Analyzing Accessibility Reviews Associated with Visual Disabilities or Eye Conditions

ALBERTO D. A. OLIVEIRA and PAULO S. H. DOS SANTOS, University of São Paulo, Brazil
WILSON E. MARCÍLIO JÚNIOR, São Paulo State University, Brazil
WAJDI ALJEDAANI, University of North Texas, USA
DANILO M. ELER, São Paulo State University, Brazil
MARCELO M. ELER, University of São Paulo, Brazil

Accessibility reviews collected from app stores may contain valuable information for improving apps accessibility. Recent studies have presented insightful information on accessibility reviews, but they were based on small datasets and focused on general accessibility concerns. In this paper, we analyzed accessibility reviews that report issues affecting users with visual disabilities or conditions. Such reviews were identified based on selection criteria applied over 179,519,598 reviews of popular apps on the Google Play Store. Our results show that only 0,003% of user reviews mention visual disabilities or conditions; accessibility reviews are associated with 36 visual disabilities or eye conditions; many users do not give precise feedback and refer to their disability using generic terms; accessibility reviews can be grouped into general topics of concerns related to different types of disabilities; and positive reviews are generally associated with high scores and negative feedback with lower scores.

CCS Concepts: • **Human-centered computing** → **Accessibility**; Accessibility design and evaluation methods; • **Software and its engineering** → Extra-functional properties; *Software creation and management*.

Additional Key Words and Phrases: accessibility, mobile application, evaluation, user review, visual disability, app store

ACM Reference Format:
Alberto D. A. Oliveira, Paulo S. H. dos Santos, Wilson E. Marcílio Júnior, Wajdi Aljedaani, Danilo M. Eler, and Marcelo M. Eler. 2023. Analyzing Accessibility Reviews Associated with Visual Disabilities or Eye Conditions. In *CHI '23: ACM Conference on Human Factors in Computing Systems, April 23–28, Hamburg, Germany*. ACM, New York, NY, USA, 21 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Modern software development processes tend to continuously deliver working software in short periods (e.g., days or weeks), which enables development teams to periodically gather feedback from stakeholders to refine and prioritize requirements. In this scenario, software tends to become more mature and stable as bugs are fixed and new features are incorporated in successive deliveries. In mobile software development, the feedback collected from stakeholders goes beyond planned revisions as mobile developers can gather users' opinions from ratings and reviews published in app stores (e.g., Google Play Store).

In fact, many studies have shown that mobile developers harness user reviews to plan for new releases because it might lead to better ratings and reputation. They usually inspect user reviews to retrieve information of bug descriptions,

Manuscript submitted to ACM

feature requests, network issues, privacy and security concerns, resource consumption problems (e.g., battery), and user interface difficulties [12, 15, 22, 23, 28, 30, 33, 36–38, 40, 43].

In particular, considering that mobile apps have poor accessibility in general [7, 17, 49] and user reviews can drive app evolution, some researchers have investigated whether and how accessibility is addressed in user reviews. Eler et al. [16] analyzed 214,053 user reviews from 701 apps to identify reviews associated with accessibility, also known as *accessibility reviews*. They concluded that only 1.24% of the reviews were, in fact, accessibility reviews, and they were concentrated in a small subset of popular apps and topics. Alshayban et al. [7] also analyzed 704 user reviews to understand accessibility concerns. AlOmar et al. [6] and Aljedaani et al. [2] used the reviews identified by Eler et al. [16] as a training dataset to propose techniques to automatically identify accessibility reviews.

Previous studies on accessibility reviews have shed light on this topic, but there is still room for valuable contributions in this research field. First, previous investigations have been conducted on apps with a small user base and, consequently, a low number of evaluations, meaning the samples might not be representative. Hence, studies on larger datasets are desirable. Second, the related works have focused on general accessibility concerns expressed by any user instead of specific concerns associated with actual barriers [7, 16].

Accessibility affects any user, but poorly designed interfaces impose barriers mostly to people with disabilities or eye conditions. Hence, in this paper, we present an investigation we set up to identify and analyze accessibility reviews that report issues that affect users with visual impairments or eye conditions because they are more likely to contain real barriers instead of user preferences. To mitigate the limitations of small datasets, our study was conducted on 179,519,598 reviews extracted from 340 Android apps.

The main goals of our study are: i) to show that accessibility issues reported in user reviews can be associated with visual disabilities or eye conditions, thus providing more evidence of how poorly designed interfaces might severely affect many users; ii) to provide an overview of the main barriers and topics addressed in accessibility reviews, as well as the main interface resources and components involved; iii) understanding how scores assigned by users are distributed among accessibility reviews; iv) providing a large dataset to foment further investigation into this topic. More specifically, we framed our investigation on five research questions:

RQ1: How many accessibility reviews are associated with visual disabilities or eye conditions?
Our aim is to provide evidence of whether and how often users associate positive or negative feedback concerning the accessibility of the app with a visual disability or eye condition. Identifying whether a review is associated with a visual disability or eye condition, however, is not a trivial task given the lack of detailed information usually found in user reviews. Therefore, in our first effort to characterize such a type of user feedback, we focused on reviews in which disabilities and conditions are explicitly mentioned by the users themselves and not according to our own interpretation.

RQ2: What are the disabilities or eye conditions associated with the accessibility reviews?
Our purpose is to characterize the accessibility reviews based on the visual disabilities or eye conditions mentioned in the user reviews. This knowledge can provide evidence that users may present different eye conditions and, therefore, might have different concerns when it comes to the app interface. To answer this question, we manually analyzed each review identified in the previous research question to assign a label based on the visual disability or eye condition mentioned by users.

RQ3: What are the main topics addressed by the accessibility reviews associated with visual disabilities or eye conditions?

Our goal is to identify users' concerns commonly expressed in the accessibility reviews. As manual content analysis is labor intensive, in this work, we resorted to topic modeling techniques that consist of automatically processing data (e.g., text) to summarize information and understand the main topics that stand out in well-defined clusters of data [20].

RQ4: What are the interface components and resources mentioned in the accessibility reviews associated with visual disabilities or eye conditions?
As many accessibility guidelines are oriented by interface components and resources, our aim is to identify which components and resources are mostly mentioned in accessibility reviews to provide some indication of which of them might be more relevant in this context.

RQ5: What are the scores of accessibility reviews associated with visual disabilities or eye conditions?
Our intent is to understand the associations between accessibility reviews and the scores assigned to the app. In particular, we want to know whether positive feedback leads to high scores and negative feedback leads to low scores, which would emphasize the difference between this type of reviews from those found in previous studies in which users express more preferences concerning accessibility than real barriers.

The main contributions of this paper are:

- We introduce a dataset of nearly 180 million user reviews from which we extracted 4,999 accessibility reviews that express concerns that affect users with visual disabilities or eye conditions. This is the largest dataset available so far with respect to accessibility review studies. This labeled dataset can be more extensively explored in further and future investigations[1].
- We provide evidence that some users emphasize how some accessibility issues may affect users with disabilities or eye conditions. In total, users associated their positive or negative feedback with 36 specific types of disabilities or eye conditions.
- We present the main concerns that stand out from accessibility reviews by using topic modeling techniques. The topics we identified show that users with different disabilities may share the same concerns when using apps of different categories. Some concerns, however, are concentrated on a few types of issues with a specific type of app (e.g., users with color blindness find it challenging to read traffic information in navigation apps).
- We show the interface resources and components (e.g. button, icon, color) that are mostly mentioned by accessibility reviews, which might be an indication of which associated guidelines can have more impact on users with disabilities or eye conditions.
- We show that positive feedback are usually associated with high scores, while negative feedback tend to be associated with low scores. In previous work, in which most reviews expressed users' preferences rather than real barriers, these associations were not observed [16].

This paper is organized as follows. Section 2 presents the related work and how our study compares to state of the art on accessibility reviews. Section 3 outlines the materials and the methods we used to conduct our investigation and to answer our research questions. Section 4 shows and discusses the results of our investigation. Section 5 discusses the threats to the validity of our investigation. Finally, Section 6 presents some concluding remarks and future directions.

[1]The dataset is available at https://github.com/marceloeler/data-paper-chi23

2 RELATED WORK

Previous studies have examined the identification of accessibility on user reviews using machine learning [2, 3, 5, 6] and analyzed how real-world professionals view accessibility during the design and development process [9, 26]. Others investigated the accessibility concerns that arise in Android applications [7, 14, 48]. However, to the best of our knowledge, there is no study analyzing user reviews in Android applications for the vision impaired people.

