Evolution may come with a price: analyzing user reviews to understand the impact of updates on mobile apps accessibility

Paulo S. H. dos Santos
Alberto D. A. Oliveira
Thais B. N. de Jesus
pauloshsantos@usp.br
albertodumont@usp.br
thais.bonjorni@usp.br
University of São Paulo
Brazil

Wajdi Aljedaani
wajdialjedaani@my.unt.edu
University of North Texas
USA

Marcelo M. Eler
marceloeler@usp.br
University of São Paulo
Brazil

ABSTRACT

Mobile applications are constantly updated to adapt to evolving user and environment requirements. As changes are successively implemented to promote user satisfaction, the complexity of mobile apps increases, and overlooked quality aspects (e.g. privacy, security, power consumption) may decline if counter-measures are not adopted. In this paper, we analyzed accessibility reviews of mobile apps to show evidence that updates may introduce barriers that make the app less accessible than its previous version according to users' perceptions. Our results show that accessibility barriers reported by users mostly include incompatibility with screen readers, the removal of accessibility features (e.g. color scheme or font customization), small font sizes and widgets, and incompatibility with the device's accessibility configuration. The accessibility barriers impact the users considering different levels: i) perception: an inability or difficulty to see, read or distinguish interface elements due to their small size or color; ii) understanding: the inability or difficulty to navigate, access and interpret information; iii) operation: the inability of difficulty to perform tasks such as adding items to the cart, reading or sending messages, and booking a ride; iv) and physical reactions, such as eyestrain, vertigo, and headache. The sentiments expressed by users are generally negative, including frustration, disappointment, and sometimes a sense of discrimination against people with disabilities. The results of our study raise an alert to organizations and developers that they should implement measures to avoid introducing accessibility barriers while they add new features to their mobile apps.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility**; *User studies*; • **Software and its engineering** → **Software evolution**.

KEYWORDS

Accessibility, mobile, app, update, evolution, user reviews

ACM Reference Format:
Paulo S. H. dos Santos, Alberto D. A. Oliveira, Thais B. N. de Jesus, Wajdi Aljedaani, and Marcelo M. Eler. 2023. Evolution may come with a price: analyzing user reviews to understand the impact of updates on mobile apps accessibility. In *XXII Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais (IHC '23), October 16–20, 2023, Maceió, Brazil.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3638067.3638081

1 INTRODUCTION

Lehman's laws describe the dynamic and the conflicting nature of software evolution [17, 18]. The first law defines that a software system must adapt to changing requirements and environments otherwise it will become progressively less useful. It seems that mobile development is a suitable illustration of this scenario as developers and organizations constantly release new and improved versions of their mobile apps to satisfy the evolving demand of their user-base and to keep up with the competition [14, 21, 25, 30].

Mobile app evolution is mostly feedback driven. Accordingly, organizations leverage user reviews published in app stores to understand the mismatch between the current version of their software and the user's expectations, which is valuable for creating or updating product roadmaps. Usually, mobile apps evolve due to many reasons [10, 14, 15, 24]: to comply with changes in the operating systems (e.g. Android and iOS); to leverage new capabilities of physical devices; to increase performance, privacy, and security aspects; to improve power consumption; to fix bugs; and to incorporate new features that make the app more competitive.

Even though software evolution is mostly driven toward user satisfaction, which can be considered the ultimate desirable quality of a product, the second Lehman's law describes the intrinsic risks of constant modifications in software: as the need for adaptation arises and changes are successively implemented, the complexity and the interdependence between system elements increase in an unstructured pattern, which can provoke an overall quality decline unless something is done to prevent or mitigate it [17, 18].

In mobile development, researchers have studied the impact of software evolution on many quality aspects, including security [28], complexity [14], and usability, resource consumption, and maintenance effort [22]. However, to the best of our knowledge, few studies have analyzed the impact of app updates on mobile accessibility.

IHC '23, October 16–20, 2023, Maceió, Brazil

ACM ISBN 979-8-4007-1715-4/23/10…$15.00
https://doi.org/10.1145/3638067.3638081

Alshayban et al. [6] and Chen et al. [9] analyzed subsequent versions of 60 and 70 mobile apps, respectively, to check whether their accessibility has improved or declined based on metrics associated with accessibility issues they found, but they have not investigated the specific changes in the mobile apps after the updates.

Therefore, in this paper, we present an overview of the impact of app updates on mobile accessibility according to the users' perspective. Accordingly, we analyzed user reviews published in the Google Play Store[1] to identify accessibility changes reported by users as a consequence of app updates. Even though many studies have been conducted to analyze accessibility reviews[2] [1, 2, 4–6, 12, 20, 23, 27], we have not found studies on the impact of app updates from the users' perspective. To make it easier to refer to accessibility reviews associated with app updates and visual disabilities or eye conditions, we call them *accessibility update reviews*.

We based our investigation on a dataset of 4,999 accessibility reviews associated with visual disabilities or eye conditions published by Oliveira et al. [23], which is the result of the analysis of nearly 180 million user reviews of 340 popular apps in the Google Play Store. We framed our investigation around six research questions:

RQ$_1$: How many accessibility reviews are associated with app updates? Our aim is to understand whether and how often users give positive or negative feedback concerning accessibility changes noticed after app updates.

RQ$_2$: What are the interface components and resources mentioned in the accessibility update reviews? Our purpose is to acquire some insight into what elements of the interface are more likely to impact accessibility after going through some changes.

RQ$_3$: What are the main issues reported in the accessibility update reviews? We intend to identify instances of accessibility barriers introduced by app updates to give evidence that even though evolution moves the app toward user satisfaction, there might be some collateral damage if certain quality aspects are overlooked. More specifically, we want to identify what rendered the interface or interaction inaccessible (e.g. small font size).

RQ$_4$: What are the accessibility improvements reported in the accessibility update reviews? We intend to show whether users recognize any accessibility improvement associated with app evolution and which enhancements were identified.

RQ$_5$: What are the main consequences for the users reported in the accessibility update reviews? Our aim is to provide evidence of the practical consequences users face when the app becomes less accessible after some update. With this question, we want to identify practical issues in effectively completing a task in the app (e.g. inability to book a meal in delivery apps).

RQ$_6$: What sentiments are expressed in the accessibility update reviews? Our purpose is to show that app updates that affect accessibility may provoke different sentiments and emotions on users, ranging from gratitude to complete dissatisfaction and rage, depending on the context.

Understanding the impact of updates on apps accessibility is relevant because they give evidence that one cannot assume that accessible apps will remain accessible indefinitely, raising an alert that, despite the fact that evolution moves the app towards a quality target, many users can be negatively affected if keeping the app accessible is not a concomitant goal. In summary, the main contributions of this paper are:

- We give evidence that even mobile apps that were accessible according to users' testimonials went through a changing process that made them less accessible
- We list the most common accessibility barriers introduced by app updates and their impact on users. Calculating and comparing accessibility metrics is important from an object perspective, but we believe that delving into the specific accessibility changes reported by users can give developers and researchers insights into which modifications are mostly affecting the usability and user experience, especially for people with visual disabilities or eye conditions.
- We show that user satisfaction is not completely achieved if app evolution does not take into consideration concomitant quality aspects such as accessibility features, as our results show that many negative sentiments and emotions are involved in users' feedback.
- We created a dataset of 694 accessibility update reviews that can be used for further investigations.

