The Journal of Systems & Software 193 (2022) 111462

Contents lists available at ScienceDirect

The Journal of Systems & Software

journal homepage: www.elsevier.com/locate/jss

What factors affect the UX in mobile apps? A systematic mapping study on the analysis of app store reviews ☆

Walter T. Nakamura [a,b,*], Edson Cesar de Oliveira [c], Elaine H.T. de Oliveira [b], David Redmiles [d], Tayana Conte [b]

[a] Federal University of Technology – Paraná (UTFPR), Academic Department of Computing (DACOM), Rua Rosalina Maria Ferreira, 1233, Campo Mourão, Brazil
[b] Federal University of Amazonas (UFAM), Institute of Computing (IComp), Avenida Rodrigo Otávio, 6200, Manaus, Brazil
[c] Secretaria de Estado da Fazenda do Amazonas – SEFAZ/AM, Avenida Andre Araujo, 150, Manaus, Brazil
[d] University of California Irvine (UCI), Department of Informatics, Irvine, CA 92697, United States of America

ARTICLE INFO

Article history:
Received 5 April 2021
Received in revised form 21 July 2022
Accepted 27 July 2022
Available online 3 August 2022

Keywords:
User experience
Influencing factors
App store reviews
Systematic mapping study

ABSTRACT

Researchers and practitioners are becoming aware of the importance of User eXperience (UX) in mobile app development. Developing merely usable apps became insufficient to meet users' needs, requiring developers to focus on promoting pleasurable experiences to get a competitive advantage. To that end, it is crucial to understand what factors can lead to positive or negative UX. In this scenario, app store reviews emerged as a valuable source to address UX issues from analyzing several self-reports of end-users experiences in the wild. Many researchers have proposed approaches to analyze such reviews and investigate the effect of factors related to the user (e.g., gender, culture) and the app (e.g., bugs, features) on UX. However, the fragmentation of the results into various studies makes it difficult to draw conclusions that can support the development process and advance the research in the field. This paper presents a systematic mapping study to address publications that analyze app store reviews and identify the factors affecting UX reflected on users' ratings and sentiment. From 25 accepted publications, we extracted 31 factors and their associated polarities. We also identified research gaps and future work opportunities with implications for practitioners and researchers.

1. Introduction

Mobile Internet users worldwide have increased in the last decades, from 738.27 million users in 2010 to over 4.3 billion users in 2020.[1] This increase in the number of users led to the rapid growth of the mobile market, which resulted in the development of a wide variety of apps that users could choose. Consequently, users became less tolerant of faulty apps (Durelli et al., 2018), which raises the importance of developers to evaluate the quality of the apps they are developing.

There are two important quality attributes to evaluate software applications: usability and User eXperience (UX). While usability is a task-oriented attribute that measures the extent to which a system, product, or service allows users to achieve their goals with efficiency and effectiveness (ISO 9241-11, 2018), UX, in turn, is more holistic. In addition to pragmatic aspects related to task accomplishment addressed by usability, UX also considers every part of user interaction, from the consequence of users' internal state, i.e., their affects, sensations, and emotions, to the characteristics of the designed system and the context where the interaction occurs (Hassenzahl and Tractinsky, 2006; Law et al., 2009). Both quality attributes play an important role in developing software applications, as more usable products induce higher levels of pleasure (Porat and Tractinsky, 2012). This pleasure, in turn, can determine users' willingness to repeat the experience and keep using the software application (Cockburn et al., 2017).

From the academic perspective, many studies have been conducted to understand how users form their judgments about their experiences. In this context, researchers started investigating which factors could affect users' perception of the experience resulting from the interaction with a software product. Researchers have been studying various factors related to UX, such as expectations (Kujala et al., 2017; Michalco et al., 2015),

☆ Editor: Nicole Novielli.

* Corresponding author at: Federal University of Technology – Paraná (UTFPR), Academic Department of Computing (DACOM), Rua Rosalina Maria Ferreira, 1233, Campo Mourão, Brazil.
E-mail addresses: waltertakashi@utfpr.edu.br (W.T. Nakamura), edson.cesar@sefaz.am.gov.br (E.C. de Oliveira), elaine@icomp.ufam.edu.br (E.H.T. de Oliveira), redmiles@ics.uci.edu (D. Redmiles), tayana@icomp.ufam.edu.br (T. Conte).

[1] https://www.statista.com/forecasts/1146312/mobile-internet-users-worldwide.

https://doi.org/10.1016/j.jss.2022.111462
0164-1212/

the mental effort expended to use the product (Hassenzahl and Sandweg, 2004), gender, and previous experience (Sagnier et al., 2020). By investigating how these factors affect UX, researchers could, for instance: (i) minimize bias in UX evaluations; (ii) create techniques that guide developers into reliable results by taking into account the influence of these factors; (iii) avoid rework in the app development process by considering the existence of these factors beforehand; (iv) support the redesign of an app by identifying the impact of the factors affecting UX. For instance, Sagnier et al. (2020) highlighted that previous experience and gender could influence UX in an immersive context. They suggest that both should be considered in future UX studies, as they can affect users' perception of the experience and influence the results. In addition to these factors related to users, there are other factors related to the software that can affect UX as well. Technical faults and bugs, for instance, were often associated with UX degradation (Kujala et al., 2011). In this sense, researchers have been studying the effect of factors related to the software itself, such as the fault proneness of the API that the app employs (Bavota et al., 2015), features (Guzman and Maalej, 2014), and interaction issues, such as the peak-end effect (Bruun and Ahm, 2015; Cockburn et al., 2015; Gutwin et al., 2016). For instance, Pagano and Maalej (2013) identified that reviews requesting new content are the least critical (4.25 stars on average). Their results indicate that users do not penalize the app so much due to the lack of content, allowing developers to focus on other more critical factors when improving the app. Researchers can also use this finding to create UX evaluation techniques that attribute weights to the evaluated items according to their impact on UX to obtain more precise indicators.

From the industrial perspective, practitioners are becoming aware of the importance of providing a good user experience when designing software applications. A survey from Lallemand et al. (2015) revealed that 83.9% of the practitioners consider UX as central or very central for their professional work, indicating that it is an essential attribute in the development and evolution of software applications. Understanding how technology can promote unique, satisfying, and enlightening experiences can bring a competitive advantage to business and industry (Alves et al., 2014). However, evaluating UX can be expensive and time-consuming, requiring highly-trained experts or several users to perform tasks (Hedegaard and Simonsen, 2014). In this context, the availability of open sources of user opinions, such as user reviews from app stores, can be a potential alternative to analyze the quality of mobile applications and identify factors that can improve or deteriorate the experience.

App stores consist of a centralized directory for mobile software applications, making installing and obtaining updates easier (Goul et al., 2012). They also store millions of reviews written spontaneously by users worldwide for various mobile apps, providing rich information for designers and analysts, such as bug reports, missing features, improvement requests, and experience reports. These reviews also have an important role in users' decisions. According to Weichbroth and Baj-Rogowska (2019), more than 70% of people read reviews before downloading an app, and 75% consider them a key driver for downloading, being considered even more trustworthy than personal recommendations by 42% of people. Moreover, negative reviews can significantly influence a company's revenue and the rate of product awareness (Hoon et al., 2012). Through user reviews, developers can identify which factors users are reporting. Moreover, they can investigate how they affect UX by analyzing, for instance, the review's rating and the score obtained from the sentiment analysis, as these two mechanisms describe users' perception of the experience with the application (Rodrigues et al., 2017). Developers can use this information to prioritize development efforts based on factors with a stronger correlation with higher and lower ratings (Nakamura et al., 2019), either for an app refinement or competitive market entry perspective (Simmons and Hoon, 2016). Moreover, the massive amount of data available makes it a valuable source of information that practitioners and researchers can use to identify patterns, tendencies, and factors that affect UX in various software applications, which would be expensive to achieve through empirical studies.

Various researchers have been working on the topic in the last decades to investigate what issues users report in their reviews and how these affect their evaluations. From the manual analysis of the reviews (Ha and Wagner, 2013; Khalid et al., 2015; Nicolai et al., 2019) to automated information extraction through machine learning techniques (Fu et al., 2013; Guzman and Maalej, 2014; Luiz et al., 2018), these works provide valuable contributions to both Software Engineering (SE) and Human–Computer Interaction (HCI) communities through methods and tools designed to analyze and classify user reviews for extracting factors that are leading to positive or negative evaluations. However, the fragmentation of these results into various studies makes it difficult for practitioners and researchers to identify factors affecting UX.

Although secondary studies have been conducted to address publications analyzing user reviews (Ebrahimi et al., 2020; Genc-Nayebi and Abran, 2017; Martin et al., 2017; Wang et al., 2019), they did not focus on identifying the effect of factors on users' ratings and sentiments. Through a systematic mapping study on the topic, we can summarize the factors that could affect UX and identify which methods have been applied to analyze the effect of these factors on UX. The broad view of a systematic mapping study allows gathering the perspective of various analyses to have a comprehensive idea of what factors have been identified and considered by different researchers when analyzing user reviews. It also allows gathering the results from several studies conducted in various datasets and contexts to obtain a more thorough analysis and draw conclusions that would be hard to get through isolated app review studies.

This paper presents a systematic mapping study to address publications that analyze user reviews from app stores. Our goal was to identify the factors that can affect the UX conveyed by mobile applications, the methods employed to analyze such reviews, the scope of these analyses, and the implications of these results for practitioners and researchers. The main contributions of this work are:

- An overview of state of the art on analyzing user reviews from app stores with a focus on UX;
- A set of factors that can affect UX and influence users' evaluations in app stores;
- An overview of the methods employed to analyze the reviews;
- Research gaps and opportunities for future work with implications to both practitioners and researchers.

The remainder of this paper is as follows: Section 2 presents related work. Section 3 describes the review protocol we followed to perform this systematic mapping study. Section 4 presents the results. In Section 5, we discuss the main findings and research opportunities. Section 6 presents the threats to validity. Finally, Section 7 concludes the paper.

2. Related work

User eXperience evaluation plays an essential role for practitioners from the industry. According to the survey from Alves et al. (2014), 93% of the practitioners considered it has a strong to decisive impact at the user interface level and 50% at the

business level. They also believed that it could lead to a new or updated user interface and implementation requirements, raising the importance of conducting such evaluations.

Several UX evaluation methods have been proposed over the years. Among them, questionnaires, interviews, and observation are the most employed (Bargas-Avila and Hornbæk, 2011; Pettersson et al., 2018). However, the main drawback of such methods is that they are often "resource-intensive and not scalable" (Chromik et al., 2020), requiring extensive experimentation (Hedegaard and Simonsen, 2013). In this scenario, the analysis of reviews from app stores emerged as a potential and cheap source of information to address UX issues, providing self-reports of several end-users experiences in the wild (Hedegaard and Simonsen, 2013). Moreover, the presence of both qualitative (the review itself) and quantitative data (ratings and sentiments) allows triangulating the results to validate the UX data and grasp it better by using qualitative data to understand patterns from quantitative data and vice-versa (Pettersson et al., 2018). In this sense, various researchers started to develop approaches to analyze such reviews.

In early works, most of the analyses had an exploratory nature, for instance, by manually analyzing a sample of reviews to investigate the vocabulary used (Hoon et al., 2012), the relationship between review length and ratings (Vasa et al., 2012), and to extract common topics (Ha and Wagner, 2013; Iacob et al., 2013). However, this manual analysis becomes limited as the number of reviews increases, making it difficult to analyze a representative sample of reviews and draw conclusions, especially for developers of popular apps. In this context, the introduction of machine learning techniques to analyze massive amounts of reviews accelerated the research on the topic (Genc-Nayebi and Abran, 2017).

From the automatic extraction of topics (Chen et al., 2014; Fu et al., 2013; Guzman and Maalej, 2014) to the investigation of specific factors, such as the influence of gender (Guzman and Paredes Rojas, 2019), culture (Guzman et al., 2018), and sentiment on users evaluations, the use of machine learning emerged as a promising and more resource-effective approach (Chromik et al., 2020) to help understanding UX by finding patterns and relationships instead of isolated data (Pettersson et al., 2018). Some researchers have been performing literature reviews to address these publications and provide a broad view of the research on user reviews mining. We present these works below.

Ebrahimi et al. (2020) aimed to address studies on mobile app privacy from the SE perspective. They performed a systematic mapping study to explore the trends on the topic, categorize the existing evidence, and identify opportunities for future research. Their categorization identified four studies that considered privacy issues reported in app reviews. One study (Khalid et al., 2015) reported that privacy was associated with the most negative impact on app ratings. However, the focus of the systematic mapping on privacy issues limits its scope.

The systematic mapping from Wang et al. (2019) focused on addressing studies that extracted SE requirements from the crowd. The authors identified which information was extracted from user opinions and in which Requirements Engineering (RE) activities this information was applied. The main limitation of this systematic mapping study is that it did not further discuss which information was obtained from user reviews for each RE activity. The authors only presented a table with examples of how some publications used the data obtained from the app stores.

In another work, Martin et al. (2017) performed a non-systematic literature review to identify publications that performed App Store Analysis for Software Engineering with a focus on the combination of technical attributes (e.g., API usage, platform version, size) and non-technical attributes (e.g., app category, number of installs, ratings). The authors identified a set of publications on app store review analysis. Among them, they described some publications that reported factors affecting users' feedback. However, the authors did not extract the factors from these publications for further analysis.

Finally, Genc-Nayebi and Abran (2017) aimed to identify proposed solutions and challenges in mining user opinions from app stores. They defined five research questions to identify the mining techniques, remedies for domain dependency challenge, criteria that make a review useful, spam detection approaches, and extracted features. The authors mention some factors during the discussion on the features extracted from app reviews. However, they did not investigate their effects on users' ratings and sentiments, as it was not the paper's focus.

Although literature reviews on the topic exist, to the best of our knowledge, no previous work focused on addressing factors that can affect UX and influence users' evaluations in app stores. The body of knowledge from our systematic mapping review would provide valuable information for practitioners and researchers. Practitioners could, for instance, identify which factors to prioritize according to their effect on UX, while researchers could develop evaluation methods and guidelines based on the addressed factors. Although researchers could analyze large datasets of user reviews to investigate these factors, the analysis of the outcomes would be limited to the focus and the context of the study. For instance, Nicolai et al. (S25) took the taxonomy from Panichella et al. (2015) as a starting point for the classification process. Then, they refined this taxonomy by including three new categories to classify the reviews according to their focus and needs. It highlights researchers looking for existing literature approaches as a starting point, either for reducing effort and time, making comparisons, validations, or extensions of existing work. Finally, the list of factors identified in this systematic mapping study can be a starting point for future research and refined by future studies. In this work, we contribute by determining the state of the art on the topic and a set of factors that practitioners and researchers can consider, for instance, to develop new software applications and evaluation methods.

3. Review protocol

Before starting our systematic literature review, we developed a review protocol according to the guidelines proposed by Kitchenham et al. (2015). This protocol defines the procedures to perform the systematic literature review, being an important document for both the validity and the practical conduct of the review (Wohlin et al., 2012). We present details of this protocol in the following subsections.

3.1. Research question

In this systematic mapping study, we aimed to answer the following main research question: "*What are the UX-related factors that influence users' evaluations in app store reviews, and how do they affect UX?*". We also defined seven sub-questions to answer specific questions related to: (i) *dataset source*: to identify the target population; (ii) *extracted information*: to identify which data were obtained for analysis; (iii) *analysis methods*: to understand how the data was analyzed; (iv) *data categorization*: to identify whether and how the data was organized; (v) *scope of the analysis*: to verify how comprehensive the study was in terms of analysis and apps sample; (vi) *identified factors and their associated polarities:* to identify whether the factor affects UX positively or negatively; and (vii) *factor influence analysis*: to investigate the extent the impact of the factor was analyzed. Table 1 presents each research sub-question.

Table 1
Research sub-questions.

Sub-question	Description
SQ1	What was the source of the analyzed reviews?
SQ2	What is the information extracted from the sources?
SQ3	Which methods were used to analyze the data extracted?
SQ4	Was the information categorized? How?
SQ5	What are the dataset size and analysis scope of the extracted publications?
SQ6	What are the identified factors and their associated polarity?
SQ7	Was the influence of the factor on user rating or sentiment analyzed?

3.2. Research scope

We carried out this systematic mapping on the IEEE Xplore, ACM, and Scopus. While Scopus is a meta-library that indexes publications from several well-known publishers (e.g., Springer, Elsevier, and Taylor & Francis), ACM and IEEE are two main digital libraries in computer science. We selected these databases as previous systematic literature reviews recommend them as adequate and relevant ones (Dyba et al., 2007; Mendes et al., 2020; Petersen et al., 2015). Additionally, we performed a backward snowballing process, following the references from each selected paper to identify other relevant ones (Wohlin et al., 2012).

3.3. Language

We selected only publications written in English, given that most international conferences and journals adopt it as the primary language. Additionally, English is the dominant language for global communication, thus making it possible to replicate or extend this systematic mapping study by other researchers.

3.4. Search terms

We first defined a set of control papers that the search engines should return to build our search string. To do so, we analyzed the publications from a systematic mapping study conducted by Genc-Nayebi and Abran (2017) that addressed studies on app store opinion mining and selected those that presented factors associated with positive, negative, or neutral evaluations.

Next, we followed the procedure described by Kitchenham and Charters (2007) to define the terms of the research. They suggest defining five parameters: population, intervention, comparison, outcome, and context. Given that our focus is not to compare interventions, we did not use the comparison parameter. Table 2 presents the set of terms for each parameter defined below:

- **Population:** user reviews from app stores;
- **Intervention:** methods/techniques employed to analyze user reviews and identify influencing factors;
- **Comparison:** not applicable, as the goal is to identify the factors from the literature;
- **Outcomes:** the effect of these factors on UX;
- **Context:** within the domain of mobile app stores.

To build the search string, we used the boolean operator "OR" between the words with similar ideas for each parameter and the boolean operator "AND" to join the four parameters (see Table 3). We tested the string several times with different combinations of words to reduce the number of publications unrelated to the research topic while ensuring that the set of control publications is returned. To define the control set, we analyzed the 24 publications from a systematic literature review (Genc-Nayebi and Abran, 2017) related to our topic. We identified four publications (S02, S03, S04, and S05) that report factors influencing users' ratings and sentiments, which comprised our control set.

Table 2
Terms selected to compose the search string.

Population	Intervention	Outcomes	Context
review	mining	experience	mobile app
opinion	analysis	UX	mobile apps
comment	processing	usability	mobile application
rating	examining	sentiment	app store*
			appstore*
			app marketplace
			app market
			app markets
			application market*

Table 3
Final search string.

