Do Android Users Write About Electric Sheep?

Examining Consumer Reviews in Google Play

Elizabeth Ha
School of Information
University of California, Berkeley
lizzy@ischool.berkeley.edu

David Wagner
Computer Science Department
University of California, Berkeley
daw@cs.berkeley.edu

***Abstract*—Consumer reviews and star ratings are integral to application markets. The content of reviews help consumers determine whether an application is "good" or not. Since consumers rely heavily on reviews when selecting applications, we wanted to know what was being written about in reviews. In particular, we wanted to know if users were discussing privacy and security risks of an application, and if not, what were they writing about instead? In our work, we manually analyzed Android users' reviews to see what they write about when reviewing Google Play applications. Overall, only 1% of our reviews mentioned application permissions. We also found that a small subset of reviews relating to preinstalled applications and applications that requested a user's rating had underlying privacy and security implications. The majority of reviews focused on the quality of applications: people often described an application using an adjective (e.g., "great app" or "horrible"), wrote about its feature/functionality, specifically said if the application worked or not, and/or put their phone or tablet model in the review. We also found that sentiment did influence reviewers' ratings of the applications. In general, the overall star rating of our sample was overwhelmingly positive, suggesting that Google Play is no different from other e-commerce sites.**

***Index Terms*—Applications, Google Play, Reviews, Social Implications of Technology**

I. Introduction

Android users—like other smartphone users—like to download applications on their phones. Since the launch of the Android Market (hereinafter Google Play) in 2008, the average number of applications installed on an Android user's phone has significantly increased [12]: in 2009 a user had an average of 22 applications on his phone [13], whereas a user had an average of 35 applications installed on his phone in 2011 [14]. While applications can provide users with entertainment or help make their lives easier, there are security and privacy risks when downloading and installing applications [6].

Like other application markets, Google Play provides different types of information to help users select applications. This information includes the application description, screenshots, user reviews and star ratings, and permissions information. Past research has shown that Google Play users rely heavily on reviews to help them determine whether an application is "good" or not since reviews can contain warnings about an application's negative qualities [8]. Felt et al. [8] found that users rely more on reviews than Android's permissions screen: 24% of participants in Felt et al.'s laboratory study relied on reviews to inform them of an application's permissions, whereas only 17% looked at the actual permission screen during application installation [8]. Since Android users rely on reviews to help them make decisions when selecting an application, we wanted to know what was being written about in reviews. In particular, we wanted to know if users were writing about privacy and security risks (i.e., permissions), and if not, what were they writing about instead? In their work, Chia, Yamamoto, and Asokan [5] found that the average rating of an Android application was not negatively correlated to the number of permissions requested, suggesting that reviews were not about privacy or security issues, but about how the application functioned and worked. In our paper, we provide data to support this assumption.

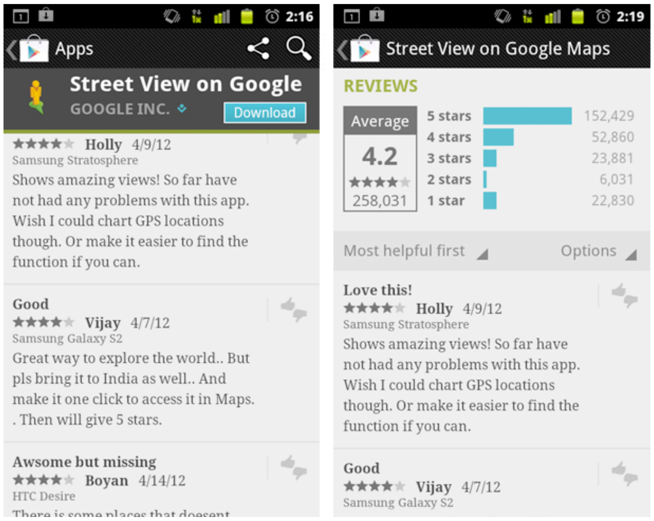
We sampled 556 reviews from 59 different applications from Google Play. From there, we developed 18 topic categories and 37 sub-topics categories, which represent positive and negative sentiments and other information people wrote about in their reviews. We used these sub-topics to manually code and classify our reviews. We also examined the star ratings in our sample and correlated ratings with our sub-topics.

Overall, the majority of reviews were informative, but only 1% of reviews mentioned permissions, with the majority of permissions-related reviews explicitly questioning them. Instead, we found people often described an application using an adjective, wrote about its feature/functionality, specifically said whether the application worked or not, and/or put their phone or tablet model in the review. While the overall star rating in our sample was overwhelmingly positive, we did find that reviewers' judgment of an application's features/functionality and whether an application worked or not appeared to have significant influence on the reviewer's star rating. We also found that negative remarks about an application's aesthetics or money and cost were associated with below-average star ratings.

