Software quality through the eyes of the end-user and static analysis tools

Kam Srisopha

USC ARR 2018
Agenda

- Motivation
- Background
- Experimental Setup
- Results
- Future work
Motivation and Background

- **Software evolves over time**, whether to adapt to a new hardware environment, to address the evolving needs of its users, or to repair defects.

- The abundance of software products on the market calls for software companies to **thrive to make the above changes and release new version of software faster**.

- In an effort to improve software quality, many source code analysis tools (e.g., PMD, FindBugs, and SonarQube) have been developed. These tools **provide numerous software metrics** that can be used to give an insight into the quality of the code such as code complexity, code smells, comment density, etc.

- While it is important for software to be developed with high code quality and within budget and schedule, **successfully achieving them is still not enough to ensure a successful software product**.
The missing piece of the puzzle

"The project was delivered on time and within budget and scope and thus was a resounding success"

What is wrong with the above claim?
John Brandon (JB) was a vice president of sales in the US and Pacific for Apple.

**JB’s Rules for Success**

- Let go of the old, make the most of the future
- Always tell the truth, we want to hear the bad news sooner than later
- The highest level of integrity is expected, when in doubt, ask
- Learn to be a good businessperson, not just a good salesperson
- Everyone sweeps the floor
- Be professional in your style, speech and follow-up
- **Listen to the customer, they almost always get it**
- Create *win/win* relationships with our partner
- Look out for each other, sharing information is a good thing
- Don’t take yourself too seriously
- Have fun, otherwise it’s not worth it
Passive Crowdsourcing through mining user feedback on app distribution platforms
User feedback plays an important role for the success of a product

- Pagano et al. [3] found that user feedback (reviews + user ratings) correlate with the number of downloads.

How do developers typically handle bad reviews?

- Implement changes/fix bugs according to what users report.
- Directly give a response to the user.

Ahmad Yusuf  July 30, 2016
★★★★★
can't sign in I've used this and it's gooodd it's like Twitter on steroid but today I can't sign in. Hope it'll be fixed. Thank you

April 4, 2015
fixed in recent update.

Matthew Bess  February 23, 2018
★★★★★
Very nearly perfect. A small bit of lag in some animations sometimes, but nothing too major. I love this app. Sadly, it looks like the last update was in 2016 (how can that be right??). I really hope this hasn't been abandoned. I'm using it as my default SMS app, but it'll be hard to do that for long if there are no more updates.

Moez Bhatti  February 23, 2018

The app has certainly not been abandoned! I've been working the last few months on completely re-writing it from scratch. The performance is much much better, and now it's absolutely beautiful. I'll be launching it in the next 2-3 weeks!
What kind of useful information that we can get from user reviews (IoT products on Amazon)

**Hardware complaint**

⭐⭐⭐⭐⭐ this was one of the worst electronic purchase I have ever mad there is almost ...
By     on July 10, 2017
Color: Black | Configuration: Echo Dot | **Verified Purchase**

I bought this a gift for someone.. this was one of the worst electronic purchase I have ever mad there is almost no volume on speaker if you are not sitting right next it you will not be able to hear it and it's sad that amazon put their name on such a product.

**Software feature request**

⭐⭐⭐⭐⭐ i think the fitbit blaze is great! love it so much more than my apple ... 
By     on March 18, 2017
Color: Black, Silver | Size: Small (5.5 - 6.7 in) | **Verified Purchase**

i think the fitbit blaze is great! love it so much more than my apple watch, as i didn't use any of the apps for it. hopefully they start integrating some kind of apps though, like My Fitness Pal or something for reminders to log food! FitBit UI has always been my favorite UI out of all of the other fitness trackers. Love that it can be used across devices (android/iOS) instead of "galaxy gear" or an "apple watch" which are only compatible with their own OS.

**Software satisfaction**
What kind of information that we can see in user reviews (Mobile applications on PlayStore)*

Software Problem Discovery

Gao Zoua  March 8, 2018
★★★★★
My news feed won't update. I'm at the top of my news feed and usually when I pull it down, it'll update. Since the last update, it'll stay stuck on someone's post from a few days ago as the first thing that shows up. I'd have to disable and re-enable the app for it to work again but it only lasts a few hours. Please fix!!

T. Coolong  November 3, 2017
★★★★★
Changing my review due to the fact that this app is by far the biggest battery drain app on my phone. I have another twitter app that doesn't do this. It also misses tweets, as it only seems to load the last 20 each time.