(review **OR** opinion **OR** comment **OR** rating) **AND** (mining **OR** analysis **OR** processing **OR** examining) **AND** (experience OR UX **OR** usability **OR** sentiment) **AND** ("mobile app" **OR** "mobile apps" **OR** "mobile application*" **OR** "app store*" **OR** appstore* **OR** "app marketplace*" **OR** "app market" OR "app markets" **OR** "application market*")

3.5. Selection criteria

We defined a set of inclusion and exclusion criteria to select publications related to our research, i.e., publications that present factors associated with positive, negative, or neutral evaluations (see Table 4). For analyzing the papers, we first needed to define the concept of UX to delimit the scope of UX-related factors. We adopted the definition from Hassenzahl and Tractinsky (2006), which, according to Lallemand et al. (2015), was the most preferred definition among practitioners and researchers. According to it, UX is "a consequence of a user's internal state (predispositions, expectations, needs, motivation, mood, etc.) the characteristics of the designed system (e.g., complexity, purpose, usability, functionality, etc.) and the context (or the environment) within which the interaction occurs (e.g., organizational/social setting, the meaningfulness of the activity, the voluntariness of use, etc.)". In this sense, we addressed factors related to the user (e.g., expectations, emotions, sentiment, demographics) and the app itself (e.g., bugs, features, functionalities). Finally, considering that we aimed to assess the effect of these factors on UX, we only included publications that reported the effect of the factor on users' ratings or reviews' sentiment, given that these two pieces of information convey the experience the user had with the application (Rodrigues et al., 2017).

The selection process comprised two steps called filters. We read each publication's title and abstract in the first filter to select those related to our research topic. First, we assessed whether the paper addressed user reviews from app stores. Then, we analyzed whether it considered UX by explicitly mentioning it or using other related terms (e.g., emotions, usability, satisfaction, sentiment). Finally, we analyzed whether the paper discussed the impact or influence of variables (factors) on UX, ratings, or sentiment. Notably, some publications did not discuss the results in the abstract, which did not allow us to know whether they identified the impact or influence of factors on UX. Thus, we

Table 4
Inclusion and exclusion criteria applied in the systematic literature review.

#	Inclusion criteria
IC1	Publications that present UX-related factors associated to negative, positive or neutral reviews from app stores.
#	Exclusion criteria
EC1	Publications that do not present UX-related factors associated to negative, positive, or neutral reviews from app stores.
EC2	Publications that are not available for reading or collecting data (publications that are accessible only through payment or are not provided by the search engine).
EC3	Publications that are not written in English.
EC4	Books, proceedings, websites, and grey literature.
EC5	Duplicated publications.

Table 5
Strength of agreement associated with kappa statistics.

Kappa statistic	Strength of agreement
<0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

decided to include such papers in the first filter to thoroughly read in the second filter to avoid missing important publications. In the second filter, we thoroughly read the publications included in the first filter. We applied the inclusion and exclusion criteria in both steps to filter the publications.

To avoid the single researcher bias, we conducted the systematic literature review involving two researchers. Before performing the first and second steps, the researchers independently classified, according to the inclusion and exclusion criteria, a sample of 20 randomly selected publications. Then, we evaluated the level of agreement between the researchers by applying Cohen's Kappa (Cohen, 1960) to ensure that the criteria were well defined and understood. The result indicated an almost perfect agreement ($k = 0.89$) according to the interpretation of Landis and Koch (1977) in Table 5.

3.6. Data extraction strategy

After selecting the publications, we started to extract the data. To do so, we created a form to extract primary studies (see Appendix A) by following the strategy proposed by Fernandez et al. (2011), which consists of defining a set of possible responses. We defined initial responses and refined them in an iterative process through screening the control set, as "important trends and ways of categorizing papers may only become evident as individual papers are read" (Kitchenham et al., 2015). Regarding analysis methods (SQ3), for instance, we identified the use of descriptive statistics (S02, S03, S04, and S05), statistical tests (S04 and S05), topic modeling (S03), sentiment analysis (S03), and manual analysis (S04 and S05). This strategy allows standardizing the extraction process, ensuring that the same data extraction criteria will be used for each sub-question, thus facilitating the classification.

For **SQ1 (dataset source)**, we defined the following responses: *(a) Google Play Store:* the dataset was obtained from the Google Play Store; *(b) Apple AppStore:* the dataset was obtained from the Apple AppStore; *(c) Windows Phone Store:* the dataset was obtained from the Windows Phone Store; and *(d) Other:* the dataset was obtained from other sources.

For **SQ2 (extracted information)**, we defined the following responses: *(a) Rating:* the star rating given by users when posting their review; *(b) User review:* the comments users made with their opinions, complaints, and suggestions for the app; *(c) App information:* app's metadata, such as release date, description, category, version, price, release/update notes, etc.; and *(d) Other:* information not directly obtained from the app store, such as app's project lifespan, number of commits, number of warnings in the source code, etc.

For **SQ3 (analysis methods)**, we defined the following responses: *(a) Topic modeling:* consists of discovering relationships between documents (as well as the terms that compose these documents) and topics, making it possible to organize textual documents according to the topics discovered (Durelli et al., 2018); *(b) Sentiment analysis:* is a method to explore the sentiment conveyed by people in textual data, determining whether the polarity of the text is positive, negative or neutral (Masrury and Fannisa Alamsyah, 2019); *(c) Descriptive statistics:* when the authors describe and graphically present an overview of the dataset being analyzed, for example, by employing techniques to measure the central tendency (median, mean, mode), dispersion (frequency, variance, standard deviation), and dependency (linear regression, Spearman, Kendall and Pearson correlations) (Wohlin et al., 2012); *(d) Statistical tests:* when the authors employed statistical tests (parametric or non-parametric) to test hypotheses and verify whether it is possible to reject a particular null hypothesis based on a sample from some statistical distribution (Wohlin et al., 2012); *(e) Manual analysis:* consists of performing a coding process, for example, by employing content analysis to extract topics manually; and *(f) Other:* when employing other types of analysis, tools, and frameworks, for example, performing static code analysis by using a tool to address potential issues in the source code.

For **SQ4 (data categorization)**, we defined the following responses: (a) *Yes:* when the authors grouped the data into categories during the analysis; (b) *No:* when there was no type of categorization. For each publication, we extracted the categories they used.

For **SQ5 (scope of the analysis)**, we defined the following responses: *(a) Individual:* the analysis was performed separately for each application; *(b) Group:* the analysis was performed for a given group or different groups of apps; *(c) General:* the analysis was performed in the whole set of applications, without separating them into groups or analyzing them individually. We also gathered information regarding the dataset size (number of apps, reviews, and categories).

For **SQ6 (identified factors and their associated polarity)**, we defined the following responses: *(a) Positive:* if the factors identified were associated with positive reviews or ratings; *(b) Negative:* if the factors identified were associated with negative reviews or ratings; and *(c) Neutral:* if the factors identified were neither associated with negative nor positive reviews or ratings. For each polarity, we reported the factors associated with it.

Finally, for **SQ7 (factor influence analysis)**, we defined the following responses: *(a) Yes:* if the publication analyzed and reported the influence of the factors on users' evaluation, for instance, by performing statistical tests, correlation analysis, or frequency distribution analysis; and *(b) No:* if the publication just reported the polarity of the factor, without assessing its influence

on users' evaluation. For each factor, we extracted the reasons behind positive and negative evaluations.

Given that the search engines could return secondary studies, we created a specific form to extract this type of publication, which can be found in Appendix B. In this form, we defined questions to address specific information inherent to secondary studies, such as the search string used, queried databases, and results.

4. Results

4.1. Selected publications

Fig. 1 presents the publication selection process. The search string returned 562 publications, of which 253 were from Scopus meta-library, 167 from IEEE, and 142 from ACM. Among them, 132 were duplicated, resulting in 430 unique publications. In the first filter (i.e., reading the title and abstract), 341 publications did not meet the inclusion criteria and were excluded. The remaining 89 publications followed the second filter to be fully read and submitted to the same inclusion and exclusion criteria of the first filter. A total of 71 publications did not meet the inclusion criteria and were excluded, resulting in 18 publications being accepted in the second filter. Notably, some of the excluded publications presented factors but did not associate them with ratings or sentiments, which resulted in their exclusion (EC1). During the conduction of this systematic mapping study, we had to deal with a tradeoff between precision (i.e., the proportion of relevant studies returned) and sensitivity (i.e., the proportion of retrieved studies that are relevant studies) (Zhang et al., 2011). We decided to adopt a broader string to increase sensitivity at the cost of some precision to avoid losing relevant papers during our search, which is usually more desired (Zhang et al., 2011). As a result, many publications were returned, but a small number of publications were included.

Regarding the backward snowballing process, we extracted 475 references from the 18 publications of the systematic mapping study. Among them, 110 were duplicated, resulting in 365 unique publications submitted to the same inclusion and exclusion criteria defined in the systematic mapping study. At the end of the process, seven publications passed in both the first and second filters. The complete list of publications can be found in Appendix C.

4.2. Publications overview

The selected publications were published between 2012 and 2019. The graph in Fig. 2 shows that the number of publications has grown since the first publication in 2012. The linear tendency, identified through simple linear regression, indicates a stable interest in investigating the reviews to understand better the application's positive, negative, and neutral aspects. Given that we performed this systematic mapping study in April 2020, the data for this year is incomplete, which may explain its lack of publications.

Most publications were published in conference proceedings (see Fig. 3). Only five publications were published in journals, and the other three were presented in conference workshops. We also analyzed in which fields these works were published. To do so, we looked for the description, scope, and call for papers section of the venue's website and categorized them. We identified 20 unique venues from nine different fields (see Fig. 4). Most of the publications were in the Software Engineering field. The most active authors on the topic are Guzman, E., Hassan, A. E., Khalid, H., and Nagappan, M., with three publications each (see Table 6). The last three authors have also worked together in all three publications (S09, S10, and S13). The number of citations of the papers in this systematic mapping study (see Fig. 5) reveals a considerable impact on the community. The publication from Guzman and Maalej (S06) has the most significant number of citations according to Google Scholar, with 576 citations, followed by the work from Pagano and Maalej (S05) and Khalid et al. (S10), with 550 and 447 citations each.

Table 6
Ranking of the most active authors in the topic.

Ranking	Authors	Publications	Total
#1	Guzman, E.	S07, S23, S24	3
	Hassan, A. E.	S09, S10, S13	3
	Khalid, H.	S09, S10, S13	3
	Nagappan, M.	S09, S10, S13	3
	Luiz, W.	S20, S22	2
#2	Maalej, W.	S05, S07	2
	Palomba, F.	S19, S25	2
	Rocha, L.	S20, S22	2
	Shihab, E.	S09, S10	2

In the following subsections, we present the results for each sub-question. It is worth mentioning that one of the publications (S17) was a secondary study. Thus, we extracted this publication using the extraction form for secondary studies and presented it separately in Section 4.10. Details of the mapping between the publications and each SQ can be found in Appendix D.

4.3. Dataset source (SQ1)

Regarding dataset source, most reviews were obtained from Google Play Store, followed by Apple AppStore (see Fig. 6). The reason to define the scope for a given app store is not informed in the publications we analyzed, except in (Pagano and Maalej [S05]), where the authors explicitly mentioned that they explored the reviews from Apple AppStore due to their prior experience with the technology and its applications. The preference to explore reviews from Google Play Store may be related to the greater number of mobile devices with the Android operating system, holding a market share of around 72.6% against 26.7% of devices with iOS, according to Statcounter GlobalStats[2] in May 2020.

The results also indicated a lack of studies involving different sources of information. Only Bano et al. [S07] and Guzman and Maalej [S18] analyzed reviews from both app stores. Two works used other data sources: Harman et al. [S02] extracted reviews from BlackBerry App World, while Kang and Park [S08] obtained the reviews from AppStoreHQ, a website that provides reviews of mobile apps from blogs, Twitter, and YouTube. Although the reviews obtained by Kang and Park [S08] were from iOS applications, we classified their dataset source as "other", given that they did not obtain the reviews directly from Apple AppStore. We did not find any work that analyzed reviews from Windows Phone Store. It may be due to its low popularity among mobile users and the small number of applications available for this operating system compared to Google Play Store and Apple AppStore. Additionally, Microsoft, responsible for maintaining Windows Phone and Windows 10 Mobile OS, discontinued them in 2017.

4.4. Extracted information (SQ2)

On the one hand, almost all publications extracted user reviews from app stores (see Fig. 7). Only one publication (Harman et al. [S02]) did not obtain user reviews but app descriptions, from where they extracted features to correlate them with other

[2] https://gs.statcounter.com/os-market-share/mobile/worldwide.

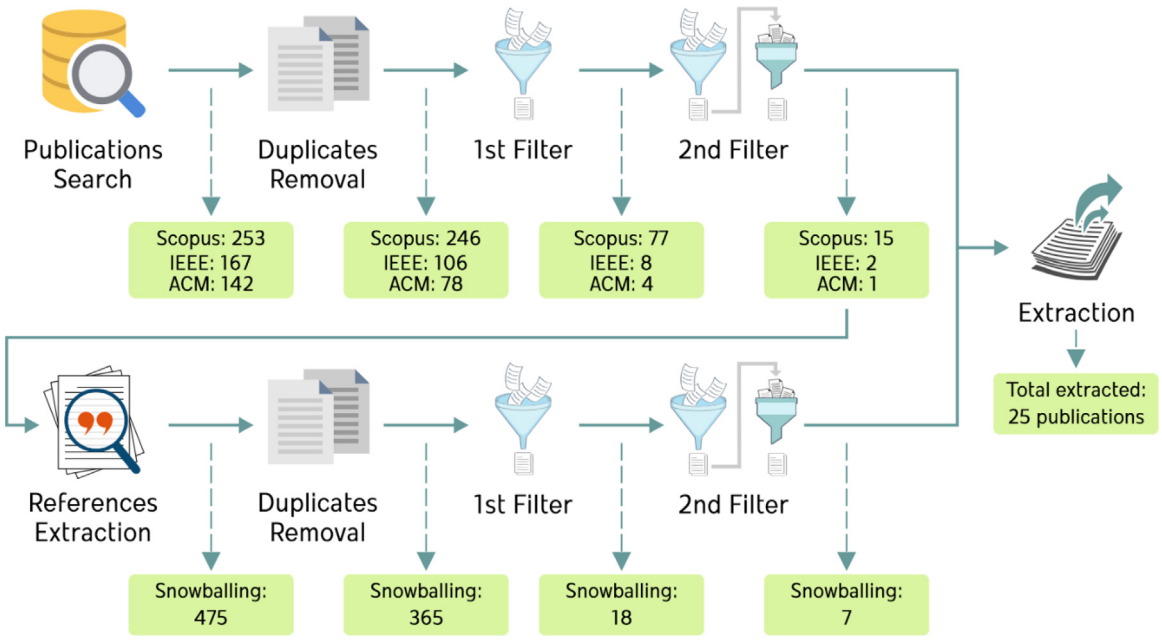

Fig. 1. Publication selection process.

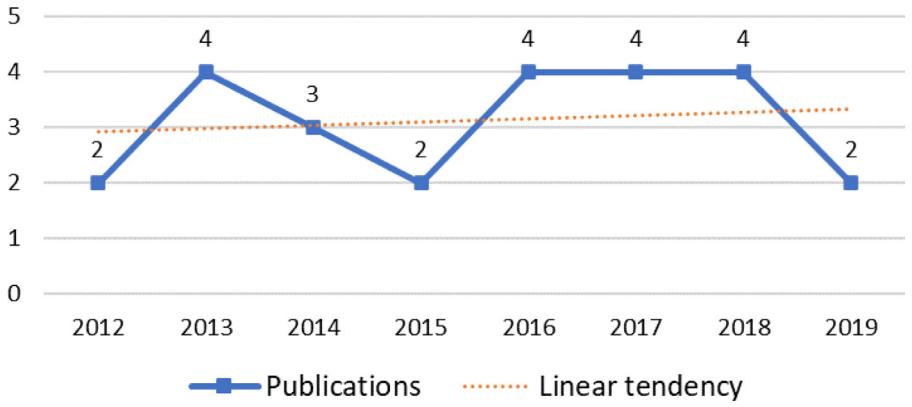

Fig. 2. Number of publications by year.

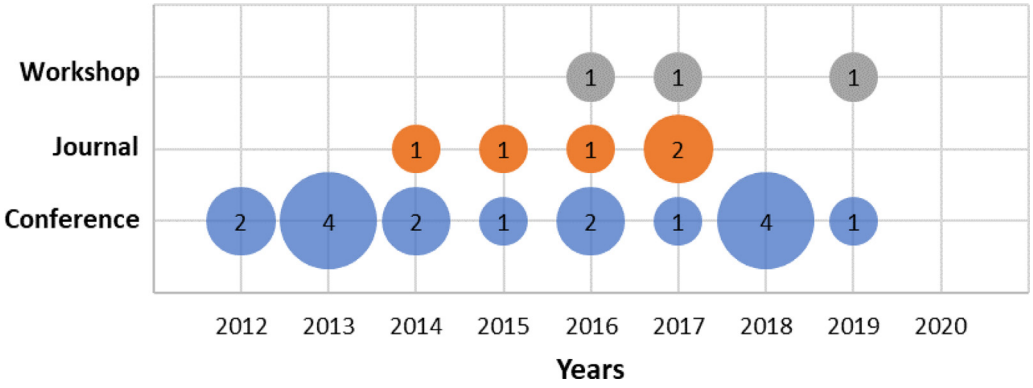

Fig. 3. Distribution of publications by year and local of publication.

variables, such as price, app rating, and the number of downloads. On the other hand, not all publications obtained user ratings for the analysis process. A possible explanation for this is that although star ratings provide a quick, direct, and objective overall evaluation of a particular app, it does not give further details about the reasons why the app has, for instance, an overall score of three stars (Luiz et al., 2018).