II. Background & Related Works

A. Google Play

Google Play has 30 different categories. Each application is placed in exactly one Google Play category. An application's reviews can be viewed on its detail page under "Reviews." Under "Reviews," the user can see the application's average star rating, total number of ratings, and the total number of ratings by stars. By default, she can only see the application's top 3 "helpful" reviews for the application (see Fig. 1). To view more reviews, she must select "See all" at the bottom of the screen. On this new page, the ratings are again sorted by "Helpfulness," but the user can sort the reviews by date, rating, or phone. She can also rate other people's reviews by selecting either the "thumbs up" or "thumbs down" icon.



Street View on Google
GOOGLE INC.

★★★★☆ Holly 4/9/12
Samsung Stratosphere
Shows amazing views! So far have not had any problems with this app. Wish I could chart GPS locations though. Or make it easier to find the function if you can.

Good
★★★★☆ Vijay 4/7/12
Samsung Galaxy S2
Great way to explore the world.. But pls bring it to India as well.. And make it one click to access it in Maps. . Then will give 5 stars.

Awsome but missing
★★★★☆ Boyan 4/14/12
HTC Desire
There is some places that doesent

Street View on Google Maps

REVIEWS

Average
4.2
★★★★☆
258,031

5 stars 152,429
4 stars 52,860
3 stars 23,881
2 stars 6,031
1 star 22,830

Love this!
★★★★☆ Holly 4/9/12
Samsung Stratosphere
Shows amazing views! So far have not had any problems with this app. Wish I could chart GPS locations though. Or make it easier to find the function if you can.

Good
★★★★☆ Vijay 4/7/12
Samsung Galaxy S2

Fig 1. On the left, a screenshot of an application's top three "Helpful" reviews. On the right, "See all" view for an application.

B. Related Works

Many researchers have looked at online reviews, but none have examined reviews in Google Play. The reviews that have been examined have been from large e-commerce sites such as Amazon.com. Chevalier and Mayzlin [4] looked at reviews of the same book at Amazon.com and Barnesandnoble.com. They found that reviews at one site did not impact sales at the other site: positive reviews led to more sales at the same site and vice versa. They also found that negative reviews had more of an impact than positive ratings. They also observed that reviews at both sites tended to be positive, and they argued that consumers were taking the time to read longer reviews.

Other researchers have examined reviews in order to better predict sales. Hu, Pavlou, and Zhang [9] argue that the average rating of a product does not accurately predict future sales. This is because the distribution of product reviews tends to be a J-curve due to two reasons: purchasing bias and under-reporting. They found that consumers who purchase a product are more likely to write a positive review. They also argue that reviewers write reviews when they are incredibly satisfied or dissatisfied with a product. Individuals who have moderate feelings are less likely to write reviews. As a result, the average rating of a product is often skewed with overly positive or negative ratings. To account for these biases, they suggest looking at other variables in order to better predict product sales.

Archak, Ghose, and Ipeirotis [2] looked at consumer reviews in the "Camera and Photo" and "Audio and Video" section of Amazon.com. They leveraged text mining and hedonic regression to predict what features consumers care about. Their technique could be used to predict what customers actually want and help increase sales.

Researchers have also examined the helpfulness of negative and positive reviews. Sen and Lerman [15] divide products into two categories: utilitarian and hedonic—products that elicit pleasure. They argue that negative reviews are found to be more useful for utilitarian products rather than hedonic products, and that the motivation for and interpretation of negative reviews differ for both products. With utilitarian products, consumers felt that reviewers' feelings and experiences were about the product itself, and thus authentic and trustworthy. With hedonic products, consumers felt that negative reviews were less about the product itself.

Zhu and Zhang [17] looked at the impact of online reviews on video game sales. They looked at both Xbox and Playstation 2 games. Zhu and Zhang found that online reviews, particularly the total number of reviews, variation of individual ratings, and overall average rating, may help bolster the sales of less popular games.

Mackiewicz [11] categorized the different ways people assert their expertise and knowledge on the web. Looking at 750 product reviews at Eopinions.com, Mackiewicz developed 10 types of assertions, which fall into 3 broader categories: "assertion of experience, familiarity with related and relevant products, and relevant role."

Much research has also focused on detecting fake reviews. Wang, Xie, Liu, and Yu [16] developed a graph to identify spammers. Rather than rely on the reviewers and their text, they looked at the relationship between the reviewer, their reviews, and the stores in which the reviews were written. They also examined the trustworthiness, honesty, and reliability of reviewers, reviews, and online stores.

Afroz, Brennan, and Greenstadt [3] developed a framework for detecting false and deceptive writing styles online. They argued that when a writer is purposefully obscuring his writing style—i.e., pretending to be someone else—certain features in his natural style changes. By detecting these changes, Afroz et al. were able to differentiate between deceptive and non-deceptive writing styles.

