Healthcare Android Apps: A Tale of the Customers' Perspective

Mariaclaudia Nicolai
University of Salerno, Italy
m.nicolai@studenti.unisa.it

Luca Pascarella
Delft University of Technology, The Netherlands
l.pascarella@tudelft.nl

Fabio Palomba
University of Zurich, Switzerland
palomba@ifi.uzh.ch

Alberto Bacchelli
University of Zurich, Switzerland
bacchelli@ifi.uzh.ch

ABSTRACT

Healthcare mobile apps are becoming a reality for users interested in keeping their daily activities under control. In the last years, several researchers have investigated the effect of healthcare mobile apps on the life of their users as well as the positive/negative impact they have on the quality of life. Nonetheless, it remains still unclear how users approach and interact with the developers of those apps. Understanding whether healthcare mobile app users request different features with respect to other applications is important to estimate the alignment between the development process of healthcare apps and the requests of their users. In this study, we perform an empirical analysis aimed at (i) classifying the user reviews of healthcare open-source apps and (ii) analyzing the sentiment with which users write down user reviews of those apps. In doing so, we define a manual process that enables the creation of an extended taxonomy of healthcare users' requests. The results of our study show that users of healthcare apps are more likely to request new features and support for other hardware than users of different types of apps. Moreover, they tend to be less critical of the defects of the application and better support developers when debugging.

CCS CONCEPTS

• **Software and its engineering** → *Software design engineering*.

KEYWORDS

Mobile Applications, Healthcare, Software Engineering

ACM Reference Format:
Mariaclaudia Nicolai, Luca Pascarella, Fabio Palomba, and Alberto Bacchelli. 2019. Healthcare Android Apps: A Tale of the Customers' Perspective. In *Proceedings of the 3rd ACM SIGSOFT International Workshop on App Market Analytics (WAMA '19), August 27, 2019, Tallinn, Estonia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3340496.3342758

WAMA '19, August 27, 2019, Tallinn, Estonia

ACM ISBN 978-1-4503-6858-2/19/08…$15.00
https://doi.org/10.1145/3340496.3342758

1 INTRODUCTION

The number of people worldwide affected by chronic diseases (such as diabetes, asthma, hypertension) has increased up to 25% in the last three decades [2]. Nevertheless, the general wellness and life expectancy of the world population are growing year by year in response to improved clinical treatment and a preventive education [3, 10, 31]. For example, Tuljapurkar et al. investigated how humans life expectancy evolved in the last centuries [48]. In this study, they caught a slow but persistent gain in the living conditions due to important changes led by technological improvements and multidisciplinary innovations. This positive trend has been constant for several decades until it has seemingly reached a saturation point in the new millennium. Later, Olshansky et al. [32] confirmed the preliminary findings of both Faber [13] and Macdonell [26] demonstrating how making medical care broadly accessible help people that go through a healthy and extended life expectancy.

Although medical scientists play a crucial role in improving overall human health, scientists in other fields contribute to the same goal by taking advantage of interdisciplinary synergies [7, 29]. One of the most popular harmonies between researchers from different fields involves medical scientists and software engineers [28, 44, 47]. Many of these practitioners focused on the ubiquity of mobile technologies as well as the diffusion of portable devices aimed at promoting people wellness [9, 31]. Following this direction, software engineers collaborate with medical scientists to improve software applications that assist people in monitoring their health state (e.g., by tracking blood pressure, body mass index, glycemic index, etc.) [17]. Even though *software as a medical device* is not a novelty, the accessibility and the popularity of such software systems make health mobile applications (from here on *apps*) a promising research branch. Indeed, this category has grown exponentially in the past ten years, involving mobile applications designed for supporting personal wellness, care administration, and medical professionals activities [31, 47]. Obiodu et al. [31] by studying the top 500 medical Apps in a European Android Market found that 45% of them are designed for promoting personal health. Similarly, Whitehead and Seaton [50] conducted an in-deep investigation aimed at clarifying the practical benefits for chronic patients who regularly use mobile apps designed for care administration. They found that mobile apps may improve symptom management through self-management interventions. Similarly, Silva et al. [47] found that mobile healthcare apps (m-Health) have a substantial impact on all healthcare services, such as hospitals, care centers, and emergency attendance.

Although researchers manifest a high interest in this field by publishing many studies in medical journals, software engineering

aspects are not in their primary focus. With our study, we want to cover this gap by investigating (i) how users of Android mobile apps interact with healthcare apps and (ii) what opinion is driven in the user reviews. More deeply, we aim at understanding what feelings are spread by users in their public comments.

To this aim, we manually analyzed 2,000 user reviews mined from two main categories healthcare and non-healthcare apps for a total of 8,431 Android mobile apps. We provide an extended taxonomy composed of 10 categories reporting user feedback and conduct a sentiment analysis on the given feedback and scores. We found that users of healthcare apps request four times more features than users of non-healthcare apps.

2 BACKGROUND AND RELATED WORK

2.1 Related Work

In this section, we discuss the related work and motivate our study.