Among the works that extracted both star ratings and user reviews, most of them (16 out of 18) performed some analysis involving star ratings and other variables. These analyses were

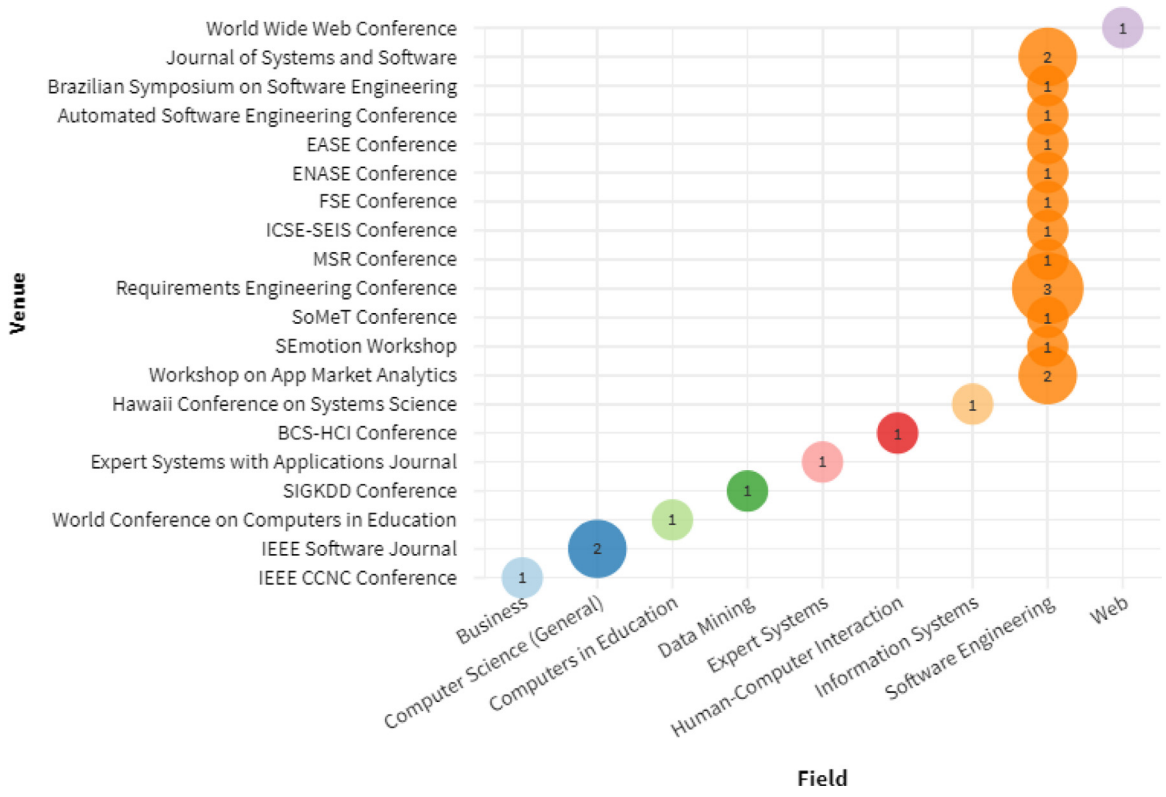

Fig. 4. Distribution of publications by venue and field.

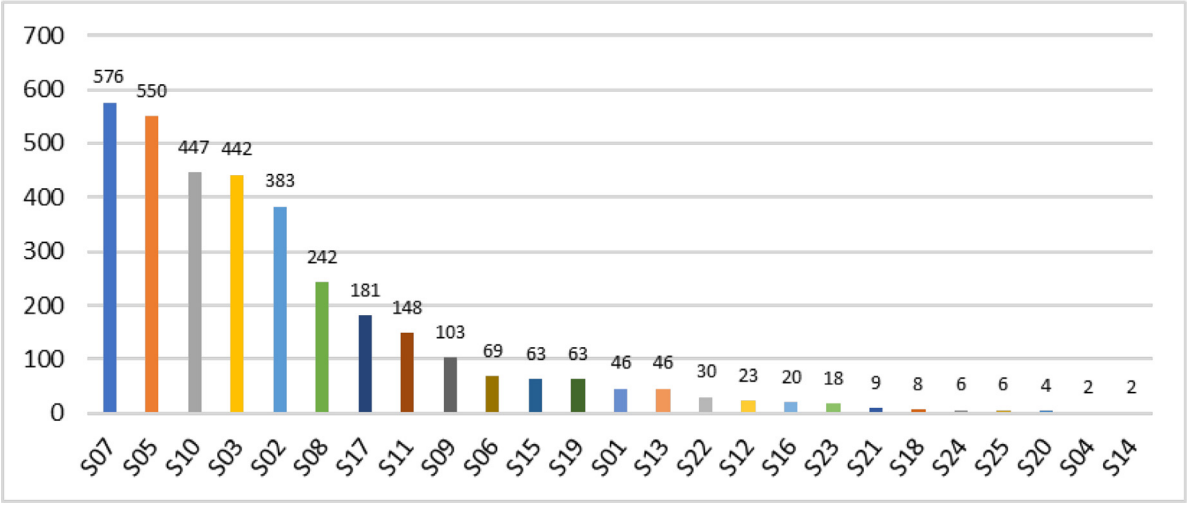

Fig. 5. Number of citations by publication.

carried out to identify inconsistent reviews (Fu et al. [S03], Luiz et al. [S22]), variations between ratings and sentiment analysis (Martens and Johann [S16], Luiz et al. [S22]), impact of different categories (Khalid et al. [S10]), type of app (healthcare vs non-healthcare) (Nicolai et al. [S25]), and features (Keertipati et al. [S15], Palomba et al. [S19]) on user ratings; as well as investigating correlations with gender (Guzman and Paredes-Rojas [S24]), culture (Guzman et al. [S23]), app version (Goul et al. [S01]), extracted topics from topic analysis (Ha and Wagner [S04], Iacob et al. [S06], Vu et al. [S11]), use of test cases (Durelli et al. [S20]), device model (Khalid et al. [S09]), and source code warnings (Khalid et al. [S13]).

Seven publications obtained app information such as app update dates, version, price, release notes, description, etc. Finally, three publications obtained other information and data, such as the Android Package Kit (APK) file to examine the code by using the FindBugs tool (Khalid et al. [S13]), app changelogs from Jira and Bugzilla (Palomba et al. [S19]), and users' device model (Khalid et al. [S09]).

4.5. Analysis methods (SQ3)

Descriptive statistics was the most used method to analyze the data (see Fig. 8). Five works (Goul et al. [S01], Harman et al. [S02], Khalid et al. [S09], Martens and Johann [S16], Guzman et al. [S23]) performed correlation analysis to *identify relationships* between different variables, such as prices, downloads, culture, and ratings. Four works (Ha and Wagner [S04], Pagano and Maalej

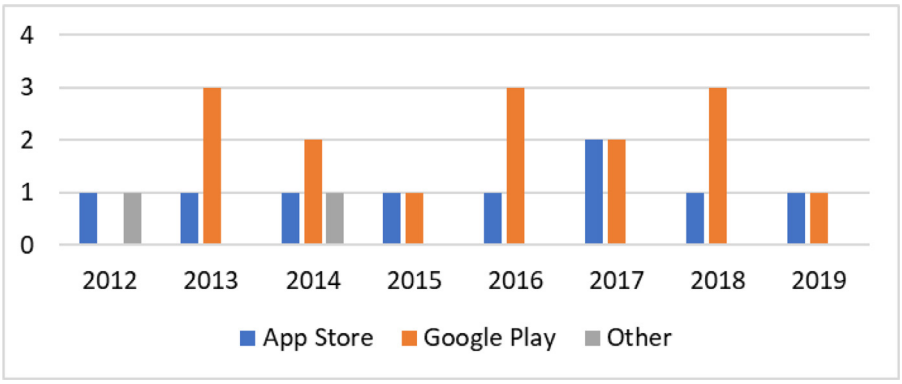

Fig. 6. Number of publications by dataset source.

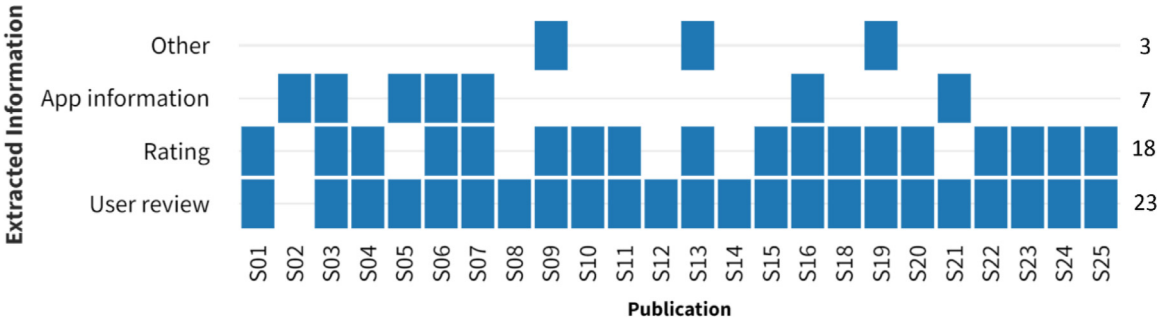

Fig. 7. Number of publications by information extracted.

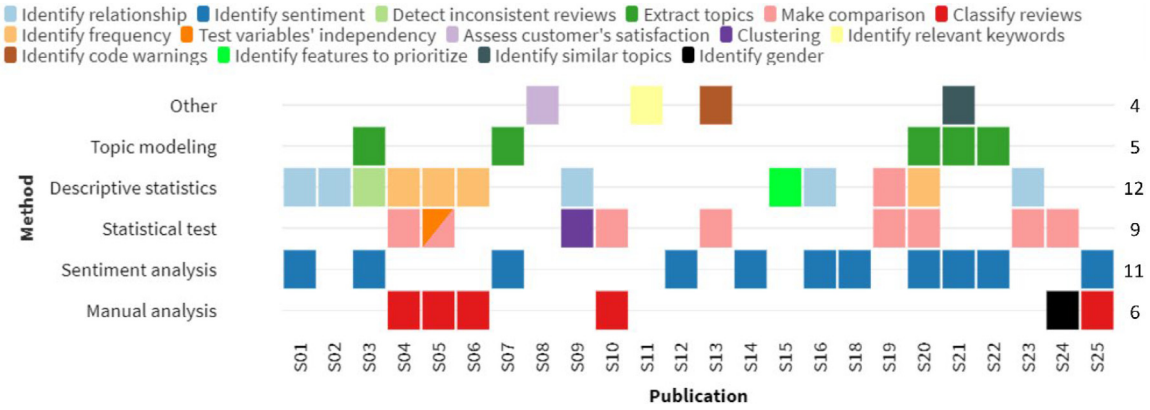

Fig. 8. Number of publications by analysis method.

[S05], Iacob et al. [S06], Durelli et al. [S20]) *identified the frequency* of different variables to identify their distribution, such as the number of reviews by star ratings and factor. Two works employed regression analysis, one of them to *detect inconsistent reviews* (Fu et al. [S03]) and the other to *identify features to prioritize* (Keertipati et al. [S15]).

Sentiment analysis was the second most employed method to analyze the data and *identify the sentiment* of the reviews. The most used technique to analyze the sentiment of user reviews was SentiStrength, employed in 3 studies (Guzman and Maalej [S07], Shah et al. [S12], Martens and Johann [S16]). Each of the remaining works employed different sentiment analysis techniques and tools, such as Stanford CoreNLP (Nicolai et al. [S25]), self-developed sentiment analysis (Durelli et al. [S20], Li et al. [S21], Fu et al. [S03]), Appbot (Bano et al. [S18]), RapidMiner (Mohan et al. [S14]), Clarabridge's tool suite (Goul et al. [S01]), SACI tool (Luiz et al. [S22]).

We identified nine publications that performed some type of **statistical test**. Most of them employed Mann–Whitney, Wilcoxon Rank-sum or Chi-square to *make comparisons* between two groups, such as positive and negative factors (Ha and Wagner [S04]), review length and price (Pagano and Maalej [S05]), high vs low rated apps (Khalid et al. [S10], Khalid et al. [S13]), apps with great vs small number of implemented suggestions from reviews (Palomba et al. [S19]), apps with vs without test cases (Durelli et al. [S20]), and gender (Guzman and Paredes-Rojas [S24]). One work tested multiple groups using Kruskal–Wallis, Chi-square, and Tukey–Kramer (Guzman et al. [S23]). The authors aimed to investigate whether there is a difference between distinct countries and the following variables: sentiments, rating, review content, review length, post-time delay since release, gender, and factors (bug report, feature request, and 'other'). There was also one publication (Pagano and Maalej [S05]) that employed Chi-square to *test variables' independency* (e.g., a given factor vs

ratings) and other that applied Scott–Knott test for *clustering* purposes (Khalid et al. [S09]), i.e., to identify groups of devices that are more prone to obtain lower ratings.

The **manual analysis** placed fourth. Five out of six publications (Ha and Wagner [S04], Pagano and Maalej [S05], Iacob et al. [S06], Khalid et al. [S10], Nicolai et al. [S25]) performed a manual coding process to tag the sentences and identify categories *to classify the reviews*. One publication performed a manual analysis to *identify users' gender* (S24).

Five publications employed **topic analysis** to *extract topics (features)* through unsupervised learning. Three works employed the Latent Dirichlet Allocation (LDA) unsupervised model (Fu et al. [S03], Guzman and Maalej [S07], Li et al. [S21]), and two works (Durelli et al. [S20], Luiz et al. [S22]) employed Non-negative Matrix Factorization (NMF) with Semantic Topic Combination (SToC).

Finally, four publications performed **other types of analysis** for various purposes. In one publication (Kang and Park [S08]), the authors proposed an approach called VIKOR that employs sentiment analysis to *assess customer satisfaction*. In another (Vu et al. [S11]), the authors applied Vector Space Model (VSM) and K-means to cluster and *identify relevant keywords*. One publication (Khalid et al. [S13]) *identified code warnings* from the app source code using the FindBugs tool to identify correlations between warning categories and low ratings. Finally, one publication (Li et al. [S21]) calculated the Jaccard similarity index to *identify similar topics*.

4.6. Data categorization (SQ4)

Half of the publications performed some type of categorization in the data they analyzed rather than presenting specific features of the application. However, we did not find any standardization. Only two studies categorized the reviews according to categories defined in previous work. Martens and Johann [S16] selected a sample of four categories from Pagano and Maalej [S05]: Bug Report, Feature Request, User Experience, and Rating. In turn, Nicolai et al. [S25] employed all the six categories defined by Panichella et al. (2015): Feature Request, Opinion Asking, Problem Discovery, Solution Proposal, Information Seeking, and Information Giving. They also included four new categories: Complaints, Compliments, Problem Reporting, and Noise.

Most of the categories were related to features/functionalities, problems/bugs, and users' positive/negative perceptions of the app. Eight publications defined categories related to features and specific functionalities, such as "feature request", "feature removal", "update", and "search" (Fu et al. [S03], Ha and Wagner [S04], Pagano and Maalej [S05], Kang and Park [S08], Khalid et al. [S10], Vu et al. [S11], Bano et al. [S18], Nicolai et al. [S25]). Other eight publications defined categories related to problems and bugs, such as "bug report", "problem reporting", and "functional error" (Ha and Wagner [S04], Pagano and Maalej [S05], Khalid et al. [S10], Vu et al. [S11], Khalid et al. [S13], Martens and Johann [S16], Guzman et al. [S23], Nicolai et al. [S25]). Seven publications defined categories related to UX aspects, such as user experience (Pagano and Maalej [S05], Martens and Johann [S16]), usability (Iacob et al. [S06]), performance (Khalid et al. [S13]), attractiveness (Fu et al. [S03]), adjective (Ha and Wagner [S04]), and complaints/compliments (Nicolai et al. [S25]). However, none of these publications analyzed the reviews through the lens of UX theory by considering pragmatic aspects related to the user's efficient and effective task achievement and hedonic aspects related to the user's emotions and sentiments. Although Pagano and Maalej [S05] and Martens and Johann [S16] defined a category named user experience, it is not related to the users' feelings and emotions as defined by UX theory, but to descriptions of the app in action, i.e., use cases where the application has proven to be useful.

4.7. Scope of the analysis (SQ5)

This subsection presents the results regarding analysis focus (individual, category, and general) and the dataset size (number of apps, reviews, and categories).

Half of the publications derived general conclusions by analyzing apps from different categories (see Fig. 9). For instance, Guzman and Paredes-Rojas [S24] analyzed user reviews from 7 countries that speak English. They obtained the reviews from 7 apps of different categories in Apple AppStore and investigated whether gender influences user's rating, time to post a comment after an app release, and review's length, sentiment, and content. In another work, Pagano and Maalej [S05] gathered 1,126,453 reviews from 25 free and paid apps from 22 categories of the Apple AppStore. They investigated which of the 14 categories they defined are associated with more positive or negative reviews.

Eight publications analyzed the reviews by groups. Some authors performed comparisons between different categories of apps, such as healthcare vs. non-healthcare apps (Nicolai et al. [S25]), with test cases vs. without test cases (Durelli et al. [S20]). Other authors focused on analyzing specific app categories such as educational applications (Bano et al. [S18]), mobile banking (Mohan et al. [S14]), social networking (Kang and Park [S08]), and productivity (Goul et al. [S01]). There was only one study that investigated a variety of categories of apps. Fu et al. [S03] analyzed 171,000 apps of 30 categories from app stores. They performed topic modeling and obtained the top-10 causes (topics) for negative reviews. For each app category, the authors identified the three topics that users complained the most.

Finally, six publications performed the analysis at the individual app level. For instance, Li et al. [S21] analyzed 1,148,032 reviews of WhatsApp from the Google Play Store. They aimed to investigate the correlation between users' positive and negative reviews before and after a sequence of apps' releases. To do so, they performed topic analysis to identify similar topics between different reviews and performed sentiment analysis over time.

The dataset size varied considerably across the studies (see Table 7). Regarding the app sample, the analysis ranged from a single app (S15, S21) to more than 170,000 apps (S03). This discrepancy is even greater when considering the number of reviews. Some publications analyzed less than a thousand reviews (S02, S04), while others analyzed millions (S03, S05, S11, S13, S16, S21).

The work of Fu et al. (S03) had the largest dataset, with more than 13 million reviews from 171 apps. They performed topic modeling to group related words and sentiment analysis to investigate their impact on ratings and identify inconsistent reviews. Through topic modeling, the authors identified the top 3 reasons for negative reviews in each of the 30 categories from Google Play Store. The large dataset comprising a variety of apps from different categories strengthens the study's findings.

Other works, in turn, had small sample sizes. For instance, Ha and Wagner (S04) analyzed 556 reviews from 59 apps. They performed manual content analysis to categorize the reviews and performed statistical tests to investigate their impact on ratings. Although the results of some tests were statistically significant, the small sample size reduces the statistical power (Wohlin et al., 2012), mainly because their analysis scope was general, that is, across all categories and apps.

4.8. Identified factor and its associated polarity (SQ6)

We identified an initial set of 118 non-unique factors from the 25 publications. First, one researcher analyzed each factor and grouped them according to their name, description, and/or keywords provided in the publication. Some publications classified the same factor into positive and negative ones, such

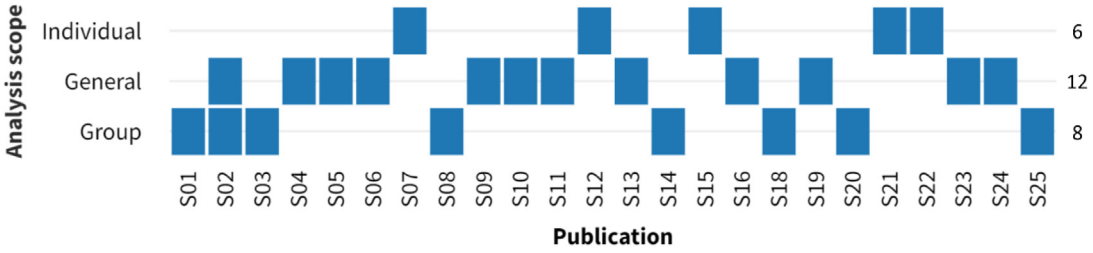

Fig. 9. Number of publications by the scope of the analysis.

