This article has been accepted for publication in IEEE Software but has not yet been fully edited.
Some content may change prior to final publication.

Is It Worth Responding to Reviews? A Case Study of the Top Free Apps in the Google Play Store

Stuart McIlroy†, Weiyi Shang‡, Nasir Ali†, Ahmed E. Hassan†
School of Computing, Queen's University, Canada†
Department of Computer Science and Software Engineering, Concordia University, Canada‡
Email: {mcilroy, nasir, ahmed}@cs.queensu.ca†, shang@encs.concordia.ca‡

***Abstract*—The value of responding to a user review of a mobile app has never been explored. Our analysis of app reviews and responses from 10,713 top apps in the Google Play Store shows that developers of frequently-reviewed apps never respond to reviews. However, we observe that there are positive effects to responding to reviews (users change their ratings 38.7% of the time following a developer response) with a median increase of 20% in the rating.**

I. Introduction

App stores provide feedback mechanisms for users by allowing a user to rate an app using a five-star rating system and to write a short review. Addressing user feedback is an important part of developing and maintaining popularity in app stores. A viable mechanism for addressing user feedback is through personally responding to a particular user review.

Developers are able to respond to a complaint or thank the user for kind remarks about the app. The response may motivate the user to change the rating of their app review or to write a more positive review. However, developers have limited time. Responding to reviews takes away time that the developer could use to enhance their app. It is not clear how often users change their rating after a response, if at all. Additionally, it would be beneficial to know which types of reviews, i.e., in terms of content, are most likely to have their rating updated if the developer were to respond to the review. A strategic choice should be made to respond to reviews, which are most likely to lead to a positive update of their ratings. To the best of our knowledge, there exist no prior research investigating how developers respond to reviews or the value of responding to reviews.

In this paper, we empirically investigate app reviews and the responses to the reviews from the perspective of developers of the top apps in the Google Play Store. Through an analysis of reviews and developer responses for reviews for the top 10,713 apps in the Google Play Store over a period of two months, we explore the following research question:

RQ: What is the value of responding to reviews?
Developers of 13.8% of the studied apps responded to at least one review during the studied time period. The most-reviewed apps never responded to a review during our study period. Users change their rating 38.7% of the time following a developer response. The median rating change is a one-star increase out of five.

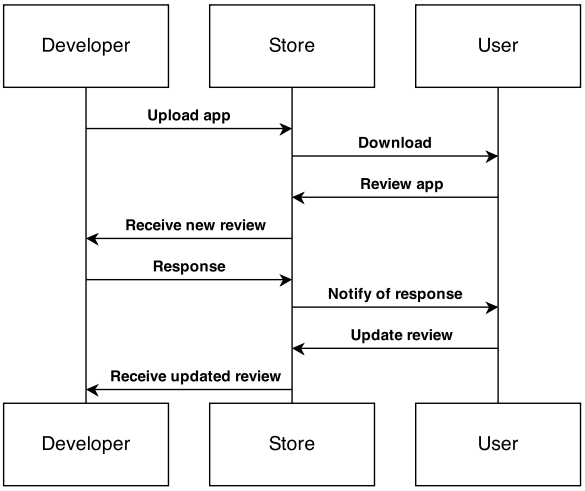

Fig. 1. The review and response process between developers and users.

TABLE I
DATASETS OF PRIOR WORK IN MINING MOBILE REVIEWS

Paper	App Store	Apps	Reviews
Iacob and Harrison [1]	Google Play Store	161	3,279
Galvis and Carreno [2]	Google Play Store	2	710
Fu et al. [3]	Google Play Store	171,493	13,286,706
Chen et al. [4]	Google Play Store	4	169,097
Pagano and Maalej [5]	Apple App Store	1,100	1,126,453

Takeaway Message: The results from our study suggest that there is value in responding to user reviews. Users are likely to update their star ratings upwards.

II. Background and Related Work

A. Rating an App

Once a user downloads an app, the user is able to leave a rating, a review or both. The rating and review can both be updated.