This paper is organized as follows. Section 2 describes the related work. Section 3 outlines the research method we adopted to answer our research questions. Section 4 presents and discusses our results by answering our research questions. Section 5 shows the threats to the validity of this study. Finally, Section 6 presents some concluding remarks and future directions.

2 RELATED WORK

This study is intended to identify changes in mobile accessibility after updates according to user reviews. Accordingly, in this section, we present related work with respect to studies that analyze accessibility reviews and studies that investigate the impact of updates on different quality aspects of mobile apps.

Many researchers have investigated accessibility reviews with different purposes: i) to characterize general accessibility issues reported by users [6, 12]; ii) to identify accessibility reviews related to specific types of disabilities [23, 27]; iii) to find accessibility associated with specific applications [1]; and iv) to propose machine learning algorithms to identify and analyze accessibility reviews [2–5]. However, to the best of our knowledge, there is a study on user reviews that investigate how users perceive the impact of app updates. Rather, there are several studies that investigate how user reviews impact software updates in general [10, 20, 24].

Previous studies have examined the impact of app evolution on different quality aspects of mobile apps, including accessibility. Gao et al. [14] found evidence that app complexity does not significantly change during app updates, probably because deletion of functionality is common according to Nayebi et al. [22], mostly motivated by unneeded functionality, poor user experience, and compatibility issues. Taylor and Martinovic [28] investigated the effects of software evolution considering security-related aspects and found out that Android apps are not getting safer as they are updated; rather, many app updates increase the number of vulnerabilities over time. Nayebi et al. [22] observed that too much functionality,

[1] Android's official app store.
[2] User reviews that comment on accessibility aspects of the evaluated app

generally the result of adding new features over time, can easily impact usability, resource consumption, and maintenance effort.

When it comes to mobile accessibility, Alshayban et al. [6] analyzed the difference of the inaccessibility rate[3] among subsequent versions of 60 mobile apps and concluded that 47% of the updates improved the app's overall accessibility while 28% of the updates rendered the app less accessible. Such results, however, cannot guarantee that the improvement observed is due to accessibility bug fixing or due to the inclusion of new interface elements with fewer accessibility problems, which would decrease the inaccessibility rate without removing any accessibility barrier. In addition, Chen et al. [9] evaluated the three last versions of 70 mobile apps and found out that the number of issues across different versions has not changed in 57 apps (82%). However, the number of issues has increased in 10 apps (14%) and decreased in 3 apps (4%).

Both studies related to accessibility evolution in mobile apps focus on specific metrics related to the number of issues found by automated tools. Furthermore, they have not investigated the specific changes in the mobile app's accessibility after the updates. In our study, instead of evaluating the (in)accessibility rate of subsequent versions of mobile apps, we investigate the impact of updates on mobile accessibility by looking at what users report when they perceive some change in the accessibility of the app they use.

3 STUDY DESIGN

The main purpose of this paper is to analyze user reviews to understand the impact of updates on the accessibility of mobile apps. This section describes the process we adopted to collect and extract information from accessibility update reviews. Figure 1 outlines the steps of our research method, namely: sampling, data analysis, prompt design, and data synthesis. Each step is detailed as follows.

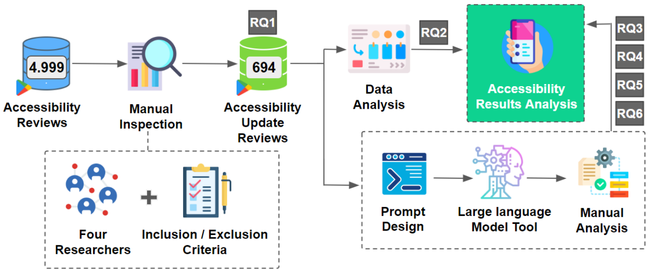

Figure 1: Overview of each step of our sampling and data analysis process.

3.1 Sampling

Our sampling process consisted of acquiring a dataset of accessibility reviews and identifying those associated with app updates.

3.1.1 Accessibility Reviews Dataset. In the first step, we resorted to a dataset of 4,999 accessibility reviews associated with visual disabilities and eye conditions published by Oliveira et al. [23]. This dataset was created based on string-matching filtering and manual inspection applied over a corpus of nearly 180 million user reviews extracted from the most popular apps from the Google Play Store. To create this initial corpus, they used an unofficial Google Play Store API[4] to extract all user reviews of the 10 most downloaded apps of each of the 34 categories of the Google Play Store.

The string-matching filtering applied over this corpus was based on more than 400 keywords associated with visual disabilities and eye conditions, resulting in more than 13,000 candidate reviews. Such reviews were manually inspected by five researchers that applied inclusion and exclusion criteria to keep only those reviews actually associated with visual disabilities and eye conditions, resulting in 4,999 accessibility reviews. All reviews have two labels: one with respect to the disability or eye condition associated with that review and one to define whether it is positive or negative feedback. Both the selection and the labeling processes are presented in detail in the study of Oliveira et al. [23].

The dataset created by Oliveira et al. [23] has 936 positive reviews (18.7%) and 4,063 negative reviews (81.3%). Positive reviews are a recognition that the app is accessible somehow while negative reviews are complaints or requests for missing accessibility features. Most apps have less than 50 reviews, while the top 20 most evaluated apps have 56% of them. The mean number of accessibility reviews per app is around 22, and the median is 6. Facebook is the most evaluated app with 376 accessibility reviews, followed by WhatsApp (233) and Amazon Kindle (203). Many apps received only one accessibility review, such as Alibaba, Sky Map, and Tripadvisor. For most apps, negative reviews are predominant. A detailed analysis of this dataset is presented by Oliveira et al. [23].

3.1.2 Accessibility reviews associated with app updates. In the second step, we conducted a manual inspection of the 4,999 accessibility reviews of this dataset to identify reviews that report changes in mobile app accessibility after some updates. The manual inspection was conducted by four researchers: one researcher analyzed all reviews while the others inspected one-third of the sample in a way that each review was inspected at least twice. To prevent exhaustion, the analysis process lasted 30 days. Each researcher had access to a subset of the reviews on a Google Spreadsheet.

For this task, we defined some inclusion and exclusion criteria. The inclusion criterion defines that we must keep in our sample only the reviews that report changes (both positive and negative) in the mobile accessibility after some update (e.g. *"Every time they get it right in terms of font size and accessibility they release an update that undoes all the good work. I am partially sighted and the latest update has almost rendered the app unusable." Facebook*).