Table 7
Dataset analysis of the publications returned in this systematic mapping study.

Paper ID	Source	Apps	Categories	Reviews	Analysis scope
S01	Apple AppStore	9	1	5036	Group
S02	BlackBerry App World	32,108	19	–	General Group
S03	Google Play Store	171	30	13,000,000	Group
S04	Google Play Store	59	30	556	General
S05	Apple AppStore	1100	22	1,126,453	General
S06	Google Play	161	6	3279	General
S07	Google Play Store Apple AppStore	7	Unspecified	32,210	Individual
S08	AppStoreHQ	8	1	1487	Group
S09	Google Play	99	4	206,751	General
S10	Apple AppStore	20	15	6390	General
S11	Google Play	95	Unspecified	2,106,605	General
S12	Apple AppStore	25	Unspecified	100,000	Individual
S13	Google Play Store	5	All Google Play categories	2,500,000	General
S14	Google Play Store	51	1	303,694	Group
S15	Google Play	1	1	4442	Individual
S16	Apple AppStore	245	23	7,396,551	General
S18	Apple AppStore Google Play Store	10	2	25,035	Group
S19	Google Play	100	18	5792	General
S20	Google Play Store	60	Unspecified	21,000	Group
S21	Google Play Store	1	1	1,148,032	Individual
S22	Google Play Store	7	Unspecified	22,815	Individual
S23	Apple AppStore	7	4	59,204	General
S24	Apple AppStore	7	Unspecified	919	General
S25	Google Play Store	8431	Health apps and non-health apps	383,758	Group

as Aesthetics-Negative and Aesthetics-Positive (Ha and Wagner [S04]), and were grouped into a single factor without polarity. In this first iteration, we ended up with 55 unique factors. One of the authors created a mind map to represent all the factors and presented it to another researcher, an expert in Software Engineering and HCI, to review it. Both discussed the merging process and refined the set of factors. Some factors were complementary to the other, such as "Recommendation" and "Dissuasion". The former is related to reviews that recommend purchasing or installing the application, while the latter advises against purchase. In this case, we grouped them into the "Recommendation" factor and merged their definition, as both situations (suggestion for acquisition and advise against purchase) are related to a recommendation. We also defined a set of keywords for each factor to characterize them and facilitate overlap identification. We merged factors with overlapping keywords into a unique and broader one. Finally, factors with generic definitions, such as "adjective", "praise", and factors that are not informative, such as "work", defined as reviews that report that the application works without technical description, were removed (factors with gray background). At the end of the process, we identified 31 unique factors. A list with all the original factors and their consolidation can be found in Appendix E.

After the consolidation process, we grouped them into categories according to their concept. First, we defined three high-level conceptual categories to group the factors according to the definition of UX presented in Section 3.5: App Factors, User Factors, and Context Factors. **App Factors** are related to the app, such as its characteristics, functionalities, features, and development. **User factors** are those related to users, such as their profile, needs, and the reasons for their positive or negative evaluations. Finally, **Context Factors** comprise factors related to the environment where the interaction occurred. Next, we refined the set of factors by analyzing the description of each factor and grouping them according to their concept. Fig. 10 presents the mapping and merging process of all identified factors with the respective categories.

In the following subsections, we present the concept behind each factor and the results of the factors' polarity analysis. We also divided the results into two subsections: (i) publications per factor polarity: to identify the polarity of the factors identified by each publication; and (ii) factors' polarity: to investigate the polarities that each factor can be associated with, regardless of the publication.

4.8.1. Factors' concept definition

We defined each factor's concept and scope to support researchers and practitioners in understanding it better. Some publications did not conceptualize the factor, providing just a set of related keywords. For instance, Fu et al. [S03] performed topic modeling by employing LDA to identify users' top complaints by analyzing negative reviews, i.e., with 1 and 2 stars. For each set of keywords from each extracted topic, they derived a name that better represents the concept behind it. For instance, the topic that comprises the words "boring, bad, stupid, waste, don't, hard, make, way, graphic, controls" they named Attractiveness. The terms imply that the factor is related to users' perceptions and judgments about the app in a given usage situation (indicated by

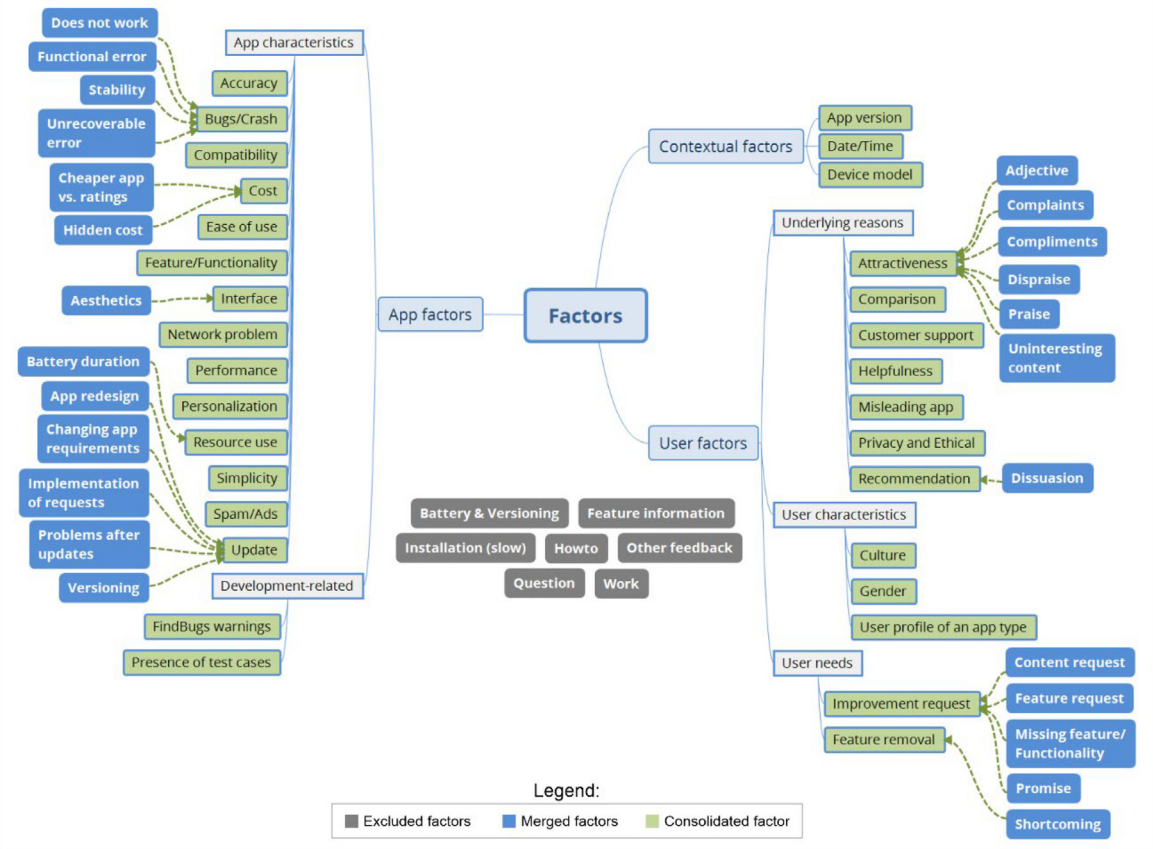

Fig. 10. Factors mapping and merging process.

features such as graphics and controls). In this sense, we looked for definitions that would fit this factor. The following definition from Hassenzahl (2018) addresses the idea of the factor: "*The user reports experiences with and feelings towards a product in a particular situation into an evaluative judgment*". Thus, we set it as the definition of the "Attractiveness" factor. There were also some factors in which the definition was vague. Khalid et al. [S10] provided the following description for the "Privacy and Ethical" factor: "The app invades privacy or is unethical". However, the concept of privacy invasion is lacking. In this sense, we searched for more complete definitions. Ebrahimi et al. (2020) defines privacy invasion in the context of mobile apps as "constant location tracking, unsolicited data collection, or any form of features that are engineered to lure users into sacrificing their privacy in exchange for more personalized services". It defines both privacy issues and app developers' unethical behaviors, reflecting the factor's concept.

To better visualize and organize the data, one of the authors created a mind map with all the available definitions for each selected factor and their respective sources. Through this mind map, we analyzed the definitions to choose the adequate one. In some cases, a publication provided a complete definition that conveys the concept of the factor. Thus, we selected it as a default. In other cases, we had to merge the description provided by multiple publications to generate a broad definition to capture all the aspects related to the factor. For instance, the 'Update' factor had seven other descriptions, as presented in Fig. 11. We abstracted non-overlapping definitions from each publication and merged them into a unique one. All the process was peer-reviewed by another researcher, an expert in Software Engineering and HCI. Table 8 presents the final definition for each factor extracted in this systematic mapping study. We did not include a definition for "App Version" and "Date/Time" factors, given that they are obtained from the reviews' metadata and not through the analysis of the reviews' content.

4.8.2. Publication by factor polarity

We identified 17 publications that reported factors associated with negative reviews (see Fig. 12). Some works, such as Fu et al. [S03], focused on investigating only negative reviews. They investigated more than 13 million user reviews of 171 thousand Android apps and extracted 10 factors associated with negative reviews: attractiveness, stability, accuracy, compatibility, connectivity, cost, telephony, picture, media, and spam. They grouped the applications into two main categories and their respective subcategories: game (arcade, puzzle, sports, etc.) and general application (communication, education, social, etc.). For each subcategory, they identified the top three factors associated with negative reviews. For games, all the subcategories had attractiveness, stability, and cost as the main factors associated with negative reviews. By contrast, the main factors for each subcategory of general applications varied, not being possible to identify a pattern.

There were 13 publications that reported factors associated with positive reviews. For instance, a study conducted by Nicolai et al. [S25] revealed that healthcare application users tend to be more positive when describing and reporting failures than users of other types of apps. The "problem discovery" and "problem reporting" categories, for instance, had 54% and 62% of positive reviews, respectively, from healthcare apps users. By contrast, these categories did not have any positive reviews from non-healthcare apps users. According to the authors, this is likely due to their willingness to (i) be proactive concerning apps that help their life and social activities and (ii) drive developers towards

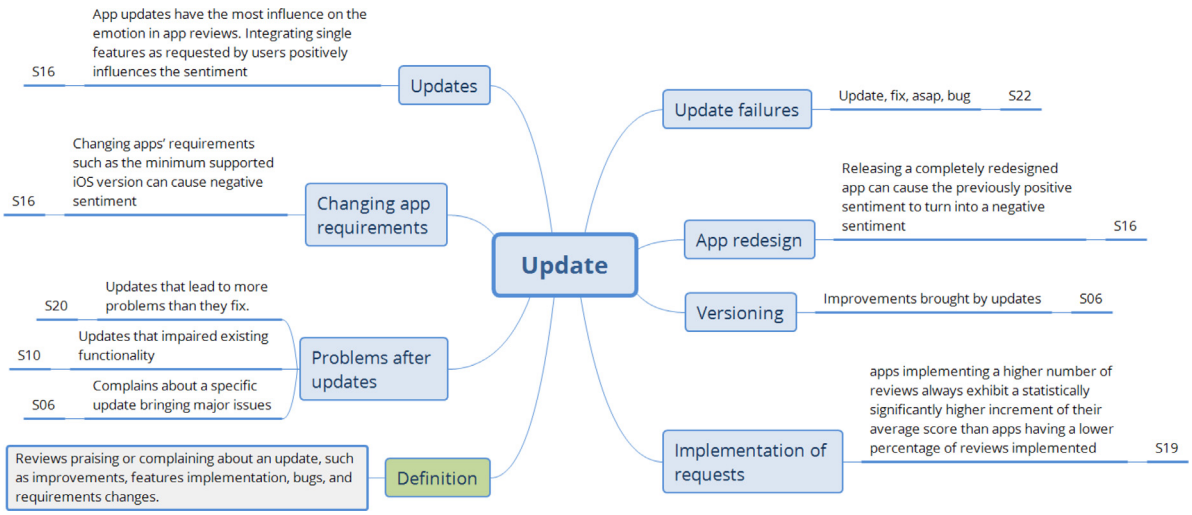

Fig. 11. Definitions for the 'Update' factor.

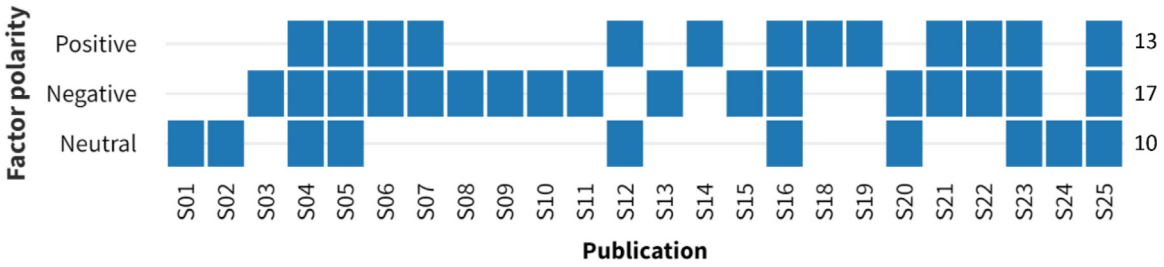

Fig. 12. Number of publications by the associated polarity of the factors they addressed.

the resolution of problems rather than blame them for missing functionalities.

Finally, 10 publications presented factors that did not affect either positively nor negatively. For instance, Guzman and Paredes-Rojas [S24] indicated that gender does not have any influence on the variables they analyzed: user's rating, time to post a comment after an app release, and review's length, sentiment, and content. In another study, Goul et al. [S20] did not find any significant difference on ratings of apps with and without test cases.

4.8.3. Factors' polarity

We analyzed each of the 31 factors by mapping their associated polarity. Figs 13, 14, and 15 present the factors according to their associated polarities. Each factor can be associated with one or more publications that analyzed it according to different scopes and produced results applicable to the individual, group, or general contexts.

Most of these factors were exclusively associated with negative reviews, such as Compatibility (Fu et al. [S03], Khalid et al. [S10], Luiz et al. [S22]), Privacy and ethical (Khalid et al. [S10]), and Spam/ads (Fu et al. [S03], Luiz et al. [S22]). Others were only associated with positive reviews, such as User profile of an app type (Nicolai et al. [S25]), Helpfulness (Pagano and Maalej [S05], Martens and Johann [S16]), Ease of Use (Guzman and Maalej [S07]), and Simplicity (Mohan et al. [S14]). A smaller portion was associated with neither positive nor negative reviews, such as the Presence of test cases (Durelli et al. [S20]), Gender (Guzman and Paredes-Rojas [S24]), and App version (Goul et al. [S01]).

There were also factors associated with more than one polarity (factors with blue background): *Attractiveness*, *Cost*, *Culture*, *Feature/Functionality*, *Improvement Request*, *Interface*, *Recommendation*, *Resource Use*, and *Update*. We analyzed these factors to understand the reasons for the contradictory results better.

The main reason is the **difference in data analysis**. Regarding *Cost*, all publications that investigated the correlation between price and user reviews found that the correlation was not significant, i.e., a cheaper or more expensive app will not necessarily lead to a more positive or negative evaluation (Harman et al. [S02], Iacob et al. [S06], Martens and Johann [S16]). In turn, some publications focused only on negative reviews (Fu et al. [S03], Khalid et al. [S10]) or divided them into positive and negative reviews (Ha and Wagner [S04]), leading to contradictory results due to the differences in the analysis. *Interface* was associated with all three polarities and addressed by three publications with different analyses. One of them analyzed the reviews at the app level (Luiz et al. [S22]), one focused only on negative reviews (Khalid et al. [S10]), and one divided it into positive and negative interface evaluations (Ha and Wagner [S04]).

Factor's neutral nature: all these factors were already associated with either negative or positive polarity according to the publications they were extracted. Some publications divided the same factor into positive and negative polarities to classify and filter the reviews. Pagano and Maalej [S05], for instance, presented two generic factors (Praise and Dispraise), both related to users' appreciation towards the app (*Attractiveness*) but with opposing polarities. They also separated reviews recommending the acquisition of the app from those dissuading users not to acquire it, both related to *Recommendation* factor. Ha and Wagner [S04] also classified the reviews related to interface overall look into Aesthetics-positive and Aesthetics-negative. In addition to these situations, the variation was more frequent when analyzing the factor at the app level. *Feature/Functionality*, for instance, was the factor associated with all three polarities. This variation is because five out of ten publications analyzed the reviews at the

Table 8
Definition of the factors extracted in the systematic mapping study.