This section highlights several prior studies that particularly shaped our methodology. Next, we divide the related work into three different aspects of accessibility on Android: mobile accessibility for the vision impaired people, where we focus on current approaches used to examine the vision impaired people; accessibility in user reviews, which focuses particularly on empirical experiments on accessibility in user reviews; topic modeling on mobile apps; which briefly discusses the significance of topic modeling in a similar context.

2.1 Mobile Accessibility for the Vision Impaired People

Previous studies have examined the design of mobile environments for the vision impaired people. For example, Barbareschi et al. [8] examined the interaction of visually challenged individuals with mobile devices in their everyday life. Their method consisted of observing the interaction and conducting semi-structured interviews. Park et al. [41] studied the difficulties and obstacles vision impaired individuals encounter when using mobile devices and applications. They presented ten accessibility heuristics for the development of mobile applications. Other researchers observed an iPhone user in order to determine the user's demand for combined visual and motor impairment [18].

Christy and Pillai [11] analyzed the rating for 57 iOS and Android applications beneficial for individuals with visual impairment. Some other works focus on the lack of requirement for the mobile application for the vision impaired people [44]. In India, Pandey et al. [39] surveyed 124 college students with visual impairment to explore the use of mobile devices and apps for visual impairment. Our work differs from theirs in both purpose and methodology. None of the prior studies analyzed Android mobile app user reviews' visual impairment reviews. In addition, previous studies relied on either observing or surveying vision impaired individuals, whereas our approach investigates around 180 million Android reviews to uncover accessibility reviews associated with vision impaired people.

2.2 Accessibility in User Reviews

Accessibility problems are commonplace, even among fully developed apps [17, 49]. Although user reviews such as Google Play Store can be an effective tool for the advancement of mobile apps, only 1.24% of mobile app users document accessibility problems to application stores [16]. In a study similar to ours, Eler et al. [16] used a string-matching method on 214,053 reviews of Android mobile apps to determine if users reported accessibility-related complaints to the app store. Furthermore, AlOmar et al. [6] performed a supervised machine learning to automatically identify user reviews as binary classification accessibility and non-accessibility related. In another study, Aljedaani et al. [2] used machine learning to categorize 2,663 reviews of accessibility apps according to four guidelines: Principles, Audio/Video, Design, and Focus. In another study, Aljedaani et al. [5] developed an automated model for classifying accessibility-related app reviews based on sentiment analysis.

Alshayban et al. [7] conducted a manual analysis of 704 user reviews from 102 Android applications. They found that users reported issues related to missing labels, text size, and color contracts. More recently, Aljedaani et al. [3] proposed a classification-based approach for automatically detecting accessibility bug reports. As a result, they found 2,567 accessibility-related bug reports across seven different open-source projects. Comparable to their method, ours

analyzes user reviews to discover those connected explicitly to the vision impairment rather than focusing solely on accessibility reviews in general. In addition, our case study was conducted on the largest datasets available, with regards to Google Play Store user reviews (nearly 180 million reviews). As a result, our dataset exceeds that of previous research in terms of size and diversity.

2.3 Topic Modeling on User Reviews

Topic Modeling has been utilized in various studies, focusing on the analysis of source code [34], issue localization [46], tweet understanding [29, 42], and mobile bug reports [1, 4]. For example, Kirilenko et al. [24] examined 14,273 visitor reviews from TripAdvisor. Nguyen and Hovy [35] analyzed 53,273 user reviews from the Best Buy US website for smart speakers. An empirical investigation was undertaken by Calheirosa et al. [10] on 3,179 reservations from Areias do Seixo hotel. Sutherland and Kiatkawsin [45] extracted meaningful and actionable insight from customers on Airbnb. Our work employs a similar methodology; however, we utilized the topic modeling technique for a different objective and type of data. Specifically, we intend to identify the different subjects in reviews associated with vision disabilities themes based on the most frequent topics and their corresponding probabilities.

3 STUDY DESIGN

The focus of our investigation is accessibility reviews that contain issues affecting people with disabilities or eye conditions. In this section, we present the cross-sectional study we designed to analyze accessibility reviews selected from a pre-defined set of reviews extracted from the Google Play Store[2]. Following, we present our sampling process in detail and briefly describe the topic modeling method we adopted in our investigation.

3.1 Sampling

We employed a purposive sampling method [47] to select accessibility reviews of the most popular apps on the Google Play Store. We selected reviews of the most download apps because they might be used by a wider range of user types, thus increasing the chances of retrieving more accessibility reviews of interest. Aiming at achieving a more diverse sample, we collected reviews of the 10 most downloaded apps of each category defined by the Google Play Store.

The API we used to extract data from the Google Play Store classifies the most popular apps at that moment into the category "APPLICATION". In our study, the following apps were classified within that category: Facebook, Messenger, Instagram, Netflix, Snapchat, Spotify, Twitter, WhatsApp Messenger, TikTok, and Telegram. Even though many of those apps could also be classified in other categories, such as "SOCIAL" and "COMMUNICATION", we followed the API classification to increase the number of categories. In total, we selected 340 apps from the Google Play Store.

The sampling process was conducted in four steps presented in Figure 1. In Step 1, we created an initial subset of 179,519,598 user reviews extracted from the 340 apps selected for our study. In Step 2, we performed string-matching filtering to identify possible reviews of interest. In Step 3, as string-matching filtering based on keywords can retrieve many false positives, we conducted a manual inspection to apply inclusion and exclusion criteria to keep only the accessibility reviews of interest, i.e., reviews that contain issues affecting people with disabilities or eye conditions. Finally, in Step 4, we labeled the accessibility reviews based on the visual disabilities or eye conditions mentioned by the users. In addition, we assigned reviews with a positive (compliments) or a negative (complaints or requests) label. The details of each step are presented as follows.

[2]https://play.google.com/

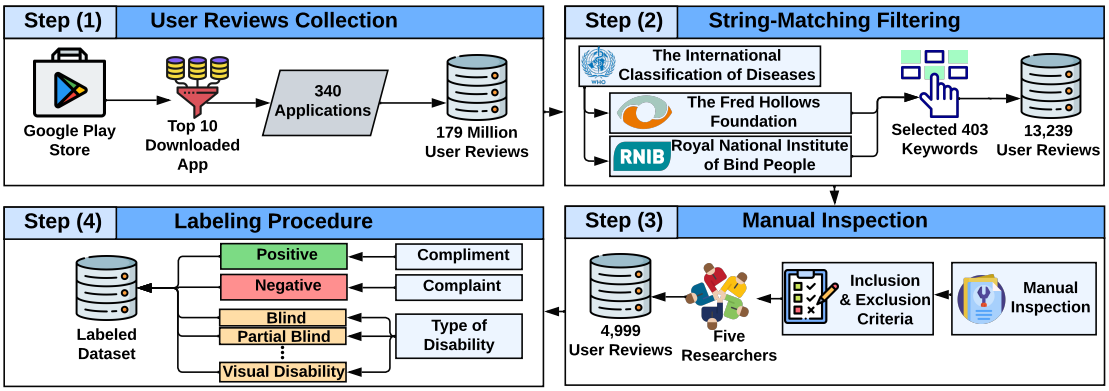

Fig. 1. Overview of each step of our sampling process.

Table 1. Some characteristics of our initial sample of 179,519,598 reviews extracted from the 340 apps

Metric	Min.	Mean	S.D.	Median	Max.
Downloads	1,000,000	429,388,235	1,311,932,997	100,000,000	10,000,000,000
Ratings	2136	4,666,658	1,693,638	1,156,199	151,032,300
Reviews collected	81	527,998	15,577,017	127,237	16,966,680
Score	2.3	4.3	0.4	4.4	5.0

3.1.1 ***Step (1): User Reviews Collection.*** We collected 179,519,598 reviews from the 340 apps selected for our study[3] using an unofficial Google Play Store API[4]. Table 1 shows some details of this dataset. The less popular apps (e.g., Maps for Minecraft PE) were downloaded by around 1 million users, and the most popular apps (e.g., Google Maps) were installed by around 10 billion users. The app with the least number of both ratings and reviews is "Maps for Minecraft PE", with 2136 ratings and 81 reviews[5]. The app with the most number of ratings is "Youtube" (151 million) and the app with the most number of reviews it the "Whatsapp" (17 million). The average score is 4.3. The lowest score belongs to the app "Google Classroom", while the highest score belongs to the app "Facemoji Emoji Keyboard".