Others have looked at star ratings and reviews to see if consumers write about privacy and security risks. Chia, Yamamoto, and Asokan [5], when studying Facebook, Chrome, and Google Play's permissions warnings, found that the average rating of an Android application was not negatively correlated to the number of permissions requested. They suggested that reviews were not about privacy or security issues, but about how the application functioned and worked. Our work supports their assumption; however, we provide a percentage for how often reviews mention privacy and security risks versus application functionality.

Felt, Greenwood, and Wagner [7], when looking the permission system for Google Chrome and AndroidOS, found that a small percentage of reviewers of Google Chrome extensions questioned these extensions' use of certain permissions.

In industry, Symantec is currently trying to find ways to analyze applications. They intend to look at an application's trustworthiness, resource usage, overall performance, and user reviews. We believe our work can be used to support them in their endeavor [10].

In our paper, we focus only on consumer reviews from Google Play. Rather than try to examine how reviews may affect downloads and/or sales or gauge trustworthiness, we attempt to understand what people do and don't write about in reviews.

III. Methodology

Data collection occurred in two phases from December 2011 to January 2012. In December 2011, we crawled Google Play to collect information about 202,264 free applications. The data scraped included the application's Google Play category, average rating, total number of ratings, and price. We chose to examine only free applications because Google Play contains more free applications than paid applications, and free applications are downloaded more than paid applications [1].

From our list of scraped applications, we selected 60 free applications whose reviews we would examine. The 60 applications were made up of two randomly selected applications from each Google Play category. We looked at only applications with reviews written in English and with at least 5 reviews. We chose to only examine 60 applications due to time constraints.

In January 2012, the reviews for our 60 applications were collected. By default, the reviews in Google Play are ordered by "Helpfulness." For our study, we looked at the first 10 reviews for each application, i.e., the 10 most helpful reviews and reviews that consumers were most likely to see by default. Some applications had less than 10 total reviews. In total, we looked at 556 reviews from 59 applications. One application was thrown out because the reviews contained illegible characters. During the time period between selecting the applications and scraping reviews, many of the selected applications changed Google Play categories. As a result, some Google Play categories have more than 2 applications and some have less.

A. Classification

After selecting our reviews, we then manually examined and classified the 556 reviews based on their content. To classify reviews, we coded them using sub-topics we iteratively developed through successive coding, validating and recoding the data. Our final classification list is made up of 37 sub-topics, which fall under 18 broader topics (Table I). These 37 sub-topics represent the content of these reviews, and they include positive and negative sentiment, and other information.

For each review, depending on what the consumer wrote about, we tagged it with our sub-topics. Regardless of the number of occurrences, each sub-topic is tagged only once for each review. Thus, each review receives a subset of the 37 possible sub-topics. Each review contained one or more of these sub-topics, though 20 reviews received "N/A," and were thrown out. Here is an example review for the British Gas app:

"Love it
Brillant [sic] *I can check my gas and electric for usage and check how much I'm spending a good way to save love it. HTC DESIRE HD"*

We coded this review with the "Adjective-Positive," "Feature/functionality-Positive," and "Model" sub-topics.

B. Validation

To validate our topics and sub-topics, we asked an outside researcher to independently recode our reviews from 10 randomly selected applications. We wanted to see if she would tag the reviews with the same sub-topics. When the researcher recoded the reviews the first time, only 38% of reviews had an exact match, 56% had a partial match, and 6% had no match. These low numbers led us to refine our topics. Reviews from 10 different applications were independently recoded. The second set of recoded reviews had a 73% exact match, 26% partial match, and 1% no match. Although the number of exact matches increased, we still felt that the number was still too low. Rather than look at exact matches within reviews, we wanted to see if there was agreement with our sub-topics. To do so, we selected reviews from 10 different applications, which were again recoded by the independent researcher. If both sets of recode either contained or didn't contain a sub-topic, then it received a score of 1. If one set had a code and the other one did not, then it received a score of 0. We then averaged the score for each sub-topic. Each topic had an average of at least 90, i.e., a 90% agreement rate. While this indicates that our classifications are not perfect, the level of consistency between reviewers appears to be sufficient for our purposes.

TABLE I. The 18 topics and 37 sub-topics we developed and used to code reviews.

Categories/Sub-categories	**Criteria**
Additional Program	Review mentions application's need of an additional program to work.
Adjective	
Adjective-Negative (Adjective-N)	Review negatively describes the entire application (rather than a feature or a functionality) using an adjective.
Adjective-Positive (Adjective-P)	Review positively describes the entire application (rather than a feature or a functionality) using an adjective.
Ads	
Ads-Negative (Ads-N)	Review complains about the number and content of ads in the application.
Ads-Positive (Ads-P)	Reviewer says that there weren't too many ads in the application or wouldn't mind having a free application that contained ads.
Aesthetics	
Aesthetics-Negative (Aesthetics-N)	Reviews negatively describes the application's overall look or interface, including images, color scheme, icons, and text.
Aesthetics-Positive (Aesthetics-P)	Reviews positively describes the application's overall look or interface, including images, color scheme, icons, and text.
Company	
Company-Negative (Co-N)	Review complains about the company who developed application. For example, reviewer complains that the company is unresponsive to emails.
Company-Positive (Co-P)	Review praises company who developed application.
Comparison	
Comparison-Negative (Comparison-N)	Review compares application A to application B, saying that application B is better.

