article alert the evaluators that much information is produced by the users themselves and we do not have to produce them. Areas such as data science, associated with artificial intelligence and conversational interfaces, have arisen for the purpose of analyzing different data formats related to the services offered to clients. The HCI area must take advantage of the capacity that these professionals are acquiring to reduce the time and effort necessary for the analysis of texts on the UX with the developed computer systems.

## 7. Conclusion and Future Work

This work carried out a comparative analysis of the Textual Evaluation method with users testing. From this comparative analysis we have identified that the User Testing technique can complement the results of a textual evaluation as long as your scope of application is limited to the problems identified by the Textual Evaluation. Thus, it avoids wasting time and effort in applying these other techniques in search of irrelevant problems. These techniques can be used to better describe the results of Textual Evaluation, since this evaluation is limited to what users report about using the System.

Problems with the app were identified in the results of each evaluation technique. The textual evaluation identified more different problems (70). Although the problems in previous textual evaluations [Mendes 2015, Freitas et al. 2016, da Silva et al. 2017] had a greater level of detail, in the evaluation performed in this paper, most problems (96%) did not present details, with other evaluation techniques being required to complement the results. On the other hand, this technique provided evaluators with a perception of the system UUX through the user's opinions.

The application of this technique in different systems is planned in future work in order to understand how the use context of the system can produce different results. In addition to evaluating the amounts of problems per posts, repeated posts, supports of the tool and other forms to support this type of evaluation. Another study to be carried out is to investigate who are the users talking about the use of systems. System improvements can be proposed based on the users' definition, enabling a better understanding to explain their praise and/or criticism. For example, if criticisms are found regarding a functionality, but 80% of users have 3 (three) years of use, one could conclude that this is a serious problem.

## Referências

Bach, C. and Scapin, D. L. (2010). Comparing inspections and user testing for the evaluation of virtual environments. *Int. J. Hum. Comput. Interaction*, 26:786–824.

Barbosa, S. and Silva, B. (2010). *Interação Humano-Computador*. Elsevier Brasil.

Bargas-Avila, J. A. and Hornbæk, K. (2011). Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2689–2698. ACM.

Bim, S. A., de Castro Salgado, L. C., and Leitão, C. F. (2016). Evaluation by inspection: Comparing methods of practical, cognitive and semiotic basis. In *Proceedings of the*

*15th Brazilian Symposium on Human Factors in Computing Systems*, IHC '16, pages 9:1–9:10, New York, NY, USA. ACM.

da Silva, T. H. O., Freitas, L. M., and Mendes, M. S. (2017). Beyond traditional evaluations: User's view in app stores. In *Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems*, IHC 2017, pages 15:1–15:10, New York, NY, USA. ACM.

de Souza Filho, J. C., Brito, M. R. F., Mendonça, A. R. R., Martins, M. V., and Sampaio, A. L. (2018). Hidrate spark: Avaliando um sistema ubíquo para motivar a ingestão de Água. In *Anais Estendidos do XVII Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais*, Porto Alegre, RS, Brasil. SBC.

Fetter, M. and Gross, T. (2014). Lilole—a framework for lifelong learning from sensor data streams for predictive user modelling. In *International Conference on Human-Centred Software Engineering*, pages 126–143. Springer.

Freitas, L. M., da Silva, T. H. O., and Mendes, M. S. (2016). Evaluation of spotify: An evaluation textual experience using the maltu methodology. In *Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems*, IHC '16, pages 50:1–50:4, New York, NY, USA. ACM.

Gray, W. D. and Salzman, M. C. (1998). Damaged merchandise? a review of experiments that compare usability evaluation methods. *Hum.-Comput. Interact.*, 13(3):203–261.

Hedegaard, S. and Simonsen, J. G. (2013). Extracting usability and user experience information from online user reviews. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2089–2098, New York, NY, USA. ACM.

Hedegaard, S. and Simonsen, J. G. (2014). Mining until it hurts: Automatic extraction of usability issues from online reviews compared to traditional usability evaluation. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, NordiCHI '14, pages 157–166, New York, NY, USA. ACM.

Korhonen, H., Arrasvuori, J., and Väänänen-Vainio-Mattila, K. (2010). Let users tell the story: Evaluating user experience with experience reports. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, pages 4051–4056, New York, NY, USA. ACM.

Mendes, M. S. (2015). *MALTU - Um modelo para avaliação da interação em sistemas sociais a partir da linguagem textual do usuário*. PhD thesis, Universidade Federal do Ceará (UFC).

MyFitnessPal (2019). Myfitnesspal - contador de calorias @ONLINE. Acesso em 02 de maio de 2019.

Nielsen, J. (2012). How many test users in a usability study? @ONLINE. Acesso em 18 de abril de 2019.

Norman, D. (2004). *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic Books.

Rogers, Y., Sharp, H., and Preece, J. (2013). *Design de Interação - 3ed*. Bookman Editora.