Developers are able to respond to a review by any user. The developer's response is public for anyone to see (not just that particular user). The user is then notified that the developer has left a response. Figure 1 shows the process of rating an app.

Digital Object Indentifier 10.1109/MS.2015.149 0740-7459/$26.00 2015 IEEE

B. Related work

Previous work confirms that reviews of mobile apps have a major impact on the success of an app [6–8]. Harman *et al*. show a strong correlation between app ratings and the total downloads of an app [6]. User reviews contain information that could help developers improve the quality of their apps, and increase their revenue. Kim *et al*. [7] conducted interviews of app buyers and shows that reviews are one of the key determinants in the user's purchase of an app. Similarly, Mudambi *et al*. [8] showed that user reviews have a major impact on the sales of online products.

The importance of user reviews motivates many recent studies on analyzing and summarizing user reviews for mobile apps as Table I shows. A recent study by Pagano and Maalej analyzed the content of reviews of both free and paid apps in the Apple App Store [5]. Guzman *et al*. [9] identify app features in the reviews using natural language processing techniques and leverage sentiment analysis to identify whether users like the features. Maalej *et al*. [10] propose an automated approach that automatically classifies reviews into four categories: bug reports, feature requests, user experiences, and ratings. Iacob and Harrison [1] built a rule-based automated tool to extract feature requests from user reviews of mobile apps – their approach identifies whether a user review contains a feature request or not. Chandy and Gu identified spam messages in reviews on the Apple App Store [11]. Carreño *et al*. [2] used opinion mining techniques and topic modelling to extract requirements from user reviews. Fu *et al*. present an approach that discovered inconsistencies in apps and analyzed the negative reviews of apps using topic analysis [3]. Khalid *et al*. manually analyzed and categorized one- and two-star reviews [12]. They manually identified the different issues that users complained about in mobile apps. Chen *et al*. propose the most extensive summarization approach to date [4]. They remove un-informative reviews, and prioritize the most informative reviews before presenting a visualization of the content of reviews. Our work differs from these studies as we focus on the value of responding to reviews, which has never been studied before.

C. Mobile App Analytics

Vision Mobile performed a survey of 7,000 developers and found that 40% of developers make use of user analytics tools [13] and 18% use crash reporting and bug tracking tools. Previous studies also highlight that app developers need analytics tools. For example, Pagano and Bruegge conducted a study on how feedback occurs after the initial release of a product [14]. The authors concluded that there is a need to structure and analyze feedback, particularly when it occurs in large quantities.

Nowadays, there exists many app analytics companies, e.g., App Annie (http://www.appannie.com/app-store-analytics/), that specialize in giving developers tools to understand how users interact with the developers' apps, how developers generate revenue (in-app purchases, e-commerce, direct buy), and the demographics of app users. These app analytics companies also provide developers with overviews of user feedback and logged crash reports. Google has promoted their own extensive analytics tools for Android developers as a key competitive differentiator relative to other mobile stores. The tools measure how users are using an app (e.g., identify the locations of users and how they reached the app). The tools also track sales data (e.g., tracking how the developer makes money through in-app purchases and calculating the impact of promotions on the sales of an app). However, other than crash reporting tools much of the analytics tools available today are mostly sales oriented instead of being software quality oriented.

III. Empirical Study Design

In this section, we present the design of our study and the data collection and processing methods used in our study.

A. Data Selection

We focus on top apps of the Google Play Store, since top apps often have a large amount of reviews and are more interested in maintaining and growing their user base (and their ratings). Our criteria for selecting an app store is based on its popularity (the Google Play Store is one of the most popular app stores), the ability to respond to reviews (iOS store does not support responding to reviews) and the availability of tools to automatically collect information from the app store. We collected the reviews and responses from 12,000 free-to-download apps from the Google Play Store. Across thirty different categories, e.g., Photography, Sports and Education, we selected the top apps in each category in the USA based on Distimos ranking of apps for a total of 12,000 (Distimo ranks the top 400 apps for each of the 30 categories). Distimo is an app analytic company (http://www.distimo.com/leaderboards/google-play-store/united-states/top-overall/free). We used Distimo's Spring 2013 top app list. We chose apps that were popular one year ago because we are interested in studying stable mature apps that had not been released recently to avoid the expected frequent burst of reviews following the early releases of an app [5].