3.1.2 ***Step (2): String-Matching Filtering.*** We employed string-matching filtering based on specific keywords to narrow down our dataset to reviews that mention visual disabilities or eye conditions. The list of keywords used in this process has large impact on the results, thus we resorted to glossaries and catalogues of technical and medical terms related to visual disabilities maintained by well-known international organizations to identify. Our main reference was the "Chapter VII - Diseases of the eye and adnexa" [6] of the "International Statistical Classification of Diseases and Related Health Problems" document, produced by the World Health Organization (WHO).We also used the "Glossary of Eye Conditions"[7] of the centenary American Foundation for the Blind (AFB), and the Eye Conditions list of the Royal

[3]All data were extracted by the end of December 2021
[4]https://github.com/facundoolano/google-play-api
[5]The number of ratings is different from the number of reviews because users can evaluate the app by solely assigning a score.
[6]https://icd.who.int/browse10/2016/en#/VII
[7]https://www.afb.org/blindness-and-low-vision/eye-conditions

National Institute of Blind People[8]. As we proceed to check glossaries and catalogs from other organizations (e.g., Fred Hollows Foundation[9]), we noticed that we have reached a saturation point where no new keyword was included.

Some conditions that are not visual disabilities appear in accessibility reviews very frequently [16], and they can severely impact the perception, understanding, and operation of an app. Therefore, we also included keywords related to visual conditions commonly associated with the use of digital products (e.g. light sensitivity, photophobia, eyestrain); and keywords related to informal expressions commonly used by older users, such as "weak eyes" and "old eyes", that represent some imprecise condition or disability. Moreover, we included keywords associated with assistive technologies commonly used by people with visual disabilities (e.g., screen reader, braillenote). In total, we defined a list of 403 keywords. By applying string-match filtering, we narrowed down our dataset from around 180 million to 13,239 reviews.

3.1.3 ***Step (3): Manual Inspection***. String-matching filtering may result in several false positives once the presence of a keyword related to a visual disability does not necessarily mean that review is related to accessibility. Manual inspection was thus conducted by five researchers to filter out reviews that are not of interest. The two less experienced researchers received 1,000 reviews to inspect while the remaining reviews were split among three more experienced researchers in the area. For this task, we defined inclusion and exclusion criteria to standardize the decision process.

The inclusion criterion defines that our final sample must contain only user reviews related to accessibility and that somehow affect users with a visual disability or eye conditions (e.g., *"Please add an option for increasing the font size. I'm visually impaired and I have a hard time reading messages." Hangouts*). In that sense, we also included reviews that refer to assistive technologies generally used by people with disabilities (e.g., *"Movies titles are not readable with talk back." MX Player*). The exclusion criteria define that reviews must be excluded basically in four cases: i) if the disability or condition mentioned is not related to a report of accessibility concerns (e.g. *"The driver very professional and kind. I'm visually impaired he was very helpful" - Uber*); ii) if the user are not clear whether they review is related to the app accessibility or just its overall functionality (e.g., "Even blind can benefit" - *Waze*). iii) if the review is not complete or does not make sense (e.g.,'textit'I'm Legally Blind" - *Pinterest*); and iv) if the a visual disability-related term does not refer to an actual visual disability or eye condition (e.g.,*"It leaves the consumer* blind *to the price until the car is already booked." - Lyft*).

During the manual inspection, whenever a researcher could not decide whether a review should be included or not, a joint decision was made. After the first round of manual inspection, two more experienced researchers read all reviews to check whether the inclusion and exclusion criteria were met. In that sense, the resultant sample, a set of 4,999 accessibility reviews, is the result of a process for which all researchers agreed 100%.

3.1.4 ***Step (4): Labeling Procedure***. We manually labeled the 4,999 accessibility reviews of our sample to include to associate each review with a visual disability or eye condition, and with a positive or negative feedback. Labels concerning visual disabilities or eye conditions were assigned based on the disability or condition mentioned by the users themselves and not by our interpretation. Aiming at reducing the number of labels applied over our sample, we used the glossaries and catalogs mentioned in Section 3.1.2 to group different reviews under the same label. For instance, "near sight" and "short sight" are synonyms of "myopia". Table 2 shows examples of how we categorized some types of disabilities.

There were only two cases we assigned labels based on our interpretation: i) we assigned the generic label "visual impairment" to reviews in which assistive technologies commonly used by visual impairment are mentioned but no

[8]https://www.rnib.org.uk/eye-health/eye-conditions
[9]https://www.hollows.org/au/eye-health/glossary-of-eye-conditions

Table 2. Examples of how disabilities were categorized throughout the labeling procedure.

Label Type	Type of Disability			
Blindness	Blind			
Partial blindness	Bit blind	Blind in one eye	Partial blind	
Legal blindness	Nearly blind	Quite blind	Severe sight impaired	
Visual impairment	Visually challenge Visually handicapped Less vision Vision issues	Eye defect Visual disability Sight issue Vision disabled	Eye disease Vision problem Sight problem Genetic eye disease	Impaired eye Eyesight problem Sight loss
Myopia	Short sight	Near sight		
Photophobia	Light sensitivity	Sensitive vision	Sensitive eye	
Partial vision	Partially sighted	Sight impaired		

disability or condition is named by the user; ii) we assigned the label "possibly photophobia" when users complain that a white background "blinds them" at night but no specific disability or condition is mentioned (e.g., actual photophobia or low vision). In the second case, the term "blind" does not refer to an actual disability, but we kept it anyway because for a specific reason: even though any user can benefit from a dark mode that reduces brightness in a low light environment, in our perception, if the user emphasizes the such negative effect of a bright screen, it might indicate a temporary effect or condition. As no condition is explicitly mentioned, we labeled such reviews as "possibly photophobia" to distinguish it from cases in which the user clearly states that the issue affects someone with photophobia.

When it comes to labeling reviews with the "positive" or "negative" label, our reasoning is that a positive review is usually a compliment on the apps accessibility (e.g., *"Calls are clear and high-quality when the connection is good accessible controls very blind friendly." Skype*), and a negative review is usually a feature request or a criticism concerning some accessibility barrier (e.g., *"Many features including edit boxes NOT accessible with screen readers." Facebook*).

We followed the procedure outlined in [27] to further validate the collected data. From among the total of 4,999 user reviews, we selected 250 to represent the 5% of user reviews we analyzed. This quantity corresponds to the sample size with a 95% confidence level and a 6 confidence interval. Three authors independently completed the labeling process. The same group of reviews was categorized as being either connected if the reviews met our inclusion and exclusion criterion and the two characteristics labeled negative and positive, and that designation was given to all three authors. The chosen reviews were not previously exposed to the authors.

To prevent exhaustion, the analysis process lasted 14 days. In order to determine the percentage of areas in which the authors agree and disagree, we cross-checked the findings of the manual labeling. In all circumstances of disagreement, a fourth author is requested to to break the tie. For the purpose of determining the extent to which the raters agreed upon the classifications, we used Cohen's Kappa coefficient [13]. We acquired a degree of agreement of 0.84. According to Fleiss et al. [19], these agreement values are nearly *perfect agreement* (*i.e.*, 0.80˘1.00).

3.2 Topic modeling

Manual content analysis of user reviews may be laborious. Thus, we leverage topic modeling techniques to automatically process text collections and summarize information, as well as to identify common themes in the specific domain of our sampled reviews. Figure 2 shows the two main steps of our procedure (Topic Modeling and Topic Analysis).

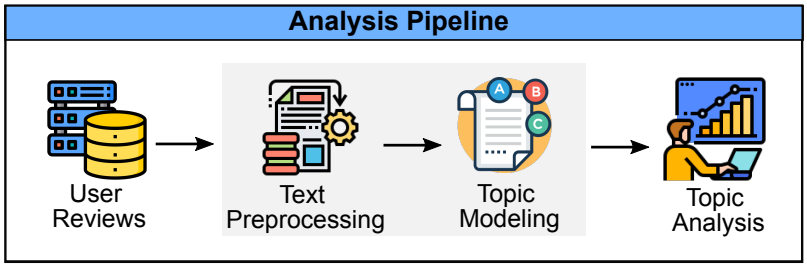

Fig. 2. Topic Analysis for understanding the reviews. The filtered reviews are preprocessed and topics are generated for analysis. Using the similarity between the topics and reviews, a detailed understanding of the topics' subjects is also considered.

First, the Text Preprocessing step removes information that could decrease the performance of the topic extraction technique or does not change the meaning of a review—we removed stopwords and concatenated common expressions (e.g., dark mode). Then, Topic Modeling computes topics composed of words that better represent a group of reviews, thus providing an overview of all user reviews and enables analysts to inspect the main terms related to each topic.

