Examining User Complaints of Wearable Apps: A Case Study on Android Wear

Suhaib Mujahid*, Giancarlo Sierra*, Rabe Abdalkareem*, Emad Shihab* and Weiyi Shang
Data-Driven Analysis of Software (DAS) Lab*
Dept. Computer Science and Software Engineering
Concordia University, Montreal, Canada
{s_mujahi, g_sierr, rab_abdu, eshihab, wshang}@encs.concordia.ca

***Abstract*—Wearable apps are becoming increasingly popular in recent years. However, to date, very few studies examined the issues that wearable apps face. Prior studies showed that user reviews contain a plethora of insights that can be used to understand quality issues and help developers build better quality mobile apps. Therefore, in this paper, we mine user reviews in order to understand the user complaints of wearable apps. We manually sample and categorize 589 reviews from 6 Android wearable apps. Our findings indicate that the most frequent complaints are related to functional errors, lack of functionality, and cost. Our results are useful to the wearable developer community since they highlight the issues that users face and care most about.**

***Keywords*-Wear Apps; Users' Reviews; Google Play Store; Empirical Studies**

I. Introduction

Mobile apps are very popular and have been the focus of numerous studies in recent years [13], [15]. A fundamental change introduced by mobile apps is the way that they are released to users, which is through app stores. App stores allow users to directly provide feedback on the mobile apps through user reviews. Although these user reviews were meant to simply provide feedback about the apps, they proved to be much more useful. For example, studies have shown that they can be used to understand user problems so that developers can avoid low ratings, which can have a major impact on the app's user base and revenues [7], [9], [16].

More recently, wearable devices have been introduced, which complement handheld devices. Wearable devices i.e., smart watches and fitness trackers, are becoming increasingly popular and are expected to reach 101 million devices by 2020 [5]. Wearable devices provide developers with access to unique sensors that can be used to enhance the user experience [2]. As such, developers began to develop apps that are specifically designed to run on these wearable devices, called wearable apps. Wearable apps are different than handheld apps that run on mobile phones since they 1) often are very lightweight (resources wise), 2) meant to run on very small screens, 3) have access to a different set of sensors, and 4) heavily depend on the mobile device to perform the majority of the heavy processing. However, wearable devices have unique characteristics that pose challenges when compared to other platforms or devices [17]. To the best of our knowledge, very few studies have focused on wearable apps to date.

Therefore, similar to the prior studies on (handheld) mobile app reviews [8], [10], [11], [19], we also investigate user complaints but our study focuses on complaints from users of wearable apps. To perform our study, we manually classify 589 reviews belonging to 6 wearable apps. The reviews were tagged by the first two authors of the paper and grouped into 15 different categories. For each category, we measured the frequency of the complaints. Our findings indicate that functional errors, lack of functionality, and cost are the three most frequent complaints.

The rest of the paper is organized as follows. Section II presents and compares related work. Section III details our study design, including our collection and selection methodology. Section IV presents our results. Section V discusses the possible threats of validity. Section VI concludes the paper and outlines potential areas for future work.

II. Related Work

The work that is most related to our study falls into two main categories: work that leveraged mobile user reviews and work focusing on wearable apps.

A. Work Leveraging Mobile User Reviews

One of the first studies to leverage mobile app reviews was by Harman *et al.* [9]. In their paper, the authors studied the correlation of user reviews with key performance metrics such as the number of downloads. They found that there is a strong correlation between app ratings and its rank based on the number of downloads, suggesting that developers should pay close attention to their user ratings.

Other studies mined user reviews to better understand the contents of these user reviews. Khalid *et al.* [11] studied low-rated user reviews from 20 free iOS apps in order to help developers understand their nature. They exposed 12 types of complaints and found that feature requests, functional errors and, crashing apps were the most frequent reasons for negative reviews. Ha *et al.* [8] manually analyzed the user reviews of 59 Android apps to examine the impact of privacy and ethical issues. They found that only around 1% of the apps contain complaints related to privacy and ethical issues.