B. Data Collection

We developed a crawler to extract the apps information such as app name, user ratings, and reviews. The crawler simulates a mobile device and interfaces with the Google Play API as a regular mobile device. We selected the Samsung Galaxy S3 phone as our simulated device since it is one of the most popular Android devices. We modified the crawler to only collect the relevant information for our study. We instituted a timer to pause the crawler to avoid issuing too many requests and we scaled the crawler over multiple machines to distribute the load.

We ran the crawler on a daily basis over a period of approximately two months beginning on January 1^{st} 2014 to March 2^{nd} 2014. 1,287 of the 12,000 top apps were not accessible during our crawl (e.g., some app were removed from the store). Hence, we collected data from 10,713 top

Digital Object Indentifier 10.1109/MS.2015.149 0740-7459/$26.00 2015 IEEE

apps. 11,047 different releases of apps were collected in the studied time period. An app can have multiple releases. In our dataset, on average, an app has 0.86 releases, i.e., most apps did not publish a new release during our studied release period. 4,073 apps published at least one release. Apps that published at least one new release had an average of 2.28 releases.

A recent study by Martin *et al*. [15] notes that all stores do not provide access to all their reviews. To ensure that we have all the available reviewers, we collected all the reviews on daily basis. A limitation of the Google Play Store is that only the 500 latest reviews per app are accessible. The crawler is unable to access any older reviews. That means if more than 500 reviews occur within the 24-hour period between runs of our crawler, then the crawler will not collect those reviews. This limitation means that we have a conservative estimate of the number of reviews for 20 (0.19%) apps that received more than 500 reviews per 24 hour time period.

IV. Approach

In this section, we present our approach to answer the research question.

A. Manual analysis

Pagano and Maalej observed 17 topics of apps reviews in App Apps store [5]. We manually labelled a statistically representative sample of 384 reviews from the Android top apps using the 17 topics that were observed by Pagano and Maalej. We also manually labelled responses from the 111,099 reviews with responses that occurred during the studied time period. Since Pagano and Maalej did not study the types of responses, we followed an iterative process to discover the types of responses until we could not find any additional types. The amount of manually examined reviews and responses is the number required for a statistical sample with a confidence level of 95% and a confidence interval of 5%. In total, we spent approximately 8 hours to manually analyze and label each review and response. The third co-author reviewed the labels for consistency. If both co-authors disagreed, they came to a consensus, which occurred for very few reviews.

B. Automated analysis

To complement our manual analysis of responses, we performed an additional automated analysis to calculate the average rating change for reviews with responses, the probability of a rating change, and the magnitude of the rating change. We then separated the reviews into 25 automatically generated topics (using LDA [16]) and examined which topics were most likely to lead to a positive change in rating. Our choice of 25 topics is motivated by a desire for general topics that are broad and that most developers would face. Moreover, Pagano and Maalej observed 17 topics for app reviews [5]. We chose 25 to make sure that we observed at least their 17 topics.

For 20 days from April 7th to April 27th we monitored if reviews changed either the rating, comment or response. We denoted a review and all subsequent changes as a review chain. The median review-chain length is 2 (meaning one review and one response). The maximum review chain is 8. We automatically analyzed 15,208 review chains in total.

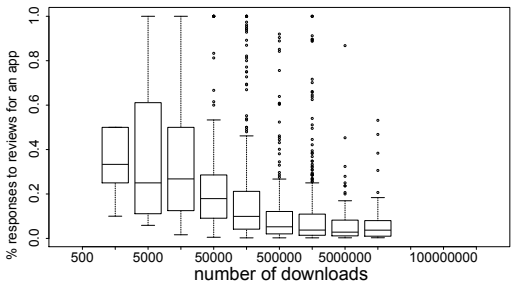

Fig. 2. Percentage of responses to the number of total reviews for each app separated by the number of downloads. Excluding apps with zero responses.

