What Are You Complaining About?: A Study of Online Reviews of Mobile Applications

Claudia Iacob

Varsha Veerappa
Oxford Brookes University
Oxford, United Kingdom
{iacob, vveerappa, rachel.harrison}@brookes.ac.uk

Rachel Harrison

In this paper, we explore the content of online reviews of mobile applications to get a better understanding of the most recurring issues users report through reviews, and the way the price and the rating of an app influences the type and the amount of feedback users report. Results show that users tend to provide positive feedback, often associating it with requirements for additional features. Also, users tend to provide more feedback for the lower rated apps and the optimal price range was found to be between £2.25 and £3.50.

Online reviews. Mobile apps.

1. INTRODUCTION

More than 60% of online customers read online reviews before buying a product, reviews being 12 times more trusted than the descriptions that are provided for the products by the sellers. In the past few years e-commerce has become more and more popular, and the number of customer reviews that a product receives tends to grow exponentially. In the UK alone, almost half of the population (47%) has reviewed a product online, leading to the situation in which one single product may have thousands of reviews associated with it. There are several consequences to this. On the one hand, customers get more and more feedback from others on products they are interested in, and so are better supported in their purchase decision process. On the other hand, the ability to read all the feedback provided for a product becomes more and more limited as the number of reviews increases. It also becomes more difficult to spot the weak points of a product reported by others, the recurring issues reported across all the reviews of a product or a class of products, and the trends (if there are any) across the reviews. Mobile application (app) stores are not an exception, apps being often associated with hundreds of reviews. Even though research has looked (with varying degrees of success) into ways to summarize online reviews (Hu, 2004; Jindal, 2008), extracting design and usability information from online reviews (Iacob, 2013; Hedegaard, 2013), and the impact online reviews have on product sales (Bounie, 2002; Chevalier, 2006; Dellarocas, 2004) and customer behaviour (Jindal, 2010), direct questions on the content and the impact of app reviews have

not been addressed by prior work. We address this gap and aim to provide a better understanding of the recurring issues that users of apps complain about and of the relationships between the users' feedback and different characteristics of the apps (such as price and rating).

2. STUDY DESIGN

The overall goal of this study is to explore what users of mobile apps report on the apps they use through online reviews. We start by looking at the recurring issues reported by users through their reviews. Further on, we explore how the apps' price and rating influence the amount and the type of feedback users give for the apps. Finally, we look at the implications these findings have on the way people provide feedback and on the way they perceive the feedback provided by others.

2.1 Data Collection

We ran a survey of the mobile app stores available online and we chose Google app store as our data source because a) the categories used for classifying apps are similar to the ones used by other stores, and b) the number of apps in each category and the number of reviews per app compare favourably to the numbers in other stores, supporting possible further generalizations across app stores. We selected the top 6 most popular categories (i.e. the categories containing the largest number of apps) and, for each category, we randomly generated a number n_k of app identifiers, where n_k is a statistically significant sample size for

the population represented by the total number of apps in that category. The categories selected and the numbers of apps for each category are: *Personalization* (54 apps), *Tools* (14 apps), *Books and References* (45 apps), *Education* (27 apps), *Productivity* (19 apps), and *Health & Fitness* (10 apps).

2.2 Data Extraction and Storing

For each randomly selected app, we automatically extracted and stored its overall rating, the total number of ratings assigned to it, its price, size, number of installs, the date of its last update, its

current version, and the reviews provided by the users. For each review, we automatically collected the date it was posted, the rating the user gave, the device associated to it, and the version of the app used by the user, as well as the title and the text of the actual review. Out of the 169 apps randomly selected, 8 apps had no reviews assigned to them which left us with 161 reviewed apps and a total of 3279 reviews. The average rating of the 161 apps considered is 4.27 (on a scale of 1 to 5), while the average number of ratings per app is 326.83 (not all ratings were associated with comments). Additionally, the average price per app is £1.92.

