A Preliminary Analysis of Mobile App User Reviews

Rajesh Vasa, Leonard Hoon, Kon Mouzakis and Akihiro Noguchi
Swinburne University of Technology
Faculty of Information and Communication Technologies
{rvasa, lhoon, kmouzakis, akihironoguchi}@swin.edu.au

ABSTRACT
The advent of online software distribution channels like Apple Inc.’s App Store and Google Inc.’s Google Play has offered developers a single, low cost, and powerful distribution mechanism. These online stores help users discover apps as well as leave a review. Ratings and reviews add value to both the developer and potential new users by providing a crowd-sourced indicator of app quality. Hence, for developers it is important to get positive reviews and high ratings to ensure that an app has a viable future. But, what exactly do users say on these app stores? And more importantly, what is the experience that compels a user to leave either a positive or a negative rating? Our analysis of 8.7 million reviews from 17,330 apps shows that users tend to leave short, yet informative reviews, and the rating as well as the category influences the length of a review. In this preliminary study, we found that users will leave longer messages when they rate an app poorly, and the depth of feedback in certain categories is significantly higher than for other.

Author Keywords
Mobile apps, rating systems, reviews, text mining, user engagement, user expectations, user issues, word-of-mouth

ACM Classification Keywords
H.2.8 [Database Applications]: Data Mining; H.3.1 [Content Analysis and Indexing]: Linguistic Processing

INTRODUCTION
In 2008, Apple introduced a catalyst for change by opening the App Store as a distribution channel for third-party apps. Consumers were granted control over the apps they could install on their mobile devices, and could publicly rate and review apps that they had installed. This feedback loop empowers users to evangelise and promote apps that they like, but also allows them to warn other users about the limitations of an app.

User reviews are an important part of the user experience, especially since subject satisfaction plays a pivotal role in acceptance (Koyani et al., 2004). Within the context of mobile apps, this model of crowd-sourced opinion is relatively new, offering users immediate access to information that can influence their purchasing decision. Chevalier & Mayzlin (2003), have found that online Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

reviews do impact sales do impact sales in other domains. Within the hotel industry, it was also found that proper analysis of reviews could yield insightful business opportunities (Vermeulen & Seegers, 2009; Ye, Law, & Gu, 2009). However, low ratings and weak reviews can influence revenues and limit the rate of growth. Given the highly competitive landscape of mobile apps, it is helpful to better understand what exactly users say in their reviews and if there are any particular aspects that would influence the depth of feedback that a user will leave.

Our longer-term research goal is to contribute to the field of Computer User Satisfaction (Delone & McLean, 2002) by gaining a deeper understanding of what users say in reviews, and how their opinion evolves over time. Our work is motivated by the hypothesis that techniques that can mine user opinions and their evolution can allow developers to prioritise and focus their efforts toward meeting user expectations.

As a preliminary step in our research journey, we start with a basic question: Is there sufficient information within users reviews to allow us to gain deep insights? And, what is the nature of these reviews? To answer these questions, in this work, we statistically analysed 8.7 million reviews of approximately 17 thousand iOS apps. These apps and reviews were mined across all 22 categories (both paid and free). To focus the study, we ask the following specific research questions: 1) What is the size of a user review and is there an average length? 2) Does the star rating affect the length of a review?, and 3) Does the category of an app influence the length of a review?

The rest of this paper is organised as follows: in our next section, we explore works related to online user reviews. Following that, we describe our dataset used and our analysis method. Our findings and implications are presented next. Finally, we discuss the limitations of this study and conclude our work with a forecast of future work.

USER REVIEWS
User reviews can behave like online Word-Of-Mouth (WOM). WOM is recognised as influential in information transmission, particularly with experience goods (Godes & Mayzlin, 2004; Granovetter, 1973). Consumer or user generated online reviews implicitly communicate user-perceived quality, from which “perceived ease of use” and “perceived usefulness” (Davis, 1989) may be inferred. This creates a feedback loop granting focused opportunity for refinement in subsequent iterations.

Broadly speaking, user reviews are a form of co-value creation (Tan & Vasa 2011). In that, the users add value.
by a) providing feedback to both the developer and the community; b) positive feedback improves an app’s chance of discovery by other users (which is good for developers), gives other users confidence (especially if there are a large number of good reviews), and can impact on the emotional well-being of the developer (providing an ego boost); and c) negative feedback informs the developer about areas that they need to focus on. It also informs the community by asking them to avoid a particular item.

