Journal of the Brazilian Computer Society, 2024, 30:1, doi: 10.5753/jbcs.2024.4321

Exploring the Influence of Software Evolution on Mobile App Accessibility: Insights from User Reviews

Alberto Dumont Alves Oliveira [**University of São Paulo** | ***albertodumont@usp.br***]
Paulo Sergio Henrique dos Santos [**University of São Paulo** | ***pauloshsantos@usp.br***]
Wajdi Aljedaani [**University of North Texas** | ***wajdialjedaani@my.unt.edu***]
Marcelo Medeiros Eler [**University of São Paulo** | ***marceloeler@usp.br***]

School of Arts, Sciences and Humanities - University of São Paulo. Rua Arlindo Béttio, 1000, Ermelino Matarazzo, São Paulo, SP, 03828-000, Brazil.

Received: 28 March 2024 • **Accepted:** 18 September 2024 • **Published:** 02 November 2024

Abstract Software systems constantly evolve to accommodate stakeholders' and environments' requirements that change over time. In this process, the frequent modifications can increase software complexity and negatively impact its global quality when conducted in an unstructured way. The evolutive nature of mobile environments led researchers to investigate how mobile app evolution impacts complexity, security, resource consumption, maintainability, usability, and accessibility. In particular, there has been limited research on the impact of app updates on mobile accessibility: most studies focused on tracking the number of accessibility violations found by automated tools across successive versions of a small set of applications. In a previous work, we made a contribution to this field by identifying accessibility reviews associated with app updates and prompting ChatGPT-4 to provide an overview of the main accessibility issues and enhancements perceived by users in the new releases of a mobile app. In this manuscript, we extend our previous work by adopting manual content analysis to delve deeper into our research questions and by adding new research questions associated with the identification of reviews linked to app updates, user characteristics, WCAG principles and guidelines, and user demands reported in accessibility reviews. Our results show that the accessibility barriers reported by users are mostly linked to the WCAG 2.2 Perceivable principle, and the Distinguishable and Adaptable guidelines, which includes poor color scheme, small font size, unlabeled elements, and lack of customization options. Accordingly, the consequences of the lack of accessibility is mainly connected to the difficult users experience to perceive elements of the interface (e.g. difficult to read and distinguish content, watch videos) and to use screen readers, in addition to feel discriminated against. The most common demand developers and organizations receive is to bring back some accessible feature or to fix accessibility bugs.

Keywords: Accessibility, mobile, app, update, evolution, user reviews.

1 Introduction

Effective software systems are inherently designed to evolve and adapt to changing requirements and environments, thereby avoiding to become less useful overtime [Lehman, 1980, 1996]. In the mobile development landscape, organizations regularly releases updated versions of their apps to address the evolving preferences of their users and to maintain a competitive edge in the market [Gao *et al*., 2019; Potharaju *et al*., 2017; McIlroy *et al*., 2016; Yang *et al*., 2022]. Numerous studies indicate that the evolution of mobile apps is largely influenced by user feedback. Alongside planned surveys, organizations also harness spontaneous feedback provided by users, including reviews posted in app stores [Gao *et al*., 2019; Palomba *et al*., 2018; Ciurumelea *et al*., 2017].

Usually, mobile apps evolve due to many reasons [Gao *et al*., 2019; Iacob and Harrison, 2013; Palomba *et al*., 2018; Ciurumelea *et al*., 2017]: 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. While evolution predominantly aims at improving software products and enhancing user satisfaction, there are inherent risks associated with continuous modifications in software: the complexity and interdependence among system elements can escalate in an unstructured manner, potentially leading to a decline in overall quality unless proactive measures are taken to prevent or mitigate such issues [Lehman, 1980, 1996].

In this context, researchers have evaluated the impact of app evolution considering different quality aspects, including security [Taylor and Martinovic, 2017], complexity [Gao *et al*., 2019], usability, resource consumption, maintenance effort [Nayebi *et al*., 2018], and accessibility [Alshayban *et al*., 2020; Chen *et al*., 2022]. In particular, there has been limited research on the impact of app updates on mobile accessibility. Alshayban *et al*. [2020] and Chen *et al*. [2022] analyzed subsequent versions of 60 and 70 mobile apps, respectively, to assess whether their accessibility had improved or declined based on metrics associated with accessibility issues they identified. However, they did not further investigate the specific changes and their impact to their users.

Therefore, in a previous study, we conducted an investigation on the impact of app updates on mobile accessibility ac-

cording to the users' perspective [Dos Santos *et al*., 2023]. In such a study, instead of evaluating the (in)accessibility rate of subsequent versions of mobile apps, we examined user reviews associated with app updates and accessibility published in the Google Play Store, Android's official app store. More specifically, we prompted ChatGPT-4[1] to analyze the user reviews and to provide us with an overview of the: i) main accessibility issues caused by the update; ii) main accessibility improvements brought by the update; iii) practical consequences of the lack of accessibility caused by the update; iv) users sentiments towards the accessibility changes perceived due to the update. To facilitate reference to accessibility reviews linked with app updates and visual disabilities or eye conditions, we call them *accessibility update reviews*.

In this manuscript, we further investigate the impact of evolution on mobile app accessibility and extend our previous work published in the Brazilian Symposium on Human Factors in Computing Systems (HCI 2023) [Dos Santos *et al*., 2023] in many ways. First, instead of relying on ChatGPT-4 to analyze accessibility update reviews and offer an overview for answering our previous and new research questions, we opted for manual content analysis to explore our inquiries more thoroughly. More specifically, we recruited three researchers to conduct a manual closed coding process on a total of 694 accessibility update reviews obtained from the previous work. Subsequently, each user review was categorized based on the WCAG 2.2 (Web Content Accessibility Guidelines)[W3C, 2023], incorporating both quantitative and qualitative analysis. Second, we included four new research questions to explore, in addition to accessibility issues, enhancements and practical implications covered in the previous study, the following: i) terms and expressions commonly used to indicate perceived app updates; ii) personal context and user characteristics; iii) WCAG principles and guidelines linked to the user reviews; and iv) general user demands for developers and organizations.

Accordingly, we framed our investigation in seven research questions, of which RQ_1, RQ_2, RQ_3, RQ_7 were introduced in this extended version of our previous work. Notably, RQ_4, RQ_5 and RQ_6, which were partially addressed in the previous work, were answered in this manuscript based on manual content analysis, resulting in different categorizations, listings, and examples of user reviews.

RQ_1: What are the terms and expressions commonly employed by users to indicate that something has changed following an update? With this RQ, we intend to create a set of keywords or expressions that convey that something in the current version of the app has changed compared to its previous release. This can be useful as a training dataset to identify reviews associated with updates in upcoming studies.

RQ_2: What are the personal context elements mentioned by users in accessibility update reviews? With this RQ, we intend to characterize the message sender with respect to its physical traits and usage context. Such knowledge can help to identify user profiles and they probable impact on user reviews.

RQ_3: What are the WCAG 2.2 principles and guidelines associated with the accessibility update reviews? With this RQ, we intend to identify which accessibility guidelines are potentially neglected by developers during software evolution and use them as underlying themes to analyze the results of our study.

RQ_4: What are the accessibility problems reported in the accessibility update reviews? With this RQ, 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). In previous work, we categorized accessibility issues based on themes emerging from user reviews and analysis using a ChatGPT-4 prompt. Here, we highlight these issues using manual content analysis, focusing on the frequency in accessibility update reviews related to interface elements.

RQ_5: What accessibility enhancements are mentioned in the accessibility update reviews? With this RQ, we intend to show whether users recognize any accessibility improvement associated with app evolution and which enhancements were identified. This knowledge can be useful to give evidence that accessibility can and should be improved over time, and that users acknowledge those enhancements in mobile apps. In prior research, only seven instances of accessibility improvements were identified, and the subsection provided very brief coverage without any examples from the reviews. Now, we present a list of ten perceived enhancements, including examples of reviews and their respective apps.

RQ_6: What are the primary outcomes or impacts experienced by users as reported in the accessibility update reviews? With this RQ, our aim is to provide evidence of the practical implications users face when the app becomes less accessible after some update. This knowledge is important to show developers and organizations the impact of the lack of accessibility in the user experience. In earlier findings, we presented the main consequences detected and reported by users, categorized into four groups based on a ChatGPT-4 prompt. Here, we delineate the practical repercussions, grouping them into eleven categories using an open coding process.

RQ_7: What are the users demands when they notice accessibility decline after an update? With this RQ, we intend to understand what are the general user demands concerning the accessibility issues they reported. This knowledge is important to understand what users expect from developers and organizations following their testimonial on their perspective regarding app accessibility.

Understanding the impact of updates on app accessibility is crucial because it provides evidence that accessible apps may not remain accessible indefinitely. This raises an alert that, although evolution aims to improve app quality, neglecting to prioritize accessibility alongside other goals can negatively affect many users. In summary, the main contributions of this paper are the following:

- A dataset[2] of accessibility update reviews manually labeled according to our content analysis process (see Sec-

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

[2] https://doi.org/10.5281/zenodo.13139793

tion 3). This dataset can be used in further studies.
- A set of keywords associated with app updates that can be used to automatically identify user reviews (e.g. using string matching or machine learning) in which users report changes in new versions of an app. Such resource can be useful to increase this study dataset or to create different datasets for future investigations, regardless of the quality aspect investigated in software evolution.
- Evidences that most accessibility update reviews are associated with the WCAG "Perceivable" principle, and the "Distinguishable", "Adaptable" and "Seizures and Physical Reactions" guidelines, suggesting that broader goals linked to them may require more attention during software evolution.
- Identification of accessibility issues introduced by app updates, which shows that accessibility compliance is not stable over time. The accessibility barriers reported by users are concentrated on the following issues: poor color schemes, inadequate font size and types, unlabeled elements, and lack of customization options.
- Identification of accessibility enhancements noticed by users following an update, giving evidence that users do recognize developers and organizations efforts to deliver more accessible products.
- Identification of the practical implications for the user experience as consequence of the lack of accessibility of the new version of the mobile apps. The practical implications are mostly related to difficulty or inability to reading or distinguishing elements on the screen; physical discomfort; difficult navigation; and inability to use screen readers. Beyond practical consequences, users report to loose autonomy and feel discriminated against.
- Identification of what users expect from developers and organizations, primarily by requesting them to address accessibility issues, incorporate accessible features, and revert the app to a more accessible state. They also pose questions to challenge organizations regarding the neglect of the needs of people with disabilities in their design processes.

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 investigation delve into user reviews aiming at understanding how users perceive the impact of updates on the mobile accessibility. Therefore, this section outlines related work associated with accessibility reviews and mobile accessibility evolution.

Numerous studies explore the influence of user reviews on software updates in general [Liu *et al*., 2023; Palomba *et al*., 2018; Ciurumelea *et al*., 2017]. However, although many studies have been undertaken to analyze accessibility reviews[3] [Eler *et al*., 2019; Aljedaani *et al*., 2022a; Alshayban *et al*., 2020; AlOmar *et al*., 2021; Aljedaani *et al*., 2021; Santiago and Marques, 2022; Oliveira *et al*., 2023; Aljedaani *et al*., 2023; Liu *et al*., 2023; Aljedaani *et al*., 2022b], there is still a notable gap in research regarding user insights on the impact of updates on the mobile app accessibility.

Considering that constant modifications can cause the decline of the overall quality of software products [Lehman, 1996], researchers have investigated the implications of app evolution on some quality aspects of mobile apps. Gao *et al*. [2019] uncovered evidence suggesting that app complexity remains relatively stable throughout updates, possibly due to frequent removal of functionality, as noted by Nayebi *et al*. [2018]. This deletion is often driven by unnecessary features, subpar user experience, and compatibility concerns. Taylor and Martinovic [2017] examined how software evolution impacts security-related factors and discovered that Android apps do not become more secure with updates. Instead, many updates tend to exacerbate the number of vulnerabilities over time. Nayebi *et al*. [2018] noted that excessive functionality, typically stemming from the continuous addition of new features, can significantly affect usability, resource consumption, and maintenance efforts.