In this paper, the Topic Modeling step uses BERTopic [20] to extract topics from the reviews. BERTopic uses a pre-trained language model to compute embeddings (dense vectors) to represent each document, structuring the text collection on a vector space model. After that, UMAP [32] computes a lower-dimensional space by preserving local and global features from the high-dimensional space. From the reduced space, BERTopic uses HDBSCAN [31] to compute clusters of documents from which a variation of TF-IDF, called class-based TF-IDF, computes the importance of terms in each cluster resulting in a topic for each cluster.

4 STUDY RESULTS

This section presents and discusses the results of our investigation. For each research question, we present the question, the results of the analysis, and we discuss the findings.

4.1 RQ1: How many accessibility reviews are associated with visual disabilities or eye conditions?

Results. We identified 4,999 accessibility reviews that express issues that affect users with visual disabilities or eye conditions, which represents only 0,003% of our initial dataset. In total, 936 reviews (18.7%) were positive, and 4,063 reviews (81.3%) were negative feedback. Only 228 out of the 340 apps we analyzed have at least one accessibility review. Table 3 shows the top 10 apps that most received accessibility reviews. Accessibility reviews are not equally or normally distributed among apps. Most apps have less than 50 reviews, while the top 20 most evaluated apps have 56% of them. The mean number of accessibility reviews per app is around 22, and the median is 6. Facebook is the most evaluated app with 376 accessibility reviews, followed by WhatsApp (233) and Amazon Kindle (203). Many apps received only one accessibility review, such as Alibaba, Sky Map, and Tripadvisor. For most apps, negative reviews are predominant.

Table 4 shows the number of accessibility reviews of the top 10 categories that most received accessibility reviews. "APPLICATION" is the category with the highest number of accessibility reviews, followed by "ANDROID WEAR" (e.g., Clock, Google Keep, Shazam, and YouTube Music) and "BOOKS AND REFERENCE" (e.g., Amazon Kindle, Google Play Books, and Wikipedia). Conversely, the category with the least number of accessibility reviews is "PARENTING" (e.g., FamilyAlbum and Pregnancy Tracker), followed by "EVENTS" (e.g., Ticketmaster and Vivid Seats). The number of negative reviews is predominant in all categories, except for "BOOKS AND REFERENCE".

Table 3. Number of accessibility reviews of the top 10 most evaluated apps

	SAMPLE		POSITIVE		NEGATIVE	
App	Reviews	Percent.	Reviews	Percent.	Reviews	Percent.
Twitter	117	2.3%	18	15.4%	99	84.6%
YouVersion Bible	129	2.6%	101	78.3%	28	21.7%
Instagram	132	2.6%	9	6.8%	123	93.2%
YouTube	169	3.4%	31	18.3%	138	81.7%
Google Chrome	187	3.7%	15	8.0%	172	92.0%
Gmail	196	3.9%	10	5.1%	186	94.9%
Gboard	199	4.0%	73	36.7%	126	63.3%
Amazon Kindle	203	4.1%	61	30.0%	142	70.0%
WhatsApp	233	4.7%	34	14.6%	199	85.4%
Facebook	376	7.5%	13	3.5%	363	96.5%

Table 4. Number of accessibility reviews of the top 10 most evaluated app categories

	SAMPLE		POSITIVE		NEGATIVE	
Category	Reviews	Percent.	Reviews	Percent.	Reviews	Percent.
MUSIC AND AUDIO	132	2.6%	26	19.7%	106	80.3%
PRODUCTIVITY	141	2.8%	30	21.3%	111	78.7%
SOCIAL	144	2.9%	5	3.5%	139	96.5%
MAPS AND NAVIGATION	145	2.9%	17	11.7%	128	88.3%
FOOD AND DRINK	173	3.5%	22	12.7%	151	87.3%
VIDEO PLAYERS	225	4.5%	47	20.1%	178	79.9%
COMMUNICATION	509	10.2%	56	11%	453	89%
BOOKS AND REFERENCE	519	10.4%	233	44.9%	286	55.1%
ANDROID WEAR	627	12.5%	161	25.7%	466	74.3%
APPLICATION	1276	25.5%	103	8.1%	1173	91.9%

Discussion. Accessibility reviews reporting issues that affect users with disabilities or eye conditions are very scarce. Even popular apps such as "Trivago", have not received any accessibility review despite their 50 million user base. Compared to previous studies, our investigation shows that this type of accessibility reviews are even more rare than general accessibility reviews, which, in previous studies, accounted for 1.2% of all reviews collected.

The scarcity of accessibility reviews may be consequence of our scope choice for this investigation. During the string-matching process, we only selected user reviews that explicitly mention some disability or condition. There may be many accessibility reviews reporting issues that affect users with disabilities but no disability is mentioned. Identifying such reviews is out of the scope of this paper.

On the other hand, the shortage of accessibility reviews might give the wrong idea that mobile apps are accessible in general, which is inaccurate since many studies have reported that even fully developed apps present trivial accessibility barriers [17, 49], and consumer silence is not necessarily indicative of satisfaction and unhappiness since customers are unlikely to share their opinions [21]. Such results motivate further investigation on why users affected by accessibility barriers do not usually give feedback, and whether evaluation mechanisms are accessible.

Table 5. Number of accessibility reviews of the top 20 most mentioned disabilities or eye conditions

	SAMPLE			POSITIVE		NEGATIVE	
Disability/Condition	Reviews	Percent.	Apps	Reviews	Percent.	Reviews	Percent.
snow blindness	4	0.1%	2	0	0.0%	4	100.0%
retinopathy	5	0.1%	4	4	80.0%	1	20.0%
dry eyes	9	0.2%	7	5	55.6%	4	44.4%
glaucoma	13	0.3%	10	5	38.5%	8	61.5%
macular degeneration	13	0.3%	10	8	61.5%	5	38.5%
myopia	21	0.4%	17	2	9.5%	19	90.5%
weak eyes	28	0.6%	17	7	25.0%	21	75.0%
astigmatism	30	0.6%	21	2	6.7%	28	93.3%
cataract	31	0.6%	17	15	48.0%	16	52.0%
partial blindness	44	0.9%	24	19	43.2%	25	56.8%
partial vision	85	1.7%	50	22	25.9%	63	74.1%
old eyes	106	2.1%	48	29	27.4%	77	72.6%
legal blindness	125	2.5%	53	51	40.8%	74	59.2%
color blindness	134	2.7%	52	9	6.7%	125	93.3%
low vision	211	4.2%	81	38	18.0%	173	82.0%
possibly photophobia	331	6.6%	69	14	4.2%	317	95.8%
photophobia	335	6.7%	72	28	8.4%	307	91.6%
eyestrain	559	11.2%	96	76	13.6%	483	86.4%
blindness	823	16.5%	138	208	25.3%	615	74.7%
visual impairment	2207	44.1%	186	430	19.5%	1777	80.5%

The fact that most accessibility reviews are negative is not surprising since people are more inclined to give negative than positive feedback [25]. It seems that negative feedback is more common in more visual apps where users have to deal with images, videos and complex interaction (e.g. social media apps), while positive feedback is more predominant in apps that are mostly based on text (e.g., Bible app, Amazon Kindle), as in the category 'BOOKS AND REFERENCE".

4.2 RQ2: What are the disabilities or eye conditions associated with the accessibility reviews?

Results. The 4,999 accessibility reviews we identified are associated with 36 different types of disabilities or eye conditions. Table 5 shows the number of accessibility reviews related to the top 20 most mentioned disabilities or eye conditions. Visual impairment is by far the most commonly mentioned disability as it is related to 2207 reviews, followed by blindness (823) and eyestrain (559). Many disabilities or conditions were mentioned less frequently (once or twice), such as lazy eyes, night blindness, retinal detachment, aniridia, blurry vision, cornea transplant, damaged retina, distorted vision, double vision, hypermetropia, keratoconus, lattice degeneration, light blindness, protanomaly, protanopia, and white blindness. For most disabilities, reviews are more negative than positive feedback.Some disabilities appears in reviews of several apps (Column 4 of Table 5). Around 81% of the apps received accessibility reviews associated with visual impairment, while 60% are related to blindness and 42% with eyestrain.

Figure 3 presents a heatmap with the distribution of the most frequently mentioned disabilities among the most evaluated apps. Each cell indicates the number of reviews associated with that app and condition. For instance, Facebook has 136 reviews associated with visual impairment, and WhatsApp Messenger has 13 reviews with low vision.