Factor	Definition	Source
Accuracy	The degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use.	Fu et al. [S03], ISO/IEC 25012 (2008)
App version	The version of the app the user was using when writing the review.	Goul et al. [S01]
Attractiveness	The user reports experiences with and feelings towards a product in a particular situation into an evaluative judgment.	Fu et al. [S03], Hassenzahl (2018)
Bugs/Crash	The reviewer writes that the application does not work and provides a technical description, such as it takes too long to load or keeps crashing.	Ha and Wagner [S04]
Comparison	Reference to other apps, e.g., for comparison	Pagano and Maalej [S05]
Compatibility	The app has problems on a specific device or an OS version	Khalid et al. [S10]
Cost	The reviewer praises the application for being free or, if it is paid, the application/service is worth/not worth the money.	Ha and Wagner [S04], Khalid et al. [S10]
Culture	Differences between the culture of each country considering six dimensions from the Hofstede model: Power Distance; Individualism vs. Collectivism; Masculinity vs. Femininity; Uncertainty Avoidance; Long-term vs. Short-term Time Orientation; Indulgence vs. Restraint.	Guzman et al. [S23]
Customer support	Users being satisfied or dissatisfied with the support they received while using apps	[S06]
Date/Time	The date/time when the user wrote the review.	Goul et al. [S01]
Device model	The device the user reported in the review or was obtained through the metadata provided by the app store.	Khalid et al. [S09]
Ease of use	Reviews reporting users' perceptions of the effort related to the usage and the experience of the interaction with the application.	[S07], Weichbroth (2020)
Feature/ Functionality	Reviews that praise or criticize an existing feature (e.g., songs, themes, video quality) or functionality (e.g., upload file, take photo).	[S04]
Feature removal	Complaint about a disliked feature that is degrading the experience	[S10]
FindBugs warnings	The warnings category from the FindBugs tool: Bad practice, Internationalization, and Performance.	[S13]
Gender	The gender of the user based on the verification of the first name in the generize.io database. Only probabilities higher than 95% for "male" or "female" are considered. Lower probabilities are assigned the "unisex" label. Names not occurring in the database are given the "unclear" label.	[S24]
Helpfulness	Comprises topics related to descriptions of the app in action. These are helpfulness, which captures use cases where the application has proven helpful, and feature information, including descriptions of application features and user interface.	[S05], [S16]
Improvement request	Requests improvement (e.g., the app is slow) or the addition of new features or content	[S05]
Interface	Reviews that describe the application's overall look or interface, including images, color scheme, icons, and text	[S04]
Misleading app	The user reports that the description of the app and all the visuals associated with it does not accurately convey the functions and features of the app	Iacob et al. [S06], Panosian (2017)
Network problem	The app had trouble with the network or responded slowly.	[S10]
Performance	The app is slow to respond to input, or laggy overall.	[S10]
Personalization	The extent to which the Web site can be customized to the needs of individual customers. Customization also provides flexibility and control regarding the content and organization of the information they want (Huang, 2002) and facilitates interactivity (Selz and Schubert, 1998)	Tarafdar and Zhang (2005)
Presence of test cases	Apps with automated tests in which the adopted test-to-code-ratio was 1 line of test code to 10 lines of production code (i.e., 1:10): a ratio of 1:10 indicates that, for every line of test code, there are 10 lines of production code.	[S20]
Privacy and Ethical	Reviews reporting app developers' unethical actions (e.g., unethical business practices or selling users' personal data) or that the app requests information that may invade users' privacy, such as personal information, constant location-tracking, unsolicited usage data collection, or any form of features that are engineered to lure users into sacrificing their privacy in exchange for more personalized services.	[S10], Ebrahimi et al. (2020)
Recommendation	The user suggests acquisition or advises against purchasing or downloading the app	[S05]
Resource Use	The app consumes or does not consume too much battery/memory.	[S10]
Simplicity	The degree of being easy to understand or being uncomplicated in form or design, described by such characteristics as the number of menu levels, the number of performed gestures to reach a destination object, and the duration of searching a button to perform a specific function.	[S14], Weichbroth (2020)
Spam/Ads	Review complains about the number and content of ads in the application or says there were not too many ads in the application or would not mind having a free application that contained ads.	[S03], [S04]
Update	Reviews that praise or complain about an update, such as improvements, features implementation, bugs, and requirements changes.	[S06], [S10], [S16], [S20]
User profile of an app type	Users' profiles can affect ratings. Healthcare app users, for instance, tend to be less critical towards errors appearing in such apps, leading to more positive ratings.	Nicolai et al. [S25]

app level, extracting specific features from each app. Each feature/functionality can be evaluated either positively or negatively by users according to the app, thus leading to many variations in the results.

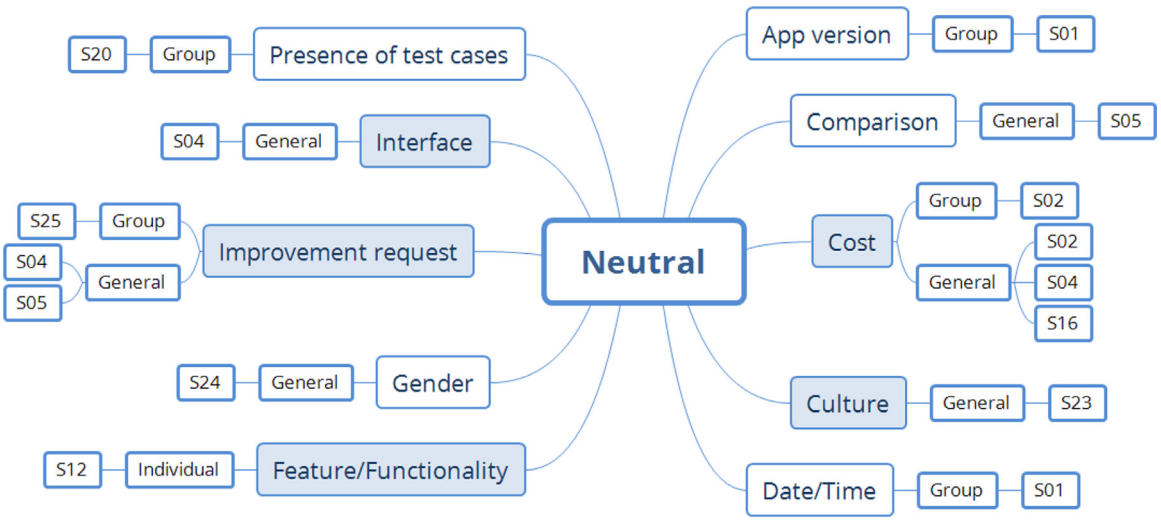

Fig. 13. Factors associated to neutral evaluations.

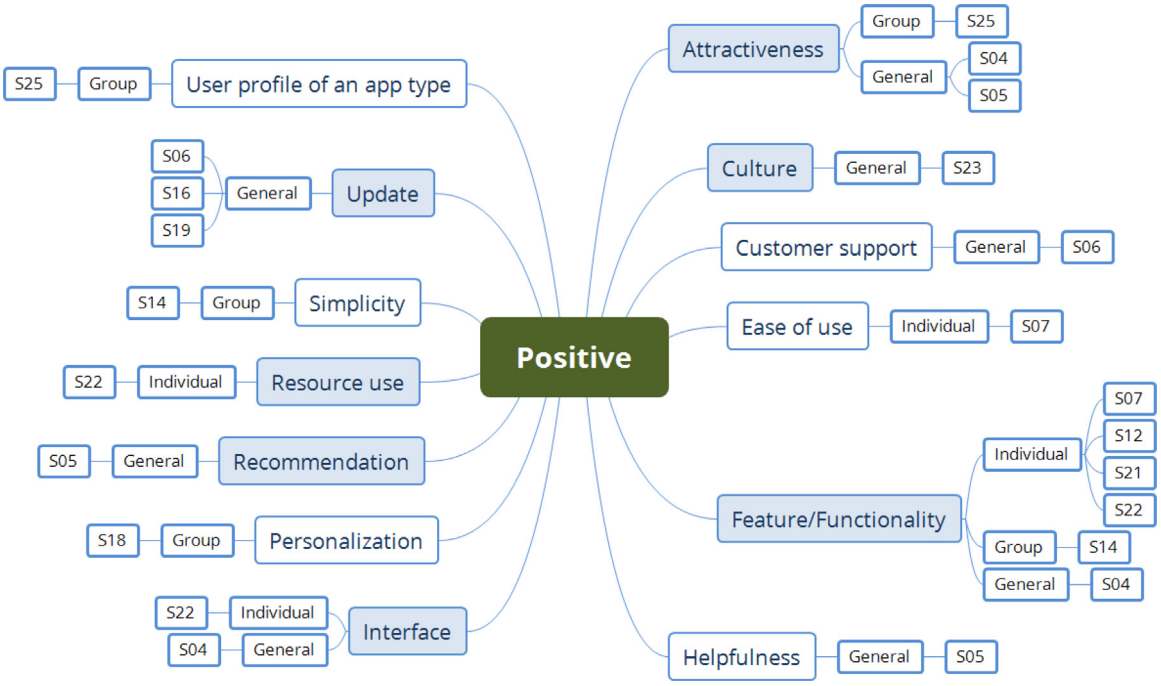

Fig. 14. Factors associated with positive evaluations.

Regarding *Resource Use*, two publications identified that battery drain is associated with negative evaluations (Khalid et al. [S10], Durelli et al. [S20]), while one publication associated the small impact on battery duration to positive evaluations in reviews of some specific apps (Luiz et al. [S22]). Regarding *Culture* [S23], the ratings and sentiments varied in each country analyzed. These variations occurred due to different cultural values from two out of six dimensions of the Hofstede's model (Geert and Jan, 1991): Power Distance (degree to which members of the country accept and expect that power is distributed equally) and Indulgence (the extent to which a society expresses their wants and impulses). Indulgence correlated positively with user ratings, while Power Distance correlated negatively. According to the authors, higher indulgent countries tended to provide more positive ratings, while countries with lower Power Distance tended to provide more negative ratings.

Regarding *Improvement Request*, four out of six publications (Pagano and Maalej [S05], Khalid et al. [S10], Li et al. [S21], Luiz et al. [S22]) associated it with negative reviews, indicating that users usually penalize the app due to the lack of features or functionalities. The other two publications (Ha and Wagner [S04] and Nicolai et al. [S25]) associated it with neutral evaluations. Regarding the work of Ha and Wagner [S04], the neutral evaluation may be due to the small sample size, as they obtained 556 reviews from 59 apps, which results in less than 10 reviews analyzed per app. Regarding the work of Nicolai et al. [S25], the neutral evaluation can be explained by the target population of the apps analyzed (healthcare users), as they tended to be more positive in general. Regarding *Updates*, most of the publications related it to negative evaluations due to problems brought by app updates, such as changes in app requirements (e.g. required a different OS version to be installed) and app redesigns that changes users' workflow (Iacob et al. [S06], Khalid et al. [S10],

Fig. 15. Factors associated to negative evaluations.

Martens and Johann [S16], Durelli et al. [S20]). However, some works also pointed out a positive perception of users on updates that added new features/functionalities or fixed bugs (Iacob et al. [S06], Martens and Johann [S16], Palomba et al. [S19]).

Dynamic nature of users' experiences, expectations, and needs: another possibility to explain the contradictory results is related to changes in users' experiences, expectations, and needs over time. According to Law and van Schaik (2010), user expectation and affect evolve dynamically in the long run. Aspects perceived as a product's differential in the past may turn into an aspect considered mandatory nowadays. In the work of Ha and Wagner [S04], for instance, the authors found that a good-looking and usable interface will not imply better ratings, but a bad interface or not usable one results in a decrease in users' ratings. It indicates that users are more demanding, requiring the software product to be good-looking and usable, increasing its weight in users' evaluations. Thus, it is essential to develop

approaches that capture the variations of the weights of these factors over time.

4.9. Factor influence analysis (SQ7)

Sixteen out of 24 publications analyzed the influence of the factors on users' sentiments or ratings. Some studies employed statistical tests to investigate whether the differences between different groups are significant, such as apps with and without test cases (Durelli et al. [S20]), device model (Khalid et al. [S09]), and the influence of gender (Guzman and Paredes-Rojas [S24]) and culture (Guzman et al. [S23]). Other studies performed some frequency/distribution analysis, for instance, to investigate the impact of each factor by analyzing their frequency according to the number of star ratings (Pagano and Maalej [S05]) and the ratio between one and two-star ratings (Khalid et al. [S10]). Finally, some publications performed correlation analysis to investigate the relationship between different variables, such as date/time and app version with ratings (Goul et al. [S01]), price and ratings (Harman et al. [S02]), and between emotions, ratings, and price (Martens and Johann [S16]). To better understand the reasons behind positive, negative, and neutral evaluations, we compiled the findings from each factor across publications (see Table 9).

4.10. Results from the secondary study

One of the accepted publications (S17) returned by the search engines was a systematic mapping study conducted by Genc-Nayebi and Abran [S17]. In this work, the authors addressed publications that proposed solutions for mining app store user reviews, reported challenges and unsolved problems in the domain, and contributions to software requirements and evolution. To do so, the authors defined five research questions to get information about: (i) the data mining techniques employed; (ii) remedies for the domain dependency challenge; (iii) review usefulness criteria; (iv) spam identification; and (v) extracted features. According to the authors, the search string returned more than 500 publications, but they did not provide a precise number. Among them, 63 publications remained as start set for the backward and forward snowballing process, which resulted in 45 research papers. However, the authors did not provide details of how many publications they excluded. At the end of the process, the authors selected 24 primary studies after applying the inclusion and exclusion criteria.

The authors listed five main findings: (i) most of the studies were exploratory, based on manual classification and correlation analysis; (ii) the approaches to extract app features do not consider the nature of app store reviews, such as short length, unstructured phrases, colloquial language, and abundant information; (iii) users and developers request a different type of information, i.e., the former is interested in the experience of other users, while the latter seeks to improve the app quality by addressing missing requirements, features, and user experience information; (iv) external sources of information, such as tweets, blogs, and code repositories could be used to enrich the data; (v) identifying useful reviews, as well as spam and fake reviews are one of the biggest problems in the domain.

The main difference between their work with ours is that they focused more on addressing general questions related to app store reviews mining, such as the domain dependency challenge (i.e., the issue related to a classifier that is trained in a given domain and performs poorly when applied into another), reviews aggregation, and spam detection. Although they mention factors in some parts of the paper, they did not address their relationship with users' ratings and sentiments. In contrast, our research was conducted from the UX perspective by investigating what factors can affect the UX conveyed by mobile applications, what analyses were carried out, and the scope of the analysis. We also analyzed the factors we extracted and aggregated them into a comprehensive set. With this study, we take one step towards understanding the factors that can affect UX, advancing the research on app store reviews mining and HCI fields.

4.11. Results summary

Table 10 presents an overview of the results based on the primary studies' counting into each sub-question. Each paper can be assigned to more than one answer in each sub-question, except in SQ4. Thus, the number of papers assigned to each sub-question can be greater than the number of primary studies included in this systematic mapping (24). The complete list of the publications and the respective mapping to each SQ is available in Appendix E.

5. Discussion

In this subsection, we present the main findings from this systematic mapping study. We also highlight and discuss the implications for researchers and practitioners.

From a set of 795 unique publications (including those from the backward snowballing process), the systematic mapping study resulted in 25 selected publications. Two facts may explain the small number of publications. First, app stores are a recent form of distributing third-party software applications, and research on the analysis of app store reviews did not start until early 2010 (Genc-Nayebi and Abran, 2017). Second, we only considered publications that associated the factor's effect with ratings or sentiment analysis results, as we aimed to identify whether the factor affected UX positively, negatively, or neutrally. Despite the few publications, we identified research opportunities and challenges that could be used as the basis for future works on the topic. In this study, we identified several factors that could influence users' perception of the experience with mobile apps. Identifying these factors could have many implications for future works, such as reducing bias in UX evaluations, supporting the software development process by highlighting the influencing factors, and developing UX evaluation methods and guidelines. However, it is noteworthy that due to the small number of publications identified, the mapping can only be considered preliminary.

What are the influencing factors on users' evaluations in app store reviews, and how do they affect the evaluation? Regarding our research question, we identified 31 unique factors that can affect users' evaluations. While factors associated exclusively with negative reviews are more related to features and functionality issues (e.g., Performance, Compatibility, Accuracy, Feature removal, Network problem), positive factors are more related to general perceptions and human aspects (e.g., Helpfulness, Ease of Use, Customer Support, User profile of app type, Culture). It indicates that dissatisfied users tend to provide details about the functionalities and aspects they are not happy with. By contrast, when giving positive reviews, they tend to describe the app's overall qualities and aspects.

Some factors can have different effects on users' evaluations according to their polarity. For instance, while negative reviews on the app's cost and the interface can decrease ratings, positive ones seem not to impact the overall rating (S04). However, the small sample weakens the strength of the outcomes, requiring further studies to support these findings. Some factors also appeared more frequently in a given type of app. Attractiveness, Stability, and Cost were the three top factors in mobile games (S03). Other findings are related to the impact of different factors.

Table 9
Findings regarding each factor identified.

Consolidated factor	Polarity			Analysis Scope			Papers	Findings
	Pos	Neg	Neu	Ind	Gro	Gen		
Accuracy		X			X		S03	**Group:** complaints related to lack of accuracy on finding information and location. It is common in the following categories: Book & Reference, Lifestyle, Productivity, Transportation, Travel, and Weather.
App version			X		X		S01	**Group:** There was no correlation between app version and ratings for business intelligence apps.
Attractiveness	X	X			X	X	S03, S04, S05, S10, S25	**General:** Reviews including emotional expressions result in more positive or negative evaluations than reviews without emotional expressions (S04). Unappealing content usually leads to negative evaluations (S10). **Group:** *healthcare user apps* are usually less satisfied in general compared to users of non-health apps (S25). Content is a key success for *mobile games* (S03).
Bugs/Crash		X				X	S03, S04, S05, S06, S10, S11, S16, S22, S25	**General:** bugs/crash was the third most negative factor (S05), with most complaints related to functional errors, especially location and authentication issues (S10). Users reporting bug-related issues tend to write why the app does not work and evaluate it very negatively (S04). Around 18% of post-update reviews complained about frequent crashing (S10). Although severe bugs greatly impact users' experience, minor bugs seem not to impact so much on user ratings (S06). **Group:** users of *healthcare apps* expect developers to improve the app and help them take care of their health. Consequently, they are less critical of app errors than non-healthcare app users (S25). All categories of *games* suffered from stability problems that resulted in negative evaluations (S03).
Comparison			X			X	S05	The authors did not further detail the factor, for example, what aspects are considered in comparisons (S05).
Compatibility		X		X	X	X	S03, S10, S22	Compatibility issues led to negative evaluations in general, category, and individual analysis (S03. S10, S22). **Individual:** the use of *PicsArt app* on tablets causes na undesired increase in users' dissatisfaction.
Cost		X	X		X	X	S02, S03, S04, S06, S10, S16	**General:** diverging results identified. Publication S04 suggests that negative reviews about the cost can decrease ratings, but positive reviews do not impact the overall rating. However, its sample size is small. Publication S02, in turn, suggests that price does not influence the number of downloads or ratings. The greater number of apps (32,108) and the rigor of the analysis using statistical tests strengthen the findings of this publication compared to S04. This finding is supported by publication S16, in which the authors did not find any correlation between price and emotions, claiming that users willing to post a review provide feedback due to intrinsic motivation. **Group:** cost is one of the three main reasons users dislike a *mobile game* (S03). However, users of such apps seem to be more tolerant regarding prices, as paid apps have more complaints about their prices than paid games (S03). *Cheap apps* are rarely reported as worth the price (S06), and users also do not like apps that claim to be free but ask them to pay to get access to some features (S10).
Culture	X	X	X			X	S23	**General:** the rating of the reviews of a specific country positively correlates with the Indulgence of the country and negatively correlates with its Power Distance.
Customer support	X					X	S06	**General:** most users (61.58%) are positive regarding customer/developer support, resulting in positive evaluations.
Date/Time			X		X		S01	**Group:** Date/Time were not correlated with sentiment of reviews of productivity apps, but it is not further analyzed.
Device		X				X	S09	**General:** the results revealed the importance of analyzing reviews according to the devices used by users, as some of them tend to receive worse ratings than others due to specific problems.
Ease of use	X			X			S07	**Individual:** for the Pinterest app, its ease of use was the most positively evaluated aspect (S07).
Feature removal		X				X	S05, S10	**General:** features users do not like highly influence user ratings (S05). It was the third most complained factor, mostly leading to reviews with a 1-star rating (S10).