In addition to those quality aspects, researchers also investigated the consequences of updates on apps accessibility, which is the focus of our study. More specifically, Alshayban *et al*. [2020] employed an accessibility rate metric, calculated from the percentage of interface elements on a screen exhibiting accessibility issues, to assess subsequent updates of 60 mobile apps. Their analysis revealed that 47% of the updates enhanced the overall accessibility of the apps, while 28% resulted in decreased accessibility and 25% had no impact on their overall accessibility levels. However, these findings do not definitively establish whether the observed improvements stem from fixing accessibility bugs or from the addition of new interface elements with fewer accessibility issues. The latter scenario could potentially decrease the inaccessibility rate without addressing existing accessibility barriers. Furthermore, Chen *et al*. [2022] assessed the three most recent versions of 70 mobile apps and discovered that the number of issues remained unchanged in 57 apps (82%). However, in 10 apps (14%), the number of issues increased, while in 3 apps (4%), it decreased.

Previous studies on the evolution of mobile app accessibility have mainly focused on specific metrics like the number of issues detected by automated tools, often overlooking the changes in accessibility that occur following updates. Our earlier work deviated from this approach by analyzing 694 user reviews to gauge the impact of updates on perceived accessibility, highlighting both improvements and issues as well as users' reactions. We used ChatGPT-4 to summarize key accessibility challenges and advancements. In this manuscript, we extend our previous research by manually analyzing the content of user reviews to gain a deeper understanding of accessibility issues and improvements, as well as examining the WCAG 2.2 principles and guidelines related to the topics discussed in each review.

[3] User reviews that comment on accessibility aspects of the evaluated app.

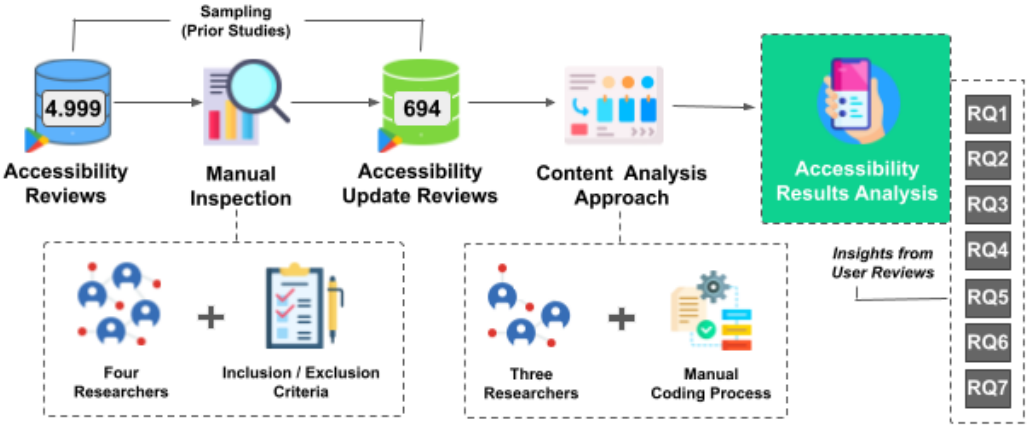

Figure 1. Overview of each step of our sampling and data analysis process.

3 Study design

This section describes the process we adopted to collect, label and extract information from accessibility update reviews. Figure 1 outlines the steps of our research method. First, we acquired a dataset of accessibility update reviews from previous studies [Oliveira *et al*., 2023; Dos Santos *et al*., 2023]. Next, we conducted the steps of a content analysis approach to code and analyze our data. Each step is detailed as follows.

3.1 Sampling

To gather insights into users' viewpoints regarding accessibility changes stemming from app updates, we opted to analyze user reviews posted on app stores. These reviews offer spontaneous opinions and assessments penned by users as they interact with the apps within their own contexts. The sampling process to produce a dataset of accessibility update reviews was conducted in two prior studies.

In a previous study, Oliveira *et al*. [2023] employed string-matching filtering and manual inspection of nearly 180 million user reviews extracted from the most popular apps from the Google Play Store to create a dataset[4] of 4,999 accessibility reviews associated with visual disabilities and eye conditions. All reviews in this dataset are labeled according to some categories: disability or eye condition associated with the review; whether the feedback is positive or negative; and the interface resources or components associated with the review. This dataset comprises 936 positive (18.7%) and 4,063 negative reviews (81.3%). Positive reviews typically indicate acknowledgment that the app is accessible to some extent, whereas negative reviews usually consist of complaints or requests for missing accessibility features.

In our most recent prior study [Dos Santos *et al*., 2023], we conducted a manual examination of the 4,999 accessibility reviews within the dataset produced by Oliveira *et al*. [2023] to select reviews that reported changes in mobile app accessibility following updates. The manual inspection involved four researchers, with one researcher analyzing all reviews, while the remaining researchers inspected one-third of the sample each, ensuring that each review was examined at least twice. Access to subsets of the reviews was provided to each researcher via a Google Spreadsheet. Researchers were instructed to retain only those reviews in which users explicitly mentioned either negative or positive outcomes regarding mobile accessibility directly caused by a recognized app update (e.g. *"The latest version (4.2) appears to have removed the ability to select a font size for reading messages. This is a major problem for those of us who are visually impaired! Please return that option!" (Gmail)*). In that sense, researchers were instructed to not keep the reviews in which users mention accessibility issues but the implications of the app update are not accessibility-related. In addition, researchers must also rule out reviews that mention updates on the Operating Systems instead of the mobile app.

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 (e.g. Android) 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 [Cohen, 1960]. We acquired a degree of agreement of 0.98. According to Fleiss *et al*. [1981], these agreement values are nearly *perfect agreement* (*i.e.*, 0.801.00). 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 Content analysis approach

Content analysis encompasses a range of systematic, rule-based methodologies employed for analyzing the informational content present within textual data. In this study, we employed both quantitative and qualitative content analysis approaches. Following a quantitative content analysis approach, we used a closed coding process to systematically as-

[4]https://github.com/marceloeler/data-paper-chi23

Table 1. Codebook.

Code	Definition	Segments example
Update keyword	Terms highlighting the occurrence of updates	"used to be able", "since the update", "last version"
User characteristics	Terms associated with the user, such as pronouns, disabilities, technology, etc.	"Some of us, eye strain" "I´m, legally blind, visually impaired fellow, like me"
Accessibility Issues	Accessibility issues users reported as a consequence of an update.	"small font size", "no dark mode options"
Accessibility improvements	Improvements and compliments highlighted by users following an update.	"The dark mode is finally here", "love the new font side update"
User requests	User demands for developers and organizations in the review.	"I hope it's fixed and back in a new update soon", "Please give us back the carousel feature!", "Stop removing options!"
Consequences/Implications	Practical implications stemmed from the lack of accessibility introduced by an update.	"that blind me", "nearly impossible to use it", "hard to distinguish between different colors and contrasts."

sign codes to segments of the accessibility update reviews according to predefined categories. Such approach allowed us to quantify and generalize findings from our dataset. The codebook for this task was developed based on the research questions we devised for this study. Table 1 shows the codebook devised in this study. In addition to assigning codes to significant segments or phrases encapsulating the core concepts within our textual data, we categorized each accessibility update reviews based on the WCAG 2.2 guidelines.

We adopted the WCAG in this process because it has been the *de facto* standard when it comes to accessibility guidelines. It was developed by the World Wide Web Consortium (W3C) and provide recommendations to enhance accessibility of digital content, particularly for people with disabilities such as those with blindness, low vision, and photosensitivity. Additionally, it has been widely used in research involving digital accessibility [Power *et al*., 2012; Reyes Arias *et al*., 2022; Haggag *et al*., 2022; Othman *et al*., 2023; Sandoval Alcocer *et al*., 2024]. The WCAG is structured around principles, which are supported by guidelines, which has specific success criteria to measure compliance and address various accessibility challenges. In total, WCAG has four main principles, thirteen guidelines and and fifty-nine success criteria. The WCAG principles are presented as follows.

1. Perceivable. Information and user interface components must be presented to users in ways they can perceive. The guidelines of this principle are: Text Alternatives, Time-based Media, Adaptable, and Distinguishable.

2. Operable. User interface components and navigation must be operable. The guidelines of this principle are: Keyboard Accessible, Enough Time, Seizures and Physical Reactions, Navigable, and Input Modalities

3. Understandable. Information and the operation of the user interface must be understandable. The guidelines of this principle are: Readable, Predictable, and Input Assistance.

4. Robust. Content must be robust enough that it can be reliably interpreted by a wide variety of user agents. This principle has a single guideline: Compatible.

3.2.1 Coding process

In the closed coding process, three researchers independently reviewed the data and assigned codes from the codebook (Table 1) to segments of the accessibility update reviews. They also classified each accessibility update reviews according to WCAG 2.2 guidelines. Each researcher had four weeks to analyze the data in order to avoid exhaustion. After coding the data independently, the inter-coder reliability of the process was ensured by requiring agreement between a minimum of two out of three researchers for the assignment of a code to a text segment. This approach helped to mitigate potential discrepancies and increase the validity and consistency of the coding process. Codes that met this criterion were considered reliable and were retained for subsequent analysis, while any instances of disagreement were resolved through discussion and consensus among the researchers.

3.2.2 Data analysis

We adopted different approaches to analyze our data and present our findings according to the RQ:

RQ1 to RQ3: update keywords, user characteristics and WCAG principles and guidelines. We used quantitative analysis to present the frequency of terms or guidelines associated with each review. We also resorted to wordclouds and wordtrees to provide visual aid to provide an overview of the topics explored within the accessibility update reviews for RQ1 and RQ2.

RQ4: accessibility issues. The list of accessibility barriers pinpointed by users following updates is enormous, hence we decided to split them into categories. Instead of creating categories based on an open coding process, we opted to organize and analyze the results based on the interface components or elements involved in each accessibility update reviews as labeled in previous work [Oliveira *et al*., 2023].

RQ5: accessibility improvements. The list of accessibility enhancements identified by users is short, therefore we opted to analyze and outline the most mentioned issues.

RQ6 and RQ7: practical implications/consequences and user demands: The list of practical implications and user demands that stem from the lack of accessibility is

shorter than the list of accessibility barriers (RQ4), but it still requires some sort of categorization to make it easier to analyze and understand. Thus, we conducted an open coding process in which we assigned categories to each accessibility update reviews based on their resemblance with respect to each RQ.

4 Results and discussion

This section presents the results of our investigation. Before answering each research question, we describe our dataset based on some of the results we obtained in our previous study [Dos Santos *et al*., 2023]. Altogether, we detected 694 accessibility reviews from the dataset created by Oliveira *et al*. [2023], detailing alterations in the mobile app accessibility subsequent to updates. The number of accessibility reviews associated with app updates represents 14% (694) of the original dataset, which is composed of 4999 accessibility reviews. The majority of these reviews (647 reviews – 93%) express negative feedback, indicating a decline in app accessibility post-update, while only 7% of reviews (47) provide positive comments, suggesting users perceived the updated version as more accessible.

The prevalence of negative reviews does not necessarily imply that all apps inevitably become less accessible as they evolve. It is plausible that numerous accessibility enhancements go unreported in user reviews due to a tendency for people to provide negative rather than positive feedback [Kraft and Martin, 2001]. However, the disproportionate amount of negative feedback serves as evidence that the evolutionary process of mobile apps, aimed at meeting user and environmental demands, may inadvertently compromise other quality aspects, such as accessibility.

Table 2 shows the top 10 interface resources or components that are mostly mentioned in accessibility update reviews. We delineated the interface components and resources associated with each accessibility update reviews by utilizing a string-matching procedure grounded in keywords derived from the design fundamentals and guidelines outlined in the Google Material Design framework[5] and in the BBC Mobile Accessibility Guidelines[6]. COLOR is by far the most comment resource (332 reviews – 48%), followed by MODE (205 reviews – 29%), SCREEN READER (96 reviews – 14%), and FONT (84 reviews – 12%).