The exclusion criterion defines that we must exclude reviews in three cases: i) the review is not related to app updates (*"Please add an option for colour blindness people. In lyrics some colors(Red) blinding my eyes please do something for people like me" Spotify*); ii) the user suggests some accessibility improvement in the next update (e.g. *"Kindly improve the accessibility of this application I am a visually impaired person it is very difficult to use this application I hope you will improve the accessibility in next update because I and my many friends prefer to use telegram instead any other platforms like WhatsApp. Please please do focus on accessibility." Telegram*); and iii) the review is associated with app updates and aspects of

[3] A metric calculated by dividing the number of elements with accessibility issues on a screen over the total number of elements on the same screen that are prone to accessibility issues.

[4] https://github.com/facundoolano/google-play-api

the app accessibility are mentioned, but the issue introduced by the update is not accessibility-related (e.g. *"Usually I love the app. Simple easy (needs a dark theme for us with light sensitivity) but after these last updates I can no longer sync anything. Says my last sync was Sept 8th. So I guess im goin back to my Gear Fit 2 smh." Fitbit*).

After the manual inspection, we cross-checked the findings of the manual classification. For every disagreement, a third researcher was requested to break the tie. In total, 100 reviews had to be further analyzed by other researchers. Consider, for example, the following user review for which we needed to reach a joint decision: *"Hellppp how to put back dark mode theme in huwaei phone? My pro30 lite just updated it system and the messages app lost it's dark theme."* (*Messages*). In this case, the third researcher decided to exclude the review because it mentions a system update that impacted a specific app, but it is not clear whether the app has also been updated.

For the purpose of determining the extent to which the raters agreed upon the classifications, we used Cohen's Kappa coefficient [11]. We acquired a degree of agreement of 0.98. According to Fleiss et al. [13], these agreement values are nearly *perfect agreement* (*i.e.*, 0.80 − 1.00). In that sense, the resultant sample, a set of 694 accessibility reviews associated with app updates, is the result of a process for which all researchers agreed on 100%.

3.2 Data analysis

We employed default quantitative analysis to answer *RQ1* and *RQ2* as they are both related to the frequency of specific types of content. For *RQ3* to *RQ6*, however, a qualitative approach must be employed. In previous work, data analysis consisted of quantifying the number of reviews in different categories based on automatic or manual labeling, such as type of feedback (e.g. bug report, feature request, complaint, compliment), type of disability (e.g. low vision, astigmatism, blindness), and associated accessibility guidelines (e.g. form, audio/video, link). In this work, we aim at identifying the main accessibility barriers and their consequences to the usability of the app and to the user experience.

Manually analyzing a large amount of data is labor-intensive. Therefore, while we understand that experienced researchers would provide us with higher-level categorization and further insights into this topic, we employed a Large Language Model (LLM) in this first study on the impact of app updates on accessibility to grasp general information we need to answer our research questions.

A LLM is a large neural network trained on massive amounts of text data to understand, summarize, classify, and generate new content [26]. Hence, an LLM can reason and answer questions about text data such as user reviews. Our main motivation to use LLMs in this study was inspired by the results of a recent investigation conducted by Byun et al. [8]. In their study, they used ChatGPT-3 to analyze text data (e.g. interview transcripts) that were available along with papers published in major Human-Computer Interaction conferences (e.g. CHI). Thus, they compared the results with qualitative analyzes conducted by human researchers. The authors concluded that, even though they were not able to accurately assess how well LLMs compare to human results because they are not experts in the many data domain they analyzed, the automated models were able to generate logical and convincing themes, discussions, and analysis of qualitative data arguably similar to those written by human researchers.

In our study, the analysis units (user reviews) are smaller than interview transcripts. In addition, our intention is to analyze user reviews to extract basic information (e.g. problems with fonts, color, and layout) that does not require deep reasoning, as most of the answers are contained in the review itself, except when some categorizations are required. Accordingly, we decided to adopt ChatGPT-4[5], an improved and paid version of ChatGPT-3, to analyze our dataset.

3.3 Prompt design

A prompt design is an input text or a query given to LLMs to guide the model in generating a desired and satisfactory answer to proposed inquiries. After a few iterations with ChatGPT-4, we realized that the same query applied to the same dataset could result in different granularities and styles of responses. Therefore, we designed a prompt to induce the model to produce answers in a more constrained structure.Following, we present our reasoning and the resultant prompt considering input data and each *RQ*.

Input data prompt. The goal is to provide ChatGPT-4 with accessibility reviews and ask it to get familiarized with the content by summarizing the data. Here is the input text we designed:

> I'm providing you with a CSV file that contains user reviews that mention accessibility aspects of mobile apps after some updates. The columns of this CSV file are identifier and user review. Here is the content of the CSV file. Please provide me with the key issues that users are consistently reporting. [CSV file content]

RQ3 prompt. We need to make sure ChatGPT-4 extracts all the accessibility problems mentioned by users, therefore we decided to outline several ways of referring to accessibility barriers. In addition, we would like ChatGPT-4 to propose some categorization to group the identified issues. Here is the input text we designed:

> Analyze the user reviews provided in the CSV file and make a list of the accessibility barriers, accessibility issues, accessibility bugs, or accessibility problems mentioned by users. Next, classify the items on the list into categories that emerge from the reviews or from other categories associated with the accessibility of a mobile app.

Here is an excerpt of ChatGPT-4 response for *RQ3* prompt[6]

> *After analyzing the reviews, the following accessibility barriers, issues, bugs, or problems have been identified:*
>
> - *Small font size*
> - *Incompatibility with device accessibility settings*
> - *Uncomfortable font style*
> - *Updates overriding previous accessibility settings*
> - *Poor readability in specific areas*

[5] https://openai.com/product/gpt-4

[6] This excerpt refers to answers provided by ChatGPT-4 considering a subset of our sample. The excerpt has been formatted and summarized to improve presentation as the model produces substantial volumes of written content.

RQ4 prompt. In opposition to *RQ3*, we would like ChatGPT-4 to identify accessibility improvements identified by users and group them into categories that emerge from the reviews themselves. Here is the input text we designed:

> Analyze the user reviews provided in the CSV file and make a list of the accessibility improvements or accessibility fixes. Next, classify the items on the list into categories that emerge from the reviews or from other categories associated with the accessibility of a mobile app.

Here is an excerpt of ChatGPT-4 response for *RQ4* prompt[7]:

> *Here are the accessibility improvements or fixes mentioned in the reviews:*
> - *Introduction of Dark Mode*
> - *Increased Font Sizes and easier-to-read font type*
> - *Improved color differentiation*
> - *Improvement in brightness control*

RQ5 prompt. We induced GPT-4 to express the consequences of using inaccessible apps as reported by users employing common expressions associated with inabilities, difficulties, troubles, and so forth. In addition, we asked GPT-4 to classify the results considering WCAG (Web Content Accessibility Guidelines) principles such as perception, understanding, and operation, in addition to any other category it could identify. Here is the input text we designed:

> Analyze the user reviews provided in the CSV file and make a list of the consequences mentioned by users, such as difficulties, inabilities, troubles, physical reactions, or other problems or situations they face due to the lack of accessibility of the mobile apps. Next, classify the items on the list into categories such as perception, understanding, operation, or any other category that emerges from the reviews.