Comparison-Positive (Comparison-P)	Review compares application A to application B, saying that application A is better.
Feature/Functionality	
Feature/Functionality-Missing (Featfunc-Missing)	Application is lacking a feature or functionality that the user would like to have or needs in order to better the experience. Feature/Functionality-Missing means that the feature/functionality currently does not exist.
Feature/Functionality-Negative (Featfunc-N)	Reviews criticize an existing feature. Typically, the feature is too slow, inaccurate, or doesn't work properly, negatively affecting the user experience.
Feature/Functionality-Positive (Featfunc-P)	Reviews praise an existing feature that works properly and is easy to use, bettering the overall user experience.
Just downloaded, don't know if it will work (Just DL-DK)	Reviewer doesn't know if the application actually works because she just downloaded it. Application asked for rating before it could be used.
Model (model)	Model of phone or tablet.
Money	
Money-Negative (Money-N)	Application claims it is free but it actually is not, or paid content/service is not worth the money.
Money-Positive (Money-P)	Review praises application for being free or if it is paid, then application/ service is worth the money.
Permissions	
Permissions-Explanation (Perms-Explanation)	Reviewer explains why the application needs certain permissions.
Permissions-Negative (Perms-N)	Reviews complain about permissions, ask why certain permissions are needed, or complain about unauthorized access to personal information (e.g., application spams contacts).
Permissions-Neutral (Perms-Neutral)	Review mentions permissions in passing, without saying if they are good or bad. Reviewer does not explain permissions.
Permissions-Positive (Perms-P)	Reviewer praises the application for not having or removing permissions.
Preinstalled	
Preinstalled-Negative (Preinstalled-N)	Review describes application as being "bloatware." Reviewer also complains about the inability to uninstall the application.
Preinstalled-Positive (Preinstalled-P)	Reviewer is glad that the application was preinstalled on her phone.
Recommendations	
Recommendations-Negative (Rec-N)	Reviewer does not like the application and does not recommend it to other users.
Recommendations-Neutral (Rec-Neutral)	Review neither positively or negatively recommends the application to user, says it is up to the user to decide if they want it or not.
Recommendations-Positive (Rec-P)	Review loves the application and recommends it to other users.
Resources	
Resources-Negative (Resources-N)	Review states that the application is a resource hog (e.g., takes up too much space, drains battery) or doesn't effectively leverage resources at all (e.g., doesn't use SD card).
Resources-Positive (Resources-P)	Review states that application doesn't hog resources (e.g., doesn't drain battery) or effectively uses resources (e.g., can save to SD card).
Tips (Tips)	Review tells other user how to install or use the application effectively.
Uninstalled (uninstalled)	Reviewer specifically says she has uninstalled the applications. Typically, user is unhappy with it.
Used to be	
Used to be-Negative (Use to be-N)	Application is negatively compared to its previous version. Before an update the application was great, however, the update had a negative effect on it.
Used to be-Positive (Use to be-P)	Application is positively compared to its previous version. The latest update made it better.
Work/Doesn't Work	
Doesn't Work with a Technical Reason (Doesn't Work-TR)	Reviewer writes that the application doesn't work and provides a technical description, such as it takes too long to load or that it keeps crashing.
Works with Technical Reason (Work-TR)	Reviewer writes the application does work and provides a technical description, such as it loads quickly, the application has no glitches, etc.
Doesn't Work with No Technical Reason (Doesn't Work-NTR)	Reviewer writes the application doesn't work, but provides no technical description. Typically, reviewer just writes, "Not working."
Work with No Technical Reason (Work- NTR)	Reviewer writes that the application works, but provides no technical description, for example, "Works." Sometimes the reviewer does provide a description, but it is not technical. For example, "this application makes me smile."

IV. RESULTS

In this section, we will present our results. In IV.A, we will look at frequently occurring topics and sub-topics. We will also look at whether reviewers wrote about permissions. This section will give us a general idea of what users frequently write about—and don't write about—in reviews. In IV.B we see if sentiment correlate with star ratings. Lastly, in IV.C, we see if there is correlation between sub-topics.

A. What Reviewers Write About

Do people mention privacy and security risks, e.g., permissions, in reviews? If not, then what do people generally write about? Figure 2 shows the percentage of how often our 18 broad topics appeared in our sampled reviews. While we developed 18 topics, only 4 topics appeared in more than 10% of reviews. The remaining 14 topics appeared in less than 10% of the reviews, creating a long tail.