Evidently, discrepancies with the color scheme or color mode/theme of mobile apps emerge as the most prevalent concern among reported updates. Notably, this issue correlates with several resources detailed in Table 2, including COLOR, MODE, BACKGROUND, and SETTING. A recurrent complaint voiced in user reviews revolves around the removal of dark mode in the latest 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 report that the update has made the app inaccessible with assistive technologies, such as screen readers, according to their feedback (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)). Additionally, another prevalent critique is that the font size has decreased in the new version, or users are unable to adjust font settings anymore (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)).

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

Resource/Component	Reviews	Proportion
COLOR	332	48%
MODE	203	29%
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%

The answer to each RQ is discussed as follows.

4.1 RQ_1: Common terms and expressions

RQ_1 aims at identifying common keywords or expressions employed by users to convey that something in the current version of the app has changed compared to its previous release. Such knowledge can help establishing a set of keywords or a training dataset to identify reviews associated with app updates from the perspective of user reviews in upcoming studies, regardless the focus of the analysis (e.g. accessibility, bug reports, security, energy consumption). Therefore, in the coding process (cf. Section 3), researchers identified specific terms and expressions related to app updates. Table 3 shows the top 20 most frequent words employed by users.

Table 3. The top 20 most frequently mentioned words when users want to express that something has changed from a previous version to a current release of the mobile app.

Word	Frequency	Word	Frequency
update	422	recent	43
new	188	since	41
now	105	used	37
version	89	updated	37
back	83	old	29
latest	75	bring	28
change	70	before	27
after	53	changed	26
was	47	removed	24
last	44	made	23

Figure 2 presents a word cloud that provide a visual presentation of the most prominent terms and expressions employed by users by displaying frequently occurring words in larger fonts. This word cloud were build considering both

[5]https://m3.material.io/components

[6]https://www.bbc.co.uk/accessibility/forproducts/guides/mobile/

isolated words and expressions (e.g. latest update, new update), that is why some words may be presented in a bigger font than some words with a higher frequency (e.g. "new" and "now").

Figure 2. Word cloud illustrating the most common terms and expressions employed by users to convey that something has changed from a previous version to a current release of the app.

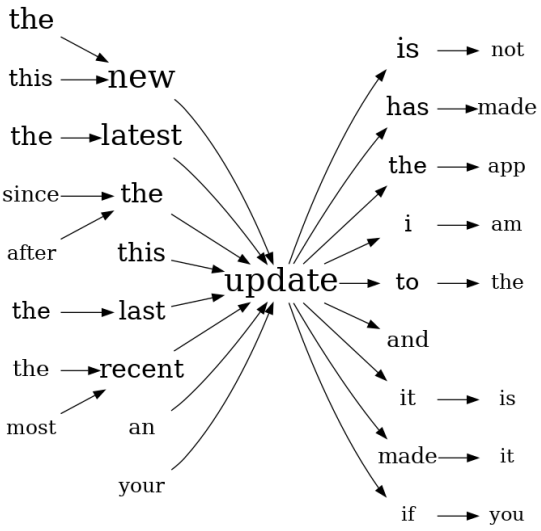

Figure 3. Word tree showing how other words frequently employed by users are connected to the keyword "update".

Discussion. Table 3 shows that most users explicitly indicate that their review refers to some characteristic of the app considering that the app has been updated. The word "update" appears 422 times, "version" is mentioned 89 times, and "updated" is emerged 37 times, suggesting that users commonly recognize and acknowledge when an update has occurred. Users also frequently use terms associated with the current status of the app, employing words such as "new", "now", "this", "latest", "after" and "recent". They also use words to refer to how things were in the past, such as "change", "was", "last", "since", "used", "old", "before", "changed", "removed" and "made". Finally, the frequently utilize words to express they desire to have previous version restored (e.g. "back" as in "bring back").

In addition to knowing the frequency of some words and visualizing most prominent terms and expressions in word cloud, understanding the context in which some vocables are used. Hence, following we show a word tree for the top most frequent terms identified in Table 3. A word tree is a visual representation that helps in data analysis by showing how phrases or words are connected to a central keyword or phrase in a hierarchical structure. The word trees that are presented as follows we considered a max length of two words before and two words after the keyword of interest.

Figure 3 shows a word tree for the keyword "udpate". Notice that high frequently mentioned words (e.g. "new", "latest", "last", "recent") appear before "update" to indicate which version of the app they are referring to; following, they characterize or present the consequence of the update: "it is", "is not", "has made" and "made".

Figure 4 shows a word tree for the keyword "new". Notice that this term is strongly associated with the word "update" and "version". It is also used to highlight that something specific within the app is different (e.g. "font", "layout", "look", "ui", "colors", "bright").

Figure 5 shows a word tree for the keyword "now". This vocable is generally used to indicate that, from that moment on (i.e. the update), some things are different from it was used to be, which can be related to the characteristics of the app interface or with the ability of the users (e.g. "now I

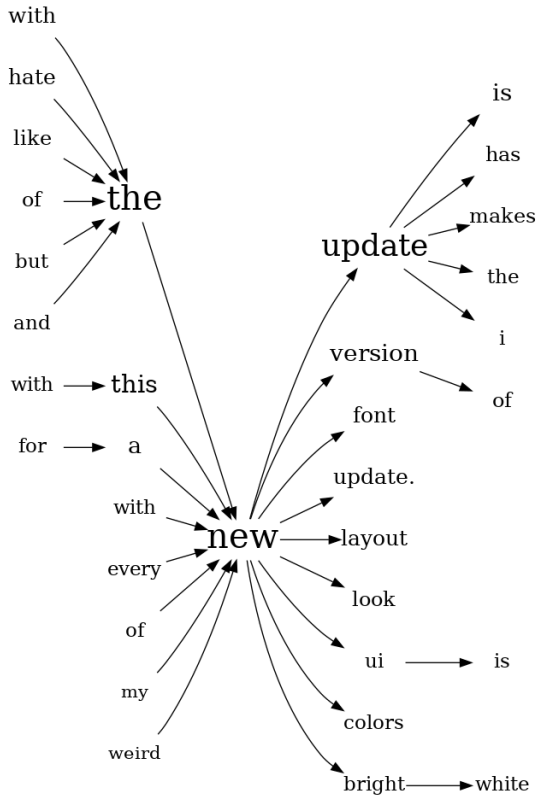

Figure 4. Word tree showing how other words frequently employed by users are connected to the keyword "new".

can't").

Figure 6 shows a word tree for the keyword "version". Users generally employ this term to refer to a previous or the current version of the app (e.g. "previous", "updated", "older", "lateste", "new") and sometimes defining the specific app or one of its characteristics they are referring to. This term is usually an indication that the user will contrast the difference from subsequent versions of the app.

Figure 7 shows a word tree for the keyword "back". This term is recurrently used by users to clearly indicate that they

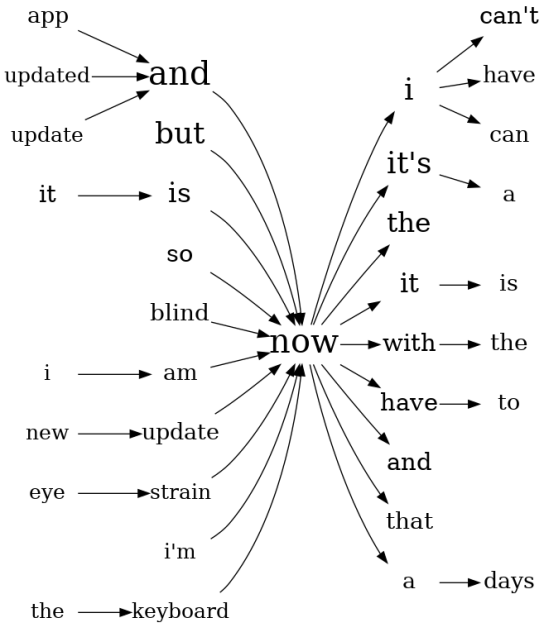

Figure 5. Word tree showing how other words frequently employed by users are connected to the keyword "now".

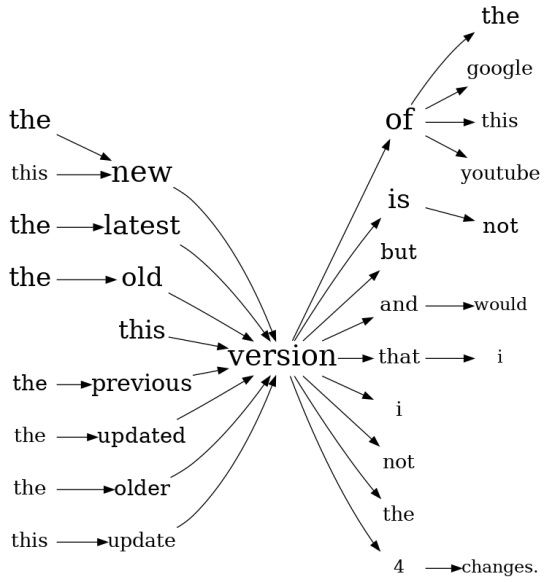

Figure 6. Word tree showing how other words frequently employed by users are connected to the keyword "version".

are not satisfied with the current version of the app as they ask developers and organizations to bring back some feature of layout. Common expressions include "bring back", "revert back", "change it back", "give us back", and so forth.

4.2 RQ$_2$: Personal context information

RQ$_2$ seeks to characterize the message sender by identifying personal context information. More specifically, researchers assigned the code *user characteristics* to segments of text that refer to: physical traits (possible disabilities or conditions), usage context (e.g. watch videos, movies, work, study), pronouns (e.g. I, he, she, us), sender's identity

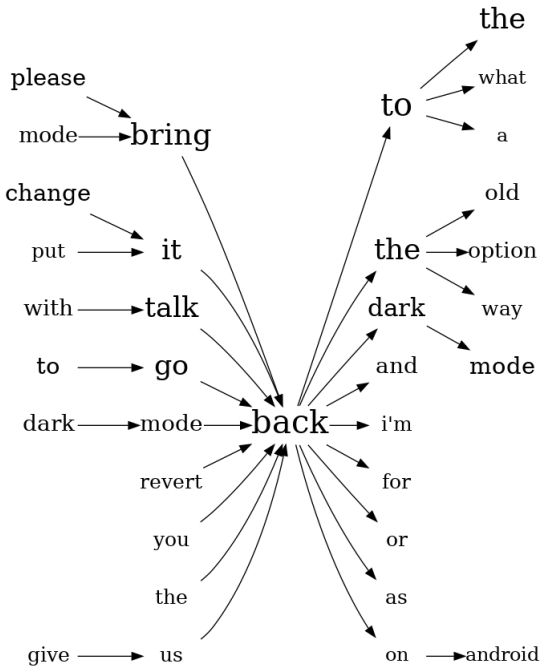

Figure 7. Word tree showing how other words frequently employed by users are connected to the keyword "back".

(whether for oneself or for others), relationships involved in the app usage (e.g. friends, family, spouse), and devices (e.g. brands, models). Figure 8 presents a word cloud that provides a visual presentation of the most prominent terms and expressions employed by users by displaying frequently occurring words and expressions in larger fonts. Table 4 shows the top 20 most frequent words employed by users.

Figure 8. Word cloud illustrating the most common terms and expressions employed by users to refer to personal perspectives, such as pronouns, personal gadgets (e.g. tablet), and disabilities.

Discussion. The word cloud clearly shows that the words that are more determinant to characterize users that wrote the accessibility reviews are related to their conditions, disabilities or physical reactions to the app interface. This result is inline with the fact that we are delving into reviews associated with mobile apps accessibility, which in general has more impact on users with disabilities. In addition, the dataset of accessibility reviews we used to extract accessibility update reviews is associated with users with visual impairments, hence it is not surprising that, in addition to using pronouns to refer to themselves (e.g. "'I", "my", "me", "us"), the most predominant expressions are associated with visual impairments, such as "visually impaired", "eye strain",

Table 4. The top 20 most frequently mentioned words when users want to express something related to themselves (e.g. personal characteristics, pronouns, devices, other people) in their review.