In effect, the information value of reviews depends on perspective. Developers enjoy value from positive reviews, while negative reviews add more value to potential future user community. Although negative reviews can be disastrous for developers, it informs the prioritisation of their efforts. The resulting decrease in downloads from bad reviews does however, limit the app's exposure, reducing the potential negativity that future reviews. While the concept of reviews for mobile apps is relatively new, other industries have incorporated user feedback into their business strategy for a number of years. Work in these domains has found that user reviews can significantly influence sales (Chevalier & Mayzlin, 2003), and that users tend to only sample reviews (Duan, Gu, & Whinston, 2008).

Platzer (2011) also states that users will leave reviews of varying length (often domain specific) and can possess abbreviations, colloquial expressions, and non-standard spelling. It is also acknowledged that reviews (Gebauer, Tang, & Baimai, 2007; Duan et al., 2008) often addressed myriad aspects within the context of the domain and the object under review. Though there is some understanding of the nature of reviews in other domains, there is mostly anecdotal evidence regarding the nature of reviews for mobile apps and our research aims to address this gap.

The structure of reviews is similar across domains; typically users can leave a numerical rating (often as stars) and a brief text or audio/video comment. Additionally, mechanisms may be employed to ensure the quality of user reviews. Kostakos (2009) showed that these quality mechanisms can be useful for displaying reviews that other users have rated as helpful, but they also have the potential to polarise the reviews (either too positive or too negative) if the users are forced to leave a quality review.

**DATASET & ANALYSIS**

For this study, we have developed a script that can parse and download all the reviews from the top 400 free and paid apps in each of the 22 categories available on the Apple App Store. The App Store currently distributes over 500,000 apps, however we had to restrict our focus to the top 400 since Apple publishes details for the top apps in a form that permits direct automation without our script being flagged as malicious. The implication of this choice is that it constrains us within the context of successful or popular apps only, based on their ranking at the time of collection. The reviews collected span over the time of 13th of July 2008 to 20th of July 2012. Our data set contains 8,701,198 reviews made by 5,530,025 users across 17,330 apps. The large size of our data set ensures that it provides us with a representative sample of the user reviews within the context of successful and popular apps.

![Figure 1. The cardinality of a Review, left by a User, per Release of an App in a Category.](image)

Each review on the App Store is comprised of a star rating between 1 and 5, a review title and a review body. Our scripts download all information available for a review and preserve the underlying data relationships as illustrated in Figure 1. To answer our research questions, we append the body of each review to its title to measure the character count of a review and analyse this size within the context of the category of the app and the star rating. We use the size of the review as a rudimentary measure of engagement, in that it is an approximation of the affect the app has on the user. We initially analysed the data using the data via the use of summary statistics, box plots, and cumulative distribution charts (to gain a visual perspective). We later confirmed our hypothesis that the rating as well as the category will have an influence on the size of the review using a one-way ANOVA test.

**FINDINGS**

We summarise our observations from analysing the 8.7 million reviews. We start by investigating the typical length of a review. We then move on to consider if the rating influences the size of a review. Finally, we analyse the reviews across categories to determine if user behaviour changes across different categories.

**Typical Size of a Review**

What is the length of a review that users leave? Is there an average and typical value?

Using a manual review analysis on a small data set Gebauer et al. (2007) determined that in general that reviews for mobile apps tend to be short. However, given the size of their data set and the manual approach used, they were not able to provide a strong numerical boundary of what to expect in terms of review length. An awareness of the expected statistical properties is useful as it may determine atypical reviews, which can assist in flagging potential spam. Our data set shows that user review length is highly skewed (see Figure 2) with an average of 117 characters, and median at 69 characters (SD: 156, and Skew: 7.24). The data confirms the prior finding that users tend to leave short messages -- in fact, nearly 75% of the reviews are short enough to fit within a tweet (140 characters). Interestingly, a small set of the reviews (2.5%) are relatively long with over 100 words or
500 characters suggesting that some users do take time to leave long reviews with potentially useful content. Nevertheless, what influences the length of a review? We now consider this specific aspect in the following sections.