Mobile Healthcare Apps. The growth of the health app market is encouraged by both medical scientists and software developers with the aim of supporting people and patients with fitness and health guidelines [30]. Therefore, practitioners of both research fields collaborate at improving health apps by experimenting with a progressive release of new features, often, driven by user exigences. Typically, this incremental approach consists of three phases. In a preliminary trial phase practitioners experiment with new functionalities, in the successive commercialization phase they engage health-interested people, and in the latter phase, doctors promote the integration of health apps into patient therapies. Choo et al. [12] tested a health app for one month in a hospital setting. In particular, their study focused on the development of an app that assisted patients following a weight loss program. They evaluated the usability and acceptability of the developed health app and how health apps mediated patient-doctor relationships. In the final result, they supported the utility of health apps to be integrated into medicaments by leading patients to a self-monitoring activity. Indeed, the two most popular app stores (i.e., Android and Apple) included more than 97,000 health apps designed to track health parameters (e.g., blood pressure, weight, blood glucose levels, etc.) [17]. Krebs et al. [20] found that more than half of the cell-phone owners surveyed in their study downloaded at least one health app (58.23% of 1,604 validated participants). Nonetheless, among those who declared to use health apps every day, the researchers noticed that 45.7% of users stopped to use these apps due to high data consumption, lack of interest, and unrevealed usage costs. In addition, by examining the information provided by the participants in their survey, Krebs et al. found that the main users of health apps are young, educated, wealthy, and healthy individuals. To give more evidence, Carroll et al. [10] performed a study aimed at analyzing the social aspect of people's daily activities and how this influenced their well-being. They studied the answers of a sample of surveyed participants who responded to HINTS on routine tasks such as physical activity, fruit and vegetable consumption, and weight loss. They found that among the social factors (i.e., sex, ethnicity, and income) that influenced the use of health apps the most representative social factors were gender, age, and education. A different perspective has been taken by Anderson et al. [5] where in a recent study investigated the role of mobile apps in helping consumers affected by chronic diseases such as diabetes, asthma, blood pressure, depression, etc.. More in deep, this study aims at understanding how the use of health apps for self-monitoring may contribute to extending the life expectancy. To this purpose, they conducted a semi-structured interview that revealed the importance of the use and the effectiveness of health and fitness apps for self-monitoring available on the market suggesting to satisfy users' needs. Similarly, with the intent to evaluate the characteristics of the most popular health apps, Sama et al. [45] found that the primary engagement method relies on a self-monitoring experience. For this study, they selected a representative sample of 400 apps (selected in the Apple iTunes marketplace). The outcomes revealed how 74,8% of the analyzed apps engage users with a self-monitoring aim.

In 2016, Whitehead and Seaton [50] conducted a comprehensive literature review, in a time frame of ten years, aimed at understanding the effectiveness of mobile apps for healthcare in supporting chronic diseases. They found empirical evidence that indicates the relief of health apps for long-term condition management specifically diabetes mellitus, cardiovascular diseases, and chronic lung diseases. Although the novelty and the importance of the research mentioned above, the primary goal is to understand how users of healthcare apps benefit from technological advisings. With our study, we want to deliver this knowledge to software developers by bringing empirical evidence emerged in different research fields (i.e., medical journals) to software engineering experts.

User-Driven Software Development. The importance of considering the user experience during software development has been thoroughly investigated in software engineering. Fu et al. [14] propose a technique to optimize the results of summarization tools filtering out the useless comments at different granularity. In addition, they create a tool that identifies reasons behind the perceived app effectiveness with the purpose of helping mobile app market operators such as Google or individual app developers. They discovered that starting from a sample of 50,000 user reviews the 0.9% of them were inconsistent with the rating. Successively, Chen et al. [11] present AR-Miner that is a computational framework for mining user reviews. This tool filters noisy and irrelevant comments, groups reviews by topic, and finally prioritizes user reports by an effective review ranking scheme to be inconsistent with the ratings. The outcome of this study highlights that 35% of user reviews labeled by their tool was informative reviews. A different research group has also done similar work, Iacob and Harrison [16] create MARA that is a tool for automatic retrieval of mobile app feature requests from user reviews. For understanding the importance of the user comments, Pagano and Maalej [33] investigated the way how users provide feedback. In their study, the authors discovered a trivial relationship between the user experience and the number of downloads. A similar result was achieved by Khalid et al. [19] by discovering 12 types of user complaints by investigating 6,390 user reviews of free iOS apps. The latter two groups of researchers highlighted that out 33% of the user reviews were related to requirements and user experience. Besides, they found that while user reviews with worst ratings express dispraise and are mostly bug reports, the top-rated user reviews are related to recommendations, helpfulness, and features information. Recently, Palomba et al. [34] combined the stat of the art in linking informal