Word	Freq.	Word	Freq.
I	250	users	24
visually impaired	72	vision problems	24
my	66	light sensitivity	19
eye strain	65	us	15
people	41	we	15
I am	41	eye straining	14
i´m	36	user	14
sensitive eyes	29	visual impairment	13
me	28	eyestrain	13
blind	28	android	12

"blind", "sensitive eyes", "vision problems", "light sensitivity".

In fact, the 4,999 accessibility reviews of the dataset from which we extracted the accessibility update reviews of our study are linked to 36 different types of visual disabilities or eye conditions [Oliveira *et al*., 2023]. The 694 accessibility update reviews we identified for this investigation are associated with 19 different types of disabilities or eye conditions. Table 5 shows the list of disabilities and the number of reviews in which they are mentioned by users. Visual impairment is by far the most commonly mentioned disability as it is related to 283 reviews (40.78%), followed by eyestrain (116 reviews – 16.71%), photophobia (94 reviews – 13.54%), low vision (25 reviews – 3.6%) and color blindness (25 reviews). Many disabilities or conditions were mentioned less frequently, such as legal blindness (14 reviews – 2.02%), old eyes (9 reviews – 1.3%), astigmatism (9 reviews), partial vision (6 reviews – 0.86%), weak eyes (5 reviews – 0.72%), glaucoma (3 reviews – 0.43%), snow blindness (2 reviews – 0.29%, cataract (2 reviews), myopia (2 reviews), macular degeneration (1 review – 0.14%), partial blindness (1 review) and blurry vision (1 review). It seems that numerous users express a desire to connect their feedback with a visual disability or eye condition to underscore the tangible barriers they encounter, thereby providing additional incentive for developers to address the reported issues. While certain users cite specific disabilities (such as myopia), a considerable number of them utilize general terms (like visual impairment).

Table 5. The list of disabilities mentioned by users in the accessibility update reviews. Column 1 shows the disability and Column 2 shows the number of reviews in which it is mentioned.

Disability/Condition	Freq.	Disability/Condition	Freq.
visual impairment	283	partial vision	6
eyestrain	116	weak eyes	5
photophobia	94	glaucoma	3
blindness	66	snow blindness	2
possibly photophobia	53	cataract	2
low vision	25	myopia	2
color blindness	18	macular degeneration	1
legal blindness	14	partial blindness	1
old eyes	9	blurry vision	1
astigmatism	9		

Besides referencing their physical conditions, but to a lesser extent, users discuss their relationships, how and why they use the apps, and provide detailed information about devices and models in order to provide data that may help resolve the problems faced. While infrequent, these observations suggest that certain users recognize potential issues tied to specific device models, potentially mentioning them in their reviews to aid developers in addressing improvement requests or problem reports. Conversely, in some instances, users express frustration with recurring issues, leading them to contemplate migrating to alternative platforms, like iOS/iPhone.

4.3 RQ_3: WCAG principles and guidelines

During the coding process, we classified the accessibility update reviewsof our dataset according to their association with the WCAG 2.2 guidelines. To answer RQ_3, we present the distribution of accessibility update reviews reviews associated with the WCAG 2.2 principles, a broad category of considerations that serve as a foundation for accessibility; and with the WCAG 2.2 guidelines, which are more specific and actionable recommendations that offer practical advice on how to achieve accessibility objectives within each principle. Breaking down the accessibility update reviews dataset into WCAG 2.2 success criteria, which are specific testable accessibility criteria, is not at the scope of this study.

Figure 9 shows the distribution of accessibility update reviews according to the WCAG principles. Each review can be associated with more than one principle since users report may be linked to different aspects of their experience when it comes to the accessibility of the app. The Perceivable principle, which defines that interface elements must be presentable to users in ways they can perceive, is by far the most frequent principle (585 reviews – 84%); followed by the Operable principle (264 reviews – 38%), which defines that users must be able to interact with all controls and interactive elements using various input methods. The Robust principle, which defines that the software and content must be robust enough to be interpreted reliably by a wide variety of user agents, including assistive technologies, is the third most frequent principle (105 reviews – 15%). The Understandable principle, which defines that information and operation of the user interface must be clear and easy to understand, is way less frequent (38 reviews – 5%). Finally, there are 29 (4%) accessibility update reviews that are too generic, thus they could not be associated with any principle (e.g. *"Updates are not much accessible so blind people facing many probs."* (Whatsapp Messenger)).

Figure 10 shows the distribution of accessibility update reviews according to the WCAG guidelines. This distribution highlights the areas where updates have been most frequently reported in terms of WCAG guidelines, with "1.4 Distinguishable" being the most common category, representing nearly 69% (477) of all reviews, followed by "2.3 Seizures and Physical Reactions", with 161 reviews (23%), and "1.3 Adaptable", with 159 reviews (23%).

Discussion. Effectively gathering information through our sensory system is a prerequisite for understanding or interacting with any interface. Therefore, it is not surprising that the majority of reviews are linked to the "Perceivable" principle since any problems in this first and fundamental

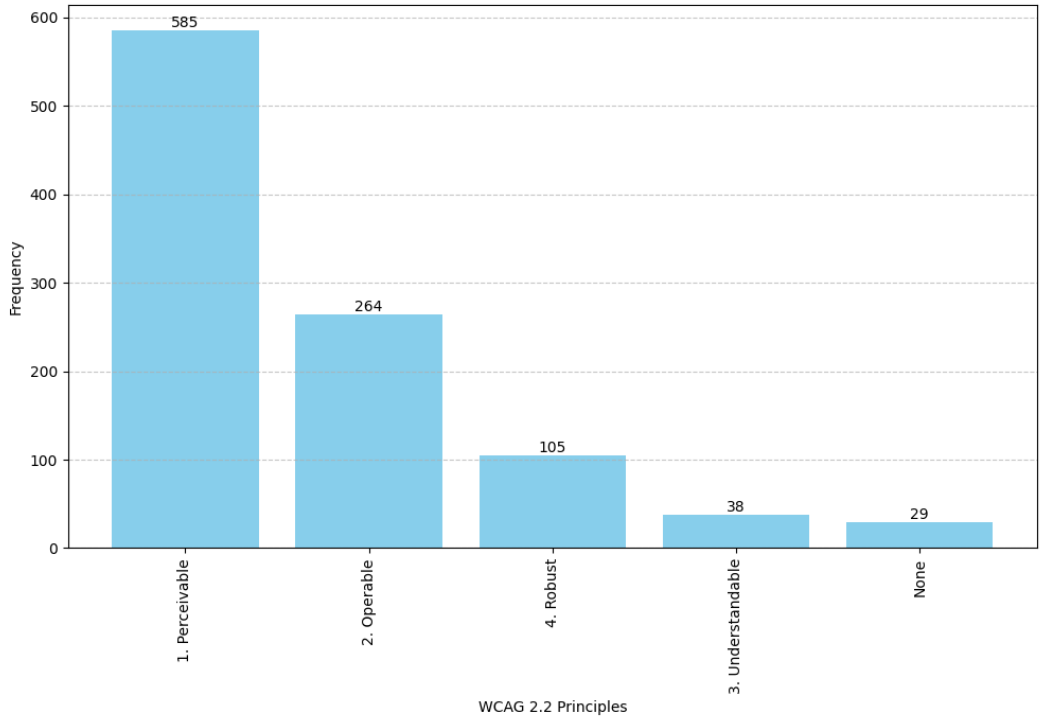

Figure 9. Frequency of WCAG 2.2 principles associated with each accessibility update reviews of our dataset.

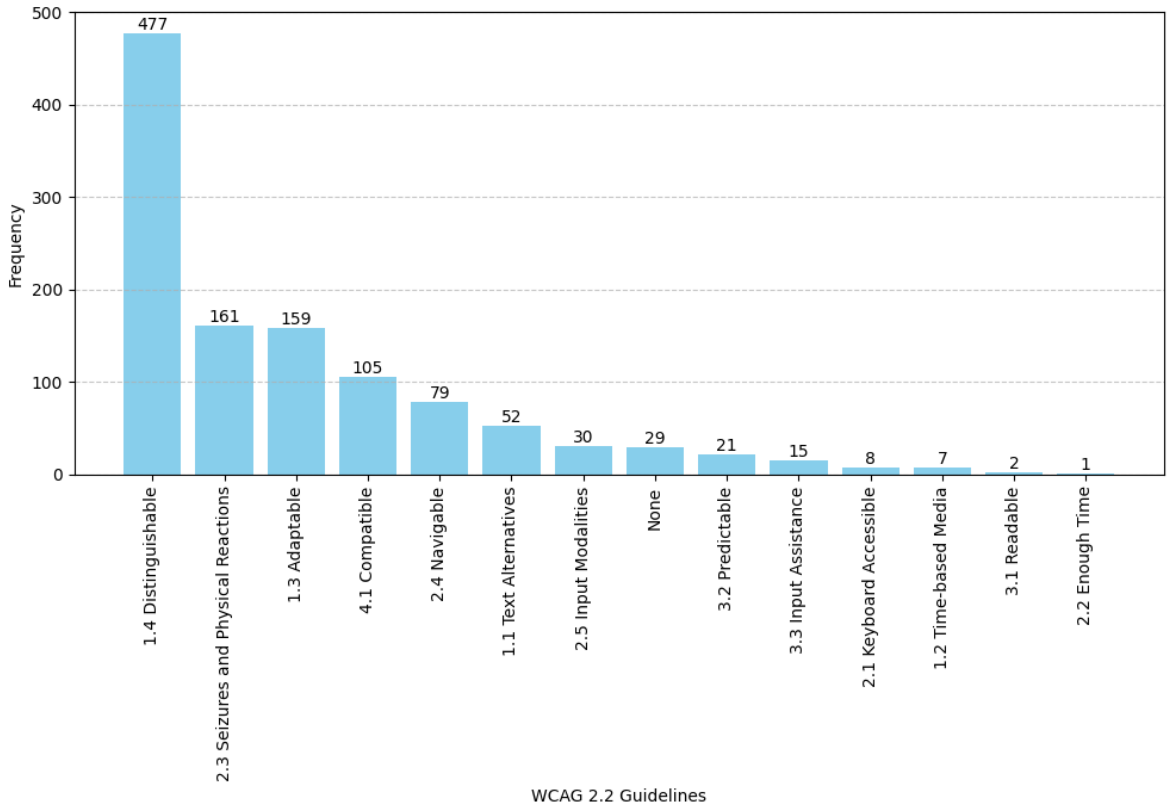

Figure 10. Frequency of WCAG 2.2 guidelines associated with each accessibility update reviews of our dataset.

layer for any successful interaction may represent a major barrier for any user. As the WCAG principles seem to be considered in a hierarchical manner to create a comprehensive framework for web accessibility, in which each principle is seen as building blocks, with some foundational principles supporting others, the Understanding principle naturally follows the Perceivable principle when it comes to their frequency. The Robust principle holds significant relevance in

this research due to its inherent nature towards assistive technology. We delve into accessibility reviews concerning individuals with visual disabilities, a demographic heavily dependent on assistive technologies. Therefore, compatibility with these technologies is paramount for mobile apps, rendering the Robust principle one of the primary focuses.

The accessibility update reviews of our dataset are associated with all 13 WCAG 2.2 guidelines, with a clear concentration on guidelines of the Perceivable and Operable principles as discussed before. The fact that most reviews are linked to the "1.4 Distinguishable" guideline is not surprising because it comprises 13 success criteria that are connected to the main accessibility barriers found in mobile applications Yan and Ramachandran [2019]; Eler *et al*. [2018]; Chen *et al*. [2022], which includes, among others, the use of color, audio control, contrast, text size configurations, text spacing, and visual presentation. This suggests a concern among users about the visual presentation of content and the ability to distinguish different elements on the screen.