- Information Seeking
- Feature Request
- Information Giving
- Other

*Categories are based off of “How Can I Improve My App? Classifying User Reviews for Software Maintenance and Evolution” - Panichella et al. ICSM 2015
Objective:
Study the complete user feedback (reviews + ratings) after each software release on Playstore to investigate whether it has any relationship with the application’s internal characteristics.
Research Questions

RQ1: *(user reviews vs static analysis tool)*

To what extent do violations reported by static code analysis tools correlate with different types of review sentences?

RQ2: *(user ratings vs static analysis tool)*

To what extent do violations reported by static code analysis tools correlate with the average user ratings?
Studies’ subjects

<table>
<thead>
<tr>
<th>App Name</th>
<th>Developer</th>
<th>Genre</th>
<th>Rating</th>
<th>Reviews</th>
<th>Releases</th>
</tr>
</thead>
<tbody>
<tr>
<td>QKSMS - Open Source Messenger</td>
<td>Moez Bhatti</td>
<td>Communication</td>
<td>★★★★★</td>
<td>9,886</td>
<td>507</td>
</tr>
<tr>
<td>Twidere for Twitter</td>
<td>Mariotaku Dev</td>
<td>Social</td>
<td>★★★★★</td>
<td>12,528</td>
<td>202</td>
</tr>
<tr>
<td>Omni Notes</td>
<td>Federico Iosue</td>
<td>Productivity</td>
<td>★★★★★</td>
<td>3,197</td>
<td>295</td>
</tr>
</tbody>
</table>

Total: 1004    46
(2536 sentences)
Manual Analysis of user reviews
Three dimensions of a review sentence

- **Sentence-level granularity** (1004 reviews => 2536 sentences)
- **Two human annotators** (one CS PhD student and one CS Master’s student)
- Use **Cohen's kappa coefficient** ($\kappa$) to access the inter-rater reliability. Resolve disagreements by discussion and voting.
- Take three weeks
- Instruct not to annotate more than **200** sentences a day as to avoid errors due to fatigue.
Intention (based on how relevant a sentence is to software developers performing maintenance tasks)

**Problem Report**

Changing my review due to the fact that this app is by far the biggest battery drain app on my phone. I have another twitter app that doesn't do this. It also misses tweets, as it only seems to load the last 20 each time.

**Inquiry**

How can I get notified when a particular person that I follow sends out a tweet?

**Other**

Simple interface and easy to use. Great for taking quick notes or typing papers. Highly recommend for download.

**Improvement Request**

Seriously my fav twitter app. Great customisation, lightweight and functional. Just wish we could see notifications like who faved and retweeted our tweets, followed, etc. Would also love an update for the new character limit.
Software Quality (ISO/IEC 25010)

ISO/IEC 25010 defines the quality of a software system as the extent to which the system satisfies the stated and implied needs of its various stakeholders [4].
Software Quality in the reviews (ISO/IEC 25010)

Dissatisfaction with Resource Utilization (Performance Efficiency)

Dissatisfaction with Functional correctness (Functional Suitability)

Satisfaction with User interface aesthetics and Operability (Usability)

Gao Zoua March 8, 2018
★★★★★
My news feed won't update. I'm at the top of my news feed and usually when I pull it down, it'll update. Since the last update, it'll stay stuck on someone's post from a few days ago as the first thing that shows up. I'd have to disable and re-enable the app for it to work again but it only lasts a few hours. Please fix!!

T. Coolong November 3, 2017
★★★★★
Changing my review due to the fact that this app is by far the biggest battery drain app on my phone. I have another twitter app that doesn't do this. It also misses tweets, as it only seems to load the last 20 each time.

Sheldon Norwood March 5, 2018
★★★★★
Simple interface and easy to use. Great for taking quick notes or typing papers. Highly recommend for download.
Software Quality in the reviews (ISO/IEC 25010)

Vivienne Lim October 19, 2017

Used to be my favourite app but EXTREMELY laggy for the past couple of months. Has happened before but resolved itself, now the issue is back. Can't scroll more than a few seconds before the app stalling and becoming unresponsive, sometimes even crashing. Takes up to A MINUTE to unfreeze or to even load the initial tweets.

Dissatisfaction with time behaviour
(Performance efficiency)

Dissatisfaction with (Reliability)

★★★★★ i think the fitbit blaze is great! love it so much more than my apple ...
By on March 18, 2017
Color: Black, Silver | Size: Small (5.5 - 6.7 in) | Verified Purchase

i think the fitbit blaze is great! love it so much more than my apple watch, as i didn't use any of the apps for it. hopefully they start integrating some kind of apps though, like My Fitness Pal or something for reminders to log food! FitBit UI has always been my favorite UI out of all of the other fitness trackers. Love that it can be used across devices (android/iOS) instead of "galaxy gear" or an "apple watch" which are only compatible with their own OS.