**Influence of Rating on Review Length**

Do users leave longer messages when they rate an app poorly? A manual, cursory scan of reviews on the App Store shows that users accompany higher ratings with a short review (often a single word like *awesome*), however such reviews do not indicate if the rating and review were due to the app meeting or exceeding user expectations. Similarly, when a user is critical of an app a manual observation suggests that they seem to leave longer reviews expressing their discontent.

Our observations are presented visually in Figure 3 as a box-plot – users tend to leave longer messages when they rate an app poorly (1 or 2 stars). Interestingly, 2 star ratings tend to receive longer reviews than 1 star reviews. Our data shows that users take time to express discontent by writing a longer review, in contrast to leaving succinct reviews when content. The median size of a 5 star review is 54 characters, while it is nearly three times longer at 144 for a 2 star review. Furthermore, an ANOVA test supports the observation that review lengths differ significantly ($p < 0.01$) across the 5 rating levels.

**Reviews across Categories**

Are users influenced by the category of the app when they leave a review? An initial manual inspection did not indicate that users would behave differently across categories (at least with respect to the review length). Given our data set size, we expected the law of large numbers to come into play and hence expected review lengths to be similar across categories. Interestingly, the ANOVA test clearly showed that review lengths differ significantly across the 22 categories ($p < 0.01$). This observation can be seen in Figure 4. We depict only 10 randomly selected categories due to space constraints, but the rest of the categories have a similar spread. We also observe that users tend to leave significantly shorter reviews for *games* than for other categories.

An unexpected finding was that the median of review length in the *health & fitness* category was 100 characters, twice as long as the median for games. This observation is summarised visually in Figure 5 and shows a significant difference across the board between the

---

**Figure 2. Histogram of Character Count of Reviews**

(cumulative values are shown on the right hand y-axis)

**Figure 3. Boxplot depicting Character Counts in relation to Ratings left by users. Outliers (top 2.5% of data) have been suppressed to improve readability.**

**Figure 4. Boxplot depicting Character Counts in relation to Categories. The following categories are depicted. Entertainment(EN), Games(GA), Health & Fitness(HE), Lifestyle(LF), Music(MU), Navigation(NV), Productivity(PR), Social Networking(SC), Sports(SP), and Utilities(UT).**

**Figure 5. Cumulative percentage of Character Count for the Health & Fitness and Games Categories.**

An unexpected finding was that the median of review length in the *health & fitness* category was 100 characters, twice as long as the median for games. This observation is summarised visually in Figure 5 and shows a significant difference across the board between the
review length of the health & fitness and the games categories. A detailed analysis that compared the categories with each other both visually and also by using a Bonferroni comparison (after the ANOVA) showed that in general, most categories tend to have different review length spreads. A few categories had similar spreads (media & video and finance), but in general the review length seems to be influenced by the category of the app.

LIMITATIONS
The nature of reviews for apps not in the top ranks may be different from our observations. This is a limitation that we are currently working to address by improving our scrapers to bypass the detectors that Apple has that are currently blocking our scripts when we scan too many app reviews within a short time period. We circumvent the introduction of human error by automating our data collection, processing, and visualisation where possible. We also intend to obtain data from Google Play Store that contains Android app reviews.

Our study focuses on the reviews left by users and does not currently extend to an analysis of each app. Some apps encourage users to leave feedback with modal dialogs, which may potentially skew the amount of feedback an app receives. However, while our dataset maintains release horizons based on review cardinality, we are presently unable to retrieve early versions of apps from the App Store to sufficiently study apps for such functionality.

CONCLUSION
Although online user reviews in different domains have been investigated for nearly a decade, there is only a minimal understanding of user reviews created for mobile apps. We have analysed 8.7 million reviews from 17,330 apps in order to better understand the nature of these reviews. Our study shows that, in general, mobile apps receive short reviews. Furthermore, both the rating and the category of an app influence the length of a review. Although we can identify the potential causes for short reviews and why poor ratings tend to elicit longer reviews we currently do not have a strong hypothesis that explains why the category of an app has an influence on the review length.

We intend to explore this gap in future work where we also intend to analyse the actual content of the review using text analysis techniques. Given our finding that most reviews tend to be short and about the size of a tweet, we are hopeful that the text analysis techniques developed for analysing Twitter and Facebook streams can be applied and hence we intend to explore that area in the near future.

REFERENCES

Davis, F.D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 13, 3 (1989), 319-340.


Granovetter, M.S. The strength of weak ties. American Journal of Sociology 78, 6 (1973), 1360-1380.