The high frequency of reviews connected to the guideline "2.3 Seizures and Physical Reactions" does not mean, in this study, that users are actively concerned about content that may trigger seizures. Rather, they are more concerned with physical reactions and discomfort that comes from white background and the absence of dark or night modes, which can include eyestrain, headaches, migraine, dizziness, and so forth. Reviews associated with this guideline are mostly associated with the "1.4 Distinguishable", which comprise the "Visual Presentation" success criteria, which defines that foreground and background colors can be selected by the user.

The third most common guideline associated with the accessibility update reviews of our dataset is the "1.3 Adaptable", which indicates that users are concerned with the requirements of the content to be adaptable across different devices, screen sizes, and user preferences. Users seem to value content that can easily adjust to their individual needs and contexts, enhancing usability and accessibility.

Another major concern for users is the fact that many apps are not compatible with assistive technologies. The specific WCAG guideline that discuss this topic is "4.1 Compatible", which outlines criteria to maximize compatibility with current and future user agents, including assistive technologies. In many cases, users mention that some app is not compatible with screen reader when they cannot adequately perceive and navigate through some interface due to the absence of descriptions of non-text content. However, in this study, the "1.1 Text Alternatives" guideline, which is concerned with providing text alternatives for any non-text content, is associated with only 10 accessibility update reviews. Even though those two guidelines may be connected, researchers only assigned a correspondent guideline to a review when users explicitly explain the reason of their complaint (e.g. *"I am currently unsatisfied with the most recent update as there is little to none accessibility features for accessing the 'playlist' section as it does not speak to me for that button that would say that it would take me to that section. All it does is say 'button double tap to activate' when it should replace 'button' with 'playlist button' so that me and potentially other blind people can tell where they are navigating."* (Amazon Music)).

The "3.1 Readable" and "2.2 Enough Time" are the guidelines that are least associated with the accessibility update reviews. The Readable guideline is concerned with making the make text content readable and understandable. More specifically, it refers to the used language, abbreviations, unusual words, reading level, and pronunciation, which seems to represent a minor concern for users in this case. The low frequency of reviews mentioning the Enough Time guidelines suggests that users may not encounter as many issues related to time limits or interruptions.

4.4 RQ_4: Accessibility problems

To answer RQ_4, researchers assigned the code *"accessibility issues"* to segments of accessibility update reviews that clearly indicates a problem in the interface that affects app accessibility according to users perception. Rather than detailing every issue within this category, we opt to organize them by different categories. Aligning with WCAG 2.2 guidelines for organization would be straightforward, given our labeling. However, WCAG guidelines provide broad objectives for making content accessible, that is why most reviews are concentrated on the top three most mentioned guidelines (cf. Figure 10), hindering more specific categorization. A more specific categorization would require breaking down each guideline in success criteria, which is out of the scope of our study. Therefore, we chose to highlight primary accessibility issues based on interface elements according to their appearance frequency in accessibility update reviews (cf. Table 2), signifying areas needing attention.

Color. Users with disabilities encountering accessibility challenges face numerous issues related to the color scheme of the user interface across various applications. The absence of options for font style, line, and character spacing, coupled with limited control over colors, severely hampers the usability of the interface, particularly for individuals with visual impairments. Moreover, the introduction of new colors that are not color-blind friendly or excessively bright and glaring exacerbates the problem, causing discomfort and making it difficult to distinguish between different elements. Users often express frustration at the lack of customization options to adjust the color scheme to their preferences or specific needs, highlighting the disregard for inclusivity in design decisions. Additionally, sudden changes in colors without warning or consideration for user feedback further compound the accessibility challenges. Following, we present excerpts extracted from accessibility update reviews associated with such problems.

> *"Why did this app change from black and white to pink and purple with the last update? I hate it! (...) I am visually impaired and the black and white was helpful to me the pink and purple makes it very difficult to see. I looked all through the app for an option to change the colors but found nothing."* (Clock)

> *"The choice of different colors for different event has disappeared (I had to get a replacement phone). I'm legally blind the colors help me identify different events on my calendar. Not happy with this change!!!"* (Google Calendar)

Mode or Theme. Users have expressed frustration over recent updates that have either removed or failed to include dark mode, night mode, or dark theme options in various applications, resulting in discomfort and usability issues, particularly for those with visual impairments or sensitivity to bright screens. In Amazon Alexa and Amazon Kindle, users lament the absence of dark mode, which they relied on to reduce eye strain. Similarly, in Duolingo, users miss the night mode feature, which is beneficial for practicing at night and reducing eye strain. Facebook users criticize the disappearance of dark mode, with some reporting unbearable brightness causing eye strain, while others express confusion over the app's failure to adhere to their device's dark mode settings. In Gmail and Google, users report dark mode disappearing after updates, causing frustration and discomfort. Snapchat users find the new design visually unpleasant, especially without a dark mode option, leading to discomfort and difficulty in using the app. These excerpts from accessibility update reviews in our dataset exemplify this finding:

> *"The newest update seems to have gotten rid of night mode. I hope it's fixed and back in a new update soon because the white background causes so much eye strain!"* (Amazon Alexa)

> *"First of all what happened to the black background when you go to your profile page? the white is eye straining and they made it so busy it's like having an anyuersim each time you glance at the screen. you should make an option for a dark mode."* (Snapchat)

Screen reader. Users with disabilities, particularly those reliant on screen readers, detail challenges in navigating apps, finding options, reading images, and encountering compatibility issues during interactions like booking or making payments. These specific grievances offer actionable insights for developers to address and improve accessibility. In contrast, generic complaints express overall dissatisfaction without specifying specific barriers, limiting their usefulness for developers seeking to enhance app accessibility. Here are some accessibility update reviews excerpts from our dataset that emphasize this:

> *"I'm using a screen reader as I'm visually impaired. In the past I could find the options for YES NO MAYBE very easily when deciding to like someone or not. Now the screen reader reads me all this BUTTONS which I haven't a clue what they are. More or less where on someone's profile does one find the YES NO MAYBE options? I'm going to deactivate and uninstall my baddoo because I now noticed I a screen reader user can't even find where to right a message to someone. So this app is not useful."* (Badoo - Dating & Meet People)

> *"Not accessible for new payment option for using screen reader for blind user I am using to phone redmi note 7 Pro and Samsung m30 both phone aa hang totally hanging using new payment option"* (Google Pay: Save- Pay- Manage)

Font. Users experiencing issues related to font size across various applications express frustration with the lack of customization options and the inability to adjust font size to accommodate their visual impairments. They report that recent updates have resulted in smaller font sizes, making text difficult to read and causing eye strain. Additionally, users highlight the importance of font size settings for individuals with low vision, emphasizing the need for accessibility features to ensure usability for all users. These are excerpts from accessibility update reviews in our dataset that highlight this point:

> *"They recently changed the font in the app to tiny and there is no way to change it back or make it large text. (...) I can no longer use the app because I've had vision loss and need the large print. (...) The app should have an option to enlarge the text. People with low vision cannot use this app."* (Calorie Counter - MyFitnessPal)

> *"really really bad new font extremely unfriendly for those with low vision or dyslexia especially since you seem to not allow a phone's UI to change it to the user's preferred font. i would give it a zero if i could."* (Twitter)

Background. Users have expressed various accessibility concerns related to the background of the app, including dissatisfaction with its white and bright appearance and the removal of dark or night mode options, making it uncomfortable for users. Additionally, users lament the inability to customize background colors, highlighting the lack of flexibility in font and background color options. Some find the light blue background problematic, while others encounter issues with background and font color clashes, such as white text on a light blue background or grey background with blue letters, which hinders readability and usability for certain users. Here are excerpts from the accessibility update reviews in our dataset that underscore this finding:

> *"This was a great app until they removed the font and background colors! Now its not customizable at all! BRING BACK COLORS!!"* (Wattpad - Read & Write Stories)

> *"With the latest upgrade my typing field in messaging is set to a grey background and blue letters. I'm sending you the bill when I get glasses due to eyestrain."* (OkCupid: Online Dating App)

Setting. Users express frustration over the lack of accessibility settings in various mobile apps, impacting their usability, especially for those with visual impairments or sensitivity to light. Complaints include the removal of dark mode options, font size settings, and color customization, causing discomfort and difficulty in reading content. Specific grievances highlight issues with keyboard functionality, screen brightness adjustments, and layout changes, emphasizing the need for developers to prioritize accessibility features and provide customization options to accommodate diverse user needs. These excerpts from accessibility update reviews in our dataset support this finding:

> *"Latest update hijacks screen brightness making reading in low light uncomfortable for people with sensitive eyes. If your devs made this change intentionally they're out of their minds. No user wants his settings hijacked. Change this."* (Amazon Kindle)

> *"I need help with this one for some reason the application changes my display setting it makes it more difficult to read... I'm visually impaired and I don't need the app making my screen dimmer... why is Google play books changing my display setting whenever I activate the app."* (Google Play Books & Audiobooks)

Text. Users encounter various accessibility problems with text in mobile apps, particularly related to text size, text-to-speech functionality, and text layout. Complaints include text being too small (similar to Font), difficulties using voice-to-text features, removal of text reflow or word wrap functionality, and issues with text-to-speech not functioning properly. Additionally, users note challenges with text box visibility when using accessibility features like magnification, highlighting the need for improved text layout and navigation options to enhance usability for blind individuals. Following, we present excerpts extracted from accessibility update reviews associated with such problems.

> *"The latest update in swiggy has made the app completely inaccessible for text to speech application for blind users. I am a blind person and I reside alone. (...) Currently the app is completely unusable for people like me and I have to depend on someone else to order which is a risk in the covid times. Please conduct an accessibility audit and make the app accessible. Meanwhile roll back the latest update."* (Swiggy : Food Delivery | Instamart Grocery | Genie)

> *"This app used to work fine with magnification but now when typing your message you lose the text box where you are typing preventing you from seeing what you are typing. The text is there it has just slipped out of view under the keyboard."* (WhatsApp Messenger)

Button. Users with disabilities encountering accessibility issues face various challenges related to buttons in the user interface of mobile applications. These challenges include buttons lacking labels necessary for assistive technology, which makes it difficult for users reliant on screen readers to understand their functionality. Changes in button placement without maintaining consistency pose navigation difficulties, particularly for blind users, affecting their ability to interact effectively with the app. Additionally, issues such as smaller buttons and the absence of clear labeling further hinder usability, especially for visually impaired individuals who rely on distinct visual cues. Furthermore, the absence of labeled essential buttons, like the confirm pick-up button in ride-sharing apps, significantly impacts the independence of visually challenged users who depend on assistive technologies like TalkBack for navigation. Here are examples from accessibility update reviews in our dataset that emphasize this:

> *"(...) the outlay is not very user-friendly for blind people. Not all buttons are clearly labeled with text tags. (...) Since last update accessibility for the blind has diminished buttons have disappeared or are not workingddp"* (YouVersion Bible App + Audio)

> *"This is a great app other than the fact that the buttons need to be labeled for a visually impaired gamer you can still see to play ps3 / ps4 however can't use a phone well enough. Please remove these buttons with the next update."* (PlayStation App)

Layout. Users express frustration with changes in app layout that negatively impact accessibility. Complaints include difficulties reading emails with new window pane layouts in Gmail, experiencing eye strain and headaches due to lack of dark mode in Gmail, adverse effects on vision caused by layout changes in Google Calendar, challenges adapting to new layouts in Instagram, and inability to adjust settings for background color in MX Player. Additionally, users criticize navigation buttons being moved to the bottom of the screen, smaller button sizes, and poor color contrast in the new layout of YouTube, highlighting concerns about decreased accessibility and user-friendliness. These accessibility update reviews excerpts from our dataset highlight this point:

> *"Used to be very happy with my Google calendar app however phone has just done automatic update and entire app has changed layout appearance. With new look it badly affects my genetic eye defect (which old version did not) and damages my vision. There is no way to change or adapt the app anywhere in any settings either. This is extremely bad set out for use of many customers who have health or disability issues and are dependant on phone apps..."* (Google Calendar)

> *"The new layout stinks! I mean seriously. 1. Now the navigation buttons are at the bottom of the screen causing us to see LESS content. 2. The buttons are smaller. Do you hate people with vision problems? The elderly? Heck even my 12 year old nephew couldn't find them on his tablet because they're so tiny and they're grey on white. If your goal is to make your application LESS accessible to all and LESS user friendly then you've succeeded. Congratulations. slow sarcastic clap "* (YouTube)

Menu. Users encounter accessibility issues with menus in various apps after recent updates. For instance, in Pandora, the new menu layout is criticized for being difficult to navigate, particularly for visually impaired individuals, as it is smaller in print and located at the top of the screen, making it challenging to see from a distance. Additionally, in Viber, users express hope for customizable font sizes in

menus and messages, especially considering the small system app font size, which poses difficulties for visually impaired users. Here are several excerpts from the accessibility update reviews in our dataset that underscore this finding:

> *"This new update is crap. (...) Also the menu is lame. It's not easy to navigate. It's no longer in big print super small and at the top of the screen so for someone who's visually impaired such as myself it's extremely difficult to see from far away. Epic fail Pandora. Looks like I'm switching to Spotify. Much easier."* (Pandora - Music & Podcasts)

> *"Hoping that updates or progress are more in line of customizable size of message fonts menus and stickers as1. System app font size display too .small for visually impaired."* (Viber - Safe Chats And Calls)

Discussion. The issues introduced by the new releases of the analyzed apps are mostly associated with the "1.4 Distinghishable" guideline (cf. Section 4). In that sense, issues found by users could be summarized in few root causes, regardless the interface resource or component: poor color choice, small font sizes and lack of customization options. Such problems can be perceived in different ways when they affect different interface elements. By reading users comments, it seems that the the absence of customization settings is the main problem reported by users once the ability to select color schemes and font configurations (size and style) would render the app more accessible, specially because users tend to have particular preferences and needs. In addition to those issues, some users reported more specific problems that require more attention, such as the absence of labels for non-textual elements, hardware or software incompatibility with screen readers, sudden layout changes and the lack of audio feedback (e.g. *"The audio confirmation - that Assistant has heard 'ok Google' and is listening - has been removed... Stop removing features... there's a reason blind people need this. Idiots."* (Google)).

4.5 RQ_5: Accessibility enhancements

The amount of accessibility update reviews that express positive feedback within our dataset is significantly lower than the amount of negative comments. Hence, to answer RQ_5, we opt to not break the result into categories according to the interface resources or components of the app interface, neither based on the WCAG guidelines. Following, we present a list of improvements perceived by users after an app update.

Introduction of color labels. Users appreciate the addition of color labels, which help those with color blindness differentiate between items. Here is an excerpt extracted from the user reviews of our dataset that underscore this finding:

> *"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))

Color scheme improvement. Enhancements to color schemes, such as differentiating brightness levels, are beneficial for users with color vision deficiencies. This user review excerpt from our dataset highlights this point:

> *"When an update finally made the green and red traffic overlays different brightness levels I was happy. I could finally see the difference between green and red without asking another person's assistance."* (Google Maps)

Addition of dark or night mode or theme. Many users emphasize the importance of these elements for reducing eye strain. Here are examples from user reviews in our dataset that emphasize this:

> *"Thank you for finally adding Dark Mode in the settings of Samsung Music on this update .. I won't be blind anymore with forced Light theme .."* (Samsung Music)

> *"The new version on WhatsApp of dark mode is good to reduce eye strain"* (WhatsApp Messenger)

Improvement of font readability. Users with visual impairments find improved font readability significantly beneficial. An excerpt from our dataset's user reviews underscores this point:

> *"The latest update seems to have a darker easier to read font. That is great for me because of my visual impairment...thanks for the new font and features."* (Dropbox: Cloud Storage Space)

Options to change background color. The ability to change background color is crucial for users who experience migraines from photophobia. This excerpt from a user review in our dataset demonstrates this finding:

> *"I like this latest update better so far. I have the option for a darker background (black) which is good especially for people like me who suffers migraine from photophobia."* (Messages)

Compatibility with TalkBack. Ensuring app compatibility with TalkBack (screen reader) greatly enhances usability for blind users. Multiple excerpts from our dataset's user reviews illustrate this finding:

> *"I am a blind person. In previous updates the app never supported with my TalkBack option but from last two updates.. The app is very well performing with my talk back.. today 1st time in my life I ordered a food dish without taking any help of others"* (Zomato - Online Food Delivery & Restaurant Reviews)

> *"Just installed the knew version of Soundcloud and I have to take back most of what I said before. Other than three unlabeled radio buttons in the login screen Talkback now sees everything that wasn't usable before. Nice work."* (SoundCloud: Play Music & Songs)

Improvement of font size settings. Fixing issues with font size settings is crucial for users with partial sight. Here is a user review excerpt from our dataset that emphasizes this:

> *"Just wanted to say thanks on behalf of the partially sighted community for fixing the font sizing bug introduced in the previous update. It was a shame it happened. Please regression test properly before releasing. Only 4 stars as there are other problems still outstanding from the last snafu."* (Gmail)

User's ability to increase font size. Allowing users to enlarge font size significantly benefits those who are legally blind. Here is a user review from our dataset that underscores this finding:

> *"I love the new font side update!!!! Thank u so much for making such an accessible app!!!! As a legally blind person it means a lot to b able to enlarge the font size as large as I like!"* (YouVersion Bible App + Audio)

Bigger and adjustable content. Enhancements that allow users to adjust content size help reduce eye strain. An excerpt from our dataset's user reviews underscores this point:

> *"I think new update is very good. The content screen is now bigger and good for students because it reduces eyestrain to some extent. Also the settings now can be adjusted during meeting only because of new update."* (Google Meet)

Compatibility with mobile accessibility settings. Fixing issues related to mobile accessibility settings ensures that users with eyesight problems can effectively use the app. Here is an excerpt from the user reviews in our dataset that underscores this finding:

> *"I've got a eyesight problem and since last update whatsapp is ignoring my accessibility setting to increase font size... Looks like issue ws fixed with latest update. Thank you"* (WhatsApp Messenger)

Discussion. 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 [Kraft and Martin, 2001]. This aspect is also emphasized by the users themselves (e.g. *"I think new update is very good. The content screen is now bigger and good for students because it reduces eyestrain to some extent. Also the settings now can be adjusted during meeting only because of new update. Pls ignore other people baseless comments because they are the only people who find mistakes in every new update and don't find it pros."* (Google Meet)). In addition, most reviews of the sample from which we identified accessibility update reviews are negative comments on the app accessibility.

Notice that several users also recognized the developers' and organizations' efforts to make apps more accessible. We estimate that there might be many accessibility improvements that are not reported by user, but the few positive feedback we found is an indication that mobile apps also tend to become more accessible as they evolve. The enhancements highlighted by users appear to address the concerns outlined in Section 4, encompassing improvements such as color scheme adjustments, increased font sizes, customizable interface settings (such as font size, type, color scheme, and brightness), and enhanced compatibility with assistive technologies like screen readers. These findings may also imply that users are keenly aware of these aspects of the user interface, suggesting that developers and organizations should prioritize addressing these particular issues.

4.6 RQ_6: Outcomes and impacts experienced

Users are in general frustrated because the app was accessible in the previous version and they become less accessible in the new release. In Section 4, we presented the specific accessibility issues introduced by an app update (e.g. small font size, white/bright background, unlabeled buttons). To answer RQ_6, we delve into the practical repercussions experienced by users as a result of these issues. For instance, a common accessibility issue is the small font size, while the consequence is that the user cannot read the content of the button. In this section, we delineate the practical repercussions stemming from the absence of accessibility, as per user testimonials, grouping them based on their resemblance. The likeness among consequences arises from an open coding process conducted on the catalog of consequences identified by researchers.

Global accessibility deterioration. One of the main consequences of the many issues introduced by app updates is the global deterioration of the mobile app accessibility. Many users simply inform developers and organizations that the app is less accessible following an update without giving a more detailed explanation of what barriers or consequences they are facing. Multiple excerpts from our dataset's user reviews illustrate this finding:

> *"I'm legally blind and this upgrade has truly discredited ebay makes it really bad for a visually impaired fellow like me."* (eBay)

> *"As someone with a vision impairment this new update is inaccessible and disappointing."* (Messages)

Difficult to read or see content. Many users reported that they were no longer able to read or see content. Some of them even specify which type of content (e.g. reading books, social network posts) and environment (e.g. in the dark) they have more difficult to adapt to. Here are some excerpts extracted from the user reviews of our dataset that underscore this finding:

> *"The new update made the bubbles so small I can barley even see it! I couldn't even tell someone had texted me until I had looked closer and saw*

> *the EXTREMELY small dot! This is very difficult for people with sight issues!"* (Messenger)

> *"The new version auto adjusts your brightness now every time you switch to the kindle app so users with rooted xooms can't read in the dark anymore."* (Amazon Kindle)

Physical discomfort or pain. Many users said the app cause them some discomfort or actual pain (e.g. eyestrain, headache, nausea, dizziness, migraine). Here are examples from user reviews in our dataset that emphasize this:

> *"I accidentally updated this app on my phone. When I re-entered the app despite my brightness being turned nearly all the way down it physically hurt to look at. Seriously I got a headache within 15 minutes of using it. The new color scheme or lack thereof is so bright that it's almost impossible to use at night due to how much eye strain it gives you. And there's not even a night mode so there's no escape. Smooth move IG team."* (Instagram)

> *"All of the white space between thumbnails now literally makes me dizzy and nauseous when scrolling through files. (...) Please enable some way to adjust the size of thumbnails they are now so hard to see for people who have some type of visual impairment like me."* (Google Drive)

Difficulty to distinguish content. Some users mentioned that they were no longer able to distinguish some content, such as colors, elements and roads in a map, for instance. These user review excerpts from our dataset highlight this point:

> *"Color blind here and previous version had this theme that helped alot with distinguishing roads. (...) Now I don't even bother to use this coz I know m gonna get lost. Very confusing."* (Google Maps)

> *"New colors are terrible looking and harder to tell apart especially with my partial colorblindness. please bring back the original colors"* (Google Keep - Notes and Lists)

Difficulty to navigate the app. Many users pointed out that the new update made the app less accessible thus more difficult to navigate, which includes the inability to interact with the app, find actionable elements or even perceive that the interactive system is ready to respond to voice commands. These excerpts from user reviews in our dataset support this finding:

> *"This new update is crap. (...) Also the menu is lame. It's not easy to navigate. It's no longer in big print super small and at the top of the screen so for someone who's visually impaired such as myself it's extremely difficult to see from far away."* (Pandora - Music & Podcasts)

> *"I'm using a screen reader as I'm visually impaired. In the past I could find the options for YES NO MAYBE very easily when deciding to like someone or not. Now the screen reader reads me all this BUTTONS which I haven't a clue what they are. More or less where on someone's profile does one find the YES NO MAYBE options? I'm going to deactivate and uninstall my baddoo because I now noticed I a screen reader user can't even find where to right a message to someone. So this app is not useful."* (Badoo - Dating & Meet People)

Difficulty to watch media content. Users find it difficult to watch different type of contents after the update. Here are excerpts from the user reviews in our dataset that underscore this finding:

> *"The new update sucks! It is pretty much impossible to watch at night in dark room as the app automatically increase your screen brightness to a level that is very uncomfortable to watch. You can go blind pretty much! It hurts my eyes so bad. Appreciate if you can remove the seperate adaptive brightness function. It is totally unnecessary!"* (Netflix)

