Weiyi Shang
EV 3.129, 1515 Ste. Catherine Street West
Department of Computer Science and Software Engineering, Concordia University
Montreal, QC, Canada H3G 1M8
1 514-848-2424 ext. 7801

April 1, 2017

Dear Editor and Reviewer:

Thank you for your insightful feedback and comments, both positive and constructive, and for allowing us the opportunity to improve our manuscript. We have taken each of your comments into consideration and made the appropriate changes and extensions to our manuscript. We made several changes throughout the paper based on the review. We feel that the paper is now much stronger.

Below, we include a description of the changes that we made to our manuscript with respect to each of the reviewer's comments. We denote the reviewer's comments in italic typeface, and our responses follow below each reviewer comment.

Reviewer 3 Comments

Comment R3.1. *One problem is that there are still no real implications mentioned other than that things have to be studied more in detail since there are no clear findings across different app stores. Fair enough, but that seems a bit vague. I was hoping for some more flesh here and some more ideas on the authors' side. Similarly regarding implications for practitioners. Currently, I don't see many - in particular new - implications for practitioners. That developers should allocate more resources for working on reviews after releases is not new, as we already know that feedback volume is higher after releases since Pagano & Maalej and Hoon et al.*

Response. Our main take-home of the paper is the demonstration of the discrepancy between different app stores. The implication of such a finding is that future research should take into consideration such discrepancy. Tooling support for app developers should also be optimized for different app stores. In addition, we find that the overhead of dealing with app reviews is not overwhelming for majority of the apps, as opposed to the findings from prior studies that apps have high amounts of reviews in average, where such high amounts are most likely due to outliers.

Comment R3.2.

Another problem I see is that the manuscript is inaccurate when it comes to the description of the results. The most prominent example for me are the apps that don't receive a large amount of reviews per day: "Our key finding is that while some apps might receive a large amount of reviews, most (over 99%) apps receive very few reviews." First, please be specific with the numbers. At one point the authors report 88% at another 99%, so which is the correct one? Second, what does "very few" mean? This is not a scientific term, and in fact 20 reviews might as well not be "very few" if you are a single developer. Please do not use such terms. The authors should report the strict observed numbers and in an interpretation section they might say they think this is "very few" or "few". This applies to many places all over the manuscript. Please only report data and separate this from any interpretation.

Response. Thanks for the suggestion. We revised our text to "only 0.19% of the apps receive more than 500 reviews per day".

Comment R3.3. *In that light, also the advise on automated approaches for user feedback analysis does not seem to be correct: "Most top apps might not benefit much from automated approaches for analysis of user reviews." I don't buy this. Where is the limit? 20 reviews? 10? I strongly believe this depends on the situation at hand. What if you are a single developer? What if you are developing 20 apps? Then you might get 400 messages a day. What do you do after a new release? You get a big spike then as we know. What if you do not have time to read and aggregate all the messages. I really believe that this statement is much too harsh. Why don't the authors take this opportunity to elaborate a bit on the different situations that might occur?*

Response. Thanks for the suggestion. We would like to clarify that we do not claim that reading reviews (potentially hundreds of them) is not an overhead. With such an overhead, automated tooling support is needed. However, there exists advanced tooling support that performs automated analysis on review contents that are based on sophisticated techniques, such as topic modeling. Such techniques may not be all optimized if the amount of text data is low. In particular, we also find that on median each review only consists of 36 characters. Automated techniques that analyze reviews need to take into consideration both the amount of review sand length of reviews.

We modify our text into "Most top apps might not benefit much from automated approaches that leverage sophisticated techniques (like topic modeling) given the a small amount of received user reviews and their limited length."

Comment R3.4. *When the authors say "The relationship between received reviews and the category of an app should be explored in further studies.", it should probably be underlined that this is needed especially because there are different results for different app stores.*

Response. Thanks for the suggestion. We modify our text accordingly.

Comment R3.5. *"However, other than crash reporting tools, many of the analytics tools available today are mostly sales-oriented instead of being software quality oriented." How do the authors define "software quality"?*

Response. We modify our text to elaborate software quality by giving examples, such as bugs, performance and reliability.

Again, we thank all of you for your valuable feedback, which has made this a stronger manuscript. We look forward to hearing your feedback on the updated manuscript.

Sincerely,
Stuart McIlroy, Weiyi Shang, Nasir Ali, Ahmed E. Hassan