(*continued on next page*)

Table 9 (*continued*).

Consolidated factor	Polarity			Analysis Scope			Papers	Findings
	Pos	Neg	Neu	Ind	Gro	Gen		
Feature/ Functionality	X	X	X	X	X	X	S03, S04, S07, S08, S11, S12, S14, S15, S21, S22	Specific features or functionality can lead to either positive, negative, or neutral ratings according to the app or category. For instance, regarding *mobile banking apps*, features like money transfer, card payments, account summary, and ease of use were associated with positive evaluations (S14). Problems with media (watching, listening, and recording) negatively affected the *Entertainment, Media & Video, and Music & Audio categories* of apps (S03). Problems with Pictures (see, save, and upload photos) negatively affected *Comics, Media & Video, Personalization, and Photography categories* (S03).
FindBugs warnings		X				X	S13	**General:** code warnings related to Bad practice (i.e., violation of essential coding practices, for example, equals problems, dropped exceptions, and misuse of finalize), Internationalization (misuse of encoding characters), and performance (slow code) are correlated with low ratings.
Gender			X			X	S24	**General:** there was no significant difference in ratings between males and females. Regarding sentiment, there was a difference in gender, but not significant.
Helpfulness	X					X	S05	**General:** the second most popular factor and the second most positive, commonly associated with reviews praising the app. Its polarity increases when associated with recommendation.
Improvement request		X	X	X	X	X	S04, S05, S10, S21, S22, S25	**General:** In general, feature requests are associated with negative reviews (S05, S10, S21, S22), indicating that missing features affect users' evaluations. However, on average, reviews associated with this factor still lie above the middle (3 stars), indicating that it does not greatly impact ratings (S05). Content request was the least critical requirements feedback, with a small impact on user ratings (S05). Missing features have less impact when users already liked the app (S04) and praise it in the reviews (S05), resulting in more positive evaluations. **Group:** healthcare app users request more features than users of non-healthcare apps. In turn, they tend to not evaluate the app negatively due to missing features (S05). **Individual:** regarding WhatsApp, negative reviews were from users requesting features related to contact options and status (S21), and upgrade of themes (S22).
Interface	X	X	X	X		X	S04, S10, S22	**General:** in general, reviews complaining about the app's interface can decrease ratings (S04, S10). Conversely, reviews praising the interface do not yield more positive evaluations (S04). **Individual:** reviews praising the app's interface resulted in positive evaluations for DropBox and PicsArt apps (S22).
Misleading app		X				X	S06	**General:** misleading app descriptions are common (21.85%) and can result in negative evaluations (S06).
Network problems		X		X	X	X	S03, S10, S11, S22	In general, network issues, such as wi-fi and mobile network problems, connection errors, login issues, and slow connections, lead to negative evaluations.
Performance		X		X		X	S10, S22	In general, problems related to slow responses to user input or overall performance can lead to negative evaluations.
Personalization	X				X		S18	**Group:** reviews reporting personalization possibilities were positive in educational apps.
Presence of test cases			X		X		S20	**Group:** The median of apps with and without automated tests did not differ significantly, indicating that it does not affect users' perceptions of the experience with the app.
Privacy and Ethical		X				X	S10	**General:** it was the most critical factor. Users are bothered by privacy invasion and the app developer's unethical actions (for example, unethical business practices or selling the user's personal data).
Recommendation	X	X				X	S05	**General:** Users can either recommend to or dissuade other users. Reviews recommending the app are positive. In turn, reviews dissuading other users are very negative. The combination of bugs and dissuasion results in the lowest average rating (S05).
Resource use	X	X		X	X	X	S10, S22	In general, resource-heavy apps (e.g., consumes much battery, memory/storage) are negatively evaluated (S10, S20, S22). In turn, users also recognize when an app is energy-efficient, leading to very positive evaluations (S22).

(*continued on next page*)

Privacy and Ethical were the most critical factor in one study (S10), with the greatest ratio of 1-to-2-star ratings, as well as the Spam/Ads factor, which led to the lowest ratings for a mobile game (S22). There was also a factor in which its effects depended

Table 9 (*continued*).

Consolidated factor	Polarity			Analysis Scope			Papers	Findings
	Pos	Neg	Neu	Ind	Gro	Gen		
Simplicity	X				X		S14	**Group:** for mobile banking apps, simplicity, friendliness, and easiness were the main reasons for the positive sentiment (S14).
Spam/Ads		X		X	X		S03, S22	In general, ads are considered annoying and negatively influence ratings (S03, S22). It was the third most complained factor in the *"Personalization" apps* category (S03). In the *Angrybirds game,* this factor resulted in the lowest ratings (S22).
Update	X	X		X	X	X	S06, S10, S16, S19, S20, S22	**General:** updates implementing user requests result in positive reviews (S06, S16, S19). Conversely, releasing a completely redesigned app can cause the previously positive sentiment to turn negative (S16). Many complaints are related to functional errors (S10) and usability issues (S06) after recent updates. **Group:** problems related to updates are the primary source of negative reviews in apps with test suites (S20). **Individual:** the worst ratings from Evernote app were due to app update failures (S22).
User profile of an app type	X				X		S25	**Group:** Healthcare apps' customers tend to be more positive when describing and reporting failures than users of other apps. Similarly, they try to recommend possible solutions to those errors more politely.

Table 10
Overview of the results for each research sub-question.

Sub-question	Answer	Number of publications
SQ1. Dataset source	Google Play Store	15
	Apple AppStore	9
	Other	2
SQ2. Extracted information	Reviews	23
	Ratings	18
	App information	7
	Other	3
SQ3. Analysis methods	Descriptive statistics	12
	Sentiment analysis	11
	Statistical tests	9
	Manual analysis	6
	Topic modeling	5
	Other	4
SQ4. Data categorization	Yes	12
	No	12
SQ5. Scope of the analysis	General	12
	Group	8
	Individual	5
SQ6. Factor's associated polarity	Negative	17
	Positive	13
	Neutral	11
SQ7. Factor's influence analysis	Yes	16
	No	8

on different variables. Regarding Update factor, while implementing small improvements requested by users may increase ratings (S16, S19), a completely redesigned interface may lead to dissatisfaction (S16). Usability issues, update problems, and broken functionalities due to a new release are also other complaints that may result in negative evaluations. In this sense, developers should be careful when updating their apps. Developers should monitor the number of positive and negative reviews related to this factor over time, especially after releasing an update.

From practitioners' perspectives, all this information provides clues to impactful factors that developers should prioritize. From the academic perspective, researchers could investigate, for instance, whether there is a significant difference in evaluating different types of apps for each identified factor. Researchers could develop approaches that automatically analyze user reviews to identify these influencing factors and compare different types of apps.

Few comparative studies between different app stores (SQ1): the reviews were obtained only from Google Play Store and Apple AppStore. This result was expected, as Android and iOS are currently the two most widely used mobile operating systems. However, we identified only two publications that analyzed reviews from both app stores. This makes it difficult for researchers and practitioners to identify particularities of applications targeted to different operating systems and evaluate whether the findings from one app store (or from an app designed for a particular OS) apply to the other. By comparing different OS, software engineers could identify whether a given bug or interaction problem may be due to the platform or the app itself. It would also make it possible, for instance, to investigate whether needs, expectations, and factors that affect UX differ between groups, allowing developers to design products that are personalized to different target populations.

Ratings are the most common variable when analyzing user reviews (SQ2): most of the studies extracted ratings from user reviews to investigate the influence of different variables, such as gender (S23) and features (S15, S19). Ratings were also employed to identify the impact of factors according to the proportion of negative reviews, i.e., with 1 and 2 stars (S05, S10). By contrast, publications S03 and S22 performed sentiment analysis to identify inconsistent reviews. The results indicated that star ratings are sometimes inconsistent with the sentiment identified

in the reviews. Such results highlight the possibility of developing approaches to filter such inconsistencies and reduce noise to get more precise results when analyzing the impact of factors on UX.

Various methods were employed to analyze user reviews (SQ3): different analysis methods were employed according to the study's goals. Descriptive statistics were primarily used to identify relationships between variables and their impact on ratings and sentiment. The considerable number of studies applying sentiment analysis techniques indicate that additional information may be necessary to understand UX better, given that the star ratings only reflect users' overall perception of the experience with the app and not of specific aspects. Moreover, inconsistencies between ratings and sentiment identified by some studies (S03 and S22) indicate that using complementary approaches may allow getting more reliable results. However, we did not find a method specifically designed to analyze app store reviews. In contrast to reviews from other online stores, mobile app store reviews are generally short in length, given that they are written and submitted from mobile devices, on which typing is not easy (Fu et al., 2013). It makes the analysis harder, as less data is available for processing per review. Moreover, according to Martens and Johann (2017), some words considered negative in the software engineering domain may not necessarily be negative by their nature. Consider, for example, the following fictitious user review: "The app has many bugs and crashes all the time on my phone". This review would probably receive a neutral sentiment when analyzed by sentiment analysis methods designed for general purposes. For instance, by analyzing this review with SentiStrength,[3] the output was a neutral sentiment. However, words like bug and crash have a very negative connotation in the software engineering domain (Martens and Johann, 2017), highlighting the importance of employing and developing methods that are adjusted to the particularities of the domain. Statistical tests were applied specifically in studies comparing groups of apps. Considering that the rating distribution is skewed among apps (Hu et al., 2018), such publications employed non-parametric tests, such as Mann–Whitney, Chi-Square, and Wilcoxon Rank-sum. Regarding Manual Analysis, most of the publications reported using coding processes to analyze a sample of reviews and group them into categories. However, they did not develop automatic approaches to classify the other reviews, missing the opportunity for deeper analyses. Publications using automatic approaches adopted topic modeling to identify groups of related terms. A drawback of the approaches employed in these studies is that the outcomes are lists of terms, which increases the cognitive load to interpret them.

Little focus on UX and much on features/functionalities (SQ4): we identified many publications that categorized the data to analyze it better, most of them considering one or more aspects related to UX, such as performance and aesthetics. However, we did not find publications that analyzed the reviews in the light of UX theory. Most of the factors were related to functionalities and use of the app, such as performance, battery, and bugs. Some publications mention the term "user experience" along with the paper, and two publications even have a category named as such (Pagano and Maalej [S05], Martens and Johann [S16]). However, they associate this category with reviews that describe a concrete feature or user interface in a scenario that the application has proven helpful. Despite being named "user experience", it does not have a relation to any of the definitions and concepts of UX, but with "descriptions of the app in action" (Pagano and Maalej, 2013). Although many publications considered users' emotions in their analysis, this information was only used to identify their opinions' polarity and extract the main topics associated with each polarity, which may explain the lack of subjective factors related to users' emotions and feelings. Moreover, we could not identify a consolidated taxonomy to categorize the reviews, which resulted in various categories, many of them without a clear definition of their concept.

Low representativeness in some studies weakens the results (SQ5): initial studies involving manual analysis were conducted in small datasets. For instance, Ha and Wagner (S04) analyzed only 556 reviews of 59 different apps from 30 categories. In turn, Iacob et al. (S06) analyzed 3279 reviews of 161 apps from 6 categories. Both works focused on classifying the reviews into categories derived through manual analysis. Although Ha and Wagner (S04) performed statistical tests to strengthen their findings and found statistically significant results, the small sample reduces the study's statistical power. The low app/reviews ratio also weakens the sample's representativeness to draw conclusions on the factors considered in such studies. Researchers could employ machine learning techniques to automatically classify and analyze larger datasets to confirm such findings.

Few studies analyzing the particularities of different groups of apps (SQ5): most of the studies analyzed the data in a general context by gathering thousands of reviews of different categories of apps. However, such analyzes only provide an overview of the reviews from the app store, hindering the identification of the particularities of each type of app. Although we identified many studies that analyzed the data in groups, most of them did not compare these different groups of apps. For instance, Nicolai et al. (S25) identified that users of healthcare applications tend to be more tolerant to bugs and lack of functionalities compared to users of other types of applications. By making comparisons, it would be possible, for instance, to identify what factors users are more concerned about according to the group the app belongs to, which may help define which factors to prioritize during the development process.

Greater number of negative factors (SQ6): the results revealed more negative factors than positives and neutrals. Many publications focused on investigating negative reviews for app improvement purposes, given that positive reviews usually do not point out problems and improvement requests. The negative factors highlight which aspects developers should consider when developing new versions or new apps to improve UX and avoid negative evaluations. By analyzing these factors, we can suggest developers prioritize the correction of bugs (Bugs/Crash), also considering compatibility issues (Compatibility) and the devices associated with them (Device). The description of the app should be clear and reflect the actual functionalities (Misleading app), avoid using advertisements (Spam/Ads), and request non-essential user information that could result in privacy invasion (Privacy and Ethical). Developers should also pay attention to the app's performance (Performance) and connectivity issues (Network problems) to provide a better experience.

Impact of the factors on users' sentiments and ratings unclear (SQ7): half of the studies only presented the factors and their associated polarity, while the other half investigated the influence of each factor on users' sentiments and ratings by performing correlation analysis, statistical tests, or frequency/distribution analysis. Although they indicate whether the factor influence or not on users' ratings and sentiments, the weight of each factor remains unclear. The publications that performed statistical tests did not present the effect size, an important measure to assess the magnitude of the effect of each factor. In contrast, other publications performed only simple descriptive statistics by analyzing the frequency and distribution of the factors. By identifying the magnitude of the effect of these factors, developers could focus their efforts on aspects that most negatively impact users' perceptions about their product while maintaining or improving the aspects that are evaluated positively in future releases.

[3] http://sentistrength.wlv.ac.uk/.

5.1. Challenges and open issues

We identified the following challenges and open issues:

Impact of the factors unclear: it remains an open issue and is also one of the main challenges. In addition to the absence of the effect size in studies performing statistical tests, there are other challenges in estimating the impact of each factor. For instance, while many factors can be present in a single review, a review has only one rating, making it hard to identify which affected more positively or negatively the user's evaluation. This issue could be mitigated by analyzing the sentiment of the reviews at the aspect level (aspect-based sentiment analysis). Most of the publications identified in our systematic mapping that performed multilabel classification (e.g., S04, S05, S10) analyzed the factors by grouping reviews assigned to the same factor and calculating the average rating. Such an approach may bias the results, as multi-labeled reviews are considered in analyzing different factors. Future research directions could include developing approaches to analyze the reviews at the aspect level to identify the sentiment associated with each factor within a review and identify their impact on users' ratings, similar to the one proposed by Wang et al. (2010) for analyzing hotel reviews.

Longitudinal perspective of the factors: only three publications analyzed the reviews from a longitudinal perspective. Xiaozhou et al. (S21) analyzed the variation of users' sentiment after different WhatsApp releases and identified polarity changes after a given update. In turn, Martens and Johann (S16) analyzed the variation in the sentiment of users of Bank of America and Gmail apps. They identified some emotion drops in both apps and found issues related to changes in requirements and features. Finally, Vu et al. (S11) performed a trend analysis to identify potential problems by comparing the number of occurrences of a given keyword over time and the moving average. Analyzing the factors from a longitudinal perspective is crucial to identify trends and understanding UX better, given that it is dynamic and changes over time. Factors that were important in the past may not be so relevant in the present. Considering the dynamic nature of UX, it is crucial to investigate the effect of the factors over time.

5.2. Future research directions

We devise the following research directions based on the contributions of this systematic mapping study:

Creation of automated approaches to analyze user reviews: the factors identified in this systematic mapping study could serve as a starting point for developing automated approaches that classify and filter user reviews to support developers in quickly identifying UX-related issues.

Development of new UX evaluation methods: researchers could develop new UX evaluation methods using these factors as references for deriving aspects to be evaluated.

Conduction of studies from a longitudinal perspective: longitudinal studies could provide insights into the evolution of these factors over time. It would allow identifying which factors gained or lost importance over time and understanding UX better to prioritize development efforts.

6. Threats to validity

Although we conducted this research following a rigorous and widely employed systematic mapping methodology, we identified some threats to validity. The researcher's bias may affect the selection and analysis of the primary studies. To minimize this threat, we reviewed the protocol and conducted the Cohen's Kappa statistical tests to analyze the level of agreement between the researchers and reduce the bias. Additionally, other experienced researcher reviewed the execution process. The research string may not cover all publications on the topic. To ensure that relevant publications are returned, we selected the most important terms related to the topic during the construction of the search string. We also performed successive iterations to refine the search string while ensuring that the reference papers were returned. Given that we conducted this research focusing on studies that analyzed user reviews from app stores, the conclusions drawn may not apply to other contexts, such as desktop and web applications. The review protocol provided in this paper may be useful to conduct similar research in other contexts and extend it in the future.

7. Conclusion

This paper presented a systematic mapping study that addressed publications that analyzed user reviews from app stores. Our goal was to identify factors that can affect UX and influence users' evaluations. From the 25 publications selected, we identified 31 factors and their associated polarities, which can be used to develop UX evaluation methods and guide the development and improvement of software applications. We also identified research opportunities that can be exploited in future work, which includes the analysis of the impact of each factor on users' ratings and sentiments; conduction of research focusing on extracting UX aspects from user reviews; comparisons between different app stores and types of apps; and the development of methods that are designed specifically to analyze user reviews from app stores by considering their specificities.

CRediT authorship contribution statement

Walter T. Nakamura: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Edson Cesar de Oliveira:** Writing – review & editing. **Elaine H.T. de Oliveira:** Writing – review & editing. **David Redmiles:** Supervision. **Tayana Conte:** Validation, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research was supported by the Brazilian funding agency FAPEAM, Brazil through processes number 062.00150/2020, and 062.00638/2019, the São Paulo Research Foundation (FAPESP), Brazil through grant #2020/05191-2, the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES), Brazil financial code 001, and CNPq, Brazil processes 204081/2018-1/PDE, 314174/2020-6, and 308513/2020-7. It was also funded by Samsung Electronics of Amazonia Ltda according to Article 48 of Decree n° 6.008/2006, under the terms of Federal Law n° 8.387/1991, through agreement n° 003, signed with ICOMP/UFAM.

Table 11
Selected publications.