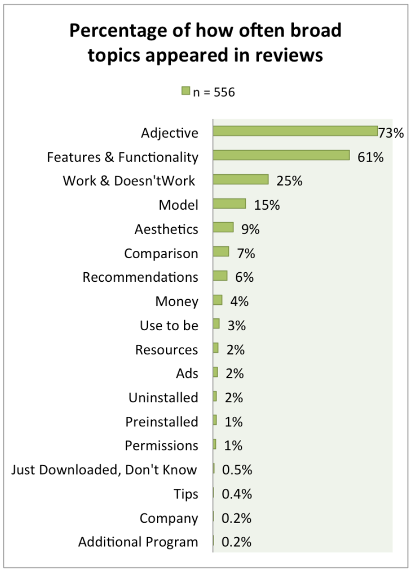


Fig 2. Percentage of how often our 18 broad topics occurred (n = 556).

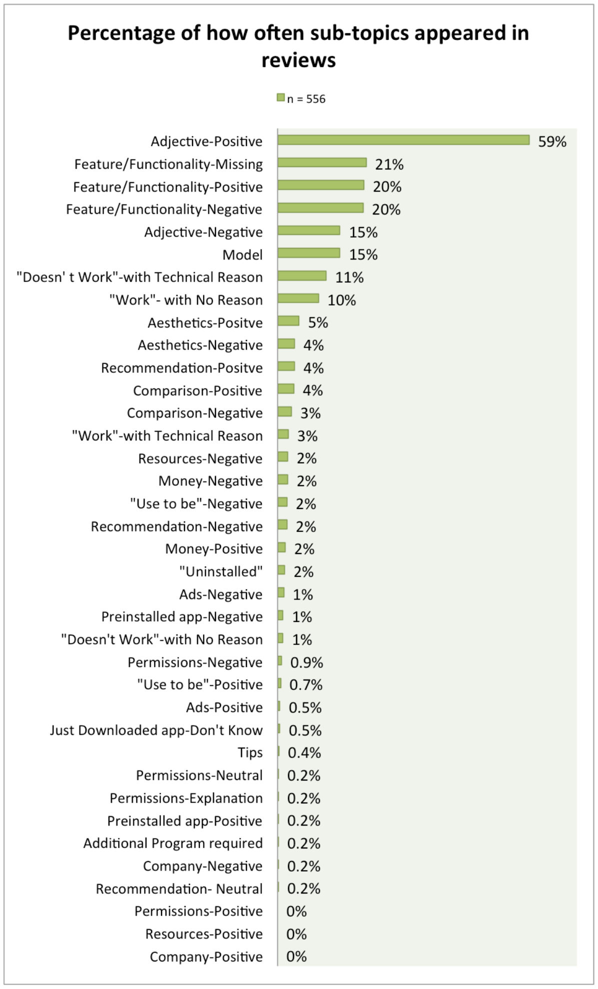


Fig 3. Percentage of how often our 37 sub-topics occurred (n = 556).

"Permissions" was mentioned only in 1% (7/556) of our reviews (Figure 2), and these reviews were about six different applications. The low occurrence of permissions-related reviews suggests that few people write about privacy and security risks. Instead, the topics "Adjectives" (73%; 408/556), "Feature/Functionality," (61%; 308/556), "Work/Doesn't Work," (25%, 137/556) and "Device Model" (15%, 81/556) were mentioned the most (Figure 2). These 4 frequently tagged topics describe how the application functions, which supports Chia et al.'s [5] assumption: consumer reviews tend to be about quality, rather than about privacy and security concerns.

Looking at Figure 3, comments about an application's feature/functionality are evenly distributed between those that are positive (20%, 112/556), negative (20%, 111/556), and refer to missing features/functionalities (21%, 115/556). Similarly, comments about whether an application worked ("Work"-No Reason, 10%, 53/556) and didn't work ("Doesn't Work- with No Technical reason, 11%, 63/556) are also evenly distributed.

Comments about permissions are not evenly distributed. Of the 1% (7/556) of reviews that mentioned permissions, 5 expressed a negative sentiment towards permissions ("Permissions-Negative," Figure 3). In these reviews, the reviewer either complained about and questioned permissions, or they complained about how an application attempted to access personal information:

"Permissions

Why does this app need to collect pictures or video from my phones camera? Great app otherwise. Uninstalling."

"Please explain to us..

Nice app, it works fine for me, but why does it need access to contact data and browsing history?? The creators must explain this.."