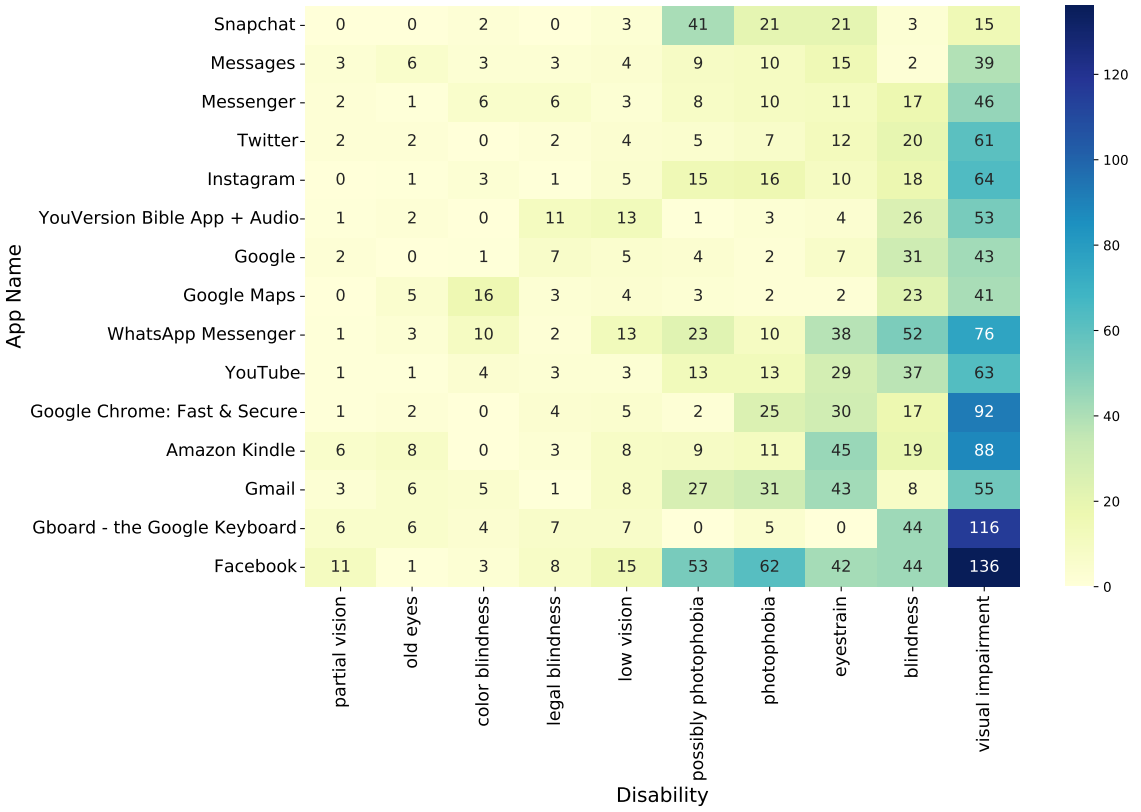

Fig. 3. Heatmap encoding the information of the number of reviews associated with a specific visual disability or eye condition and app.

Discussion. It appears that many users are concerned with linking their feedback with some visual disability or eye condition to make it clear they are facing real barriers, thus further motivating developers to fix the reported issues. While some users mention specific disabilities (e.g. myopia), many of them employ generic terms (e.g., visual impairment).

The analysis of the reviews of predominant disabilities might give evidence of common barriers faced by users in different conditions. For instance, Figure 3 shows that the number of reviews associated with color blindness and Google Maps is higher than for other apps, which can indicate that this app might be using colors to convey meaningful information. In particular, many Google Maps reviews refer to problems in visualizing traffic routes (e.g., "*needs a differentiation between traffic routes & satellite image similar to the previous versions. 60% of us are color-blind*", "*needs to have a traffic option for red/green color blindness. right now good and bad traffic look the same, perhaps a color picker.*").

In addition, we noticed that some conditions (e.g., eyestrain, photophobia) are more common in apps that users tend to spend more time on and at any time of the day, such as Facebook, Kindle, YouTube, WhatsApp Messenger, and Gmail. In such cases, organizations should design interfaces and interactions to prevent some effects, such as eyestrain, and enable different themes (e.g., dark or night mode) for those with photophobia or related conditions.

4.3 RQ3: What are the main topics addressed by the accessibility reviews associated with visual disabilities or eye conditions?

Results. We employed topic modeling to provide a general overview of the underlying subjects associated with the 4,999 reviews we processed and Topic Analysis to understand the meaning each topic conveyed. Table 6 presents twelve topics we automatically extracted from our sample. Each topic has a readable name (Column 1), which was manually defined based on common terms (Column 2) or the predominant visual disability or eye condition associated with that topic. For instance, the topic ***Font Size*** comprises several reviews concerning the font size of the app, and most reviews of the topic ***Color Blindness*** mention that particular disability. We also present a general explanation of what each topic conveys (Column 3), as well as the number of reviews associated with that topic (Column 4).

Discussion. Topic Analysis is valuable for revealing interesting subjects of accessibility reviews, guiding a broader qualitative analysis. We illustrate the benefits of the approach by analyzing some topics presented in Table 6.

- **Visual Impairment (I and II)**. This topic comprises positive and negative feedback expressing how vision impaired users can easily or poorly use the app, but no specific detail is provided (e.g., *"should be more accessible for visually challenged." - Instagram*). For understanding the meaning of these reviews, one needs to rely on the context and other information to identify weaknesses and strengths of the app accessibility. It is not surprising that two major generic topics emerged as almost half of our sample has the "visual impairment" label.
- **Eyestrain**. This topic summarizes reviews related to requests for dark mode features to reduce eyestrain (e.g., *"please give dark mode for this app it's so eye straining to watch in meet." Google Meet*). This topic differs from the ***Night Sensitivity*** because, in the latter, reviews are associated with specific disabilities and conditions. It is worth mentioning that, even though eyestrain is not a visual disability, we include it because it is a recurrent effect reported by many users. In fact, it is the third most frequent condition found in our sample.
- **Color Blindness**. This topic summarizes reviews in which users mention their inability to differentiate some colors. The topic also presents the terms *red*, *green*, and *blue*, which might show some evidence of the most common types of color-blindness—*red-green* is more usual than *blue-green* color blindness [10]. Another interesting term in this topic is *traffic*. For example, the heatmap presented in Figure 3 may indicate it refers to Google Maps reviews. In fact, the reviews *"please offer a color blindness mode for traffic. a pattern instead of yellow/green/red would be greatly appreciated!"* and *"needs to have a traffic option for red/green color blindness. right now good and bad traffic look the same. perhaps a color picker"* confirms that using color blindness-friendly color encoding or customization features must be considered when developing apps for a broader audience.
- **Photophobia**. This topic has terms (e.g., *eyes*, *sensitive*) that gives evidence that some features usually requested due to user's preferences, such as dark mode, may also be related to some types of disabilities (e.g., *"great app one question when will there be a dark theme because I have very sensitive eyes." Google Drive*, *"i have sensitive eyes, and the dark mode was the only thing that made this app tolerable. please give back dark mode." Facebook*).

Some of the topics found in our analyzes have also been identified in previous work. For instance, the most predominant topic identified in the work of Eler et al. [16] was "Theme/Mode", which usually refers to features such as dark or night mode. This topic is strongly related to some topics identified in Table 6, such as Photophobia, Night Sensitivity and Eyestrain. Such results seem to indicate that the presence of a dark mode can be very beneficial for both users with and without disabilities or eye conditions, either due to a preference or due to a real necessity. "Font" and "Size" were two common topics in previous studies. On the other hand, there are topics we identified that were not

[10] https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/color-blindness/types-color-blindness

Table 6. Topics retrieved from the whole filtered dataset. We defined the topic name and provide the terms pertaining to each topic, as well as accompanied explanation.