Here is an excerpt of ChatGPT-4 response for *RQ5* prompt[8]:

> *These consequences can be classified into the following categories:*
> - *Perception Problems: color recognition issues, Eye strain, difficulty reading*
> - *Operation issues: navigation difficulties, inability to use in different light settings*
> - *Understanding Problems: unclear features due to the lack of contrast or distinguishable colors*
> - *Physical Reactions: vertigo or headaches, eye damage concern*

RQ6 prompt. We decided to provide the model with examples of sentiments the users might experience and suggest the classification into negative, positive, and neutral. Here is the input text we designed:

> Analyze the user reviews provided in the CSV file and make a list of the sentiments associated with the reviews. Examples of sentiments are: frustration, satisfaction, Disappointment, gratitude, and so forth. Next, classify each sentiment according to some categorization, such as negative, positive, and neutral.

Here is an excerpt of ChatGPT-4 response for *RQ6* prompt[9]:

> *Based on the reviews provided in the CSV file, here are the sentiments expressed by users:*
> - *Negative Sentiments: Frustration, Disappointment, Dissatisfaction, Anger*
> - *Positive Sentiments: Gratitude*
> - *Neutral Sentiments: Pleading (While this sentiment indicates dissatisfaction, it also reflects hope for improvement, making it somewhat neutral.)*

3.4 Data synthesis

The results we present in this paper are not the direct output produced by the LLM we used as we performed further analysis to synthesize the results produced by ChatGPT-4 due to many reasons. First, we could not submit our whole sample at once because ChatGPT-4 has an input limit. Thus, we submitted several subsets of accessibility update reviews to ChatGPT-4 separately and then we merged the results we obtained.

Second, we evaluated the categorization suggested by ChatGPT-4 to organize the results of the accessibility barriers identified (*RQ3*). As the accessibility reviews were not analyzed at once, ChatGPT-4 suggested many different categories or themes because they were based on different subsets of reviews. When we merged all the results, we noticed that ChatGPT-4 suggested almost 70 categories closely related to each other. For instance, the categories *Font-related Issues*, *Text Accessibility*, *Text Legibility Issues*, *Text Readability* and *Text Size Problems* are used to classify very similar accessibility issues. Therefore, we merged those categories that were closely related.

In addition, many categories suggested by ChatGPT-4 are based on specific resources, such as *Text* and *Color*, and many accessibility problems are related to more than one category. For instance, the inadequate contrast of a text can fall into many categories, such as *Text*, *Color*, *Brightness*, and even *Mode/theme*. Thus, we decided to summarize the categories into broader themes, namely: Assistive technology, Consistency, Compatibility, Customization options, Display and Visual Design, Input and Feedback, and Layout.

Furthermore, we validated the classification proposed by ChatGPT-4 of the consequences of using inaccessible apps (*RQ5*) considering the categories we proposed: perception, understanding, operation, and physical reactions. For each item classified by ChatGPT-4, we found a WCAG guideline associated with that particular issue. Accordingly, we verified whether the classification proposed by the tool was the same as the WCAG principle associated with that guideline. For instance, ChatGPT-4 classified the issue "Difficulty Reading or Seeing Content due to lack of contrast" into the *understanding* category. However, that accessibility issue is related to the guidelines 1.4.3 - Contrast (Minimum) and 1.4.6 - Contrast (Enhanced) of the *Perceivable* principle. Hence, we reclassified that consequence into the *perception* category. This validation process was carried out by two researchers in a joint effort, therefore there was no cross-validation to verify intercoder reliability in this case.

[7] See footnote 6

[8] See footnote 6

[9] See footnote 6

4 STUDY RESULTS

In this section, we answer our research questions and discuss the results of our investigation.

4.1 RQ$_1$: How many accessibility reviews are associated with app updates?

We manually inspected the dataset provided by Oliveira et al. [23] and we identified 694 accessibility reviews that report changes in the mobile app accessibility after some updates. Most reviews are negative feedback (93%), which means the app became less accessible after the update (e.g. *"Why did the update make the print so small!!! Seriously this is ridiculous. There are many people who have vision issues who need the option for larger text."Samsung Health app*). Only 7% of the reviews are positive feedback in which users find the new version of the app more accessible (e.g. *"The first and by far the best feature to me on this update is the color labels. Being color blind is now easier to deal with on Google calendar."Google Calendar*). The reviews we identified report accessibility changes in 107 distinct mobile apps.

Table 1 shows the top 10 apps that most received accessibility update reviews. The first column shows the name of the app, the second column shows the number of accessibility update reviews, the third column shows the number of general accessibility reviews of the original dataset [23], and the last column shows the ratio between columns two and three. Facebook is the app that most received accessibility update reviews with 119 reviews, followed by Gmail (57), WhatsApp (43), Google Chrome (39), and Twitter (30).

Table 1: The top 10 apps that most received accessibility reviews associated with app updates

App	Update Reviews	All Reviews	Proportion
Facebook	119	376	31,65%
Gmail	57	195	29,23%
WhatsApp	43	233	18,45%
Chrome	39	187	20,86%
Twitter	30	117	25,64%
Gboard	28	199	14,07%
Google	25	104	24,04%
Messenger	21	116	18,10%
Kindle	17	203	8,37%
YouTube	17	169	10,06%

Discussion. The number of accessibility reviews associated with app updates represents 14% of the original dataset, which is composed of 4999 accessibility reviews. Almost half of the apps of the dataset have at least one review that reports accessibility changes after some update (107 out of 228). For some of the apps, however, the number of accessibility update reviews may represent a larger proportion. The accessibility update reviews of Facebook, for instance, represents 31.5% of its accessibility reviews. Considering the whole dataset, there are some apps that only received accessibility update reviews, usually those apps with only one or two accessibility reviews, such as *Pregnancy App & Baby Tracker*, *theScore: Sports News & Scores* and *Google Fit: Activity Tracking*. For most apps, the number of accessibility update reviews goes up to 30%.

Most accessibility update reviews report a decline of the app accessibility after some update. This information alone is not evidence that all apps become less accessible as it evolves. There might be many accessibility improvements that are not reported by user reviews because people are more inclined to give negative than positive feedback [16]. However, the larger proportion of negative feedback is evidence that the evolution of mobile apps to accommodate user and environmental needs may negatively impact other quality aspects such as accessibility.

This result is a warning that one cannot assume that, over time, the software will become or remain accessible. If accessibility is not a requirement at the beginning of the process, it will not probably be during software evolution. Furthermore, even mobile apps that are concerned with accessibility should implement measures and adopt accessibility requirements during evolution to avoid introducing accessibility barriers.