documents to source code to create Cristal that is a tool for tracing informative crowd user reviews back to source code changes. It enables users to measure to what extent developers accommodate user requests. Researchers proposed several approaches for linking informal documentation (i.e., emails, IRC, forums, etc.) onto source code or other artifacts [4, 8, 25, 39]. For example, Bacchelli et al. [8] used lightweight textual grep-based analysis and IR techniques [46] to link the email content to source codes. Parnin et al. [39] built a tool to reconstruct traceability links between Stack Overflow discussions and API classes. These links allow researchers to measure the coverage of APIs in Stack Overflow discussions. Pascarella et al. laid the basis to link code comments with source code by training a machine learning tool with the purpose of categorizing comments in natural language into a double layer taxonomy [40, 41]. Similarly, Linares-Vásquez et al. [25] reconstruct links between Stack Overflow questions and Android APIs to identify how developers react to API changes. Recently, Alkadhi et al. [4] propose a solution to extract the rationale content discussed by developers in Internet Relay Chat (IRC) channels. With a manual analysis of 7,500 messages, they create a model based on a machine learning binary classifier to automatically extract rationale discussions achieving a 0.76 precision and 0.79 recall. Although the research is progressively covering different fields, none of the above techniques identifies user experiences with the purpose of assisting developers of a specific category such as developers of healthcare mobile apps.

2.2 Study Motivation

In the following, we discuss three examples in which users report their experiences with mobile apps as a motivation for this study.

Although software researchers already highlighted the effectiveness of user reviews in software development [11, 14, 16, 19, 33] and, at the same time, several clinical scientists observed a positive benefit brought by virtual assistants (such as health mobile apps) [10], medical researchers reported a premature abandon trend in medical journals [20]. Krebs and Duncan [20], in a recent study, highlight a negative trend of how users of health apps tend to abandon health apps after a short try prematurely. They foun,d that up to 58% of surveyed people in the United States have successfully downloaded and installed health apps exploring apps for specific diseases dedicated or generic committed apps for fitness or nutrition. However, even if the number of users that experienced health apps is pretty high (934 users of 1,604 interviewed), only 55% of them continue to use downloaded apps. This premature abandonment is due to several causes such as high data consumption, loss of interest, missed features, and hidden costs. Aimed by these preliminary motivations, we inspected user reviews to find additional evidence that supports and expands Krebs and Duncan findings such as feature missing, battery leakage, and software issues. For example, in the following review, a user reports a missing feature, but even this missing, s/he gives a five stars grade for the quality of the app.

> "*Great app. Helps you to be fit. Small request.. Can you add a feature in which fitness band vibrate at specified time (just once) like a reminder not like an alarm.. Rest it's amazing device and app.. Thank you.*"

Another enthusiast user gave a five stars grade despite a battery problem and some missing features.

> "*I have been using the mi band 2 for over an year now. The app surely has come a long way. Can you also add accepting calls and putting it on speaker mode??? Also can we get a toggle for the 'phone disconnection' alert, because my house has relatively thicker walls and it keeps on disconnecting thus reducing battery life.*"

Finally, another user reports a positive feeling, thus rating the app with five stars even though s/he experiences a continuous reboot during software updates.

> "*This makes the Fitbit so much more useful. I don't want a smartwatch because it seems way too expensive for a notification extension. But getting that functionality on my Charge 2 is great! Only annoying thing is having to reboot on updates. But a small price to pay for extra functionality.*"

In line with Krebs and Duncan results, Li et al. discovered that many users of mobile apps are likely to uninstall fresh installed apps within two days [22]. Nonetheless, we believe that users of healthcare mobile apps tend to be less critical with respect of discovered issues trying to support developers of open-source code with a high rating. In the following, we want to better understand how users of healthcare apps interact with their developers.

3 METHODOLOGY

The *goal* of this study is to provide a deeper understanding of how users of healthcare apps interact with software developers and what do they request to them, with the *purpose* of evaluating whether exists peculiar characteristics that would require the development process of healthcare apps to be different from the one of standard applications (e.g., by improving privacy protection and reliability for health). The *perspective* is of both practitioners and researchers: the former are interested in understanding how they can provide better support to users and prevent negative users' experiences; the latter are interested in assessing the feasibility of specialized methodologies easing the development process of healthcare apps. In this following subsections, we describe our research questions and the methodology adopted to address them.

3.1 Research Questions

Our work is structured around two main research questions. In the first place, we aim at investigating how users of healthcare apps interact with software developers. To this aim, we analyze and classify what users suggest within user reviews, i.e., an instrument that is widely adopted by users to report failures, suggest new features, etc. [36, 38, 42]. Hence, we ask our first research question:

> **RQ$_1$.** *What do users of healthcare apps report into user reviews?*

After classifying which types of information users report, we analyze user reviews with the aim of understanding their content, particularly the kind of sentiment that is shared. This leads to our second research question:

> **RQ$_2$.** *What is the sentiment of the user reviews reported by healthcare apps' users?*

Addressing the aforementioned research questions, we aim at improving our scientific understanding on how developers interact with healthcare apps. Specifically, with $\mathbf{RQ}_1$ we understand **what** types of information users report, while $\mathbf{RQ}_2$ allows us to understand **how** users report opinions.