Figure 1: Recurring classes of codes and refined codes across mobile apps reviews

2.3. Data Analysis

For analysing the reviews collected, we defined and used a coding scheme able to capture the recurring issues users report through reviews. For defining the coding scheme, we randomly selected 125 reviews and had two coders independently annotate them. Once this process was completed, the annotations made by the two were further discussed and categorized into *classes of code*. All together, we found 9 classes of codes, $c_1, \dots, c_9$. To assess the reliability of the classes of codes, another sample of reviews was selected for analysis and annotation. Following discussions, the classes of codes were further refined into 75 *refined codes*, each refined code belonging to a class of codes. Having elicited the classes of codes and the refined codes, the coding process (Figure 1) of a review, R , consists of the following steps: 1) the text of R is divided into significant snippets of text $\{s_1, s_2, \dots, s_n\}$, each snippet of text being considered a *snippet raw code*, 2) each snippet code, s_i , is associated with a *refined code*, r_j , belonging to a class of codes, c_m , 3) we form the *R-tuple* $= (c_m, r_j, s_i)$ which is associated with the

review R . Consequently, a review may be associated with many R-tuples, and may have more than one refined code assigned to it. In the example described in Figure 1, R is associated with the following R-tuples: (positive feedback, functionality, s_1), (comparative feedback, positive, s_2), (requirements, missing GUI feature, s_3), (reporting, medium bug, s_4).

3. RESULTS

Users tend to provide positive feedback. Out of the entire feedback provided through reviews (i.e. all the R-tuples), 49.02% reported positive feedback for the apps (Figure 2). Negative feedback was rarer: only 6.47% of the R-tuples reported negative feedback, the major issues users reported as negative being the use of apps on specific devices, the apps' functionality, their GUIs, the apps' speed, and the apps' size. Users were also prone to provide comparative feedback (comments which compared the app reviewed with a similar app already existing on the market), 3.32% of all the feedback collected providing comparative remarks between various apps, and

81.9% of these remarks were favourable to the app reviewed as opposed to another one.

Usually, apps are considered worth their price and are rarely uninstalled. Money is a recurring theme across the reviews, 7.80% of the feedback collected reported on the apps' value and similar price-related issues. Notably, almost half of the feedback related to pricing (47.80%) indicated that the apps are worth their price. In addition to reporting the app's value, users also made remarks on whether they asked for a refund and whether they got refunded; 13.59% of the price-related feedback suggested that the users asked for a refund. Reports on uninstalling the app were rarer (7.91%), but were still more than those reporting switching from a free version of the app to a priced one (5.07%).

Figure 1: Distribution of the classes of codes across the feedback collected

Reviews are used for expressing requirements and bugs. 23.3% of the feedback analysed constitutes requirements from users. They can indicate the need for completely new functionality or GUI features or they can express the need for a different preference for existing functionality or GUI features. Other types of requests were for more updates where the users ask the developer to continuously improve the app, more features, more options to improve customization, and versions for new devices. Also, we found that users like to report problems encountered while using apps, 11.51% of the feedback pointing out various issues users had faced. The majority of these reports described major bugs (38.10%), severe problems

which prevent the users from using the app. Medium severity bugs accounted for 28.33% of the entire feedback reporting usage issues. Lastly, minor bugs accounted for 20.49% of the reporting feedback. Some of the bugs reported received answers for fixing them from other users through reviews, such fixes accounting for 4.95% of the reporting feedback.

A quarter of the reviews recommended apps to other users. The feedback related to usability represented 5.11% of the entire feedback considered. Out of the usability feedback, 26.62% reported that apps are easy to use whereas only 4.64% of such reports expressed difficulty in using apps. As far as learnability is concerned, 3.71% users considered apps difficult to learn whereas 0.61% considered them easy to learn. Moreover, 8.66% of the usability-related feedback reported the fact that apps are difficult to setup. A quarter of all the feedback related to usability issues consisted of recommending the app reviewed. Customer/developer support was addressed by 2.39% of the feedback considered. A large majority of this feedback (61.58%) was positive, users being satisfied with the support they received while using apps. However, complaints on the misleading description of apps were also common (21.85%). Additionally, we identified versioning as a recurring theme across users' feedback, reports on such issues accounting for 2.24% of the entire feedback considered. While most of the version-related references were positive remarks related to the improvements brought by updates (49.29%), complains about a specific update bringing major usability issues to the whole app represented 23.94% of the version-related feedback.