V. Results

A. Automated analysis

Most apps do not respond to reviews. Only 13.8% of 10,713 apps responded to reviews during the studied time period. As Figure 2 shows, apps with greater number of downloads never respond, but some apps in the mid-range of number of downloads responded often. The apps with high response percentages have a low number of reviews (possibly indicating that apps with a large number of reviews are overwhelmed already with their reviews).

Responses often lead to a positive change in review rating. Looking at all review chains, we find that 38.7% of the users changed their rating after a response. We also find that the median change in rating was a positive increase of one star (20% increase). This finding demonstrates that developers can benefit from responding to reviews. However, since only 13.8% of the apps have responded to their reviews, developers may not realize that responding to reviews has such a large benefit (a positive median increment of one star). Moreover, we find that the average star rating for the apps that do not respond to reviews is only 1.7. Such low star rating indicates that these apps indeed need a chance to increase their rating by responding to reviews. Some users even updated their review to notify the developer that the response had solved their problem or that the user was thankful that the developer had directly responded to them. We also find that most reviews with responses are low-rated reviews with an average of 2.2 stars. This finding supports prior research, which targeted negative reviews (1 and 2 star reviews [17–19]) as being reviews of great interest and concern to developers of mobile apps (over higher star-rating reviews).

The most common review topic that received responses, at 8% was about crashing. The issue of crashing is a serious one and has a great impact on the experience of the user. It is understandable that developers focus on these reviews. We find that the chance of a rating change for each review topic was distributed between 15% and 42%. The two topics with highest chance of a rating change were about notifications and not being able to connect to the app.

Digital Object Indentifier 10.1109/MS.2015.149 0740-7459/$26.00 2015 IEEE

TABLE III
TYPES OF DEVELOPER RESPONSES TO REVIEWS IN OUR SAMPLE.

Response Type	Description	Amount
Instructions	A developer provides assistance on how to use the feature or steps to fix the problem the user is having	120
Request Contact	A developer asks the user to contact the developer through a given email.	119
Thanks	A developer thanks the user for a positive review.	49
Next Release Fix	A developer states that a fix for the user problem will appear in the next release.	31
Current Release Fix	A developer states that the problem is already fixed in the current release. The user is asked to update their software.	26
Next Release Feature	A developer states that the requested user feature will appear in the next release.	12
Question	A developer asks for clarification.	9
Other	A response is in another language or does not address the user review.	6
Rating-review mismatch	A developer asks why the positive review does not match the negative rating or vice-versa.	6
Current Release Feature	A developer states the current release contains the feature. The user is asked to update their software.	3

TABLE II
TYPES OF REVIEWS THAT ARE WRITTEN BY USERS IN OUR SAMPLE.

Review Type	Description	Amount
Praise	A user states they are pleased with the app.	129
Bug report	A user states there is an error or unexpected behavior occurring.	126
Dispraise	A user complaints about the content or features of the app.	120
Request	A user makes a request for a new feature or addition to the app.	89
Update issue	A user preferred the previous release of the app.	36
Other app	A user mentions a competitor to the app.	10
Other	Another language.	10

These are specific issues that can be addressed by developers. Users that wrote reviews concerning problems associated with a specific Samsung phone were the most likely to change their review following a response. The responses gave specific advice about the phone.

The most common response topic, at 7%, was concerned with developers notifying the user that a requested feature is either in development or is planned. Once again, the most common response topic is not the one associated with the greatest rating change. The response topic that had the highest chance of changing a rating (with 39.1% chance) was a topic on notifying the user that the issue, about which the user had originally complained, had been resolved. The users may not have known that the issue was fixed and being told personally resulted in a rating change.

B. Manual analysis

We next looked at the review and response types that occur in our manually labelled data. In our manual analysis, more than one type can occur in the same review or response. Any review or response that does not conform to one of the types is considered as 'other'. The 'other' reviews and responses were usually written in a language other than English or were not written in coherent English.