Topic	Terms	Comment	Size
Font Size	font, size, small, text, app, large, visually impaired, zoom, fonts, larger	Reviews on the necessity of increasing font size, as well as updating the apps for zoom features.	419
Visual Impairment - I	blind, visually impaired, accessible, people, good, person, easy, please, use, friendly	General reviews with feedback about the accessibility of the apps for the visually impaired.	355
Visual Impairment - II	app, blind, visually impaired, accessible, person, people, make, application, good, use	General reviews with feedback about the accessibility of the apps for the visually impaired.	421
Eyestrain	night, theme, dark, mode, eye, strain, add, eyes, would, please	Reviews that request dark mode features mainly to avoid eyestrain because they use the app at night.	99
Screen Reader	screen, reader, readers, accessible, app, talkback, users, use, work, read	General reviews related to positive and negative feedback on the use of screen readers.	144
Reading Apps	books, kindle, read, book, reading, app, audible, love, amazon, text	Reviews in which users mainly give positive feedback to the reading apps (e.g., Kindle) due to the features that make them more accessible to visually impaired.	132
Talkback	talkback, talk, back, accessible, app, use, please, working, support, smule	Reviews related to TalkBack, a screen reader that assist visually impaired on using smartphones. It comprises compliments when apps fully support TalkBack and complaints when apps do not support such a function.	143
Night Sensitivity	dark mode, snapchat, blind, android, night, app, add, eyes, sensitive, us	Users requests for night mode. Snapchat received a handful of reviews, which include user specific mentioning eye sensitivity and users complaining about eyestrain due to bright colors.	201
KeyBoard	keyboard, typing, gboard, use, type, like, blind, visually impaired, emojis, google	Reviews associated with keyboards, where the reviews are mainly related to the app Gboard.	97
Bible Reading App	bible, god, read, love, app, reading, word, audio, listen, plans	A specific topic where users mostly compliment the speech-to-text feature of the Bible Reading App.	97
Color Blindness	color, colour, red, green, colors, blind, see, blue, color blindness, traffic	Reviews related to color blindness, interestingly more frequent in the Google Maps and WhatsApp reviews, in which users complain they cannnot differentiate between traffic and read-confirmation color encodings.	98
Photophobia	dark mode, light, sensitivity, eyes, sensitive, eye, strain, option, update, please	Reviews in which users comment about the necessity of dark mode. The reviews are mainly related to request for dark mode due to eye sensitivity or the bright colors causing eyestrain.	97

common in previous work. For instance, reviews related to screen reader are scarce in previous work, as well as topics associated with particular disabilities or eye conditions.

It is noticeable that the predominance of generic feedback makes it difficult to automatically identify topics concerning specific accessibility issues. Many text mining techniques require a certain amount of text associated with a set of

Table 7. Number of accessibility reviews and disabilities or eye conditions associated with interface components and resources

Component/Resource	Reviews	Conditions	Component/Resource	Reviews	Conditions
EMOJI	25	6	PRINT	81	14
MEDIA	32	9	LABEL	105	6
BAR	32	8	AUDIO	119	16
LINK	44	9	VIDEO	146	14
LIST	45	9	IMAGE	154	14
NOTIFICATION	53	8	BUTTON	229	12
LAYOUT	57	9	BACKGROUND	271	16
ICON	60	10	COLOR	302	18
MENU	64	8	FONT	373	17
MAP	67	15	TEXT	388	20

common terms or topics to identify a particular group. Therefore, extracting more details on specific concerns reported in accessibility reviews may require manual content analysis in some cases.

The analysis of topics and the specific reviews might give an overview of accessibility issues commonly found in mobile apps. In many cases, users might suggest a possible solution. Removing some accessibility barriers, however, is not a trivial task when the solution might clash with the app patterns or with metaphors used during the design. Many users seem to be aware that a single solution for everyone is not straightforward, thus many of them asks for features that allow some sort of customization. Customization was also a common topic found in previous studies [16].

4.4 RQ4: What are the interface components and resources mentioned in the accessibility reviews associated with visual disabilities or eye conditions

Table 7 shows the number of accessibility reviews for the top 20 most mentioned interface components or resources. It also shows how many distinct disabilities or eye conditions are associated with those reviews. The list of interface components and resources we used as reference was built based on the design foundations and guidelines presented in the Google Material Design[11] and in the the BBC Mobile Accessibility Guidelines[12]. The frequency of each component or resource was calculated by string-matching. Noticeable, TEXT and FONT are the most frequently mentioned resources, followed by COLOR, BACKGROUND and BUTTON.

Figure 4 shows the distribution of reviews among visual disabilities or eye conditions depending on the interface resource or component. Each cell shows how many reviews are associated with that resource or component and that visual disability or eye condition. For instance, there are 6 accessibility reviews associated with FONT and myopia (Row 1, Column 1). Not surprisingly, for most resources, reviews are more concentrated in the generic "visual impairment". For some interface resources or components, some visual disabilities are mostly frequent mentioned by users, such as FONT and low vision (41), BUTTON and blindness (67), and COLOR and color blindness (96).

Discussion. Many design guidelines (e.g. Google Material Design, BBC mobile accessibility guidelines) that include accessibility recommendations are driven by interface components and resources. Understanding how accessibility reviews, components and disabilities or conditions are connected can provide some insight into which guidelines might be more relevant depending on the context. Here are some examples:

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

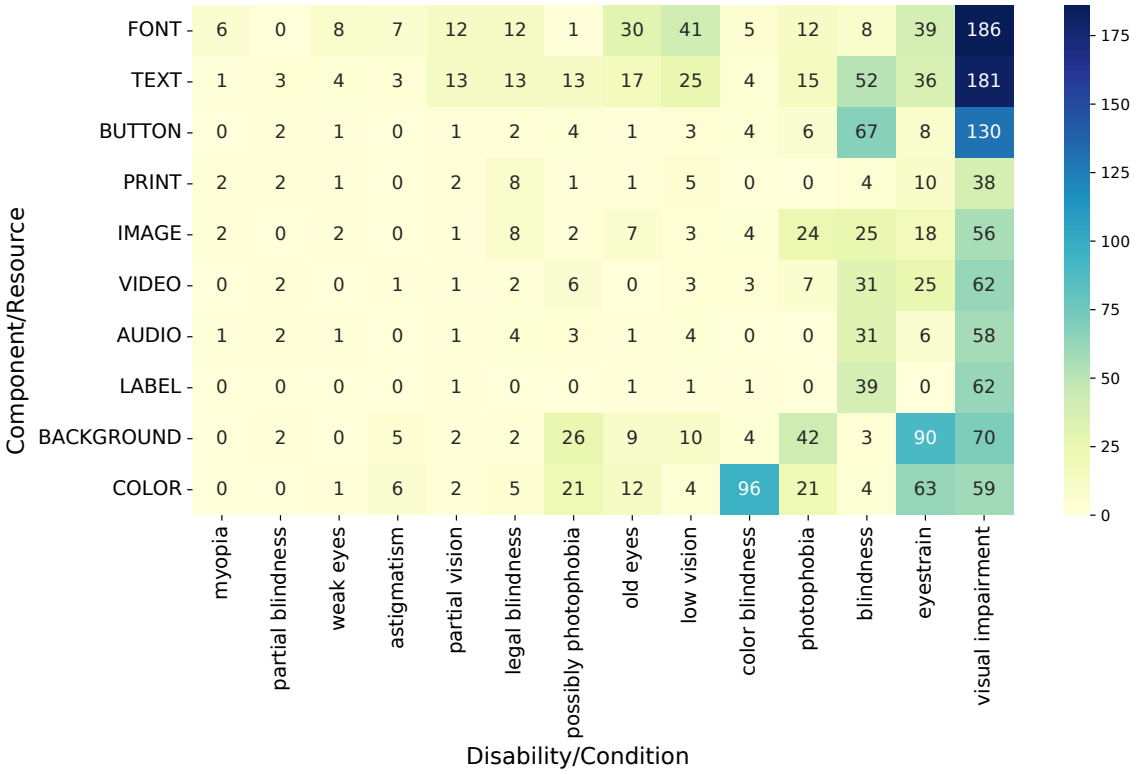

Fig. 4. Heatmap encoding the information of the number of reviews associated with a specific visual disability or eye condition and an interface component or resource.

- FONT: decisions regarding the font (size, color) will mostly impact users that have some difficulty seeing (e.g. low vision), but not blind users and those that depend on screen readers.
- LABEL: labeling elements will mostly impact blind and vision impaired users. In this case, reviews predominantly mention screen readers, for which labels are extremely relevant.
- BACKGROUND: background choices will mostly impact users with light sensitivity and may cause eyestrain.
- COLOR: as expected, color choices will mostly impact color blind users, but it will also impact users with light sensitivity as brighter colors and low contrast ratio would have many negative effects.

4.5 RQ5: What are the scores of accessibility reviews associated with visual disabilities or eye conditions?

Results. Figure 5 shows the distribution of scores (1 to 5) associated with positive reviews (936) and negative reviews (4,063). Negative reviews are slightly concentrated in score 1 (33%) followed by 3 (20%) and 4 (19%). Positive reviews, on the other hand, are clearly concentrated in the highest scores: 5 (82%) and 4 (14%).