Reviews often report connected issues. It is common for one review to report several issues and so we looked at the most recurring issues users report when reviewing apps. For that, we considered all the pairs (R', R'') , where $R' = (c_x, r_a)$, and $R'' = (c_y, r_b)$ are partially defined R-tuples. For each pair, we computed the percentage of reviews associated with both R' and R'' over the entire number of reviews considered; the most significant fragment of our results is described in Figure 3. Often, users associate overall positive feedback with requirements for missing functionality for apps. Even if users were happy with the app overall, they would always look for improvements. Moreover, when identifying missing logic features of the app reviewed, users often provide examples of several such features within a single review. Similarly, providing positive feedback on the functionality of an app is often broken into several bits such that various features are commented on independently.

	positive overall	positive functionality	positive GUI	positive customization	positive device	negative overall	negative device	comparative positive	price not worth	price worth	requirements missing logic ft	requirements missing GUI ft	requirements diff pref for GUI ft	reporting major bug	reporting minor bug	usability recommend it	usability easy to use	customer support positive
positive overall	0.00%	5.85%	5.36%	1.79%	1.40%	0.00%	0.39%	2.28%	0.54%	5.12%	6.40%	2.31%	1.76%	0.97%	2.80%	1.92%	1.86%	1.95%
positive functionality	5.85%	1.73%	1.64%	0.42%	1.28%	0.06%	0.15%	2.13%	0.12%	1.64%	1.67%	0.27%	0.36%	0.24%	0.33%	0.45%	0.70%	0.54%
positive GUI	5.36%	1.64%	0.00%	0.64%	0.27%	0.00%	0.06%	0.51%	0.09%	0.70%	0.94%	0.27%	0.30%	0.21%	0.21%	0.27%	0.36%	0.24%
positive customization	1.79%	0.42%	0.64%	0.06%	0.06%	0.03%	0.03%	0.18%	0.00%	0.21%	0.21%	0.00%	0.15%	0.00%	0.06%	0.24%	0.18%	0.18%
positive device	1.40%	1.28%	0.27%	0.06%	0.00%	0.00%	0.24%	0.51%	0.03%	0.39%	0.42%	0.06%	0.21%	0.15%	0.15%	0.09%	0.09%	0.21%
negative overall	0.00%	0.06%	0.00%	0.03%	0.00%	0.00%	0.33%	0.00%	0.88%	0.00%	0.00%	0.06%	0.09%	1.64%	0.03%	0.03%	0.00%	0.03%
negative device	0.39%	0.15%	0.06%	0.03%	0.24%	0.33%	0.00%	0.03%	0.21%	0.03%	0.21%	0.00%	0.00%	1.49%	0.00%	0.00%	0.00%	0.00%
comparative positive	2.28%	2.13%	0.51%	0.18%	0.51%	0.00%	0.03%	0.03%	0.00%	0.82%	0.57%	0.00%	0.09%	0.06%	0.21%	0.36%	0.42%	0.30%
price not worth	0.54%	0.12%	0.09%	0.00%	0.03%	0.88%	0.21%	0.00%	0.00%	0.00%	0.42%	0.00%	0.03%	0.00%	0.18%	0.00%	0.03%	0.00%
price worth	5.12%	1.64%	0.70%	0.21%	0.39%	0.00%	0.03%	0.82%	0.00%	0.03%	0.57%	0.00%	0.27%	0.03%	0.24%	0.30%	0.39%	0.24%
missing logic ft	6.40%	1.67%	0.94%	0.21%	0.42%	0.00%	0.21%	0.57%	0.42%	0.57%	1.58%	0.73%	0.60%	0.18%	0.57%	0.12%	0.21%	0.12%
missing GUI ft	2.31%	0.27%	0.27%	0.00%	0.06%	0.06%	0.00%	0.00%	0.00%	0.00%	0.73%	0.12%	0.15%	0.00%	0.21%	0.00%	0.00%	0.00%
diff pref for GUI ft	2.31%	0.36%	0.30%	0.15%	0.21%	0.09%	0.00%	0.09%	0.03%	0.27%	0.60%	0.15%	0.06%	0.15%	0.24%	0.03%	0.00%	0.00%
reporting major bug	0.97%	0.24%	0.21%	0.00%	0.15%	1.64%	1.49%	0.06%	0.00%	0.03%	0.18%	0.00%	0.15%	0.18%	0.15%	0.00%	0.00%	0.06%
reporting minor bug	2.80%	0.33%	0.21%	0.06%	0.15%	0.03%	0.00%	0.21%	0.18%	0.24%	0.57%	0.21%	0.24%	0.15%	0.18%	0.06%	0.09%	0.06%
usability recommend it	1.92%	0.45%	0.27%	0.24%	0.09%	0.03%	0.00%	0.36%	0.00%	0.30%	0.12%	0.00%	0.03%	0.00%	0.06%	0.03%	0.18%	0.24%
usability easy to use	1.86%	0.70%	0.36%	0.18%	0.09%	0.00%	0.00%	0.42%	0.03%	0.39%	0.21%	0.00%	0.00%	0.00%	0.09%	0.18%	0.00%	0.09%
support positive	1.95%	0.54%	0.24%	0.18%	0.21%	0.03%	0.00%	0.30%	0.00%	0.24%	0.12%	0.00%	0.00%	0.06%	0.06%	0.24%	0.09%	0.00%