4.2 RQ$_2$: What are the interface components and resources mentioned in the accessibility update reviews?

Table 2 shows the top 10 interface resources or components that are mostly mentioned in accessibility update reviews. We identified the interface components and resources for each review employing a string-matching process based on keywords extracted from the design foundations and guidelines presented in the Google Material Design[10] and in the BBC Mobile Accessibility Guidelines[11]. COLOR is by far the most comment resource (332), followed by MODE (205), SCREEN READER (96), and FONT (84).

Most reviews refer to one (33.5%) or two (34.5%) resources. Most reviews that mention only one resource are associated with COLOR, SCREEN READER, and FONT. Reviews that mention two resources are mostly related to COLOR and MODE, and BACKGROUND and COLOR. We found out that 81 reviews (11.5%) do not specify any resource or component (e.g. *"I love the app but with each update, it becomes less accessible to blind users." Pinterest*).

Table 2: Top 10 interface resources or components that are mostly mentioned in accessibility update reviews

Resource/Component	Accessibility Reviews	Proportion
COLOR	332	48%
MODE	205	29.5%
SCREEN READER	96	14%
FONT	84	12%
BACKGROUND	55	8%
SETTING	48	7%
TEXT	45	6,5%
BUTTON	42	6%
LAYOUT	16	2%
MENU	14	2%

Discussion. Noticeably, updates clashing with the color scheme (or color mode/theme) of mobile apps are the most commonly reported issue. In fact, this problem is connected to many resources

[10] https://m3.material.io/components
[11] https://www.bbc.co.uk/accessibility/forproducts/guides/mobile/

outlined in Table 2, such as COLOR, MODE, BACKGROUND, and SETTING. One frequent complaint among user reviews is that dark mode has been removed in the new version of the app (e.g. *"Worst update ever they just removed dark mode. My eyes are blind now thanks a lot facebook." Facebook*).

Many users also denounce that the update has rendered the app inaccessible with assistive technologies, such as screen readers (e.g. *"After this update I'm not able to book can it was working fine till last update I can't confirm ride with TalkBack fix it" Uber*). Furthermore, another common criticism is that the font size became smaller in the new version or that users are no longer able to change font settings (e.g. *"The new font is terrible. It's hard to read and gives me eye strain. For some reason it overrides my phone's font settings. That's really bad for accessibility. Please give us the option to disable it and use our system font." Twitter*). A more detailed account of the accessibility issues mentioned by users is discussed in Section 4.3

The fact that most reviews are associated with one or two resources seems to indicate that updates have affected a specific aspect of the accessibility of the app or that the user is focused on denouncing the most relevant issue. Either way, many specific issues affect the overall usability of the app. On the other hand, many users are not concerned with providing developers and organizations with specific details of the problems introduced by updates. In many cases, the reviews only state that the app became more or less accessible (e.g. *As someone with a vision impairment this new update is inaccessible and disappointing. At least give the option of reverting to the old look like you do with Gmail Messages*).

Identifying the resources that are mostly associated with accessibility problems introduced by app updates is relevant because it indicates that developers should be very careful with them during app evolution. Changing color schemes, and fonts or removing customizing options should not be treated lightly, which reinforces the need for policies and procedures to make sure that accessibility requirements are part of the trade-offs and impact analysis usually conducted before implementing software updates.

4.3 RQ$_3$: What are the main issues reported in the accessibility update reviews?

The analysis of the accessibility update reviews shows that users consistently report many accessibility problems introduced in the new releases of mobile applications. Many accessibility update reviews, however, are not specific, as the users only comment that the app is less accessible than the previous version. When it comes to more specific reviews, most issues are related to the removal of accessibility-related features (e.g. dark mode, font or color settings, text-to-speech), changes in layout that makes understanding and navigating the app harder, and lack of support for assistive technologies or alternative input devices. Furthermore, users frequently ask to have the option to revert to a previous version of the app that they found more accessible and user-friendly.

Table 3 outline many of the accessibility problems detected and reported by users according to the categories we propose (cf. Section 3). Following we briefly describe each category:

Assistive technology. Accessibility barriers in this category are mostly related to the lack of proper labels or because the app behaves abnormally when screen readers are in use.

Table 3: Accessibility issues perceived by users after updates

Categories	Main issues perceived after update
Assistive technology	Buttons missing labels Talkback stopped reading out emojis Images missing descriptions Issues with scrolling using screen readers Talkback cannot read links/input field Talkback reads out sensitive information Lack of Text-to-Speech Support Text-to-speech cannot read video titles
Consistency	No dark mode in some Android versions No dark mode depending on the app version No dark mode only in some areas of the app Undeclared changes in design or layout Inconsistent updates across platforms Voice input feature requires a double tap Changes in color coding
Compatibility	Font does not follow the device settings Mobile app overrides device font settings Small font size even at max setting App is crashing/lagging with Talkback
Customization options	Removal of font size and type settings No option to change the font style No option to configure the color scheme No option to personalize layout/themes No option to configure notification/display No option to set line or character spacing No option to personalize shortcuts
Display/Visual Design	Low color contrast in different elements Colors are not color-blind friendly High saturation colors Inadequate background color Inadequate auto-brightness feature Absence of Dark Mode Inadequate color choices/combinations Unwanted automatic brightness increase Small text size and light text color Inadequate font style Removal of text reflow (Word Wrap) Inadequate design of widgets (size, font)
Input and feedback	Talkback announcing incorrectly typed words The new keyboard is not user-friendly Inaccessibility of password input Keyboard layout changes (smaller keys) Lack of Tactile Feedback Malfunction of voice input features Removal of the microphone button Voice recognition does not work properly
Layout	Lack of distinguishable colors, lines, borders, or hierarchy to separate and distinct elements Icons/buttons moved to inconvenient places Removal of cascading tabs Tiny navigation buttons Overlap of the app on buttons to answer calls Removal of readable direction arrow in map

Consistency. Accessibility problems in this category are related to the fact that users expect certain features and settings to remain consistent, and sudden changes can cause confusion and discomfort.

Compatibility. Issues in this category are related to incompatibility between mobile apps and environment (e.g. Android).

Customization options. Accessibility barriers in this category refer to the lack of features to customize the appearance of the apps according to the user's preferences or needs, including the ability to change back to previous versions.

Display and Visual Design. Accessibility issues within this category are predominant as they are associated with visual aspects of the app, such as color schemes and font size.

Input and feedback. Problems in this category are related to the lack of support or availability of suitable input or feedback features.

Layout. Issues within this category refer to changes to the layout and interface that made the apps harder to understand and navigate.

In addition to accessibility issues related to the user interface, some users complain that, despite multiple updates, previously reported accessibility problems are not fixed. In that sense, users also say it appears that new versions of the app have not been adequately tested for accessibility before release. In addition, a few users express frustration about not being able to get help when they experience problems due to their visual impairments.

Discussion. The small number of accessibility update reviews was enough to identify numerous distinct characteristics of interface elements that changed after some updates and rendered them inaccessible. It seems that most of the reported issues could have been avoided once users considered the app accessible to some degree before the new release. It means that, at some point, accessibility requirements were considered during software development; however, over time, accessibility was overlooked during app evolution.