ID	Title
S01	Goul, M., Marjanovic, O., Baxley, S., & Vizecky, K. (2012, January). Managing the enterprise business intelligence app store: Sentiment analysis supported requirements engineering. In 2012 45th Hawaii International Conference on System Sciences (pp. 4168–4177). IEEE.
S02	Harman, M., Jia, Y., & Zhang, Y. (2012, June). App store mining and analysis: MSR for app stores. In 2012 9th IEEE working conference on mining software repositories (MSR) (pp. 108–111). IEEE.
S03	Fu, B., Lin, J., Li, L., Faloutsos, C., Hong, J., & Sadeh, N. (2013, August). Why people hate your app: Making sense of user feedback in a mobile app store. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1276–1284).
S04	Ha, E., & Wagner, D. (2013, January). Do android users write about electric sheep? examining consumer reviews in google play. In 2013 IEEE 10th Consumer Communications and Networking Conference (CCNC) (pp. 149–157). IEEE.
S05	Pagano, D., & Maalej, W. (2013, July). User feedback in the appstore: An empirical study. In 2013 21st IEEE international requirements engineering conference (RE) (pp. 125–134). IEEE.
S06	Iacob, C., Veerappa, V., & Harrison, R. (2013, September). What are you complaining about?: a study of online reviews of mobile applications. In 27th International BCS Human–Computer Interaction Conference (HCI 2013) 27 (pp. 1–6).
S07	Guzman, E., & Maalej, W. (2014, August). How do users like this feature? a fine grained sentiment analysis of app reviews. In 2014 IEEE 22nd international requirements engineering conference (RE) (pp. 153–162). IEEE.
S08	Kang, D., & Park, Y. (2014). Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. Expert Systems with Applications, 41(4), 1041–1050.
S09	Khalid, H., Nagappan, M., Shihab, E., & Hassan, A. E. (2014, November). Prioritizing the devices to test your app on: A case study of android game apps. In Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering (pp. 610–620).
S10	Khalid, H., Shihab, E., Nagappan, M., & Hassan, A. E. (2015). What do mobile app users complain about?. IEEE software, 32(3), 70-77.
S11	Vu, P. M., Nguyen, T. T., Pham, H. V., & Nguyen, T. T. (2015, November). Mining user opinions in mobile app reviews: A keyword-based approach (t). In 2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE) (pp. 749–759). IEEE.
S12	Shah, F. A., Sabanin, Y., & Pfahl, D. (2016, November). Feature-based evaluation of competing apps. In Proceedings of the International Workshop on App Market Analytics (pp. 15-21).
S13	Khalid, H., Nagappan, M., & Hassan, A. E. (2016). Examining the relationship between FindBugs warnings and app ratings. IEEE Software, 33(4), 34–39.
S14	Mohan, L., Mathur, N., & Reddy, Y. R. (2015, April). Improving Mobile Banking Usability Based on Sentiments. In International Conference on Evaluation of Novel Approaches to Software Engineering (pp. 180–194). Springer, Cham.
S15	Keertipati, S., Savarimuthu, B. T. R., & Licorish, S. A. (2016, June). Approaches for prioritizing feature improvements extracted from app reviews. In Proceedings of the 20th international conference on evaluation and assessment in software engineering (pp. 1–6).
S16	Martens, D., & Johann, T. (2017, May). On the emotion of users in app reviews. In 2017 IEEE/ACM 2nd International Workshop on Emotion Awareness in Software Engineering (SEmotion) (pp. 8–14). IEEE.
S17	Genc-Nayebi, N., & Abran, A. (2017). A systematic literature review: Opinion mining studies from mobile app store user reviews. Journal of Systems and Software, 125, 207–219.
S18	Bano, M., Zowghi, D., & Kearney, M. (2017, July). Feature based sentiment analysis for evaluating the mobile pedagogical affordances of apps. In IFIP World Conference on Computers in Education (pp. 281–291). Springer, Cham.
S19	Palomba, F., Linares-Vásquez, M., Bavota, G., Oliveto, R., Di Penta, M., Poshyvanyk, D., & De Lucia, A. (2018). Crowdsourcing user reviews to support the evolution of mobile apps. Journal of Systems and Software, 137, 143–162.
S20	Durelli, V. H., Durelli, R. S., Endo, A. T., Cirilo, E., Luiz, W., & Rocha, L. (2018, September). Please please me: does the presence of test cases influence mobile app users' satisfaction?. In Proceedings of the XXXII Brazilian Symposium on Software Engineering (pp. 132–141).
S21	Li, X., Zhang, Z., & Stefanidis, K. (2018, September). Mobile App Evolution Analysis Based on User Reviews. In 17th International Conference on Intelligent Software Methodologies, Tools, and Techniques (SoMeT) (pp. 773–786).
S22	Luiz, W., Viegas, F., Alencar, R., Mourão, F., Salles, T., Carvalho, D., Gonçalves, M. A., & Rocha, L. (2018, April). A feature-oriented sentiment rating for mobile app reviews. In Proceedings of the 2018 World Wide Web Conference (pp. 1909–1918).
S23	Guzman, E., Oliveira, L., Steiner, Y., Wagner, L. C., & Glinz, M. (2018, May). User feedback in the app store: a cross-cultural study. In 2018 IEEE/ACM 40th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS) (pp. 13–22). IEEE.
S24	Guzman, E., & Rojas, A. P. (2019, September). Gender and user feedback: An exploratory study. In 2019 IEEE 27th International Requirements Engineering Conference (RE) (pp. 381–385). IEEE.
S25	Nicolai, M., Pascarella, L., Palomba, F., & Bacchelli, A. (2019, August). Healthcare Android apps: a tale of the customers' perspective. In Proceedings of the 3rd ACM SIGSOFT International Workshop on App Market Analytics (pp. 33–39).

Table 12
Primary studies mapping to each sub-question.

ID	Venue			Sample		SQ1				SQ2				SQ3						SQ4		SQ5			SQ6			SQ7	
	a	b	c	Apps	Reviews	a	b	c	d	a	b	c	d	a	b	c	d	e	f	a	b	a	b	c	a	b	c	a	b
S01	X			9	5036		X			X	X				X	X					X		X				X	X	
S02	X			32,108	–				X			X				X					X		X	X			X	X	
S03	X			171,000	13,000,000+	X				X	X	X		X	X	X				X			X			X		X	
S04	X			59	556	X				X	X					X	X	X		X				X	X	X	X	X	
S05	X			1100	1,126,453		X				X	X				X	X	X		X				X	X	X	X	X	
S06	X			161	3279	X				X	X	X				X		X		X				X	X	X	X	X	
S07	X			7	32,210	X	X			X	X	X		X	X						X	X			X	X			X
S08		X		8	1487				X		X								X	X			X			X			X

(*continued on next page*)

Table 12 (*continued*).

ID	Venue			Sample		SQ1				SQ2				SQ3						SQ4		SQ5			SQ6			SQ7	
	a	b	c	Apps	Reviews	a	b	c	d	a	b	c	d	a	b	c	d	e	f	a	b	a	b	c	a	b	c	a	b
S09	X			+99	206,751	X				X	X		X			X	X				X			X		X		X	
S10		X		20	6390		X			X	X						X	X		X				X		X		X	
S11	X			95	2,106,605	X				X	X								X	X				X		X			X
S12			X	25	100,000+		X				X				X						X	X			X		X		X
S13		X		5000	~2,500,000	X				X	X		X				X		X	X				X		X		X	
S14	X			51	303,694	X					X				X						X		X		X				X
S15	X			1	4442	X				X	X					X					X	X				X		X	
S16			X	245	7,396,551		X			X	X	X			X	X				X				X	X	X	X	X	
S18	X			10	25,035	X	X			X	X				X					X			X		X				X
S19		X		100	5792	X				X	X		X			X	X				X			X	X			X	
S20	X			60	21,000+	X				X	X			X	X	X	X				X		X			X	X	X	
S21	X			1	1,148,032	X					X	X		X	X				X		X	X			X	X			X
S22	X			7	22,815	X				X	X			X	X						X	X			X	X			X
S23	X			7	59,204		X			X	X					X	X			X				X	X	X	X	X	
S24	X			7	919		X			X	X						X	X			X			X			X	X	
S25			X	8431	383,758	X				X	X				X			X		X			X		X	X	X	X	

Appendix A. Extraction form for primary studies

TITLE: **AUTHORS:** **PUBLISHED IN:** **VENUE:** ( ) Conference ( ) Journal ( ) Workshop **YEAR:**	
Table for data extraction	
Publication summary	Overview of the publication (What is the goal of the publication? What is the motivation?)
Research sub-questions	**Answers**
SQ1. From what source were the reviews obtained?	(a) Google Play Store (b) Apple App Stor (c) Windows Phone Store (d) Other
SQ2. Which information was extracted from the source?	(a) Rating (b) User review (c) App information (d) Other (specify)
SQ3. Which methods were used for analyzing the data extracted?	(a) Topic modeling (b) Sentiment analysis (c) Descriptive statistics (d) Statistical tests (e) Manual analysis (f) Other (specify)
SQ4. Was the information categorized? How?	(a) Yes (describe) (b) No
SQ5. How was the data organized during the analysis?	(a) Individual (b) Group (c) General
SQ6. What polarity is the factor associated to?	(a) Positive (b) Negative (c) Neutral
SQ7. Was the influence of the factor on user rating or sentiment analyzed?	(a) Yes (b) No

Appendix B. Extraction form for secondary studies

TITLE: **AUTHORS:** **PUBLISHED IN:** **VENUE:** ( ) Conference ( ) Journal ( ) Workshop **YEAR:**	
Extraction table for systematic reviews and mappings	
Q1. What is the purpose of the research?	Description of the research goals.
Q2. What are the research questions?	Description of the research questions that the SRL/SML sought to answer.
Q3. Which string was used?	String used in the search.
Q4. In what fields has the string been searched?	Description of the fields in which the string was searched, such as title, abstract, or full-text.
Q5. Which databases have been queried?	Listing of the databases in which the string was run.
Q6. What are the inclusion criteria?	Description of the inclusion criteria.
Q7. How many articles are included?	Total number of articles included after the 2nd filter.
Q8. What information is extracted from the articles?	Description of the fields used in the extraction form.
Q9. Describe the analysis of results.	Description of how the analysis was performed and its results.
Q10. What are the limitations of this SRL/SML?	Description of SRL/SML limitations.

Appendix C

See Table 11.

Appendix D

See Table 12.

Appendix E

See Table 13.

Table 13
Factors' consolidation and polarity mapping.

Original Factor	First iteration	Consolidated factor	Pos	Neg	Neu	Ind	Gro	Gen	Publication
Accessible	Unknown (vague)	–	X			X			S22
Accuracy	Accuracy	Accuracy		X			X		S03
Adjective (negative)	Adjective	Attractiveness		X				X	S04
Adjective (positive)	Adjective	Attractiveness	X					X	S04
Ads in the paid version	Spam/Ads	Spam/Ads		X		X			S22
Aesthetics (negative)	Aesthetics	Interface		X				X	S04
Aesthetics (positive)	Aesthetics	Interface			X			X	S04
App crashing	Bugs/Crash	Bugs/Crash		X				X	S10
App improvement from reviews	Implementation of requests	Update	X					X	S19
App interface	Interface	Interface	X			X			S22
App redesign	App redesign	Update		X				X	S16
App version	App version	App version			X		X		S01
Attractiveness	Attractiveness	Attractiveness		X			X		S03
Authentication	Feature/Functionality	Feature/Functionality		X				X	S11
Battery and Versioning	Battery and Versioning	–		X				X	S11
Battery drain	Battery duration	Resource use		X			X		S20
Bug report	Bugs/Crash	Bugs/Crash		X				X	S16
Bugs	Bugs/Crash	Bugs/Crash		X				X	S05
Bugs	Bugs/Crash	Bugs/Crash		X		X			S22
Calorie counter, workout tracker	Feature/Functionality	Feature/Functionality	X			X			S12
Changing app requirements	Changing app requirements	Update		X				X	S16
Cheaper app	Cheaper app	Cost			X			X	S06
Compatibility	Compatibility	Compatibility		X			X		S03
Compatibility	Compatibility	Compatibility		X				X	S10
Complaints	Complaints	Attractiveness		X			X		S25
Compliments	Compliments	Attractiveness	X				X		S25
Connection	Network problem	Network problems		X				X	S11
Connectivity	Network problem	Network problems		X			X		S03
Connectivity	Network problem	Network problems		X		X			S22
Content request	Content request	Improvement request			X			X	S05
Cost	Cost	Cost		X			X		S03
Cost (negative)	Cost	Cost		X				X	S04
Cost (positive)	Cost	Cost			X			X	S04
Culture	Culture	Culture	X	X	X			X	S23
Customer support (positive)	Customer support	Customer support	X					X	S06
Date/Time	Date/Time	Date/Time			X		X		S01
Design	Interface	Interface		X		X			S22
Device model	Device model	Device		X				X	S09
Difficulty evaluating the hotels in the app	Unknown (vague)	–		X		X			S22
Dispraise	Dispraise	Attractiveness		X				X	S05
Dissuasion	Dissuasion	Recommendation		X				X	S05
Does not work-TR	Does not work	Bugs/Crash		X				X	S04
Easiness to use	Ease of use	Ease of use	X			X			S07
Ethical aspects (misleading app)	Misleading app	Misleading app		X				X	S06
Extended time (time, long, slow)	Performance	Performance		X		X			S22
Facebook	Not a factor	–		X				X	S11
Feature information	Feature information	–	X					X	S05
Feature removal	Feature removal	Feature removal		X				X	S10
Feature request	Feature request	Improvement request		X				X	S05
Feature request	Feature request	Improvement request		X				X	S10
Feature request	Feature request	Improvement request			X		X		S25
Feature/Functionality (negative)	Feature/Functionality	Feature/Functionality		X				X	S04
Feature/Functionality (positive)	Feature/Functionality	Feature/Functionality	X					X	S04
Features	Feature/Functionality	Feature/Functionality		X		X			S15
FindBugs warnings	FindBugs warnings	FindBugs warnings		X				X	S13
Functional error	Functional error	Bugs/Crash		X				X	S10
Gender	Gender	Gender			X			X	S24
Helpfulness	Helpfulness	Helpfulness	X					X	S05
Hidden cost	Hidden cost	Cost		X				X	S10
How to	How to	–			X			X	S05
Improvement request	Improvement request	Improvement request		X				X	S05
Instantaneous	Unknown (vague)	–	X			X			S22
Interface design	Interface	Interface		X				X	S10
Interface style	Interface	Interface	X			X			S22
Bugs	Bugs/Crash	Bugs/Crash		X				X	S06
Media	Feature/Functionality	Feature/Functionality		X			X		S03
Messaging	Feature/Functionality	Feature/Functionality		X				X	S11
Missing Feature/Functionality	Missing feature/Functionality	Improvement request			X			X	S04
Network problem	Network problem	Network problems		X				X	S10
Other app	Comparison	Comparison			X			X	S05
Other feedback	Other feedback	–			X			X	S05
Personalization	Personalization	Personalization	X				X		S18
Photo editing	Feature/Functionality	Feature/Functionality	X			X			S22
Picture	Feature/Functionality	Feature/Functionality		X			X		S03

(*continued on next page*)

Table 13 (*continued*).

Original Factor	First iteration	Consolidated factor	Pos	Neg	Neu	Ind	Gro	Gen	Publication
Pin thing, find thing, view file, open file	Feature/Functionality	Feature/Functionality	X			X			S07
Praise	Praise	Attractiveness	X					X	S05
Presence of test cases	Presence of test cases	Presence of test cases			X		X		S20
Price X Downloads Price X Ratings	Cost	Cost			X		X	X	S02
Price/Cost	Cost	Cost			X			X	S16
Privacy and Ethical	Privacy and Ethical	Privacy and Ethical		X				X	S10
Problem discovering	Bugs/Crash	Bugs/Crash		X			X		S25
Problem reporting	Bugs/Crash	Bugs/Crash		X			X		S25
Problems after updates	Problems after updates	Update		X			X		S20
Problems after updates	Problems after updates	Update		X				X	S10
Problems after updates	Problems after updates	Update		X				X	S06
Promise	Promise	Improvement request		X				X	S05
Question	Question	–			X			X	S05
Recommendation	Recommendation	Recommendation	X					X	S05
Request for features regarding contact and status options	Feature request	Improvement request		X		X			S21
Resource heavy	Resource use	Resource use		X				X	S10
Search: Ease of searching of various information Touch: Ease of clicking and dragging of various contents	Feature/Functionality	Feature/Functionality		X			X		S08
Shortcoming	Shortcoming	Feature removal		X				X	S05
Show pin, search something, update time, want upload, take photo	Feature/Functionality	Feature/Functionality		X		X			S07
Simplicity, friendly, ease of use	Simplicity	Simplicity	X				X		S14
Snapchat	Not a factor	–		X				X	S11
Songs	Feature/Functionality	Feature/Functionality	X			X			S22
Spam/Ads	Spam/Ads	Spam/Ads		X			X		S03
Stability	Stability	Bugs/Crash		X			X		S03
Storage	Resource use	Resource use		X		X			S22
Support for loading large videos	Feature/Functionality	Feature/Functionality	X			X			S22
Telephony	Feature/Functionality	Feature/Functionality		X			X		S03
Theme upgrades (update, stickers, themes, wallpaper)	Feature request	Improvement request		X		X			S22
Themes	Feature/Functionality	Feature/Functionality	X			X			S22
Time battery	Battery duration	Resource use	X			X			S22
Time consuming installation and finding packages	Install (slow)	–		X		X			S22
Track calorie, track weight, exercise activity	Feature/Functionality	Feature/Functionality			X	X			S12
Transfer of money, ability to make card payments, getting account summary, ease of access, etc.	Feature/Functionality	Feature/Functionality	X				X		S14
Uninteresting content	Uninteresting content	Attractiveness		X				X	S10
Unrecoverable error	Unrecoverable error	Bugs/Crash		X				X	S11
Unresponsive app	Performance	Performance		X				X	S10
Update failures	Update	Update		X		X			S22
Update the application	Unknown (vague)	–		X		X			S22
Updates (positive)	Update	Update	X					X	S16
Use in tablets	Compatibility	Compatibility		X		X			S22
User profile of an app type	User profile of an app type	User profile of an app type	X				X		S25
Versioning	Versioning	Update	X					X	S06
Video and voice call quality of the app	Feature/Functionality	Feature/Functionality	X			X			S21
Work-NTR	Work	–	X					X	S04

References

Alves, R., Valente, P., Nunes, N.J., 2014. The state of user experience evaluation practice. In: Proceedings of the 8th Nordic Conference on Human-Computer Interaction Fun, Fast, Foundational - NordiCHI '14. Presented At the 8th Nordic Conference. ACM Press, Helsinki, Finland, pp. 93–102. http://dx.doi.org/10.1145/2639189.2641208.

Bargas-Avila, J.A., 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, CHI '11. ACM, Vancouver, BC, USA, pp. 2689–2698. http://dx.doi.org/10.1145/1978942.1979336.

Bavota, G., Linares-Vasquez, M., Bernal-Cardenas, C.E., Penta, M.D., Oliveto, R., Poshyvanyk, D., 2015. The impact of API change- and fault-proneness on the user ratings of android apps. IEEE Trans. Softw. Eng. 41, 384–407. http://dx.doi.org/10.1109/TSE.2014.2367027.