Table 8 shows the scores associated with the top 10 most evaluated apps. The mean score of negative reviews is 2.6, and the median is 3, while the mean score of positive reviews is 4.7, and the median is 5. Column 2 shows the general score of that app as it appears in the Google Play Store (GPS), Column 3 shows the mean score for positive reviews, and Column 5 shows the mean score for negative reviews. The scores associated with positive reviews are generally higher

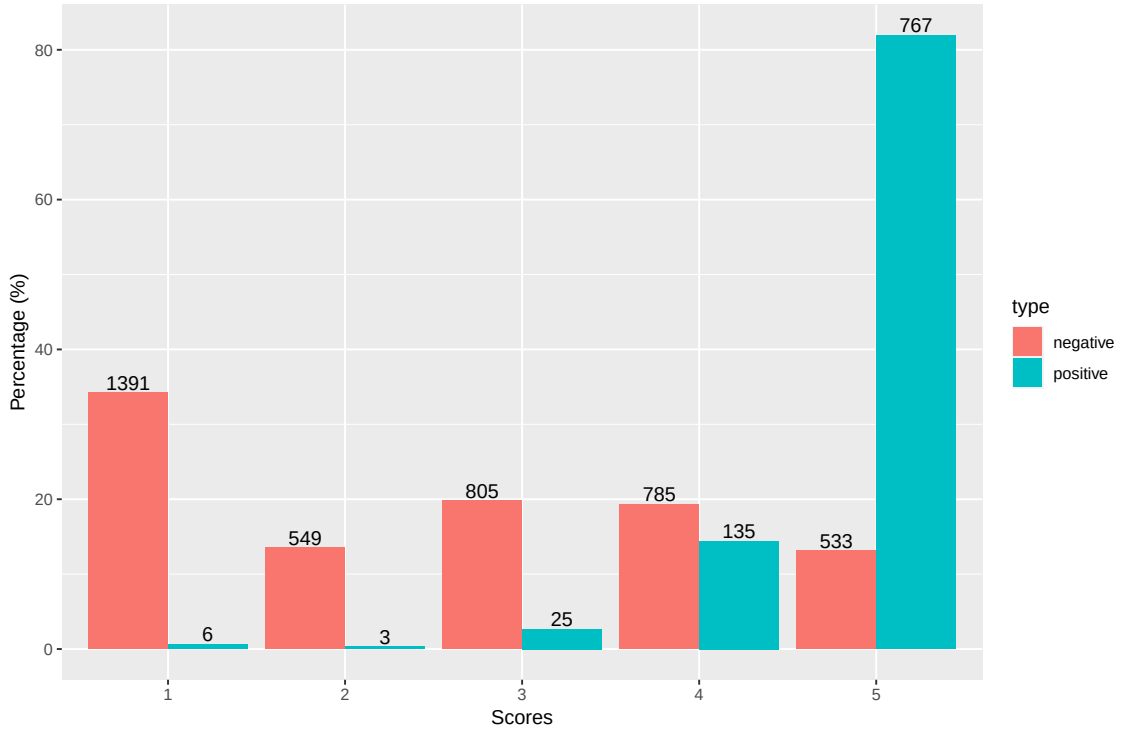

Fig. 5. A bar plot that shows the concentration of reviews of different groups (negative and positive) in each score (1 to 5)

Table 8. Scores associated with accessibility reviews of the top 10 most evaluated apps

		POSITIVE		NEGATIVE	
App	App store score	Mean score	Median	Mean score	Median
All apps	–	4.8	5	2.6	3
Twitter	3.7	4.7	5.0	2.7	3.0
YouVersion Bible	4.9	4.9	5.0	3.3	3.0
Instagram	4.1	4.4	5.0	3.0	3.0
YouTube	4.3	4.5	5.0	2.9	3.0
Google Chrome	4.0	4.7	5.0	2.3	2.0
Gmail	4.1	4.1	4.5	2.0	2.0
Gboard	4.5	4.7	5.0	2.7	3.0
Amazon Kindle	4.7	4.8	5.0	2.5	2.0
WhatsApp	4.3	4.5	5.0	3.4	4.0
Facebook	3.3	4.4	4.0	2.1	2.0

than the general score of the app in the app store. On the other hand, the mean score associated with negative reviews is significantly lower than the mean score of positive reviews and the score registered at the app store.

Discussion. Despite some exceptions, positive feedback in accessibility reviews leads to high scores and negative feedback to relatively low scores in general. The same pattern was not observed in previous studies [16], where the scores of both negative and positive reviews followed almost the same distribution. In fact, in their work, negative feedback was associated with low scores only when a real barrier, usually associated with some sort of visual disability or eye condition, was mentioned.

We noticed that, in some cases, scores are high even when the user faces an accessibility barrier and asks for some fix or feature enhancement (e.g. *"Improvement needed for visually impaired users. Score 5."(Google Chrome)*). In many cases, users are satisfied with the app in general, but face issues regarding its accessibility. It seems their scores are based on the whole scenario rather than the accessibility barrier in those cases (e.g. *"I have no problem with the app. I just wish you have a DARK MODE version. I'm suffering from a Proliferative Diabetic Retinopathy (PDR). I can clearly see and read when it's in DARK MODE just like your Messenger. Hear me please(...). Score 5." ((Facebook)*).

On the other hand, there are positive reviews associated with low scores. This happens for the same reasons presented above. Users compliment the app for its accessibility and give a low score for no apparent reason (e.g. *"Application is an accessible for blind user and visually impaired. thankyou. Score 1."* (*KineMaster*)), or they complain about any general feature besides accessibility that seems to lead to the low scores.

Organizations should know that accessible apps will be well evaluated by users with disabilities or eye conditions, which might increase their reputation in app stores or communities. Unfortunately, the number of reviews written by users with disabilities is so low that their scores might not affect the general scores at the app store. In that sense, users affected by accessibility issues should be more active by providing feedback on the apps they use so organizations and developers can be aware of the importance of a good and inclusive design.

5 THREATS TO VALIDITY

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

Sampling Bias. Each step of our sampling process may introduce bias to our study. The selected apps may not be used by users with disabilities. To mitigate this threat, we selected the most well-known and widely used apps of the Google Play Store. The accuracy of the string-matching selection depends on the set of keywords defined for the process. To ensure we would identify most reviews of interest, we resorted to technical and popular terms related to visual disabilities and eye conditions extracted from glossaries organized by official organizations such as the World Health Organization. In addition, researchers may fail to classify reviews correctly during the manual inspection. To mitigate this threat, we performed cross-validation among researchers to make sure all parts involved agreed upon the decisions of the including, excluding and labeling process.

External Validity. We conducted our investigation over a large dataset extracted from the Google Play Store. Thus, it includes only reviews of Android applications. In addition, one limitation of our investigation is that we only selected reviews in which some visual disability or eye condition is mentioned. We understand that many other reviews might report issues that affect people with disabilities, but identifying such reviews was not within this study's scope. Finally, we would be able to produce more generalized results if we included reviews of iOS apps.

6 CONCLUDING REMARKS AND FUTURE WORK

This paper presented an investigation on accessibility reviews associated with issues that affect users with disabilities or eye conditions. Our sample was identified within nearly 180 million reviews collected from 340 popular apps of the Google Play Store. As far as we know, this is the largest study on accessibility reviews and the first to dive into the

specific reviews associated with visual disabilities or eye conditions. Investigating reviews in this perspective helps identifying real barriers and more urgent matters with respect to the app accessibility, either by using automatic text mining or by manual content analysis.

Our results and the related work show that, unfortunately, accessibility reviews are scarce, thus they may not be enough to drive the accessibility evolution of mobile apps, even when results show that happy users are prone to give high scores and unhappy users tend to give low scores to the evaluated app. Such scarcity of accessibility reviews might imply that users do not leverage feedback mechanisms either because they do not know or believe that their feedback could help making apps more accessible, or because the feedback mechanisms are not accessible for many users.

Further investigation is required to understand this phenomenon and devise strategies to leverage this powerful way to communicate with the organizations and push for some changes, especially because, when it comes to accessibility requirements, organizations rarely specify accessibility needs [9, 26]. In addition, users need to use the opportunity to be more precise with respect to the accessibility issues they are reporting, otherwise their feedback might not be useful.

In future work, we intend to (i) perform manual content analysis to extract valuable information from our sample; (ii) devise an interactive guide in which practitioners can explore accessibility topics and navigate through accessibility reviews to understand the evidence for each accessibility recommendation; (iii) leverage machine learning models to extract features from the accessibility reviews we identified and find similar reviews in our whole dataset; (iv) leverage topic modeling to automatically identify the main topics addressed by users depending on the declared disabilities or type of interface element mentioned (e.g., font, color, organization); Moreover, (v) investigate whether apps accessibility has increased or decreased over time according to users' perceptions.