"Be aware

This keep [sic] *freezing my phone i* [sic] *got a bunch of error messages and it keep trying to access my email this is spam junk"*

Our finding is similar to past studies where researchers found that a small number of users were questioning the use of permissions [7, 8]. Of our sample, only 1 of 7 "Permissions" reviews provided an explanation ("Permissions-Explanation," Figure 3):

"@Omar- It needs to read your contacts, because barcodes can contain contact information, and it needs to read your browsing history, because barcodes can contain URL information…"

There are a number of reasons why there are few reviews that explain why permissions are required. One reason is that the number of knowledgeable users who may be able to provide explanations is much smaller than the number of expert users who just notice permissions. A potential way to educate "permissions conscious" users is to require application developers to address permissions in the application's description. While Felt et al. [7] showed that reviewers could pressure developers to provide an explanation at a later time, developers can prevent initial negative reactions by providing an explanation during application submission. We provide more suggestions in our "Discussion" section.

We also saw other potential privacy and security concerns in our data. 1% (7/556, Figure 3) of reviews negatively criticized pre-installed applications. Pre-installed applications are applications that already come installed on the phone, and they cannot be uninstalled. Pre-installed applications differ by phone and carrier, and they also have their own set of permissions. There are implications to pre-installed applications: if a user dislikes or is uncomfortable with a pre-installed application's permissions, she does not have the ability to uninstall the application. However, in our study, none of the "Pre-installed-Negative" reviews complained about permissions. This could be because the applications in question did not have suspicious permissions or because people were unaware of permissions. Instead, the majority of these reviews (6/7) specifically complained about the inability to uninstall the applications:

"Its horrible for me because i [sic] *can't uninstall it. If you use photobucket, then i* [sic] *recommend downloading it tov* [sic] *your phone if you don't plan on uninstalling it.... EVER...."*

"Never downloaded it, dont want it, cant seem to get rid of it. Every time i restart my g2, the photobucket logo is in the top left. How do i get rid [sic]"

"Good app. Works well. I use it occasionally. 1 star because it comes preloaded with no option to uninstall."

This suggests that there are no privacy concerns with pre-installed applications. Instead, Android users are more concerned with their inability to uninstall these types of applications.

We also found that 0.5% (3/556) reviews asked users to rate and review the application before it was actually used ("Just DL-DK"). While these reviews were not common, this still suggests a possible abuse or manipulation of the review system that Google Play may find useful to automatically detect.

B. Sub-Topics and Star Ratings

Now we look at our reviews' overall star ratings. Then, we look at the top 4 topics' sub-topics and their overall star rating, including "Aesthetics" and "Money." We do not analyze other topics (e.g., "permissions") due to their small sample size.

1) Overall Star Ratings

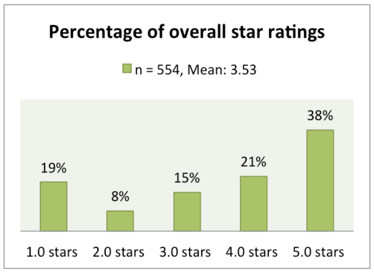


Fig 4. Our dataset's overall star ratings. Our graph shows a J-curve distribution, and reviews were generally positive (n = 554).

We recorded the number of stars the reviewers gave in each review (i.e., the "star rating"). In Google Play, reviewers can only rate an application based on a scale of 1 to 5. Every review comes with a star rating for the application. Although our data set contained 556 reviews, we threw out two ratings because they had a rating of 0.0. We believe the 0.0 ratings are a mistake caused by an error in our software.

Figure 4 shows the overall distribution of the star ratings in our dataset. Similar to Hu et al.'s work [9], the ratings in our dataset show a J-curve distribution and are generally positive. In this respect, it can be argued that Google Play consumers are no different from consumers of other online markets. Hu et al. suggest that consumers suffer from purchasing bias, in that they are more likely to view the product more positively since they committed the time and money to purchase it [9]. Furthermore, since there is a spike for 1.0 star rating and 5.0 star ratings, it appears that (similar to Hu et al.'s finding) consumers who are reviewing applications are only doing so when they are either incredibly satisfied or dissatisfied [9].

2) Adjectives

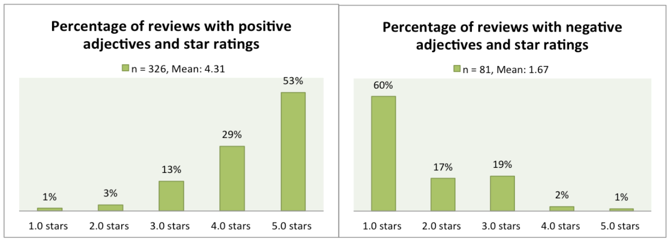


Fig 5. Left: Distribution for reviews containing positive adjectives (n = 326). Right: Reviews containing negative adjectives (n = 81)

Reviews that have a positive adjective (e.g., "love it," "awesome," "great app," etc.) were rated significantly higher (4.31) than those that did not (2.42; z = -13.453, p = 0.000; Figure 5, Left). The star rating for this sub-topic is overwhelmingly positive: 53% (174/326) of reviews labeled received a five-star rating.