Satisfaction with Usability

Satisfaction with Compatibility
**Sentiment** *(tone of a sentence)*

*mel2k*  January 5, 2018

⭐⭐⭐⭐⭐

Very well-done and intuitive memo app. The ability to hyperlink is great. My only quibble is lack of ability to save the note to an external file format.

*Neutral*

Positive

Negative

*Neutral*

M Arif  January 31, 2018

⭐⭐⭐⭐⭐

Please add finger print sensor to unlock password protected note.
Why manual analysis?
Challenges in investigating these reviews?

1). Investigating the complete reviews after each software release for popular application is laborious and extremely difficult.
   - Facebook can receive over 4000 reviews on average in one day [3].


3). All purpose sentiment analysis tool does not work well for software engineering domain.
   - “Unpredictable” → positive for movies, negative for SE.

4). It is like looking for a black cat in a coal cellar
   ● Our previous work (CIBSE-WER 2018) ~ 15% of all review sentences on IoT products on Amazon contain information that could help accelerate software evolution and maintenance.
   ● Pagano et al. [3] (RE 2013) ~33% of all review sentences of mobile applications on Google Play store contain information that could help accelerate software evolution and maintenance.
   ● Guzman et al. [2] (RE 2016) ~ 19% of all tweets contain information relevant to technical stakeholders.
Not all aspects of the application’s software quality can be perceived by the users. For example, Security is least observable by the users.
Twidere

Average of user ratings for each release
Twidere

Average of user ratings for each release
Results for manual classification (Twidere):
### Static Analysis tools and metrics

| 1 | ![SonarQube Logo](sonarqube.png) | **Code Smell**  
(Maintainability related issue in the code) |
|---|----------------------------------|-------------------------------------------------|
| 2 | ![PMD Logo](pmd.png) | **Java Rulesets:**  
Empty code, Naming, Braces, Import statements, Coupling, Unused code, Unnecessary, Design, Optimization, String and StringBuffer |
| 3 | ![FindBugs Logo](findbugs.png) | **Standard bug pattern:**  
Dodgy Code, Bad Practice, Malicious code, Performance, Correctness, Security, Multithreaded correctness, Internalization |
Results for static analysis tools (Twidere app)

Twidere
Results for static analysis tools (Twidere app)

Twidere

Software quality metrics/size analyzed

Releases

- Code Smells (SonarQube)
- PMD
- FindBugs
**RQ1:** We pair these 11 sentence types with each quality metric (33 pairs)

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Sentence Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intention</td>
<td>Problem report</td>
</tr>
<tr>
<td>Sentiment</td>
<td>Negative, Positive</td>
</tr>
<tr>
<td>Software Quality</td>
<td>Dissatisfaction with {functional suitability, performance efficiency, compatibility, usability, reliability, security, maintainability, portability}</td>
</tr>
</tbody>
</table>

**RQ2:** We pair each quality metric with the average user ratings for each release.
Results

RQ1:

- Of all 33 pairs, only 7 pairs show a significantly strong correlation with one another (p < 0.05)
- 2 strong positive correlations found for SonarQube (Dis-reliability and Dis-performance)
- 5 strong positive correlations found for PMD (Negative sentiment, Problem report, Dis-functional, Dis-reliability, and Dis-performance)
- No correlation found for FindBugs
- **Positive sentiment** shows a negative correlation with PMD and SonarQube, (less code quality, more positive reviews) but not statistically significance.

RQ2:

- All test pairs (3 pairs) exhibit a negative correlation (less violation, better user ratings)
- Only the pair of PMD and user ratings show significantly strong negative correlation (p<0.05)
Conclusion

● To some extent, user feedback correlates with code quality

● Not all aspects of SQs can be perceived by end-users. Likewise, not all can be observed by static analysis tools.

● User reviews contain a wealth of information that can accelerate software evolution and maintenance.
  ○ Information can be used to help us understand how software evolves.
  ○ Information can help us understand the product’s software quality.

● More work in this area is needed to help solve several challenges in investigating the reviews.

● Code quality only represents one side of the coin. The voice of the end-users matters!
This work can be extended to several directions.

- Investigate the feasibility of incorporating the state-of-the-art automated review analysis tool to help with data collection.

- Conduct a large scale study involve more apps and at least 15 releases per app to test if our preliminary analysis holds.

- Find a traceability link between bug reported by users and bug records in the application’s issue-tracking tools.

- Incorporate static analysis tools that are specifically tailored for Android application.

- **In progress:** Using NLP and ML to create a tool that can prioritize bugs reported by users, recommend the appropriate developers to fix it, and locate the defect in the source code (bug localization).
Thank you for your attention!

Questions?
Reference