3.2 Context Selection

The *context* of the study consists of 236 healthcare open-source Android mobile apps. We consider the publicly available dataset developed by Geiger et al. [15] as a starting point: It provides a graph-based database composed of 8,431 verified open-source Android apps whose source code has been cloned in a private and freely accessible GitLab repository.[1] Moreover, this dataset also provides verified metadata information that includes a reference to the Google Play[2] location of each app. The selection of this dataset is driven by two main reasons: (i) it contains the largest collection of real mobile apps available in the literature [15]; and (ii) it contains open-source healthcare apps, thus enabling the possibility to perform our analyses.

From the initial list of 8,431 apps, we query the graph-based dataset and extract all the apps referring to health, fitness, and medical categories: this procedure outputs 236 results, which represent the healthcare apps of our study. For each app, we then mine the corresponding Google Play store location and extract the list of user reviews. Unfortunately, such reviews are permanently stored on Google Play only for a limited amount of time [35]: this is the reason why we decide to mine all those available during the last three months. We are able to extract a total of 23,085 user reviews.

In our study, we aim at verifying if there are peculiarities that characterize healthcare mobile apps. To reach this goal, we need a baseline with which to compare the findings achieved on healthcare apps. Thus, from the dataset of Geiger et al. [15] we also extract the information related to 8,195 non-healthcare apps, that we use to mine the corresponding user reviews. This procedure leads to the mining of an additional number of 360,673 user reviews: so, globally we reach 383,758 of them.

As our study requires manual inspections, the analysis of such amount of user reviews is prohibitively expensive. Thus, from the set of 383,758 reviews, we define a stratified random sample composed of 2,000 of them. This represents a 95% statistically significant sample of the total number of user reviews, with a confidence interval of 2.2% (assuming a 50% population proportion). This statistically significant sample represents the final context of our study.

3.3 $\mathbf{RQ}_1$. Methodology

To address $\mathbf{RQ}_1$, we conduct a manual analysis aimed at classifying the review feedback left by users into user reviews. More specifically, we applied a three-step iterative *content analysis approach* [23, 43] involving two of the authors of this paper, who have complementary expertise in line with the goal of our analyses: the first one is a software engineer having more than ten years of mobile programming experience; the second one is a Master student in medical disciplines who has more than two years of experience in healthcare-related user studies. From now on, we refer to both of them as *inspectors*; they classify a total of 1,100 user reviews each following the procedure reported below:

Iteration 1. In the first phase, the inspectors analyze an initial set of 300 user reviews; 100 of them are in common and are used to control the agreement of the inspectors. As an existing taxonomy of users' requests is already available from work proposed by Panichella et al. [38], the inspectors firstly rely on that even though they are allowed to refer to other taxonomies (e.g., Khalid et al. [19] propose 12 categories of complaints of iOS apps) or add new items in the taxonomy if needed (i.e., if no item in the previous taxonomy matches at least one of the analyzed user reviews). Indeed, the initial investigation highlights the need for extending the provided taxonomy with a fine-grained classification that gives more expressiveness to users' feedback of both healthcare and non-healthcare apps. The output—that represents a side contribution of this paper—is a draft taxonomy that partially overlaps the definitions of Panichella et al. [38] and extends such categories with a fine-grained description.

Iteration 2. In the second phase, the inspectors opened a discussion about the names and types of the categories assigned so far. In this discussion, also the other authors of the paper participated with the aim of stimulating a consensus. Afterward, the two inspectors firstly re-categorize the 300 user reviews according to the taxonomy emerged from the discussion. Then, the inspectors classified other 200 user reviews. This phase validates the categories coming from the first step by confirming some of them and redefining others. After the completion, the inspectors opened a new discussion with the other authors aimed at refining the draft taxonomy, merging overlapping categories or characterizing better the ones already existing previously.

Iteration 3. In the third phase, the inspectors re-categorize the 500 user reviews previously analyzed. At the same time, they also classify the remaining 600, but this time they have an overlap of 100 user reviews, which is needed to measure how the agreement evolved. During this step, the inspectors try to apply the draft taxonomy coming from the second iteration to verify it on an unseen set of data. As a result, they do not find any further mismatch.

The output of the iterative content analysis is assessed using the Krippendorff's α inter-rater agreement metric [21], which is equal to 98%, thus indicating an excellent agreement and further suggesting the reliability of the context analysis sessions performed.

3.4 $\mathbf{RQ}_2$. Methodology

To address $\mathbf{RQ}_2$, we conduct sentiment analysis [37] of the user reviews belonging to the two sets previously built, i.e., healthcare and non-healthcare apps. To this aim, we rely on the Stanford CoreNLP natural language processing toolkit [27], which provides APIs for preprocessing and analyzing the sentiment contained in informal texts. Among all the approaches available for sentiment analysis [24], we select CoreNLP because of the results achieved in comparative studies, where the tool reached performance similar, if not higher, than other existing techniques and tools [6, 49]. It is important to note that, according to recent findings [18], sentiment analysis tools are generally not suitable for software engineering

[1] https://about.gitlab.com/
[2] https://play.google.com/store

Table 1: User review categories manually inspected. The categories denoted with * come from Panichella et al. [38].