0.0-10 11-0.5 0.51-1 1-2 2-7

Figure 3: Distribution of code sequences across the reviews considered

Another trend we noticed is that users tend to associate overall positive feedback with more specific positive feedback, namely positive remarks on the apps' functionality and/or GUI. Other times, the positive overall feedback came as a consequence of the way a particular feature of the app reviewed worked. Similarly, positive overall feedback was also associated with remarks on an app being worth its price, users often relating overall high satisfaction with an app's value. To a lesser extent, rating an app negatively was associated with comments suggesting that the app is not worth its price. Surprisingly, providing positive feedback for an app does not necessarily imply recommending the app to other users, such associations being rare. Moreover, positive overall feedback does not necessarily point to a bug-free app, reports on minor bugs often being associated with positive overall comments. However, once an app was associated with a report of a major bug, it was rare that the app was also positively commented on. On the other hand, negative overall feedback goes hand in hand with reports of major bugs, users usually providing such feedback after struggling to either make the app work in the first place or use it without the app closing or freezing unexpectedly. More specifically, negative feedback on the specific device being used is mainly associated with reports of major bugs, indicating that most of the users' frustration comes from the device they are using and its limitations.

Apps rated lower get more feedback. As expected, positive overall feedback increases with the ratings, apps rated higher getting a larger

amount of positive feedback (Figure 4A). Similarly, users' reports on the apps' value increase as the ratings increase, none of the apps we considered with a rating in (2, 3] being considered worth its price. On the other hand, reports on apps not worth their price decrease as the ratings increase. Users identified a larger number of missing features for apps rated lower, while the number of reports of medium complexity bugs decrease with the rating. Major bug reports decrease in number as ratings increase, as well. Additionally, users provide a larger amount of feedback for apps rated lower than 3, the average number of R-tuples per review being 2.13 for apps rated lower than 3 in comparison to 1.91 and 1.94 for apps rated between (3, 4] and (4, 5] respectively.

Inexpensive apps are seldom considered worth their price. We identified 6 price ranges of interest to our findings, namely [0.50, 0.60), [0.60, 1.20), [1.20, 2.25), [2.25, 3.50), [3.50, 5.0), [5.0, 20.0). Apps costing between £2.25 and £3.50 received the maximum amount of feedback indicating they are worth their price (Figure 4B). Moreover, value increases with the price in this range, whereas it suddenly decreases to lower values. Additionally, in the same range ([2.25, 3.50)) apps were mostly considered easy to use and were the most highly recommended ones among users. Surprisingly, minimum costs were associated with minimum value, users rarely mentioning cheap apps being worth their price. Moreover, the cheapest apps attracted the largest number of comments on missing functionality from users. In contrast, apps costing more than £5.00 were rarely associated

with feedback indicating missing functionality. Reports of major bugs are more often at the extremes, i.e. among the cheapest and the most expensive apps. On the one hand the complexity of the more expensive apps may be responsible for triggering a larger number of bugs and, on the other hand, cheaper apps tend to be less rigorously tested before being released on the market, leading to users identifying a larger number of bugs.

Figure 4: Feedback distribution across rating ranges (A) and price ranges (B)

The apps which were associated with the maximum value and the most expensive apps were also associated with larger amounts of comparative feedback favourable to them. In other words, users often remarked that an app worth its price was better than another similar one and/or they

favoured the most expensive app between two apps with similar features. In terms of the amount of feedback provided by users for each price range, the most commented on were the cheapest apps with an average of 2.61 R-tuples associated with each review. Following the cheapest apps were the apps costing between £2.25 and £3.50, i.e. the price range in which the apps were the most recommended, and in which the apps were voted as being worth their price.