TABLE I: Statistics of Studied Android Wearable Apps

Wear App Name	Rating	Low Reviews	Sampled Reviews
Odyssey Watch Face	4.4	125	94
Wear Mini Launcher	4.4	133	99
InstaWeather for Android Wear	4.2	141	103
Watch Faces for Android Wear	4.1	154	110
WatchMaster - Watch Face	4.0	124	94
Luxury Watch Faces for Wear	3.9	115	89
Total		**791**	**589**

There are also a plethora of other works on mobile apps, that leverage users reviews for their techniques. Due to space limitations, we only discuss the most relevant studies in this section, however, we refer the reader to a recent survey by Martin et al. [13] for a more comprehensive list of studies on mobile apps.

B. Work Focusing on Wearable Apps

Very few studies have focused on the study of wearable apps, but many different paths are beginning to get explored in the domain. Recently in [20], Zhang *et al.* presented a formal semantics to statically model the notification mechanism of Android Wear, and contributed with the development of two domain-specific tools, one for test case execution and another for automated test generation. Ahola [4] exposed 3 issues and limitations in the Android Wear platform found during wearable app development that are better wear Internet connectivity, virtual button support for watch faces, and software configurable language support for voice input. From a different perspective, Lyons [12] did a study on the user perceptions of functionality and design of smartwatches, including android wearable devices. Based on user feedback and contrast to traditional watches, possible features for future wearable app are suggested. Min et al. [14] explored the battery usage of wearable apps and performed an online survey to get direct feedback and concerns from users. They found that most users do not complain about the battery usage of their wearable devices.

Chauhan *et al.* [5] did a previous categorization of smart watch apps from Samsung, Apple, and Android Wear. They used data from Android Wear Centre and GoKO [1], [3] as sources to get the wearable app identifiers for crawling their information; we applied the same approach to initialize our crawling phase. To the best of our knowledge, this is the first work focusing on the study of wearable app user complaints.

III. Study Design

The goal of our study is to determine the most frequent and negatively impacting user complaints of wearable apps. To do so, we mine the Google Play Store for the reviews of wear apps. In the following sections we describe our data selection and collection, as well as detail our review classification methodology.

A. Data Collection and Selection

For the purpose of our study, we select a number of wearable apps that have user reviews. First, we obtained the available Android Wear apps on Google Play Store by collecting their identifiers from two alternative app markets: *Android Wear Center* [1] and GoKo [3]. The two aforementioned sources have been used in prior work focusing on wearable apps [5]. Then, we mined the wearable apps using a data scrapper that we developed. The scrapper collected various information about each wear apps, including: the user review's text, its rating, the developer's reply to the review, if any, and the apps' overall rating. To enhance performance of the scrapper, it was deployed on a cluster of machines in order to distribute the requests sent.

In total, we mined the data for 4,722 wear apps that are developed by 2,732 unique developers. The 4,722 apps had 1,284,349 user reviews, which we mined. Since we are interested in user complaints, and similar to the prior study by Khalid *et al.* [11], we select the low-rated reviews (i.e., 1 and 2 star reviews) since they are most likely to contain the user complaints. Since we need a reasonable amount of reviews to perform our analysis, we only considered apps that had more than 100 low-rated reviews. After performing these steps, we randomly select 6 apps that have 791 reviews.

Since this is the first study to examine user complaints for wear apps, we opt to perform our analysis of the user complaints manually. Given that this manual classification is a time and resource intensive task, we selected a random statistically representative sample of complaints from each application. The sample sizes were selected to attain a 5% confidence interval and a 95% confidence level in the population being sampled. This random sampling process resulted in 589 total reviews varying from 115 to 154 reviews per app. The list of the studied wear apps, their overall rating on Google Play store, the number of total low rated reviews and the number of examined reviews is shown in Table I.

B. Manual Classification of User Reviews

Once we obtained all of the reviews, we categorize them in order to come up with the different types of complaints. To do so, we used a Grounded Theory type of technique [6], [18], where we took a random sample of reviews from all the selected apps and did a simple manual classification of them. This step was done mainly to come up with an initial set of categories that the reviews can be grouped into. In the end of this step, we came up with 15 different categories, which we call complaint types.

Once we came up with the initial 15 complaint types, we proceeded to categorize the sampled user reviews (in total 589 reviews). To facilitate the categorization of the reviews, we built a web-based tool that enabled the categorization of the review - presenting the review details and the respective developer reply, if a developer posted a reply to the review. The tool also had the option to add a new category in case a review belonged to a category that was not listed. Every review was tagged with all suitable categories, i.e., one review can

TABLE II: User Complaint Types Rank and Median Percentage for the Most Frequent Complaints.