We only found six out of 17 topics that are observed by Pagano and Maalej (shown in Table II). The six observed topics are praise, bug report, dispraise, request, update issue and other apps. We think the reason is that we focus on top free apps from Google Play Store, while Pagano and Maalej focus on both paid and free apps in Apple App Store. Some of the topics, such as dissuasion, would only appear for paid apps and some topics, such as howto, are more likely to appear for non-top apps. In addition, we could not clearly differentiate between a feature request and an improvement request, since we are not domain experts of the apps. We combined both topics as one topic, i.e., request. We observed that reviews contained a mixture of topics between praise and dispraise. The most common issues were praise for the app, followed by the reporting of bugs and dispraise about the content of the app.

We also find ten common response types in developer responses, as Table III shows. The ten types are instructions, request contact, thanks, next release fix, current release fix, next release feature, question, rating review mismatch, current release feature and other. The most common response type is instructions on how to solve the user problem, the second most common being a canned request to email the developers. The other types include reassurances that a problem is already fixed or that the problem would be fixed in an upcoming release. The same reassurances are provided to users who complained about a lack of a feature. The last three types are either thanking the user for leaving a kind review, asking why the negative rating of the review does not match the positive review or inquiring for further information about the user's complaint.

Finally we matched the reviews and responses in a table to show which response types occur most often with review types. As Table IV shows, most responded-to-reviews are praise, requests, and bug reports reviews. The majority of the

Digital Object Indentifier 10.1109/MS.2015.149 0740-7459/$26.00 2015 IEEE

responses to these reviews are either direct instructions from the developer on how to solve the problem or the developer asking the user to contact them by email. We find that often a user would leave praise for an app but then either have a request or a minor problem to which the developer would respond.

VI. Threats to Validity

Some threats could potentially limit the validity of our results. We now discuss such threats and how we control or mitigate them.

Construct Validity. We investigate the issues raised in reviews using LDA. LDA may not have optimized results when applied on short text, such as tweets and app reviews. However, prior research has shown that LDA can successfully extract topics from tweets [20]. Since, we manually labelled our dataset of reviews with the different issue types, some reviews may have been incorrectly labelled. To mitigate this threat, we performed this labelling in an iterative manner, went over each review multiple times to ensure correct labelling of the reviews.

Internal Validity. There may be reviews that are spam reviews. To prevent spam in app reviews, Google requires users to log in using their Google Id before reviewing. We believe that the impact of spam reviews is likely minimal, nevertheless, future studies should evaluate the impact of spam reviews on recent research that mines reviews.

Threats to External Validity. The selection of the top apps could bias our results. Given the large number of unsuccessful and spam apps in the store, we feel that our study of top apps is warranted instead of blindly studying all apps. We only studied apps that are free. Paid apps may exhibit different reviewing and developer response patterns in comparison to free apps. However, many paid apps have free versions available to download and there are considerably more free apps than paid apps in the Google Play Store. Moreover, many free apps have in-app purchase features. Such apps need to consider the value of reviews for financial reasons. Future studies should carefully tag apps based on whether they are truly free or not. Unfortunately such information is not easily accessible in an automated fashion.

VII. Conclusion

Most top apps do not respond to reviews, however responding can lead to a positive change in rating. Addressing specific issues and notifying the users that requested features are available are most likely to lead to a change in the review rating.

References

[1] C. Iacob and R. Harrison, "Retrieving and analyzing mobile apps feature requests from online reviews," in *Proceedings of the Tenth International Workshop on Mining Software Repositories*. IEEE Press, 2013, pp. 41–44.

[2] 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*, ser. ICSE '13. Piscataway, NJ, USA: IEEE Press, 2013, pp. 582–591.

[3] B. Fu, J. Lin, L. Li, C. Faloutsos, J. Hong, and N. Sadeh, "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*, ser. KDD '13. New York, NY, USA: ACM, 2013, pp. 1276–1284.