> *"I love youtube but there is a weird new update where I cant just click the fullscreen button to make it fullscreen it has a 'rotate' button that does nothing and I have vision problems so I cant watch my videos):. I have tried turning 9n my phones rotate screen option but this doesnt do anything on youtube at all let alone while watching a video."* (eBay: Buy- sell & save money)

Incompatibility with screen reader. Some users reported software and hardware incompatibility with screen reader after an update, which includes delays on the screen reader response, crashes and inability to operate the interface. These are excerpts from user reviews in our dataset that highlight this point:

> *"Since two versions the app is very very laggy it's less responsive With screen readers like TalkBack and voice assistant. The problem still persists this version too"* (Facebook)

> *"Now with the latest Google update there is a conflict with Talkback which makes texting impossible."* (Google)

Difficulty to perceive and operate the app using screen readers. Some users reported that they were no longer able to perceive some content using screen readers due to the lack of accessibility of the new version of the app. Here are some user review excerpts from our dataset that emphasize this:

> *"I am a visually challenged person using Uber with TalkBack. I am not able to find confirm pick up button after this recent update. Please make it accessible at the earliest. I am a daily user."* (Uber - Request a ride)

> *"We visually impaired use ola with screen reader. This updated app sucks. Very difficult to book apply promo or rate the driver. Not accessible for us. Loved the previous versions. 1 star for the updates."* (Ola- Safe and affordable rides)

Difficult using virtual keyboard. Many users were not longer ablet o use the virtual keyboard provided by the Operating System, which is an app itself. Here are some excerpts extracted from the user reviews of our dataset that underscore this finding:

> *"once it was my favorite keyboard but now the new update is made my typing more difficult choosing emoji is also very very tough... please keep the visually challenged people in mind and design further update..."* (Gboard - the Google Keyboard)

> *"The keyboard is too confusing. It's not good for the visually impaired. And it changed on it's own on my phone. I need a keyboard that is easy to use and not microscopic. I want my old keyboard back. It was larger and easier to use. The explanation to change to another keyboard is in technical speak and I'm afraid to change. This keyboard covers what you are typing and that's a really big problem for me."* (Gboard - the Google Keyboard)

Feeling discriminated against. Some users, in addition to detailing the accessibility issues they encountered and the practical implications of the app's lack of accessibility, expressed feeling discriminated against by developers and organizations when they introduce updates that make the app less accessible in a new version. These excerpts from user reviews in our dataset exemplify this finding:

> *"I'd rate lower than zero I hate updates and constant changes. Hard enough being deaf and almost blind now chrome took away cascading tabs now harder to see I call this discrimination against disabled."* (Google Chrome: Fast & Secure)

> *"The new layout stinks! I mean seriously.1. Now the navigation buttons are at the bottom of the screen causing us to see LESS content.2. The buttons are smaller. Do you hate people with vision problems? The elderly? (...) If your goal is to make your application LESS accessible to all and LESS user friendly then you've succeeded. Congratulations. slow sarcastic clap "* (YouTube)

Lose of autonomy. In addition to feeling discriminated against, users report that they lose their autonomy when the app is not accessible once they need to ask for help to complete a task. Here is an excerpt from the user reviews in our dataset that underscores this finding:

> *"Currently the app is completely unusable for people like me and I have to depend on someone else to order which is a risk in the covid times."* (Swiggy : Food Delivery | Instamart Grocery | Genie)

Discussion. Even though the issues reported by users (cf. Section 4) falls in similar categories (e.g. poor color schemes, small font sizes, lack of customization settings, unlabeled buttons), the practical consequences may vary depending on the application and user context. Indeed, it is crucial for users to report the practical consequences resulting from the lack of accessibility, as they serve as compelling evidence of the detrimental effects caused by inadequate accessibility implementation. This underscores the importance of making developers and organizations aware of their responsibility in promoting digital inclusion. Furthermore, some users appear to recognize the significance of user reviews and challenge developers and organizations by accusing them of discrimination. They also emphasize feeling a loss of autonomy in certain situations, which is frustrating for most users especially because the last version seem to be more accessible for them.

4.7 RQ_7: Users demands

RQ_7 delves into the nature of requests articulated by users within their reviews. User reviews can be used to identify bugs to fix and enhancements to implement in future versions of mobile app. When presented in details, the accessibility issues and their impact on users experience described in accessibility update reviews may be enough to generate requirements or demands for next releases. However, some users explicitly call organizations and developers to action. Some of those demands are presented as follows.

Stop removing accessibility features. Users continually asks developers and organizations to stop removing features that make the app more accessible, such as the dark mode or audio alert, for instance. This is an excerpt from a user review in our dataset that highlights this point:

> *"The audio confirmation - that Assistant has heard 'ok Google' and is listening - has been removed… A 2ND TIME. No I'm not going through the rigmarole of filling in a bug report. Last time we all received a shed load of 'oh it must be your phone's fault contact the manufacturer' rubbish only to have the audible alert reinstated after a month… by Google. STOP REMOVING FEATURES... there's a reason blind people need this. Idiots."* (Google)

Fix the accessibility issues. One of the most common explicit demand users make is to fix the accessibility issues introduced by a new update. Here are excerpts from the user reviews in our dataset that underscore this finding:

> *"OMG this new update sucks! I HATE the all white background! Now it's an eye strain viewing my emails! I like the old setup. Please fix this NOW!"* (Gmail)

> *"Latest update almost impossible to use by visually impaired person. Font size at max setting is ridiculously small for someone with low vision - renders app completely unusable. Please fix font size."* (Twitter)

Reward promising. Some users promise to improve their rating (e.g. give it 5 stars) if their request is attended. Here is an excerpt from the user reviews in our dataset that underscores this finding:

> *"Htc Evoshift 4g. 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)

New features. Users commonly ask developers and organizations to include features that will make the app more accessible. These are excerpts from user reviews in our dataset that highlight this point:

> *"Hello I am a Capital One customer who is totally blind. But I have friends who are visually impaired who are also a customer. I would like to know on your next update can you put a dark screen mode button because it is a lot of strain on their eyes with a white screen background"* (Capital One Mobile)

> *"Last update blind me with stupid white status bar. Everything is white. Let us choose background colours. It is just one option in the menu and few lines in its code. Come on."* (Google Chrome: Fast & Secure)

Revert back some feature. Many users ask developers and organizations to give them back some specific feature in future releases. Here are some user review excerpts from our dataset that emphasize this:

> *"Dark mode is gone. My eyes can't take the white background. Please put dark mode back for us with sensitive eyes!"* (Facebook)

> *"This was a great app until they removed the font and background colors! Now its not customizable at all! BRING BACK COLORS!!!"* (Wattpad - Read & Write Stories)

Revert back an entire update. Many users ask developers and organizations to revert back an entire update, possibly making an over statement. These excerpts from user reviews in our dataset exemplify this finding:

> *"Not friendly to old eyes just changed it and I don't like bar graphs and. B S update it still sucks go back to old format"* (AccuWeather: Weather Radar)

> *"PLEASE GET RID OF THE NEW UPDATE. EVERYONE HATES IT! Screen now looks cluttered and is more difficult to use. Less accessible especially for people with eyesight problems. Put it back"* (Snapchat)

Asking for help. Many users explicitly ask for help in their user reviews, so they can effectively perceive and operate the application. Multiple excerpts from our dataset's user reviews illustrate this finding:

> *"Took away dark mode....I have cataracts...I can no longer see the page...HELP!"* (Facebook)

> *"Updated this app and now the text is very small Im sight impaired! Any suggestions on how I can enlarge the text? Prevuous app was fine."* (eBay: Buy- sell & save money)

A warning to the community. Some users use their reviews to warn other users of the accessibility problems introduced by a particular update. Here are examples from user reviews in our dataset that emphasize this:

> *"Every time Google updates anything it sucks worse. Do not update to the April 24 2020 version!!!!!!! Now it's almost unusable."* (Gboard - the Google Keyboard)

> *"DO NOT UPDATE!!! UNINSTALL TO GO BACK TO THE FACTORY VERSION. THE BLINDING APPEARANCE IS HORRIBLE!!! If you do and do not like it you must uninstall to take you back to the factory version. I HATE the new look!!! I have a vision impairment and need more colors."* (Gmail)

Accessibility awareness increase. Some users explicitly ask developers and organizations to be aware that people with disabilities are also users of their mobile apps, hence they should be aware of the accessibility requirements that should be implemented. These user review excerpts from our dataset highlight this point:

> *"Ouch. App now has pure white backgrounds throughout and is very hard on the eyes. (...) They've confused 'simplicity' with 'functionality' and in the process they remove tons of useful UI cues and features. Very frustrating for power users as well as users with visual impairment. Google PLEASE consider users with disabilities by providing low contrast and dark themes at the very least."* (Gmail)

> *"(...) the new color scheme is just idiotic. Don't you think there are people with vision problems out there? Using 3 shades of green seriously? The new sleep score is also completely useless the sleep efficiency made sense. Bring back weight management!!!!!"* (Samsung Health)

Questioning. In addition to making it clear that accessibility bugs needs to be fixed and accessibility features must be implemented, many users ostensibly interrogate developers and organizations upon their actions regarding their decisions on the last update. Some of the questions are rhetorical questions considering design decisions, but some of them refers to their accessibility awareness and discrimination against people with disabilities. Here are some excerpts extracted from the user reviews of our dataset that underscore this finding:

> *"Why remove the custom transparency setting? Why remove the 3-day forecast? Why make the font size so small? Do you hold a grudge against visually impaired people?"* (AccuWeather: Weather Radar)

> *"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. Dumbest update ever. How about making the app usable for people? With the population aging why do something this stupid?"* (Amazon Prime Video)

Advice offering. Some users go beyond the description of the accessibility barriers they found to offer advise on how they should act given the lack of accessibility scenario, which ranges from checking accessibility standards from implementing regression tests. These excerpts from user reviews in our dataset support this finding:

> *"Just wanted to say thanks on behalf of the partially sighted community for fixing the font sizing bug introduced in the previous update. It was a shame it happened. Please regression test properly before releasing. Only 4 stars as there are other problems still outstanding from the last snafu."* (Gmail)

> *"This is a great app other than the fact that the buttons need to be labeled for a visually impaired gamer you can still see to play ps3 / ps4 however can't use a phone well enough. Please remove these buttons with the next update. You can check Google's accessibility development guidelines."* (PlayStation App)

Discussion. The fact that some users explicitly demand concrete actions (e.g. fix bug, revert update, add new feature) from developers and organizations may be an indication that they are aware of the relevance of user reviews on some product's release planning. In addition, users show some knowledge on the development process as they specifically asks for organizations to check accessibility guidelines and run accessibility tests. Furthermore, they confronting organizations concerning their accessibility awareness and discrimination against people with disabilities.

Considering software evolution and the influence of user reviews, it is crucial for users to urge developers and organizations to take responsible action regarding the electronic services and products they create, particularly since these are utilized by a diverse user base, including individuals with disabilities. When users assert their entitlement to accessible products and demonstrate awareness of accessibility standards and development procedures, they actively advocate for improved accessibility in products and services. This stands in contrast to users who choose not to voice their perspectives through user reviews, emails, or social media platforms such as Facebook.

5 Threats to Validity

This section presents the threats to the validity of our study, namely sampling bias and researcher bias.

Sampling bias. Our study inherits the sampling bias of the studies that produced the dataset of accessibility reviews we used [Oliveira *et al*., 2023; Dos Santos *et al*., 2023]. 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. All disagreements were resolved by a joint decision of three or four researchers.

Researcher bias. The coding process is based on manual analyzes conducted by researchers, who can inadvertently impose their own perspectives on the analysis process. To mitigate this threat, we conducted a closed coding process with pre-defined categories and we utilized multiple coders to independently analyze the data. Each reviewer received the complete dataset in a Google Spreadsheet along with detailed instructions and coding samples. In addition, the coding and analysis process is the result of the agreement of at least two out of three researchers.