Complaint Type	Description	Rank	Median (%)
Functional Error	A bug related to the functionality of the wear app	1	31.44
Lack of Functionality	Absence or deficiency of features in the wear app	2	17.02
Cost	Complaint about the wear app costs or business model	3	16.32
Connection & Sync Issues	Problems in connectivity with the wearable	4	15.21
Device Compatibility	The wear app is not compatible with a given device	5	12.74
Battery Drainage	The wear app is draining the battery excessively	6	9.09
UI Problems	Complaints about the interface design	7	8.72
Installation Problems	Issue while pushing the wear app to the wear device	8	5.16
Privacy & Ethical	Invasion of privacy or ethical concerns complaint	9	4.10
Feature Request	The user requires a specific new feature	10	3.96
Spam Notifications	The wear app generates many unwanted nottifications	11	3.58
App Crashing	The wear app stops completely, goes idle or restarts	12	3.54
Performance Issue	The wear app slows or over use the resource	13	2.73
Feature Removal	A feature has been removed after an update	14	1.06
Missing Notifications	The wear app lost or delayed notifications	15	0.91

have one or multiple tags based on its content. For example: if a user complaint mentions a battery drainage problem and also a connection issue, the review will be classified with the 'Connection & Sync Issues' and 'Battery Drainage' tags. In some instances, the user provided uninformative content in his/her review (e.g., "Just nonsense, I hated this game..."), in which case we put them in the 'Uninformative' category. Table II provides a list of the 15 complaint types, along with a brief description and an example review that falls in the complaint types.

IV. Results

Once all the reviews in our dataset are categorized into the different complaint types, we proceed to answer our research question. In particular, we are interested in knowing what issues users complain about.

Since wearable apps are an emerging trend, our goal is to understand the types of user complaints so that developers can anticipate potential problems and plan their quality assurance efforts accordingly. Similar to prior studies on user complaints for handheld device apps [11], we start by examining the different types of complaints based on the low-rated reviews of wearable apps. To come up with the different complaint types, we manually categorized the different wearable app reviews as mentioned earlier in Section III. We then rank the different complaint types based on their frequency in the examined reviews.

Table II shows the 15 different complaint types that we discovered from the wearable app reviews. For each category, we provide a brief description and an example review. From the table, we observe that many of the complaints types are related to the features provided by the wearable apps (e.g., feature removal, feature request), the behaviour of the wearable apps (e.g., app crashing, notifications, battery drainage) and external factors (e.g., the cost of the app, privacy & ethical issues).

Next, to distinguish between the the different complaint types, we measured the frequency of each complaint type. To do so, we follow the same approach used by Khalid *et al.* [11], where we measure the percentage of reviews that belong to each complaint type on a per app basis. We calculate the percentage per app since different apps can have a different number of reviews, and if we do not normalize per app, then apps with more reviews could bias our results. Once we calculate the percentage of reviews for each complain type, we take the median percentage (from all the wearable apps) and assign it to the complaint type. Finally, we rank all of the complaint types from 1 - 15, where 1 is the highest (i.e., most frequent rank) and 15 is the least ranked.

The third and fourth columns of Table II show the rank and median percentage of user reviews per complaint type. From the table we observe that complaints related to functional errors (i.e., bugs related to the functionality of the wearable app), cost (i.e., issues related to the business model of the wearable app) and lack of functionality (i.e., deficiencies in the functionality of the app) are the most frequent complaints for Android Wear apps.

The most frequent complaints from the wearable app users are related to functionality errors, cost and lack of functionality.

V. Threats to validity

Our study is subject to a number of internal and external threats to validity.

Internal validity: Due to our manual classification phase being time consuming, we did not cover all of our data set, instead we took a sample of our dataset. Also, our categorization is heavily dependent on the quality of the reviews provided by the users. As shown in prior studies, most user reviews contain useful information, however, in some cases different levels of details may lead to different complaint types.

External Validate: We found over 17,000 wearable app related user reviews but we filtered them down to 791, and hence, our data set can be considered small. On the same line of thought, the filtering phase for the wearable app related reviews may have discarded some useful information that did not match our filtering rules. Moreover, our study is performed on Android Wear apps, hence our findings may not generalize to wearable apps from other platforms.