For instance, many users complain that accessibility features or facilities were removed after some update, like the user's ability to change font styles and colors. Removing such a feature must be part of an evolution plan rather than an arbitrary barrier introduced by changing a single property of an interface component (e.g. button label). In addition, users also denounce that many apps are no longer complying with accessibility settings (e.g. font size, color contrast) defined by the device or operating system, which also characterizes a deliberated decision to keep interface elements unchanged and therefore not adaptable to different user needs.

In that sense, considering that small changes can cause practical, physical, and emotional consequences for users (cf. Sections 4.5 and 4.6), it is clear that developers and organizations must adopt countermeasures to avoid the decline of the overall accessibility of their apps as new functionalities are implemented.

4.4 RQ$_4$: What are the accessibility improvements reported in the accessibility update reviews?

A small subset of accessibility update reviews emphasizes that the app accessibility has improved after some updates. Considering that the list of improvements is far less than the accessibility retrogression, we have not grouped them into categories. Here are instances of accessibility improvements mentioned by users:

- Ability to change the background color and font size
- Better compatibility with Assistive Technologies
- Introduction of Dark Mode
- Increased Font Sizes and easier-to-read font types
- Improved color differentiation and color labels options
- Improvement in brightness control
- More adaptable layout and interface

Discussion. As expected, most accessibility improvements reported by users are quite the opposite of the general complaints listed in Section 4.3. Note that the most commented aspects are also related to color and brightness; hence, developers should pay special attention to this particular aspect of the interface during the development and evolution of their mobile apps. It is important to emphasize that users also recognize the developers' and organizations' efforts to make apps more accessible.

The fact that the majority of accessibility update reviews are negative feedback does not necessarily mean that most apps are becoming less accessible because people are more likely to give negative than positive feedback [16]. In addition, most reviews of the sample from which we identified accessibility update reviews are negative comments on the app accessibility.

4.5 RQ$_5$: What are the main consequences for the users reported in the accessibility update reviews?

The analysis of the accessibility update reviews shows that the many reported accessibility problems have direct consequences for users. The most commonly reported consequences involve difficulties in using applications, reading content, viewing videos and subtitles, and utilizing screen readers. In more specific cases, users have reported being unable to complete operations, such as purchasing air tickets, products in online stores, or even ordering a meal for delivery. Users have also mentioned physical impacts, such as eye strain and headaches. Most consequences come from problems related to font size and settings, color settings, including dark mode absence, color background, brightness, and changes in the interface.

Table 4 shows the main consequences detected and reported by users according to the categories we proposed (cf. Section 3). For each category, we refer to both generic and app-specific problems reported by users. Following we briefly describe each category:

Perception. Issues in this category involve the user's abilities to perceive information and interface components (e.g., *"This used to be a good app. After the last update they made the text so small that it's almost impossible to read for people with vision problems. There's also no way to make the text bigger." Amazon Prime Video*). The main consequences related to perception involve difficulty in recognizing and interpreting sensory information and the user's ability to perceive the information and interface components presented. Some users reported difficulties in distinguishing elements, inability to use the app at night, and reading difficulties.

Understanding. Problems in this category involve the user's ability to understand and process the information being presented (e.g., *"Before the last update everything was well if not completely accessible for the blind the majority was. Now its not. FIX IT PLEASE and I'll give it 5 stars again. None of the buttons are labeled." Pandora - Music & Podcasts*). This involves cognitive load or difficulty in understanding the information due to a visual impairment or eye condition.

Table 4: Consequences of the accessibility issues introduced by app updates

Category	Consequences
Perception	Reading content (emails, social media) Distinguishing location flags/arrow) Distinguishing roads in locomotion apps Recognizing lyrics in descriptions Seeing content Following color schemes Accessing info using screen reader
Understanding	New updates/layout changes Button functions due to poor labeling Various functions and unclear features Specific not described sections If words were typed correctly Order of notifications
Operation	Interacting with Multimedia Content Editing, receiving, or sending texts Using magnification features Adding items to a cart Booking a tax/ride or meals Commenting/Liking posts Applying promo codes Downloading content Watching videos Playing music in streaming apps Deleting multiple emails at once Reading e-books Scrolling with screen readers Navigating tabs/menus w. screen readers Inputting passwords w. screen readers
Physical Reactions	Blurred vision and eye strain Physical discomfort Headaches, migraines, and vertigo Physical pain Potential long-term eye damage

Operation. Consequences in this category involve the user's ability to operate or interact with the app effectively (e.g., *"Highly inaccessible application especially after recent update visually impaired customers cannot book meal on their own option of incriment /decriment quantity has disappeared and developpers are not paying attention to genuine concerns of visually impaired persons." Swiggy*). The main consequences related to operation involve the difficulty or the inability to perform specific tasks.

Physical Reactions. This category encompasses physical discomfort or adverse reactions experienced by users due to the lack of accessibility (e.g., *"The new color change needs an off switch. I'm sure it's very helpful to some as it was meant to be. But others including myself now are dealing with eyestrain and headaches after just minutes of being on tumblr. Please give us options so no one has to struggle to look at this site." Tumblr*). The main consequences involve visual discomfort due to brightness or lack of Dark Mode.

Discussion. We believe that showing the problems users have to interact with specific apps to complete specific tasks emphasizes the importance of creating or maintaining accessible apps. The consequences range from minor setbacks in perceiving some content to the inability to effectively complete a task or feeling physical or emotional pain. It is also clear that most accessibility barriers fit under one of the three first principles defined by WCAG (perceivable, understandable, and operable) and some of them may have an impact in all categories (e.g., missing or inadequate labels). In addition, some consequences may also be related to the robustness principle once it is associated with compatibility with user agents such as assistive technologies.

Table 5: Sentiments associated with accessibility user reviews

Category	Sentiment	Reason/Circumstance
Negative	Anger	Feels discriminated against
	Annoyance	Bugs/glitches/lack of accessibility
	Concern	Eye damage due to bright mode
	Confusion	New features or layout change
	Despair	Unusable app
	Discontent	Inaccessibility
	Fear/hesitation	Changing complex settings
	Frustration	Cannot access certain features
	Regret	Upgrading to a new version
	Sadness	Unable to use apps effectively
Positive	Appreciation	Accessibility feature available
	Excitement	App became more usable
	Gratitude	Accessibility improvement
	Hope	Requests will be answered
	Pleasure	Accessibility improvement
	Relief	Dark mode is back
	Satisfaction	Features started working
Neutral	Acceptance	Reverts to older versions
	Indifference	Reports issues without emotions
	Pleading	Asks for improvements

4.6 RQ$_6$: What sentiments are expressed in the accessibility update reviews?

Sentiment analysis is commonly used to comprehend what customers feel about products. In this study, we report on user sentiments concerning updates that affect app accessibility. We present the most common sentiments inferred or expressed by users considering three categories: positive, negative, and neutral sentiments. Table 5 shows the sentiment in each category and the common circumstances or reasons behind that sentiment. Following we explain each category of sentiment we defined:

Negative. These sentiments were often associated with updates that made the apps less accessible or disregarded user needs and preferences. They showcase users' negative reactions to the lack of accessibility in the apps and indicate areas where improvements are needed (e.g., *"I can no longer use the app because I've had vision loss."*, *"I HATE the new background color."*).