Conversely, 60% (49/81) of reviews labeled with "Adjective-Negative" received a 1-star rating (Figure 5, Right). Reviews containing negative adjectives (e.g., "horrible," "sucks," "boring," etc.) were rated significantly lower (1.67) than those that did not have negative adjectives (3.85; z = 11.237, p = 0.0000, Mann-Whitney). On average, the difference between "Adjective-Negative" and "Adjective-Positive's" ratings is 2.64 stars.

3) Features/Functionalities

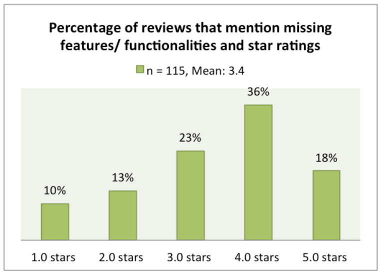


Fig 6. 115 of our reviews mentioned a missing feature or functionality. Surprisingly, 36% of those reviews received a 4-star rating (n = 115).

Figure 6 shows the distribution for sub-topic "Feature/Functionality-Missing." While reviews labeled with this sub-code are generally positive, 4-star ratings are more common than 5-star ratings. We interpret the spike in 4-star ratings to mean that applications tagged with this sub-topic were generally pretty good, but not perfect. Overall, customers liked the application, but because it lacked some sort of feature they wanted or expected, they did not give it a 5-star review.

The average star rating for reviews labeled "Feature/Functionality-Missing" is 3.4, whereas the mean for the rest of reviews is 3.56. The difference is statistically significant (z = 2.314, p = 0.0207, Mann-Whitney). The difference is small, which suggests that reviewers do not penalize applications for missing features and functionalities, or reviewers are more likely to mention missing features and functionalities if they already like the application.

The average star rating for reviews labeled with "Feature/Functionality-Positive" is 4.47, whereas the mean for those without is 3.29. The difference is statistically significant (z = -7.156, p = 0.0000, Mann-Whitney; Figure 7, Left). Reviews tagged "Feature/Functionality-Positive," will have a positive overall star rating.

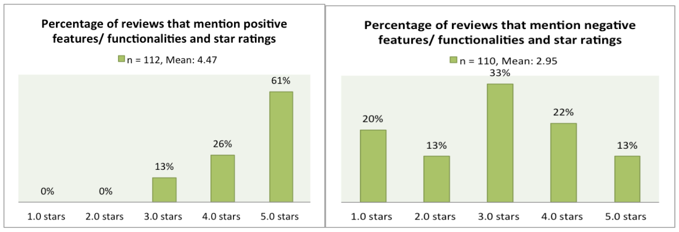


Fig 7. Left: Distribution of reviews that positively described a feature or functionality (n = 112). Right: Reviews that negatively described a feature or functionality (n = 110).

The mean star rating for reviews labeled "Feature/Functionality-Negative" (Figure 7, Right) is 2.95, whereas the mean for the rest is 3.67. The difference is statistically significant (z = 2.734, p = 0.0063, Mann-Whitney).

In general, reviews that positively describe features and functionality were rated about 1.52 stars higher (on average) than reviews that negatively mention features and functionality. This suggests that an application's feature and functionality is important to users and makes a big difference to their overall evaluation of an application.

4) Doesn't Work/ Work

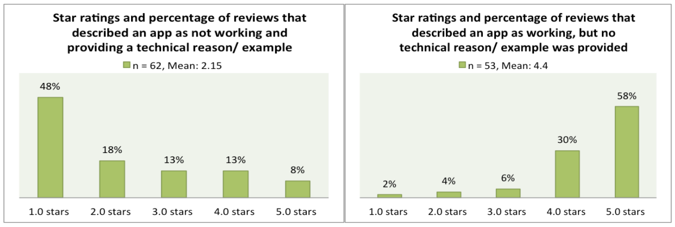


Fig 8. Left: Star distribution for reviews that stated an application didn't work and provided a technical reason or example (n = 62). Right: Reviews that described an application as working, but no technical reason or explanation was provided (n = 53).

We examined 2 of 4 sub-topics for the "Work/Doesn't Work" topic: "Doesn't Work-with a Technical Reason " and "Work-with no Technical Reason." We do not look at the others because of their small sample sizes. Most reviews (40%, 30/62) tagged with "Doesn't Work-with a Technical Reason" ("Doesn't Work-TR;" Figure 8, Left) have a 1.0-star rating. The mean for reviews labeled with "Doesn't Work-TR" is low: 2.15, vs. 3.70 for all other reviews. The difference is statistically significant (z = 7.247, p = 0.0000, Mann-Whitney).