VI. Conclusion and Future Work

Users provide direct feedback on their experience of mobile apps through user reviews. Prior work showed that user reviews can be mined to effectively determine user complaints to help developers understand the issues that users of handheld apps face the most, so they can be mitigated.

Given that wearable apps are a new trend that is only increasing in popularity, in this paper, we mine user reviews in order to understand the user complaints of wearable apps. We manually sample and categorize 589 reviews from 6 wearable apps. We find 15 unique complaint types that wearable users report in user reviews. Our findings indicate that the most frequent complaints are related to functional errors, cost and lack of functionality.

In the future, we plan to perform our study on more wearable apps and also examine some of the reasons for the most frequent and impacting user complaints.

References

[1] Android wear center. http://www.androidwearcenter.com. (Accessed on 09/09/2016).

[2] Creating wearable apps — android developers. https://developer.android.com/training/wearables/apps/index.html. (Accessed on 10/02/2016).

[3] Goko. http://goko.me. (Accessed on 09/09/2016).

[4] J. Ahola. Challenges in android wear application development. In *International Conference on Web Engineering*, pages 601–604. Springer, 2015.

[5] J. Chauhan, S. Seneviratne, M. A. Kaafar, A. Mahanti, and A. Seneviratne. Characterization of early smartwatch apps. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pages 1–6. IEEE, 2016.

[6] J. M. Corbin and A. Strauss. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, 13(1):3–21, 1990.

[7] L. V. Galvis Carreño and K. Winbladh. Analysis of user comments: An approach for software requirements evolution. In *Proceedings of the 2013 International Conference on Software Engineering*, ICSE '13, pages 582–591, Piscataway, NJ, USA, 2013. IEEE Press.

[8] E. Ha and D. Wagner. Do android users write about electric sheep? *Examining consumer reviews in google*.

[9] M. Harman, Y. Jia, and Y. Zhang. App store mining and analysis: Msr for app stores. In *Proceedings of the 9th IEEE Working Conference on Mining Software Repositories*, pages 108–111. IEEE Press, 2012.

[10] L. Hoon, R. Vasa, J.-G. Schneider, and K. Mouzakis. A preliminary analysis of vocabulary in mobile app user reviews. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*, pages 245–248. ACM, 2012.

[11] H. Khalid, E. Shihab, M. Nagappan, and A. E. Hassan. What do mobile app users complain about? *IEEE Software*, 32(3):70–77, 2015.

[12] K. Lyons. What can a dumb watch teach a smartwatch?: Informing the design of smartwatches. In *Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pages 3–10. ACM, 2015.

[13] W. Martin, F. Sarro, Y. Jia, Y. Zhang, and M. Harman. A survey of app store analysis for software engineering. *IEEE Transactions on Software Engineering*, PP(99):1–1, 2016.

[14] C. Min, S. Kang, C. Yoo, J. Cha, S. Choi, Y. Oh, and J. Song. Exploring current practices for battery use and management of smartwatches. In *Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pages 11–18. ACM, 2015.

[15] M. Nagappan and E. Shihab. Future trends in software engineering research for mobile apps. In *Proceedings of the 23rd IEEE International Conference on Software Analysis, Evolution, and Reengineering (SANER'16)*, 2016.

[16] D. Pagano and W. Maalej. User feedback in the appstore: An empirical study. In *Proceedings of the 21st IEEE International Requirements Engineering Conference*, RE '13, pages 125–134. IEEE Press, 2013.

[17] R. Rawassizadeh, B. A. Price, and M. Petre. Wearables: has the age of smartwatches finally arrived? *Communications of the ACM*, 58(1):45–47, 2015.

[18] C. B. Seaman. Qualitative methods in empirical studies of software engineering. *IEEE Transactions on Software Engineering (IST)*, 25(4):557–572, 1999.

[19] R. Vasa, L. Hoon, K. Mouzakis, and A. Noguchi. A preliminary analysis of mobile app user reviews. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*, pages 241–244. ACM, 2012.

[20] H. Zhang and A. Rountev. Analysis and testing of notifications in android wear applications. In *Proceedings of the 39th International Conference on Software Engineering*, 2017.