Positive. These sentiments reflect the appreciation for the efforts made by developers and their hopes for future improvements. Positive sentiment was expressed, for example, regarding a feature (dark mode) that enhanced the app's accessibility (e.g., *"Thank u so much for making such an accessible app!"* and *"God bless u!"*).

Neutral. In a broader context, a neutral sentiment might be expressed if a user feels indifferent to the changes or if the changes do not significantly impact their user experience (e.g., *"Please undo this update and go back to what it used to be please."*).

Discussion. In the context of using a mobile application, where the user aims to successfully perform tasks such as reading a message, sending an email, or ordering a meal, the sentiments expressed at the end of these tasks are fast and instantaneous reactions that can arise from the success or failure of the operation. As noticed in previous work, user reviews related to digital accessibility tend to be more negative than positive, resulting in a greater identification of negative sentiments expressed by users.

It is noticeable that these sentiments, especially the negative ones, are reactive and lack thoughtful filtering. Some are even expressed with capitalized words (e.g., *"I HATE the new background color."*) and excessive exclamation points (e.g., *"every update getting worse!!!!!! understanding!? to update means to do better not the opposite."*).

In general, positive sentiments are expressed when accessibility is improved or when the user can successfully utilize accessibility features, such as screen readers. On the other hand, neutral sentiments merely indicate the user's need to express accessibility issues without reporting specific consequences or seeking improvements. Furthermore, neutral sentiments also reflect users' resilience as they wait for potential enhancements in the app's accessibility.

5 THREATS TO VALIDITY

This section presents the threats to the validity of our study.

Sampling bias. Our study inherits the sampling bias of the study that produced the dataset of accessibility reviews we used [23]. In short, the dataset is the result of string-matching filtering followed by manual inspection, but specific measures were taken to mitigate bias by carefully selecting keywords based on official glossaries and having each review inspected by at least two researchers. Our study also introduces some bias in the process of narrowing down the initial dataset from 4,999 accessibility reviews to 694 accessibility update reviews. To mitigate this threat, each review was inspected by at least two reviewers to decide whether it was related to the impact of app updates on app accessibility. All disagreements were resolved by a joint decision of three or four researchers.

Researcher bias. The researcher's knowledge or assumptions of the knowledge domain may influence the results. In our study, we need to consider the bias introduced by ChatGPT-4 as it was used to extract and categorize the information from the accessibility update reviews. To mitigate this threat, we designed a prompt to induce the model to produce outputs following a certain template or pattern. In addition, we do not believe the bias introduced by ChatGPT-4 invalidates or weakens our results because user reviews consist of small and simple phrases containing specific information regarding app functionality, interface elements, and accessibility. Therefore, we believe the model was able to accurately extract the information we needed from each review. In any case, even if ChatGPT-4 could not extract information from each and every review of our sample, our study is not intended to identify and quantify all instances of accessibility issues reported by users; rather, we aimed at providing a general overview of the consequences of updates to the app accessibility.

When it comes to the classification of the accessibility issues and the consequences reported by users in the accessibility update reviews, we mitigated the threat by having human researchers analyze, reclassify, and synthesize the information produced by ChatGPT-4 (cf. Section 3.4). To mitigate the human research bias, we also had more than one researcher analyzing all the data to jointly classify and organize all the information extracted from the accessibility update reviews.

6 CONCLUDING REMARKS AND FUTURE WORK

This paper shows evidence that app evolution impacts mobile accessibility. In particular, we listed many barriers and improvements introduced by app updates, as well the practical consequences and user sentiments associated with the accessibility update reviews we analyzed. Most reviews were negative feedback reporting that the new version of the app is no longer accessible; few reviews reported accessibility improvements. The accessibility decline reported by users may have a severe impact on the users' perception, understanding, and operation of mobile apps, which sometimes may lead to physical pain, negative sentiment, and emotional states. To the best of our knowledge, this is the first investigation of mobile app accessibility evolution from the user's perspective.

It is important to make it clear that we are not claiming that an intrinsic characteristic of mobile apps such as the evolution *per se* is harmful to accessibility. We are also not claiming that apps tend to become less accessible over time. In fact, there are no conclusive studies about this subject once the related work conducted small studies with 60 or 70 apps and reached different results [6, 9].

Our aim is to emphasize and give detailed evidences that, once some level of accessibility has been achieved at any stage of development, we should not presume the apps will remain accessible as they evolve. In that scenario, it is understandable the frustration of many users that, having found an accessible app in a scenario in which accessibility is scarce due to many reasons [6, 7, 19, 29], they realize it is no longer accessible after some update. Hence, unless accessibility quality goals are intentionally defined for the evolution process, modifications will probably introduce barriers that will affect the experience, especially for users with disabilities.

In future work, we intend to perform manual content analysis to extract more detailed information from accessibility update reviews and eventually compare the results that we have achieved by using ChatGPT-4. In addition, we intend to provide separate analyzes of the impact of app updates considering each interface component and resource and to link each accessibility barrier or improvement to specific accessibility guidelines (e.g., WCAG).

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) and the São Paulo Research Foundation (FAPESP - grant 2022/10910-3).

REFERENCES

[1] Wajdi Aljedaani, Mohammed Alkahtani, Stephanie Ludi, Mohamed Wiem Mkaouer, Marcelo M. Eler, Marouane Kessentini, and Ali Ouni. 2023. The State of Accessibility in Blackboard: Survey and User Reviews Case Study. In *Proceedings of the 20th International Web for All Conference* (Austin, TX, USA) *(W4A '23)*. Association for Computing Machinery, New York, NY, USA, 84–95. https://doi.org/10.1145/3587281.3587291

[2] Wajdi Aljedaani, Mohamed Wiem Mkaouer, Stephanie Ludi, and Yasir Javed. 2022. Automatic Classification of Accessibility User Reviews in Android Apps. In *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)*. IEEE, 133–138.

[3] Wajdi Aljedaani, Mohamed Wiem Mkaouer, Stephanie Ludi, Ali Ouni, and Ilyes Jenhani. 2022. On the identification of accessibility bug reports in open source systems. In *Proceedings of the 19th International Web for All Conference*. 1–11.

[4] Wajdi Aljedaani, Furqan Rustam, Stephanie Ludi, Ali Ouni, and Mohamed Wiem Mkaouer. 2021. Learning sentiment analysis for accessibility user reviews. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*. IEEE, 239–246.

[5] Eman Abdullah AlOmar, Wajdi Aljedaani, Murtaza Tamjeed, Mohamed Wiem Mkaouer, and Yasmine N El-Glaly. 2021. Finding the needle in a haystack: On the automatic identification of accessibility user reviews. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–15.