Category	Description
Complaints	Users express a negative feeling or an abandon of the given app.
Complimentes	Users express a positive perception appreciating the app.
Feature requests*	Users are typically satisfied but need more features.
Information giving*	Generic sentences used to inform or update others about something.
Information seeking*	Sentences related to attempts to obtain information or receive help from developers.
Opinion asking*	Sentences used for requiring someone to express her/his point of view about something explicitly.
Problem discovering*	Sentences related to issues and unexpected behaviors.
Problem reporting	Users describe the scenario that caused a malfunction.
Solution proposal*	Users suggest workarounds or temporary fixes.
Noise	The content of the review does not bring a valid meaning.

research, as they are machine learners not trained on data coming from technical contexts such as software development. Nevertheless, in our case we aim at analyzing user reviews, that are informal by nature: as such, we argue that the exploited tool can provide us with accurate information.

More in detail, by the text composing each user review, CoreNLP estimates its sentiment by giving as output a value ranging between 0 (very negative sentiment) and 4 (very positive sentiment). We use the scores output by the tool to address our research question and understand whether and how different the healthcare users are with respect to other users.

4 ANALYSIS OF THE RESULTS

4.1 RQ_1. What do users of healthcare apps report into user reviews?

Table 1 presents the results of the categorization of user reviews. With respect to the base taxonomy adopted for classification [38], we were able to refine it and find three new categories such as *'Complaints'*, *'Compliments'*, and *'Problem Reporting'*. The first refers to reviews where users just complaint about the functionalities of an app; the second represents the opposite situation, where customers just report their gratitude for the developed product; finally, the third one is related to those reviews that report and describe problems appearing in the app. It is worth noting that the latter category differs from *'Problem Discovery'*, as it describes errors rather than just signaling their presence. Moreover, a developer engaged in improving defective apps may be more interested in *'Complaints'* rather than *'Compliments'* feedbacks because the second category is less informative in term of software needs.

We also deeper analyzed the distribution of user reviews in both healthcare and non-healthcare apps. Table 2 shows the percentage of the user reviews across 10 categories and the average star rating for each of this category. On average non-healthcare apps have higher star rating: this is 3.9 (on a five stars scale) against 3.1 for healthcare apps. Some categories such as *Information seeking*, *Opinion asking*, and *Solution proposal* receive comparable attention between the two groups of apps. In both cases, these categories only contain a marginal number of reviews (less than 4%). It is worth noting that even if *Opinion asking* contributes only marginally, we kept this category because this makes our study aligned with the study of Panichella et al. [38]. On the contrary, a consistent difference between the two app classes is related to the *Feature Requests* category. In such case users of healthcare apps are more prone to ask new features (15.3% against 4.0%). This behavior may have two origins. On the one hand, users of healthcare apps appear to be less satisfied (7.3% of comments belong to *Complaints* category), and thus they ask more features. On the other hand, even generically satisfied users are encouraged to demand more features because they feel that developers are not really experts of the field. For instance, let consider the following user review, which belongs to the com.pacoapp.paco app:

> "*[...] I realize that this is made by programmers and that the main focus is the behavioral analyses, but the app itself is really confusing. A few suggestions I'd make, that personally would keep me around [...]*"

In this case, the user does not only provide his/her opinion on the app but s/he also tries to recommend suggestions for possible improvements that would help developers in better supporting customers.

Table 2: Frequency of user review categories.

Category	Healthcare		Non-healthcare	
	Perc.	Stars	Perc.	Stars
(C1) Complaints	7.3%	1.6	4.9%	1.5
(C2) Compliments	5.5%	4.8	8.4%	4.9
(C3) Feature requests	12.4%	3.4	4.0%	4.1
(C4) Information giving	42.0%	3.2	56.0%	4.3
(C5) Information seeking	3.6%	2.8	2.2%	3.8
(C6) Opinion asking	0.1%	1.0	0.1%	1.0
(C7) Problem discovering	15.3%	2.3	10.5%	2.3
(C8) Problem reporting	5.4%	2.4	2.4%	2.6
(C9) Solution proposal	0.6%	3.7	0.6%	3.5
(C10) Noise	8.0%	3.9	10.9%	4.0

Finding 1: We found three categories of user reviews, i.e., *'Complaints'*, *'Compliments'*, and *'Problem Reporting'* not reported in previous taxonomies. Moreover, our findings reveal that healthcare apps' customers tend to ask the introduction of more features than users of other apps.

Table 3: Sentiment of the considered user reviews. Results are reported in percentage. H refers to healthcare and NH stands for non-healthcare.

Cat.	Very neg.		Neg.		Neutral		Pos.		Very pos.	
	H	NH	H	NH	H	NH	H	NH	H	NH
C1	76	78	14	13	8	6	2	3	0	0
C2	0	0	3	5	37	45	35	40	25	10
C3	1	1	24	27	65	63	7	8	3	1
C4	0	0	2	2	95	97	3	1	0	0
C5	0	0	4	7	89	88	6	3	1	2
C6	1	3	13	25	55	57	31	15	0	0
C7	11	44	33	38	31	18	23	0	2	0
C8	22	37	16	32	45	31	17	0	0	0
C9	3	8	5	7	62	72	24	10	6	3

4.2 RQ_2. What is the sentiment of the user reviews reported by healthcare apps' users?

Table 3 reports the results achieved when running the chosen sentiment analyzer, i.e., CoreNLP, on the user reviews of healthcare and non-healthcare mobile apps in our dataset.