6 Concluding Remarks and Future Work

This study explored accessibility update reviews to glean insights into how users perceive the influence of app evolution on mobile accessibility. Numerous users highlight the conditions that diminish the app's accessibility when comparing the current version to the previous one, and the practical implications of that, providing evidence that accessibility can regress. Occasionally, users go a step further and challenge developers and organizations, questioning their design decisions and their apparent disregard for users with disabilities. Their frustration expressed in their confrontation is understandable especially because finding an accessible app in an environment where accessibility is scarce, as documented in several studies [Eler *et al*., 2018; Bi *et al*., 2021; Yan and Ramachandran, 2019; Alshayban *et al*., 2020], can be a rare and valuable discovery. Therefore, when these users discover that the app is no longer accessible after an update, it can be deeply disappointing and exacerbate feelings of exclusion and frustration. In expressing their feelings and demands, many users also demonstrate a keen awareness of accessibility standards and practices in software development, such as regression testing, which may strengthen their claims and their impact on software evolution.

Despite the predominance of negative feedback over positive opinions, it is crucial to clarify that this study does not assert a tendency for mobile accessibility to decline as apps evolve. As mentioned earlier, users are more inclined to provide negative feedback than positive remarks. Indeed, there is a lack of conclusive studies on this subject. Previous related works have conducted small-scale studies involving only 60 or 70 apps, yielding disparate results [Alshayban *et al*., 2020; Chen *et al*., 2022].

Rather, our objective with this investigation is to sound a warning and provide evidence that, once a certain level of accessibility is achieved at a particular stage of development, we cannot assume that apps will maintain their accessibility as they evolve. If accessibility is not prioritized from the outset, it is unlikely to be prioritized during the software evolution process. Furthermore, even apps initially designed with accessibility in mind should implement ongoing measures and adhere to accessibility requirements throughout their evolution to prevent the introduction of accessibility barriers.

While our overall understanding of the impact of app evolution on mobile accessibility remains consistent with our prior study, this recent research has provided deeper insights through manual analysis of accessibility reviews. Although ChatGPT-4 can effectively extract specific information from accessibility update reviews due to their concise nature, the manual coding process allowed us to empathize with users' perspectives and grasp the nuances of their specific issues, practical implications and requests. Even though the focus of this paper is in the evolution process, the results of our analysis (e.g. accessibility issues, enhancements, implications and results) can contribute to the body of knowledge of studies that investigate accessibility reviews.

In future work, we intend to further analyze the accessibility update reviews of our dataset to label each review according to the WCAG 2.2 success criteria, which are more specific testable statements that specify requirements for achieving accessibility. That would allow a more precise mapping between specific issues and implications with accessibility standards objectives. Moreover, we intend to label our dataset based on the BBC's Mobile Accessibility Guidelines[7], an accessibility standard more focused on mobile apps and devices. Additionally, we aim to juxtapose the findings from our manual analysis with those generated by employing a Large Language Model (LLM) like ChatGPT-4, assessing its suitability for exploring the intricacies within the context of user review analysis.

Declarations

This study was financed in part by the São Paulo Research Foundation (FAPESP - grant 2022/10910-3).

Authors' Contributions

Alberto Dumont Alves OIiveira: Conceptualization, Methodology, Manual Coding Process, Investigation, Data curation, Writing Original Draft, Writing - Review Editing. Paulo Sergio Henrique dos Santos: Conceptualization, Manual Coding Process, Investigation, Writing Original Draft. Wajdi Aljedaani: Formal analysis, Writing - Review Editing. Marcelo Medeiros Eler: Conceptualization, Methodology, Manual Coding Process, Investigation, Data curation, Writing - Original Draft, Writing - Review Editing, Supervision, Funding acquisition.

[7]https://www.bbc.co.uk/accessibility/forproducts/guides/mobile/

Competing interests

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.

References

Aljedaani, W., Alkahtani, M., Ludi, S., Mkaouer, M. W., Eler, M. M., Kessentini, M., and Ouni, A. (2023). The state of accessibility in blackboard: Survey and user reviews case study. In *Proceedings of the 20th International Web for All Conference*, W4A '23, page 84–95, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3587281.3587291.

Aljedaani, W., Mkaouer, M. W., Ludi, S., and Javed, Y. (2022a). Automatic classification of accessibility user reviews in android apps. In *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)*, pages 133–138. IEEE. DOI: 10.1109/CDMA54072.2022.00027.

Aljedaani, W., Mkaouer, M. W., Ludi, S., Ouni, A., and Jenhani, I. (2022b). On the identification of accessibility bug reports in open source systems. In *Proceedings of the 19th International Web for All Conference*, pages 1–11. DOI: 10.1145/3493612.3520471.

Aljedaani, W., Rustam, F., Ludi, S., Ouni, A., and Mkaouer, M. W. (2021). Learning sentiment analysis for accessibility user reviews. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*, pages 239–246. IEEE. DOI: 10.1109/ASEW52652.2021.00053.

AlOmar, E. A., Aljedaani, W., Tamjeed, M., Mkaouer, M. W., and El-Glaly, Y. N. (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*, pages 1–15. DOI: 10.1145/3411764.3445281.

Alshayban, A., Ahmed, I., and Malek, S. (2020). Accessibility issues in android apps: state of affairs, sentiments, and ways forward. In *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*, pages 1323–1334. IEEE. DOI: 10.1145/3377811.3380392.

Bi, T., Xia, X., Lo, D., Grundy, J., Zimmermann, T., and Ford, D. (2021). Accessibility in software practice: A practitioner's perspective. *ACM Transactions on Software Engineering and Methodology*. DOI: 10.1145/3503508.

Chen, S., Chen, C., Fan, L., Fan, M., Zhan, X., and Liu, Y. (2022). Accessible or not? an empirical investigation of android app accessibility. *IEEE Transactions on Software Engineering*, 48(10):3954–3968. DOI: 10.1109/TSE.2021.3108162.

Ciurumelea, A., Schaufelbühl, A., Panichella, S., and Gall, H. C. (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)*, pages 91–102. IEEE. DOI: 10.1109/SANER.2017.7884612.

Cohen, J. (1960). A coefficient of agreement for nomi-

nal scales. *Educational and psychological measurement*, 20(1):37–46. DOI: 10.1177/001316446002000104.

Dos Santos, P. S. H., Oliveira, A. D. A., De Jesus, T. B. N., Aljedaani, W., and Eler, M. M. (2023). Evolution may come with a price: analyzing user reviews to understand the impact of updates on mobile apps accessibility. In *Proceedings of the XXII Brazilian Symposium on Human Factors in Computing Systems*, IHC '23, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3638067.3638081.

Eler, M. M., Orlandin, L., and Oliveira, A. D. A. (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*, pages 1–11. DOI: 10.1145/3357155.3358477.

Eler, M. M., Rojas, J. M., Ge, Y., and Fraser, G. (2018). Automated accessibility testing of mobile apps. In *2018 IEEE 11th International Conference on Software Testing, Verification and Validation (ICST)*, pages 116–126. IEEE. DOI: 10.1109/ICST.2018.00021.

Fleiss, J. L., Levin, B., Paik, M. C., *et al*. (1981). The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23. DOI: 10.1002/0471445428.ch18.

Gao, J., Li, L., Bissyandé, T. F., and Klein, J. (2019). On the evolution of mobile app complexity. In *2019 24th International Conference on Engineering of Complex Computer Systems (ICECCS)*, pages 200–209. DOI: 10.1109/ICECCS.2019.00029.

Haggag, O., Grundy, J., Abdelrazek, M., and Haggag, S. (2022). Better addressing diverse accessibility issues in emerging apps: A case study using covid-19 apps. In *2022 IEEE/ACM 9th International Conference on Mobile Software Engineering and Systems (MobileSoft)*, pages 50–61. DOI: 10.1145/3524613.3527817.

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

Kraft, F. B. and Martin, C. L. (2001). Customer compliments as more than complementary feedback. *Journal of Consumer Satisfaction, Dissatisfaction & Complaining Behavior*, 14:1–13. Available at: `https://www.jcsdcb.com/index.php/JCSDCB/article/view/101`.

Lehman, M. (1980). Programs, life cycles, and laws of software evolution. *Proceedings of the IEEE*, 68(9):1060–1076. DOI: 10.1109/PROC.1980.11805.

Lehman, M. M. (1996). Laws of software evolution revisited. In Montangero, C., editor, *Software Process Technology*, pages 108–124, Berlin, Heidelberg. Springer Berlin Heidelberg. DOI: 10.1007/BFb0017737.

Liu, T., Wang, C., Huang, K., Liang, P., Zhang, B., Daneva, M., and van Sinderen, M. (2023). Rosematcher: Identifying the impact of user reviews on app updates. *Information and Software Technology*, page 107261. DOI: https://doi.org/10.1016/j.infsof.2023.107261.

McIlroy, S., Ali, N., and Hassan, A. E. (2016). Fresh apps: an empirical study of frequently-updated mobile apps in the google play store. *Empirical Software Engineering*, 21(3):1346–1370. DOI: 10.1007/s10664-015-9388-2.

Nayebi, M., Kuznetsov, K., Chen, P., Zeller, A., and Ruhe, G. (2018). Anatomy of functionality deletion: An exploratory study on mobile apps. In *Proceedings of the 15th International Conference on Mining Software Repositories*, MSR '18, page 243–253, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3196398.3196410.

Oliveira, A. D. A., Dos Santos, P. S. H., Marcílio Júnior, W. E., Aljedaani, W. M., Eler, D. M., and Eler, M. M. (2023). Analyzing accessibility reviews associated with visual disabilities or eye conditions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3544548.3581315.

Othman, A., Dhouib, A., and Nasser Al Jabor, A. (2023). Fostering websites accessibility: A case study on the use of the large language models chatgpt for automatic remediation. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '23, page 707–713, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3594806.3596542.

Palomba, F., Linares-Vásquez, M., Bavota, G., Oliveto, R., Penta, M. D., Poshyvanyk, D., and Lucia, A. D. (2018). Crowdsourcing user reviews to support the evolution of mobile apps. *Journal of Systems and Software*, 137:143–162. DOI: 10.1016/j.jss.2017.11.043.

Potharaju, R., Rahman, M., and Carbunar, B. (2017). A longitudinal study of google play. *IEEE Transactions on Computational Social Systems*, PP:1–15. DOI: 10.1109/TCSS.2017.2732167.

Power, C., Freire, A., Petrie, H., and Swallow, D. (2012). Guidelines are only half of the story: accessibility problems encountered by blind users on the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 433–442, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2207676.2207736.

Reyes Arias, J. E., Kurtzhall, K., Pham, D., Mkaouer, M. W., and Elglaly, Y. N. (2022). Accessibility feedback in mobile application reviews: A dataset of reviews and accessibility guidelines. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3491101.3519625.

Sandoval Alcocer, J. P., Merino, L., Fernandez-Blanco, A., Ravelo-Mendez, W., Escobar-Velásquez, C., and Linares-Vásquez, M. (2024). A developer's guide to building and testing accessible mobile apps. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, FSE 2024, page 713–715, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3663529.3663821.

Santiago, M. T. and Marques, A. B. (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, IHC '22, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3554364.3559114.

Taylor, V. F. and Martinovic, I. (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*, ASIA CCS '17, page 45–57, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3052973.3052990.

W3C, W. W. W. C. (2023). Web content accessibility guidelines – wcag 2.2. Available at: `https://www.w3.org/TR/WCAG22/`.

Yan, S. and Ramachandran, P. (2019). The current status of accessibility in mobile apps. *ACM Transactions on Accessible Computing (TACCESS)*, 12(1):1–31. DOI: 10.1145/3300176.

Yang, A. Z. H., Hassan, S., Zou, Y., and Hassan, A. E. (2022). An empirical study on release notes patterns of popular apps in the google play store. *Empirical Software Engineering*, 27(2):55. DOI: 10.1007/s10664-021-10086-2.