REFERENCES

[1] Wajdi Aljedaani, Yasir Javed, and Mamdouh Alenezi. 2020. Lda categorization of security bug reports in chromium projects. In *Proceedings of the 2020 European symposium on software engineering*. 154–161.

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

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

[4] Wajdi Aljedaani, Meiyappan Nagappan, Bram Adams, and Michael Godfrey. 2019. A comparison of bugs across the ios and android platforms of two open source cross platform browser apps. In *2019 IEEE/ACM 6th International Conference on Mobile Software Engineering and Systems (MOBILESoft)*. IEEE, 76–86.

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

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

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

[8] Giulia Barbareschi, Catherine Holloway, Katherine Arnold, Grace Magomere, Wycliffe Ambeyi Wetende, Gabriel Ngare, and Joyce Olenja. 2020. The social network: How people with visual impairment use mobile phones in kibera, Kenya. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

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

[10] Ana Catarina Calheiros, Sérgio Moro, and Paulo Rita. 2017. Sentiment classification of consumer-generated online reviews using topic modeling. *Journal of Hospitality Marketing & Management* 26, 7 (2017), 675–693.

[11] Beula Christy and Aishwarya Pillai. 2021. User feedback on usefulness and accessibility features of mobile applications by people with visual impairment. *Indian Journal of Ophthalmology* 69, 3 (2021), 555.

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

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

[14] Henrique Neves da Silva, Andre Takeshi Endo, Marcelo Medeiros Eler, Silvia Regina Vergilio, and Vinicius HS Durelli. 2020. On the Relation between Code Elements and Accessibility Issues in Android Apps. In *Proceedings of the 5th Brazilian Symposium on Systematic and Automated Software Testing*. 40–49.

[15] A. Di Sorbo, S. Panichella, C. V. Alexandru, C. A. Visaggio, and G. Canfora. 2017. SURF: Summarizer of User Reviews Feedback. In *2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C)*. 55–58. https://doi.org/10.1109/ICSE-C.2017.5

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

[17] Marcelo Medeiros Eler, José Miguel Rojas, Yan Ge, and Gordon Fraser. 2018. Automated accessibility testing of mobile apps. In *2018 IEEE 11th International Conference on Software Testing, Verification and Validation (ICST)*. IEEE, 116–126.

[18] Silvia B Fajardo-Flores, Laura S Gaytán-Lugo, Pedro C Santana-Mancilla, and Miguel A Rodríguez-Ortiz. 2017. Mobile Accessibility for People with Combined Visual and Motor Impairment: A case Study. In *Proceedings of the 8th Latin American Conference on Human-Computer Interaction*. 1–4.

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

[20] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[21] Chris Hydock, Zoey Chen, and Kurt Carlson. 2020. Why unhappy customers are unlikely to share their opinions with brands. *Journal of Marketing* 84, 6 (2020), 95–112.

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

[23] Claudia Iacob, Rachel Harrison, and Shamal Faily. 2014. Online Reviews as First Class Artifacts in Mobile App Development. In *Mobile Computing, Applications, and Services*, Gérard Memmi and Ulf Blanke (Eds.). Springer International Publishing, Cham, 47–53.

[24] Andrei P Kirilenko, Svetlana O Stepchenkova, and Xiangyi Dai. 2021. Automated topic modeling of tourist reviews: does the Anna Karenina principle apply? *Tourism Management* 83 (2021), 104241.

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

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

[27] Stanislav Levin and Amiram Yehudai. 2019. Towards software analytics: Modeling maintenance activities. *arXiv preprint arXiv:1903.04909* (2019).

[28] Xiaozhou Li, Zheying Zhang, and Kostas Stefanidis. 2018. Mobile App Evolution Analysis Based on User Reviews. In *SoMeT*.

[29] Wilson E. Marcílio, Danilo M. Eler, and Rogério E. Garcia. 2021. Contrastive analysis for scatterplot-based representations of dimensionality reduction. *Computers & Graphics* (2021). https://doi.org/10.1016/j.cag.2021.08.014

[30] Stuart Mcilroy, Nasir Ali, Hammad Khalid, and Ahmed E. Hassan. 2016. Analyzing and Automatically Labelling the Types of User Issues That Are Raised in Mobile App Reviews. *Empirical Softw. Engg.* 21, 3 (June 2016), 1067–1106. https://doi.org/10.1007/s10664-015-9375-7

[31] Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software* 2, 11 (2017), 205. https://doi.org/10.21105/joss.00205

[32] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software* 3, 29 (2018), 861.

[33] M. Nayebi, B. Adams, and G. Ruhe. 2016. Release Practices for Mobile Apps – What do Users and Developers Think?. In *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, Vol. 1. 552–562. https://doi.org/10.1109/SANER.2016.116

[34] Anh Tuan Nguyen, Tung Thanh Nguyen, Jafar Al-Kofahi, Hung Viet Nguyen, and Tien N Nguyen. 2011. A topic-based approach for narrowing the search space of buggy files from a bug report. In *2011 26th IEEE/ACM International Conference on Automated Software Engineering (ASE 2011)*. IEEE, 263–272.

[35] Hanh Nguyen and Dirk Hovy. 2019. Hey Siri. Ok Google. Alexa: A topic modeling of user reviews for smart speakers. In *EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text*. Association for Computational Linguistics.

[36] D. Pagano and W. Maalej. 2013. User feedback in the appstore: An empirical study. In *2013 21st IEEE International Requirements Engineering Conference (RE)*. 125–134. https://doi.org/10.1109/RE.2013.6636712

[37] F. Palomba, M. Linares-Vásquez, G. Bavota, R. Oliveto, M. Di Penta, D. Poshyvanyk, and A. De Lucia. 2015. User reviews matter! Tracking crowdsourced reviews to support evolution of successful apps. In *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. 291–300. https://doi.org/10.1109/ICSM.2015.7332475

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

[39] Yogendra Pandey, Jaehoon Lee, Devender R Banda, Nora Griffin-Shirley, The Nguyen, and Vitalis Othuon. 2022. A survey of mobile app use among university students with visual impairment in India. *British Journal of Visual Impairment* (2022), 02646196211067358.

[40] S. Panichella, A. Di Sorbo, E. Guzman, C. A. Visaggio, G. Canfora, and H. C. Gall. 2015. How can i improve my app? Classifying user reviews for software maintenance and evolution. In *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. 281–290. https:

//doi.org/10.1109/ICSM.2015.7332474

[41] Kyudong Park, Taedong Goh, and Hyo-Jeong So. 2014. Toward accessible mobile application design: developing mobile application accessibility guidelines for people with visual impairment. In *Proceedings of HCI Korea*. 31–38.

[42] Michael J. Paul, Mark Dredze, and David A. Broniatowski. 2014. Twitter Improves Influenza Forecasting. *PLoS Currents* 6 (2014).

[43] L. Pelloni, G. Grano, A. Ciurumelea, S. Panichella, F. Palomba, and H. C. Gall. 2018. BECLoMA: Augmenting stack traces with user review information. In *2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER)*. 522–526. https://doi.org/10.1109/SANER.2018.8330252

[44] Clauirton Siebra, Tatiana Gouveia, Jefte Macedo, Walter Correia, Marcelo Penha, Marcelo Anjos, Fabiana Florentin, Fabio QB Silva, and Andre LM Santos. 2016. Observation based analysis on the use of mobile applications for visually impaired users. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. 807–814.

[45] Ian Sutherland and Kiattipoom Kiatkawsin. 2020. Determinants of guest experience in Airbnb: a topic modeling approach using LDA. *Sustainability* 12, 8 (2020), 3402.

[46] Stephen W Thomas, Bram Adams, Ahmed E Hassan, and Dorothea Blostein. 2011. Modeling the evolution of topics in source code histories. In *Proceedings of the 8th working conference on mining software repositories*. 173–182.

[47] Ma. Dolores C. Tongco. 2007. Purposive Sampling as a Tool for Informant Selection. *Ethnobotany Research and Applications* 5 (Dec. 2007), 147–158. https://ethnobotanyjournal.org/index.php/era/article/view/126

[48] Christopher Vendome, Diana Solano, Santiago Liñán, and Mario Linares-Vásquez. 2019. Can everyone use my app? an empirical study on accessibility in android apps. In *2019 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 41–52.

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