As it is possible to observe, in most of the cases the sentiment remain stable between the two categories of apps. As an example, the category *'Complaints'* has a similar distribution of sentiments, with a much higher percentage of reviews falling under the *'Very Negative'* sentiment. There are, however, three exceptions to this general discussion: these relate to the categories *'Problem Discovery'*, *'Problem Reporting'*, and *'Solution Proposal'* (last three rows of Table 3).

While one can expect a large majority of negative sentiments for user review categories related to issues raised during the execution of mobile apps, we observe that this is not necessarily true in case of healthcare mobile apps. Indeed, the percentage of positive or very positive reviews is 54% and 62%, respectively, for the categories *'Problem Discovery'* and *'Problem Reporting'*. As opposed, these percentages are 0 in the case of non-healthcare apps. This clearly highlights that users of healthcare apps tend to be less critical toward errors appearing in these apps; this is likely due to their willingness to (i) be proactive with respect to apps that help their life and social activities and (ii) drive developers toward the resolution of problems rather than blame them for missing functionalities.

As an example, in the following, we report the text of a user review received by com.xiaomi.hm.health, an app that helps users in keeping under control their sleeping time.

> *"Everything is great about this but one thing i must mention is that when i receive a call my mi2 band keeps vibrating even after i have received the call. I will rate 5star after this is fixed."*

As shown, even though the user reports a problem, s/he does it politely, also explaining well what his/her problem is. A similar discussion can be done for the *'Solution Proposal'* category, where we observe a much higher percentage of positive reviews for healthcare apps (30% versus 13%). Also, in this case, customers tend to be as proactive as possible and suggest solutions to the problems they discovered.

Finding 2: Healthcare apps' customers tend to be more positive when describing and reporting failures than users of other apps. Similarly, they try to recommend possible solutions to those errors in a more polite way.

5 THREATS TO VALIDITY

Many factors could have influenced our study. In the first place, we manually analyzed 2,000 user reviews to classify them are build a taxonomy of users' comments that represents an extension of previously defined one [38]. We are aware that such a taxonomy extension may still be incomplete when applied to a different set of mobile apps. Nevertheless, to ensure both correctness and completeness of the categories of user reviews identified, we iteratively built the taxonomy by merging and splitting categories if needed. As an additional validation, we kept 100 user reviews as overlapping sample with the aim of verifying the agreement between the two involved inspectors. The high agreement reached (98%) indicates the stability and reliability of the classifications made.

In the second part of our study, we performed sentiment analysis. In so doing, we relied on the Stanford CoreNLP natural language processing toolkit [27]. Our selection was based on the results of previous empirical comparisons, which showed that the performance of CoreNLP is high and similar to other sentiment analyzers [6, 49]. Of course, we cannot exclude possible imprecision in the way the toolkit has computed the sentiment of the considered user reviews. Further analyses, conducted on a different sample, might be beneficial to corroborate our findings.

6 CONCLUSION

Although the popularity of mobile apps is growing [1] and the interest of software engineers and medical scientists is notably high, only a few studies merge these two fields to bring evidence across domains [17, 31, 47]. In this paper, we started looking at the intersection between mobile apps and healthcare mechanisms, by analyzing what the users of healthcare apps ask in their user reviews and whether they do that differently from non-healthcare users. To this purpose, we first manually analyzed 2,000 user reviews with the aim of classifying the types of comments left for healthcare and non-healthcare apps. Secondly, we assessed how the sentiment of these user reviews is and whether there are differences between healthcare and non-healthcare apps.

The main results of the study indicate the existence of ten categories of user reviews: while most of them are similar to those previously discovered in the literature [38], we found three additional ones. By analyzing them, we found that users of healthcare apps tend to ask more feature requests concerning other users, and this is likely because the developers of those apps are not aware of the specific customers' needs. Moreover, we found that healthcare users tend to be more proactive in the case of app's failures and try to propose solutions to developers.

Based on our findings, we claim that the development process of healthcare apps should be further supported by our research community using specific tools and methodologies able to provide developers with insights into the customers' needs. Our future research agenda is oriented to the definition and investigation of

those novel methodologies. At the same time, we plan to corroborate the findings observed in this paper by analyzing more user reviews. In addition, we plan to compare the development processes of those two categories through the analysis of the version control system guaranteed by the open access of the selected apps.

ACKNOWLEDGMENTS

Bacchelli and Palomba acknowledge the support of the Swiss National Science Foundation through the SNF Project PP00P2_170529.

REFERENCES

[1] [n.d.]. App Download and Usage Statistics . http://www.businessofapps.com/data/app-statistics/. [Online; accessed 04-February-2019].

[2] [n.d.]. World Health Organization. http://www.who.int/mediacentre/factsheets/fs312/en/. [Online; accessed 28-March-2018].