Bruun, A., Ahm, S., 2015. Mind the gap! comparing retrospective and concurrent ratings of emotion in user experience evaluation. In: Abascal, J., Barbosa, S., Fetter, M., Gross, T., Palanque, P., Winckler, M. (Eds.), Human-Computer Interaction – INTERACT 2015. Springer International Publishing, Cham, pp. 237–254.

Chen, N., Lin, J., Hoi, S.C.H., Xiao, X., Zhang, B., 2014. Ar-miner: mining informative reviews for developers from mobile app marketplace. In: Proceedings of the 36th International Conference on Software Engineering - ICSE 2014. Presented At the 36th International Conference. ACM Press, Hyderabad, India, pp. 767–778. http://dx.doi.org/10.1145/2568225.2568263.

Chromik, M., Lachner, F., Butz, A., 2020. Ml for ux? - an inventory and predictions on the use of machine learning techniques for UX research. In: Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society. Presented At the NordiCHI '20: Shaping Experiences, Shaping Society. ACM, Tallinn Estonia, pp. 1–11. http://dx.doi.org/10.1145/3419249.3420163.

Cockburn, A., Quinn, P., Gutwin, C., 2015. Examining the peak-end effects of subjective experience. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15. Presented At the 33rd Annual ACM Conference. ACM Press, Seoul, Republic of Korea, pp. 357–366. http://dx.doi.org/10.1145/2702123.2702139.

Cockburn, A., Quinn, P., Gutwin, C., 2017. The effects of interaction sequencing on user experience and preference. Int. J. Hum.-Comput. Stud. 108, 89–104. http://dx.doi.org/10.1016/j.ijhcs.2017.07.005.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46. http://dx.doi.org/10.1177/001316446002000104.

Durelli, V.H.S., Durelli, R.S., Endo, A.T., Cirilo, E., Luiz, W., Rocha, L., 2018. Please please me: Does the presence of test cases influence mobile app users' satisfaction? In: Proceedings of the XXXII Brazilian Symposium on Software Engineering, SBES '18. Association for Computing Machinery, New York, NY, USA, pp. 132–141. http://dx.doi.org/10.1145/3266237.3266272.

Dyba, T., Dingsoyr, T., Hanssen, G.K., 2007. Applying systematic reviews to diverse study types: An experience report. In: First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007). Presented At the First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007). IEEE, Madrid, Spain, pp. 225–234. http://dx.doi.org/10.1109/ESEM.2007.59.

Ebrahimi, F., Tushev, M., Mahmoud, A., 2020. Mobile app privacy in software engineering research: A systematic mapping study. Inf. Softw. Technol. 106466. http://dx.doi.org/10.1016/j.infsof.2020.106466.

Fernandez, A., Insfran, E., Abrahão, S., 2011. Usability evaluation methods for the web: A systematic mapping study. Inf. Softw. Technol. 53, 789–817. http://dx.doi.org/10.1016/j.infsof.2011.02.007.

Fu, B., Lin, J., Li, L., Faloutsos, C., Hong, J., Sadeh, N., 2013. Why people hate your app: making sense of user feedback in a mobile app store. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '13. Presented At the 19th ACM SIGKDD International Conference. ACM Press, Chicago, Illinois, USA, p. 1276. http://dx.doi.org/10.1145/2487575.2488202.

Geert, H., Jan, H.G., 1991. Cultures and Organizations: Software of the Mind. McGaw-Hill, London.

Genc-Nayebi, N., Abran, A., 2017. A systematic literature review: Opinion mining studies from mobile app store user reviews. J. Syst. Softw. 125, 207–219.

Goul, M., Marjanovic, O., Baxley, S., Vizecky, K., 2012. Managing the enterprise business intelligence app store: Sentiment analysis supported requirements engineering. In: Proceedings of the 2012 45th Hawaii International Conference on System Sciences, HICSS '12. IEEE Computer Society, USA, pp. 4168–4177. http://dx.doi.org/10.1109/HICSS.2012.421.

Gutwin, C., Rooke, C., Cockburn, A., Mandryk, R.L., Lafreniere, B., 2016. Peak-end effects on player experience in casual games. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16. Presented At the 2016 CHI Conference. ACM Press, Santa Clara, California, USA, pp. 5608–5619. http://dx.doi.org/10.1145/2858036.2858419.

Guzman, E., Maalej, W., 2014. How do users like this feature? A fine grained sentiment analysis of app reviews. In: 2014 IEEE 22nd International Requirements Engineering Conference (RE). Presented At the 2014 IEEE 22nd International Requirements Engineering Conference (RE). IEEE, Karlskrona, Sweden, pp. 153–162. http://dx.doi.org/10.1109/RE.2014.6912257.

Guzman, E., Oliveira, L., Steiner, Y., Wagner, L.C., Glinz, M., 2018. User feedback in the app store: a cross-cultural study. In: Proceedings of the 40th International Conference on Software Engineering Software Engineering in Society - ICSE-SEIS '18. Presented At the 40th International Conference. ACM Press, Gothenburg, Sweden, pp. 13–22. http://dx.doi.org/10.1145/3183428.3183436.

Guzman, E., Paredes Rojas, A., 2019. Gender and user feedback: An exploratory study. In: Damian, D. L.S.-W., Perini, A. (Eds.), Proceedings of the IEEE International Conference on Requirements Engineering. IEEE Computer Society, pp. 381–385. http://dx.doi.org/10.1109/RE.2019.00049.

Ha, E., Wagner, D., 2013. Do android users write about electric sheep? Examining consumer reviews in google play. In: 2013 IEEE 10th Consumer Communications and Networking Conference (CCNC). Presented At the 2013 IEEE 10th Consumer Communications and Networking Conference (CCNC). IEEE, Las Vegas, NV, pp. 149–157. http://dx.doi.org/10.1109/CCNC.2013.6488439.

Hassenzahl, M., 2018. The thing and I (summer of '17 remix). In: Blythe, M., Monk, A. (Eds.), Funology 2: From Usability To Enjoyment. Springer International Publishing, Cham, pp. 17–31. http://dx.doi.org/10.1007/978-3-319-68213-6_2.

Hassenzahl, M., Sandweg, N., 2004. From mental effort to perceived usability: transforming experiences into summary assessments. In: Extended Abstracts of the 2004 Conference on Human Factors and Computing Systems - CHI '04. Presented At the Extended Abstracts of the 2004 Conference. ACM Press, Vienna, Austria, p. 1283. http://dx.doi.org/10.1145/985921.986044.

Hassenzahl, M., Tractinsky, N., 2006. User experience - a research agenda. Behav. Inf. Technol. 25, 91–97. http://dx.doi.org/10.1080/01449290500330331.

Hedegaard, S., 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. Presented At the SIGCHI Conference. ACM Press, Paris, France, p. 2089. http://dx.doi.org/10.1145/2470654.2481286.

Hedegaard, S., 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. Presented At the 8th Nordic Conference. ACM Press, Helsinki, Finland, pp. 157–166. http://dx.doi.org/10.1145/2639189.2639211.

Hoon, L., Vasa, R., Schneider, J.-G., Mouzakis, K., 2012. A preliminary analysis of vocabulary in mobile app user reviews. In: Proceedings of the 24th Australian Computer-Human Interaction Conference on - OzCHI '12. Presented At the 24th Australian Computer-Human Interaction Conference. ACM Press, Melbourne, Australia, pp. 245–248. http://dx.doi.org/10.1145/2414536.2414578.

Hu, H., Bezemer, C.-P., Hassan, A.E., 2018. Studying the consistency of star ratings and the complaints in 1 & 2-star user reviews for top free cross-platform android and iOS apps. Empir. Softw. Eng. 23, 3442–3475. http://dx.doi.org/10.1007/s10664-018-9604-y.

Huang, Albert, 2002. A research taxonomy for e-commerce system usability. AMCIS 2002 Proceedings 94.

Iacob, C., Veerappa, V., Harrison, R., 2013. What are you complaining about?: A study of online reviews of mobile applications. In: Presented At the 27th International BCS Human Computer Interaction Conference (HCI 2013). http://dx.doi.org/10.14236/ewic/HCI2013.38.

ISO 9241-11, 2018. Ergonomics of human–system interaction — Part 11: Usability: Definitions and concepts.

ISO/IEC 25012, 2008. Software engineering — Software product quality requirements and evaluation (square) — Data quality model.

Khalid, H., Shihab, E., Nagappan, M., Hassan, A.E., 2015. What do mobile app users complain about? IEEE Softw. 32, 70–77. http://dx.doi.org/10.1109/MS.2014.50.

Kitchenham, B.A., Budgen, D., Brereton, P., 2015. Evidence-Based Software Engineering and Systematic Reviews. CRC Press.

Kitchenham, B., Charters, S., 2007. Guidelines for performing systematic literature reviews in software engineering.

Kujala, S., Mugge, R., Miron-Shatz, T., 2017. The role of expectations in service evaluation: A longitudinal study of a proximity mobile payment service. Int. J. Hum.-Comput. Stud. 98, 51–61. http://dx.doi.org/10.1016/j.ijhcs.2016.09.011.

Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., Karapanos, E., Sinnelä, A., 2011. Ux curve: A method for evaluating long-term user experience. Interact. Comput. 23, 473–483. http://dx.doi.org/10.1016/j.intcom.2011.06.005.

Lallemand, C., Gronier, G., Koenig, V., 2015. User experience: A concept without consensus? Exploring practitioners' perspectives through an international survey. Comput. Hum. Behav. 43, 35–48. http://dx.doi.org/10.1016/j.chb.2014.10.048.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (159), http://dx.doi.org/10.2307/2529310.

Law, E.L.-C., Roto, V., Hassenzahl, M., Vermeeren, A.P.O.S., Kort, J., 2009. Understanding, scoping and defining user experience: A survey approach. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09. ACM, Boston, MA, USA, pp. 719–728. http://dx.doi.org/10.1145/1518701.1518813.

Law, E.L.-C., van Schaik, P., 2010. Modelling user experience – an agenda for research and practice. Interact. Comput. 22, 313–322. http://dx.doi.org/10.1016/j.intcom.2010.04.006.

Luiz, W., Viegas, F., Alencar, R., Mourão, F., Salles, T., Carvalho, D., Gonçalves, M.A., Rocha, L., 2018. A feature-oriented sentiment rating for mobile app reviews. In: Proceedings of the 2018 World Wide Web Conference, WWW '18. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva. CHE, pp. 1909–1918. http://dx.doi.org/10.1145/3178876.3186168.

Martens, D., Johann, T., 2017. On the emotion of users in app reviews. In: Proceedings of the 2nd International Workshop on Emotion Awareness in Software Engineering, SEmotion '17. IEEE Press, Buenos Aires, Argentina, pp. 8–14.

Martin, W., Sarro, F., Jia, Y., Zhang, Y., Harman, M., 2017. A survey of app store analysis for software engineering. IEEE Trans. Softw. Eng. 43, 817–847. http://dx.doi.org/10.1109/TSE.2016.2630689.

Masrury, R.A., Fannisa Alamsyah, A., 2019. Analyzing tourism mobile applications perceived quality using sentiment analysis and topic modeling. In: 2019 7th International Conference on Information and Communication Technology (ICoICT). pp. 1–6. http://dx.doi.org/10.1109/ICoICT.2019.8835255.

Mendes, E., Wohlin, C., Felizardo, K., Kalinowski, M., 2020. When to update systematic literature reviews in software engineering. J. Syst. Softw. 167, 110607. http://dx.doi.org/10.1016/j.jss.2020.110607.

Michalco, J., Simonsen, J.G., Hornbæk, K., 2015. An exploration of the relation between expectations and user experience. Int. J. Hum.-Comput. Interact. 31, 603–617. http://dx.doi.org/10.1080/10447318.2015.1065696.

Nakamura, W.T., de Oliveira, E.H.T., Conte, T., 2019. Negative emotions, positive experience: What are we doing wrong when evaluating the ux? In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems - CHI EA '19. Presented At the Extended Abstracts of the 2019 CHI Conference. ACM Press, Glasgow, Scotland Uk, pp. 1–6. http://dx.doi.org/10.1145/3290607.3313000.

Nicolai, M., Pascarella, L., Palomba, F., Bacchelli, A., 2019. Healthcare android apps: A tale of the customers' perspective. In: Proceedings of the 3rd ACM SIGSOFT International Workshop on App Market Analytics, WAMA 2019. Association for Computing Machinery, New York, NY, USA, pp. 33–39. http://dx.doi.org/10.1145/3340496.3342758.

Pagano, D., Maalej, W., 2013. User feedback in the appstore: An empirical study. In: 2013 21st IEEE International Requirements Engineering Conference (RE). Presented At the 2013 IEEE 21st International Requirements Engineering Conference (RE). IEEE, Rio de Janeiro-RJ, Brazil, pp. 125–134. http://dx.doi.org/10.1109/RE.2013.6636712.

Panichella, S., Sorbo, A.Di., Guzman, E., Visaggio, C.A., Canfora, G., Gall, H.C., 2015. How can i improve my app? Classifying user reviews for software maintenance and evolution. In: 2015 IEEE International Conference on Software Maintenance and Evolution (ICSME). Presented At the 2015 IEEE International Conference on Software Maintenance and Evolution (ICSME). IEEE, Bremen, Germany, pp. 281–290. http://dx.doi.org/10.1109/ICSM.2015.7332474.

Panosian, H., 2017. Submitting your app. In: Learn IOS Application Distribution. Springer, pp. 153–158.

Petersen, K., Vakkalanka, S., Kuzniarz, L., 2015. Guidelines for conducting systematic mapping studies in software engineering: An update. Inf. Softw. Technol. 64, 1–18. http://dx.doi.org/10.1016/j.infsof.2015.03.007.

Pettersson, I., Lachner, F., Frison, A.-K., Riener, A., Butz, A., 2018. A bermuda triangle?: A review of method application and triangulation in user experience evaluation. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18. ACM, Montréal, QC, Canada, pp. 461:1–461:16. http://dx.doi.org/10.1145/3173574.3174035.

Porat, T., Tractinsky, N., 2012. It's a pleasure buying here: The effects of web-store design on consumers' emotions and attitudes. Hum.-Comput. Interact. 27, 235–276. http://dx.doi.org/10.1080/07370024.2011.646927.

Rodrigues, P., Silva, I.S., Barbosa, G.A.R., Coutinho, S., Mourão, F., 2017. Beyond the stars: Towards a novel sentiment rating to evaluate applications in web stores of mobile apps. In: Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17 Companion. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva. CHE, pp. 109–117. http://dx.doi.org/10.1145/3041021.3054139.

Sagnier, C., Loup-Escande, E., Valléry, G., 2020. Effects of gender and prior experience in immersive user experience with virtual reality. In: Ahram, T., Falcão, C. (Eds.), Advances in Usability and User Experience, Advances in Intelligent Systems and Computing. Springer International Publishing, Cham, pp. 305–314. http://dx.doi.org/10.1007/978-3-030-19135-1_30.

Selz, D., Schubert, P., 1998. Web assessment: A model for the evaluation and assessment of successful electronic commerce applications. In: Proceedings of the Thirty-First Hawaii International Conference on System Sciences, Vol. 4. pp. 222–231. http://dx.doi.org/10.1109/HICSS.1998.655278.

Simmons, A., Hoon, L., 2016. Agree to disagree: On labelling helpful app reviews. In: Proceedings of the 28th Australian Conference on Computer-Human Interaction, OzCHI '16. Association for Computing Machinery, New York, NY, USA, pp. 416–420. http://dx.doi.org/10.1145/3010915.3010976.

Tarafdar, M., Zhang, J., 2005. Analyzing the influence of web site design parameters on web site usability. Inf. Resour. Manage. J. (IRMJ) 18 (4), 62–80. http://dx.doi.org/10.4018/irmj.2005100104.

Vasa, R., Hoon, L., Mouzakis, K., Noguchi, A., 2012. A preliminary analysis of mobile app user reviews. In: Proceedings of the 24th Australian Computer-Human Interaction Conference on - OzCHI '12. Presented At the 24th Australian Computer-Human Interaction Conference. ACM Press, Melbourne, Australia, pp. 241–244. http://dx.doi.org/10.1145/2414536.2414577.

Wang, C., Daneva, M., Sinderen, M., Liang, P., 2019. A systematic mapping study on crowdsourced requirements engineering using user feedback. J. Softw.: Evol. Process 31. http://dx.doi.org/10.1002/smr.2199.

Wang, H., Lu, Y., Zhai, C., 2010. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Presented at the ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 783–792.

Weichbroth, Pawel, 2020. Usability of mobile applications: A systematic literature study. IEEE Access 8, 55563–55577. http://dx.doi.org/10.1109/ACCESS.2020.2981892.

Weichbroth, P., Baj-Rogowska, A., 2019. Do online reviews reveal mobile application usability and user experience? The case of WhatsApp. In: 2019 Federated Conference on Computer Science and Information Systems (FedCSIS). pp. 747–754. http://dx.doi.org/10.15439/2019F289.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A., 2012. Experimentation in Software Engineering. Springer Science & Business Media.

Zhang, H., Babar, M.A., Tell, P., 2011. Identifying relevant studies in software engineering. Inf. Softw. Technol. 53, 625–637. http://dx.doi.org/10.1016/j.infsof.2010.12.010.

Walter Nakamura holds a Ph.D. in Informatics from the Federal University of Amazonas (UFAM) and is currently an assistant professor at the Federal University of Technology – Paraná (UTFPR). His research interests include the intersection between Human–Computer Interaction, Software Engineering, and Machine Learning.

Edson Oliveira earned his Ph.D. in Informatics from the Federal University of Amazonas (UFAM) in 2017. He is currently a researcher at Amazonas State Secretariat of Finance (SEFAZ). He has experience in Computer Science, with an emphasis on Software Engineering, Information Retrieval, Data Mining, and Machine Learning.

Elaine H. T. de Oliveira is an Associate Professor at the Institute of Computing at the Federal University of Amazonas (UFAM) and earned her Ph.D. in Informatics in Education from the Federal University of Rio Grande do Sul (UFRGS). Her main research interests include Informatics in Education, Computer Education, Artificial Intelligence applied to Education, Cognitive Science, and Adaptive Systems.

David Redmiles is a Professor in the Department of Informatics at the University of California, Irvine (UCI) in the Donald Bren School of Information and Computer Sciences (ICS). His current research focuses on distributed and collaborative software engineering, especially the aspects of awareness and trust among collaborators.

Tayana Conte is an associate professor at the Federal University of Amazonas (UFAM), Institute of Computing (IComp), heading the Usability and Software Engineering (USES) lab. Her research interests include the intersection between Software Engineering and Human–Computer Interaction, Software Quality, Human-Centered Computing, and Empirical Software Engineering.