[4] N. Chen, J. Lin, S. C. H. Hoi, X. Xiao, and B. Zhang, "Ar-miner: Mining informative reviews for developers from mobile app marketplace," in *Proceedings of the 36th International Conference on Software Engineering*, ser. ICSE 2014. New York, NY, USA: ACM, 2014, pp. 767–778.

[5] D. Pagano and W. Maalej, "User feedback in the appstore: An empirical study," in *Proceedings of the 21st. IEEE International Requirements Engineering Conference*. IEEE, 2013.

[6] M. Harman, Y. Jia, and Y. Z. Test, "App store mining and analysis: Msr for app stores," in *Proceedings of the 9th Working Conference on Mining Software Repositories (MSR '12)*, Zurich, Switzerland, 2-3 June 2012.

[7] H.-W. Kim, H. L. Lee, and J. E. Son, "An exploratory study on the determinants of smartphone app purchase," in *The 11th International DSI and the 16th APDSI Joint Meeting*, Taipei, Taiwan, July 2011.

[8] S. M. Mudambi and D. Schuff, "What makes a helpful online review? a study of customer reviews on amazon.com," *MIS Quarterly*, vol. 34, no. 1, pp. 185–200, 2010.

[9] E. Guzman and W. Maalej, "How do users like this feature? a fine grained sentiment analysis of app reviews," in *Proceedings of the 2014 IEEE 22nd InternationalRequirements Engineering Conference (RE)*, Aug 2014, pp. 153–162.

[10] H. N. Walid Maalej, in *Proceedings of the 23rd IEEE International Requirements Engineering Conference*, 2015, p. to appear. [Online]. Available: https://mobis.informatik.uni-hamburg.de/wp-content/uploads/2015/06/review_classification_preprint.pdf

[11] R. Chandy and H. Gu, "Identifying spam in the ios app store," in *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*. ACM, 2012, pp. 56–59.

[12] H. Khalid, E. Shihab, M. Nagappan, and A. Hassan, "What do mobile app users complain about? a study on free ios apps," *IEEE Software*, vol. PP, no. 99, pp. 1–1, 2014.

[13] V. mobile, "Developer Economics Q1 2014: State of the Developer Nation," Tech. Rep., 05 2014.

[14] D. Pagano and B. Bruegge, "User involvement in software evolution practice: a case study," in *Proceedings of the 2013 International Conference on Software Engineering*. IEEE Press, 2013, pp. 953–962.

[15] W. Martin, M. Harman, Y. Jia, F. Sarro, and Y. Zhang, "The app sampling problem for app store mining," in

Digital Object Indentifier 10.1109/MS.2015.149 0740-7459/$26.00 2015 IEEE

TABLE IV
COUPLING OF REVIEW TYPES WITH RESPONSE TYPES.

	Instructions	Request contact	Thanks	New Release fix	New Release Feature	Current Release Fix	Other	Rating and Review Mismatch	Question	Total Responses
Praise	33	24	45	8	6	4	1	6	3	134
Request	35	31	3	7	10	4	2	0	1	130
Bug report	39	60	3	10	4	12	0	0	6	124
Dispraise	52	39	2	16	2	10	1	0	2	93
Update issue	11	15	1	3	1	7	0	0	0	38
Other	3	4	1	0	0	0	2	0	0	10
Other apps	3	5	1	0	0	0	0	0	1	10

Proceedings of the 12th Working Conference on Mining Software Repositories (MSR), Florence, Italy, 2015.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[17] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 142–150.

[18] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2004, p. 271.

[19] H. Khalid, E. Shihab, M. Nagappan, and A. E. Hassan, "What do mobile app users complain about? a study on free ios apps," in *IEEE Software*. IEEE Press, 2014.

[20] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the First Workshop on Social Media Analytics*, ser. SOMA '10. New York, NY, USA: ACM, 2010, pp. 80–88.

Digital Object Indentifier 10.1109/MS.2015.149 0740-7459/$26.00 2015 IEEE