[3] [n.d.]. World Health Organization. http://www.who.int/publications/10-year-review/dg-letter/en/. [Online; accessed 28-March-2018].

[4] Rana Alkadhi, Manuel Nonnenmacher, Emitza Guzman, and Bernd Bruegge. [n.d.]. How Do Developers Discuss Rationale? ([n. d.]).

[5] Kevin Anderson, Oksana Burford, and Lynne Emmerton. 2016. Mobile health apps to facilitate self-care: a qualitative study of user experiences. *PLoS One* (2016).

[6] Michelle Annett and Grzegorz Kondrak. 2008. A comparison of sentiment analysis techniques: Polarizing movie blogs. In *Conference of the Canadian Society for Computational Studies of Intelligence*. Springer, 25–35.

[7] Arnold Elvin Aronson. 1990. *Clinical voice disorders: An interdisciplinary approach*. Thieme New York.

[8] Alberto Bacchelli, Michele Lanza, and Romain Robbes. 2010. Linking e-mails and source code artifacts. In *Proceedings of the 32Nd ACM/IEEE International Conference on Software Engineering-Volume 1*. ACM, 375–384.

[9] Jeffrey Boase and Rich Ling. 2013. Measuring mobile phone use: Self-report versus log data. *Journal of Computer-Mediated Communication* 18, 4 (2013), 508–519.

[10] Jennifer K Carroll, Anne Moorhead, Raymond Bond, William G LeBlanc, Robert J Petrella, and Kevin Fiscella. 2017. Who uses mobile phone health apps and does use matter? A secondary data analytics approach. *Journal of medical Internet research* 19, 4 (2017).

[11] Ning Chen, Jialiu Lin, Steven CH Hoi, Xiaokui Xiao, and Boshen Zhang. 2014. AR-miner: mining informative reviews for developers from mobile app marketplace. In *Proceedings of the 36th International Conference on Software Engineering*. ACM.

[12] Seryung Choo, Ju Young Kim, Se Young Jung, Sarah Kim, Jeong Eun Kim, Jong Soo Han, Sohye Kim, Jeong Hyun Kim, Jeehye Kim, Yongseok Kim, et al. 2016. Development of a weight loss mobile app linked with an accelerometer for use in the clinic: usability, acceptability, and early testing of its impact on the patient-doctor relationship. *JMIR mHealth and uHealth* 4, 1 (2016).

[13] Joseph F Faber. 1982. Life tables for the United States: 1900-2050. (1982).

[14] Bin Fu, Jialiu Lin, Lei Li, Christos Faloutsos, Jason Hong, and Norman Sadeh. 2013. Why people hate your app: Making sense of user feedback in a mobile app store. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1276–1284.

[15] Franz-Xaver Geiger, Ivano Malavolta, Luca Pascarella, Fabio Palomba, Dario Di Nucci, and Alberto Bacchelli. 2018. A Graph-based Dataset of Commit History of Real-World Android apps. In *Proceedings of the 15th International Conference on Mining Software Repositories, MSR*.

[16] Claudia Iacob and Rachel Harrison. 2013. Retrieving and analyzing mobile apps feature requests from online reviews. In *Proceedings of the 10th Working Conference on Mining Software Repositories*. IEEE Press, 41–44.

[17] RG Jahns and P Houck. 2013. Mobile Health Market Report 2013-2017. *Research2Guidance* (2013).

[18] Robbert Jongeling, Proshanta Sarkar, Subhajit Datta, and Alexander Serebrenik. 2017. On negative results when using sentiment analysis tools for software engineering research. *Empirical Software Engineering* 22, 5 (2017), 2543–2584.

[19] Hammad Khalid, Emad Shihab, Meiyappan Nagappan, and Ahmed E Hassan. 2015. What do mobile app users complain about? *IEEE Software* (2015).

[20] Paul Krebs and Dustin T Duncan. 2015. Health app use among US mobile phone owners: a national survey. *JMIR mHealth and uHealth* 3, 4 (2015).

[21] Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. (2011).

[22] Huoran Li, Xuan Lu, Xuanzhe Liu, Tao Xie, Kaigui Bian, Felix Xiaozhu Lin, Qiaozhu Mei, and Feng Feng. 2015. Characterizing smartphone usage patterns from millions of android users. In *Proceedings of the 2015 Internet Measurement Conference*. ACM, 459–472.

[23] William Lidwell, Kritina Holden, and Jill Butler. 2010. *Universal Principles of Design*. Rockport Publishers.

[24] Bin Lin, Fiorella Zampetti, Gabriele Bavota, Massimiliano Di Penta, Michele Lanza, and Rocco Oliveto. 2018. Sentiment Analysis for Software Engineering: How Far Can We Go?. In *Conference on Software Engineering*.

[25] Mario Linares-Vásquez, Gabriele Bavota, Massimiliano Di Penta, Rocco Oliveto, and Denys Poshyvanyk. 2014. How do api changes trigger stack overflow discussions? a study on the android sdk. In *proceedings of the 22nd International Conference on Program Comprehension*. ACM, 83–94.