4. IMPLICATIONS AND LIMITATIONS

This work impacts two types of professionals: app developers who wish to get a better understanding of the issues users usually report on and app stores maintainers who would like greater app specific feedback from users or synthetic summaries of the feedback provided. We identified 9 recurring themes across the feedback considered, suggesting possible ways to customize the retrieval and the presentation of app specific feedback. Most of the feedback analysed was positive. Moreover, both positive and negative feedback reported on the same app specific characteristics such as size, speed, functionality, the GUI, and the device used. Users also suggested additional functionality for the apps they were using by either expressing preferences for existing functionality or suggesting the development and integration of new functionality, the latter being a valuable source of innovation for app developers. Users provide such suggestions after using the app, appropriating it to their own needs and carefully considering its limitations. User feedback may also indicate the presence of bugs. Moreover, fixes for bugs are also sometimes indicated. Thus our work can also be used to support app testing.

We also found that major bugs usually trigger additional negative feedback and that positive feedback goes hand in hand with notifications of missing functionality. Users would mostly provide negative feedback after having struggled to make an app work. However, users who were happy with an app were often motivated to support its improvement. Also, the true spike in positive feedback only occurs for apps with a rating higher than 4, whereas negative feedback is proportionally distributed across ratings. Price-wise, the optimal price range for users' feedback is [£2.25, £3.50]. This is an indication for app developers as to how much users are willing to pay for an app in order for them to consider it to be of good value. Surprisingly, cheaper apps are not necessarily commented on favourably.

Users tend to provide more feedback for either lower rated apps or cheaper apps. In both cases, they provide more detailed descriptions of the problems they faced when using the apps, the

missing features of the app, and the usability issues they encountered. From the developers' perspective, this is an efficient and reliable way to perform usability testing on their apps. The quantity of feedback provided per app is a measure of the level of development the app has reached and can be used for evaluating both the impact the app has already made and the limitations and bugs yet to be addressed. We looked at reviews from one app store (all written in the English language) and the results of our study are drawn from these reviews only. We believe that the findings we identified may be recurrent over other app stores. However, more empirical work is needed for generalizations of our results over various app stores.

5. CONCLUSIONS

We looked at 3279 reviews of 161 apps available online with the aim of understanding what recurring issues users like to report. We identified 9 classes of feedback, namely positive, negative, comparative, price related, request for requirements, issue reporting, usability, customer support, and versioning. Each class of feedback was further analysed for in-depth understanding. Additionally, we identified relationships between a) the overall rating of apps and the type and amount of feedback they receive, and b) the price of apps and the amount and type of feedback they receive. Our findings are of use to app developers and testers, maintainers of repositories and also HCI designers. These results are related to the way the app ratings reflect the users' feedback, and the price range users perceive as the best value for money.

6. REFERENCES

- Bounie, D., Bourreau, M., Gensollen, M., Waelbroeck, P. Do Online Customer Reviews Matter? Evidence from the Video Game Industry, Telecom ParisTech Working Paper No. ESS-08-02
- Chevalier, J.A., Mayzlin, D. The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*, 43(3), 345-354, (2006).
- Dellarocas, C., Awad, N.F., and Zhang, X. Exploring the Value of Online Product Ratings in Revenue Forecasting: The Case of Motion Pictures, *ICIS2004*, 379-386, (2004).
- Hu, M., Liu, B. Mining and summarizing customer reviews. *KDD '04*, 168-177, (2004).
- <http://www.itu.int/ITU-D/ict/facts/2011/material/ICTFactsFigures2011.pdf>
- Jacob, C., Harrison, R. Retrieving and analyzing mobile apps feature requests from online reviews. In *Proceedings of the 10th Working Conference on Mining Software Repositories (MSR '13)*. 41-44, (2013).
- Jindal, N., Liu, B. Opinion spam and analysis. *WSDM-2008* (2008).
- Jindal, N., Liu, B., Lim, E. Finding unusual review patterns using unexpected rules. *CIKM '10*, (2010), 1549-1552.
- Hedegaard, S., Simonsen, J. G. Extracting usability and user experience information from online user reviews. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. 2089-2098, (2013).