[6] Abdulaziz Alshayban, Iftekhar Ahmed, and Sam Malek. 2020. Accessibility issues in android apps: state of affairs, sentiments, and ways forward. In *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*. IEEE, 1323–1334.

[7] Tingting Bi, Xin Xia, David Lo, John Grundy, Thomas Zimmermann, and Denae Ford. 2021. Accessibility in software practice: A practitioner’s perspective. *ACM Transactions on Software Engineering and Methodology* (2021).

[8] Courtni Byun, Piper Vasicek, and Kevin Seppi. 2023. Dispensing with Humans in Human-Computer Interaction Research. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA ’23)*. Association for Computing Machinery, New York, NY, USA, Article 413, 26 pages. https://doi.org/10.1145/3544549.3582749

[9] Sen Chen, Chunyang Chen, Lingling Fan, Mingming Fan, Xian Zhan, and Yang Liu. 2022. Accessible or Not? An Empirical Investigation of Android App Accessibility. *IEEE Transactions on Software Engineering* 48, 10 (2022), 3954–3968. https://doi.org/10.1109/TSE.2021.3108162

[10] Adelina Ciurumelea, Andreas Schaufelbühl, Sebastiano Panichella, and Harald C Gall. 2017. Analyzing reviews and code of mobile apps for better release planning. In *2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 91–102.

[11] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[12] Marcelo Medeiros Eler, Leandro Orlandin, and Alberto Dumont Alves Oliveira. 2019. Do Android app users care about accessibility? an analysis of user reviews on the Google play store. In *Proceedings of the 18th Brazilian Symposium on Human Factors in Computing Systems*. 1–11.

[13] Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions* 2, 212-236 (1981), 22–23.

[14] Jun Gao, Li Li, Tegawendé F. Bissyandé, and Jacques Klein. 2019. On the Evolution of Mobile App Complexity. In *2019 24th International Conference on Engineering of Complex Computer Systems (ICECCS)*. 200–209. https://doi.org/10.1109/ICECCS.2019.00029

[15] C. Iacob and R. Harrison. 2013. Retrieving and analyzing mobile apps feature requests from online reviews. In *2013 10th Working Conference on Mining Software Repositories (MSR)*. 41–44. https://doi.org/10.1109/MSR.2013.6624001

[16] Frederic B. Kraft and Charles L. Martin. 2001. Customer Compliments as More than Complementary Feedback. *Journal of Consumer Satisfaction, Dissatisfaction & Complaining Behavior* 14 (March 2001), 1–13. https://www.jcsdcb.com/index.php/JCSDCB/article/view/101

[17] M.M. Lehman. 1980. Programs, life cycles, and laws of software evolution. *Proc. IEEE* 68, 9 (1980), 1060–1076. https://doi.org/10.1109/PROC.1980.11805

[18] M. M. Lehman. 1996. Laws of software evolution revisited. In *Software Process Technology*, Carlo Montangero (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 108–124.

[19] Manoel Victor Rodrigues Leite, Lilian Passos Scatalon, André Pimenta Freire, and Marcelo Medeiros Eler. 2021. Accessibility in the mobile development industry in Brazil: Awareness, knowledge, adoption, motivations and barriers. *Journal of Systems and Software* 177 (2021), 110942.

[20] Tianyang Liu, Chong Wang, Kun Huang, Peng Liang, Beiqi Zhang, Maya Daneva, and Marten van Sinderen. 2023. RoseMatcher: Identifying the impact of user reviews on app updates. *Information and Software Technology* (2023), 107261. https://doi.org/10.1016/j.infsof.2023.107261

[21] Stuart McIlroy, Nasir Ali, and Ahmed E. Hassan. 2016. Fresh apps: an empirical study of frequently-updated mobile apps in the Google play store. *Empirical Software Engineering* 21, 3 (June 2016), 1346–1370. https://doi.org/10.1007/s10664-015-9388-2

[22] Maleknaz Nayebi, Konstantin Kuznetsov, Paul Chen, Andreas Zeller, and Guenther Ruhe. 2018. Anatomy of Functionality Deletion: An Exploratory Study on Mobile Apps. In *Proceedings of the 15th International Conference on Mining Software Repositories* (Gothenburg, Sweden) *(MSR ’18)*. Association for Computing Machinery, New York, NY, USA, 243–253. https://doi.org/10.1145/3196398.3196410

[23] Alberto Dumont Alves Oliveira, Paulo Sérgio Henrique Dos Santos, Wilson Estécio Marcílio Júnior, Wajdi M Aljedaani, Danilo Medeiros Eler, and Marcelo Medeiros Eler. 2023. Analyzing Accessibility Reviews Associated with Visual Disabilities or Eye Conditions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI ’23)*. Association for Computing Machinery, New York, NY, USA, Article 37, 14 pages. https://doi.org/10.1145/3544548.3581315

[24] Fabio Palomba, Mario Linares-Vásquez, Gabriele Bavota, Rocco Oliveto, Massimiliano Di Penta, Denys Poshyvanyk, and Andrea De Lucia. 2018. Crowdsourcing user reviews to support the evolution of mobile apps. *Journal of Systems and Software* 137 (2018), 143 – 162. https://doi.org/10.1016/j.jss.2017.11.043

[25] Rahul Potharaju, Mizanur Rahman, and Bogdan Carbunar. 2017. A Longitudinal Study of Google Play. *IEEE Transactions on Computational Social Systems* PP (08 2017), 1–15. https://doi.org/10.1109/TCSS.2017.2732167

[26] Emma Sabzalieva and Arianna Valentini. 2023. ChatGPT and artificial intelligence in higher education: quick start guide. *UNESCO International Institute for Higher Education in Latin America and the Caribbean* 5 (April 2023). https://unesdoc.unesco.org/ark:/48223/pf0000385146

[27] Mara Taynar Santiago and Anna Beatriz Marques. 2022. Are User Reviews Useful for Identifying Accessibility Issues That Autistic Users Face? An Exploratory Study. In *Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems* (Diamantina, Brazil) *(IHC ’22)*. Association for Computing Machinery, New York, NY, USA, Article 6, 11 pages. https://doi.org/10.1145/3554364.3559114

[28] Vincent F. Taylor and Ivan Martinovic. 2017. To Update or Not to Update: Insights From a Two-Year Study of Android App Evolution. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security* (Abu Dhabi, United Arab Emirates) *(ASIA CCS ’17)*. Association for Computing Machinery, New York, NY, USA, 45–57. https://doi.org/10.1145/3052973.3052990

[29] Shunguo Yan and PG Ramachandran. 2019. The current status of accessibility in mobile apps. *ACM Transactions on Accessible Computing (TACCESS)* 12, 1 (2019), 1–31.

[30] Aidan Z. H. Yang, Safwat Hassan, Ying Zou, and Ahmed E. Hassan. 2022. An empirical study on release notes patterns of popular apps in the Google Play Store. *Empirical Software Engineering* 27, 2 (March 2022), 55. https://doi.org/10.1007/s10664-021-10086-2