[26] WR Macdonell. 1913. On the expectation of life in ancient Rome, and in the provinces of Hispania and Lusitania, and Africa. *Biometrika* 9, 3/4 (1913), 366–380.

[27] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 55–60.

[28] Vivien Marx. 2013. Biology: The big challenges of big data.

[29] Anita Mehta. 2012. *Granular Matter: an interdisciplinary approach*. Springer Science & Business Media.

[30] Chilukuri K Mohan and Dayaprasad Kulkarni. 2016. The role of health informatics in volunteer supported healthcare for underserved populations. In *Global Humanitarian Technology Conference (GHTC), 2016*. IEEE, 660–665.

[31] Vivian Obiodu and Emeka Obiodu. 2012. An empirical review of the top 500 medical apps in a European Android market. *Journal of Mobile Technology in Medicine* 1, 4 (2012), 22–37.

[32] S Jay Olshansky, Douglas J Passaro, Ronald C Hershow, Jennifer Layden, Bruce A Carnes, Jacob Brody, Leonard Hayflick, Robert N Butler, David B Allison, and David S Ludwig. 2005. A potential decline in life expectancy in the United States in the 21st century. *New England Journal of Medicine* 352, 11 (2005), 1138–1145.

[33] Dennis Pagano and Walid Maalej. 2013. User feedback in the appstore: An empirical study. In *Requirements Engineering Conference (RE), 2013 21st IEEE International*. IEEE, 125–134.

[34] Fabio Palomba, Mario Linares-Vasquez, Gabriele Bavota, Rocco Oliveto, Massimiliano Di Penta, Denys Poshyvanyk, and Andrea De Lucia. 2015. User reviews matter! tracking crowdsourced reviews to support evolution of successful apps. In *Software Maintenance and Evolution (ICSME), 2015 IEEE International Conference*.

[35] 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.

[36] Fabio Palomba, Pasquale Salza, Adelina Ciurumelea, Sebastiano Panichella, Harald Gall, Filomena Ferrucci, and Andrea De Lucia. 2017. Recommending and localizing change requests for mobile apps based on user reviews. In *Proceedings of the 39th international conference on software engineering*. IEEE Press, 106–117.

[37] Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2, 1–2 (2008), 1–135.

[38] Sebastiano Panichella, Andrea Di Sorbo, Emitza Guzman, Corrado A Visaggio, Gerardo Canfora, and Harald C Gall. 2015. How can i improve my app? classifying user reviews for software maintenance and evolution. In *Software maintenance and evolution (ICSME), 2015 IEEE international conference on*. IEEE, 281–290.

[39] Chris Parnin, Christoph Treude, Lars Grammel, and Margaret-Anne Storey. 2012. Crowd documentation: Exploring the coverage and the dynamics of API discussions on Stack Overflow. *Georgia Institute of Technology, Tech. Rep* (2012).

[40] Luca Pascarella. 2018. Classifying code comments in Java mobile applications. In *2018 IEEE/ACM 5th International Conference on Mobile Software Engineering and Systems (MOBILESoft)*. IEEE, 39–40.

[41] Luca Pascarella, Magiel Bruntink, and Alberto Bacchelli. 2019. Classifying code comments in Java software systems. *Empirical Software Engineering* (2019), 1–39.

[42] Luca Pascarella, Franz-Xaver Geiger, Fabio Palomba, Dario Di Nucci, Ivano Malavolta, and Alberto Bacchelli. 2018. Self-Reported Activities of Android Developers. In *International Conference on Mobile Software Engineering and Systems*.

[43] Luca Pascarella, Davide Spadini, Fabio Palomba, Magiel Bruntink, and Alberto Bacchelli. 2018. Information needs in contemporary code review. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 135.

[44] Steve Pieper, Bill Lorensen, Will Schroeder, and Ron Kikinis. 2006. The NA-MIC Kit: ITK, VTK, pipelines, grids and 3D slicer as an open platform for the medical image computing community. In *Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on*. IEEE, 698–701.

[45] Preethi R Sama, Zubin J Eapen, Kevin P Weinfurt, Bimal R Shah, and Kevin A Schulman. 2014. An evaluation of mobile health application tools. *JMIR mHealth and uHealth* 2, 2 (2014).

[46] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Vol. 39. Cambridge University Press.

[47] Bruno MC Silva, Joel JPC Rodrigues, Isabel de la Torre Díez, Miguel López-Coronado, and Kashif Saleem. 2015. Mobile-health: A review of current state in 2015. *Journal of biomedical informatics* 56 (2015), 265–272.

[48] Shripad Tuljapurkar, Nan Li, and Carl Boe. 2000. A universal pattern of mortality decline in the G7 countries. *Nature* 405, 6788 (2000), 789.

[49] SM Vohra and JB Teraiya. 2013. A comparative study of sentiment analysis techniques. *Journal JIKRCE* 2, 2 (2013), 313–317.

[50] Lisa Whitehead and Philippa Seaton. 2016. The effectiveness of self-management mobile phone and tablet apps in long-term condition management: a systematic review. *Journal of medical Internet research* 18, 5 (2016).