In comparison, reviews tagged with "Work-with No Technical Reason" ("Work-NTR;" Figure 8, Right) are overwhelmingly positive. The mean for reviews labeled with "Work-NTR" is 4.4, much more positive than those not coded with "Work-NTR" (3.44). The difference is statistically significant (z= -4.302, p = 0.0000, Mann-Whitney). Reviews with a comment indicating the application works are associated with a rating that is 2.25 stars higher on average, than reviews commenting that the application doesn't work.

5) Model

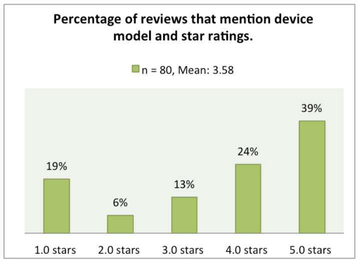


Fig 9. 80 reviews mentioned a device model. There is no correlation between mentioning a device model and star ratings (n = 80).

The distribution of ratings for reviews that mention a device model (Figure 9) is similar to the overall distribution for all reviews—a J-curve. The difference in means (3.58 vs. 3.52) is not statistically significant (p = 0.7855, Mann-Whitney).

6) Aesthetics & Money

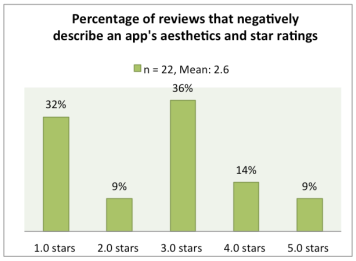


Fig 10. Star distribution for reviews that negatively described an application's aesthetics (n = 22).

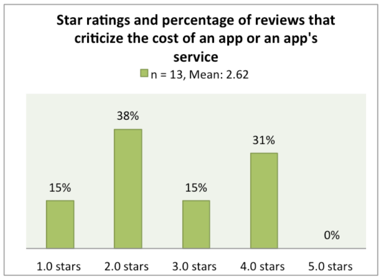


Fig 11. Star distribution for reviews that negatively described the cost of an application or service (n = 13).

For the topics "Aesthetics" and "Money," we found no statistically significant differences between the average ratings of reviews that praised an application's aesthetics ("Aesthetics-Positive") or its cost ("Money-Positive") and reviews that did not praise aesthetics or cost. However, there were statistically significant differences between ratings for reviews tagged with "Aesthetics-Negative" or "Money-Negative" versus reviews that were not tagged with either sub-topics: reviews tagged with "Aesthetics-Negative" (Figure 10) had a lower average rating than others (2.6 vs. 3.57, z = 3.190, p = 0.0014, Mann-Whitney), and reviews tagged with "Money-Negative" (Figure 11) had a mean of 2.62 stars vs. 3.55 for other reviews (z = 2.504, p = 0.0123, Mann-Whitney). For "Money-Negative," we observed that most reviews received either a 2-star (5/13, 38.46%) and 4-star rating (4/13, 30.77). We interpret 2-star ratings to mean that the sampled applications still had some decent qualities and were not bad enough to receive a 1-star rating. We interpret 4-star ratings to mean that the application would have been perfect and would have otherwise received a 5-star rating if it were not for financial reasons.

Overall, our data suggests that positive reviews for an application's aesthetics or cost will not impact its overall star rating. However, negative reviews complaining about an application's aesthetics or cost can decrease ratings.

C. Correlation Between Sub-Topics

Is there a correlation between sub-topics? Does one sub-topic frequently appear with another and vice versa? We also looked at the top 8 sub-topics to see if they have a relationship with each other. We computed the correlation coefficient for all pairs, and used Fisher's exact test (two-tailed) to test for the existence of a statistically significant correlation.

We observed both unsurprising and interesting results. For example, a review that praised an application's feature or functionality ("Feature/Functionality-Positive") was positively correlated with a positive adjective (0.2472, p = 0.0000). Conversely, a negative adjective is negatively correlated with a positive adjective (-0.4417, p = 0.0000). Our most interesting finding is that in reviews that describe the application as a working or not working, the reviewer is more likely to also mention the model of the reviewer's Android phone or tablet (0.1258, p = 0.0030; 0.2479, p = 0.0000, respectively). This suggests that many reviewers are trying to be helpful to other readers in identifying compatibility with different devices.

V. DISCUSSION

Our research suggests that consumer reviews in Google Play are informative, though few reviewers explicitly expressed privacy and security concerns. In our study, 1% of reviews (7/556) mentioned permissions, and 5 of 7 "Permissions" reviews complained about or questioned the use of permissions. Similar to past studies, this indicates that there are a small number of users who are aware of and pay attention to permissions [7, 8]. Only 1 of 7 "Permissions" reviews (and 1 out of all 556 reviews) explained why specific permissions were required. There are a number of possible reasons why we found low occurrences of community-based explanations: there are few users who understand why permissions are required, knowledgeable users do not see people's questions because they are buried between other reviews, or because knowledgeable users are not responding to such questions. One recommendation is to require application developers to explain *why* permissions are required in the application description page when the application is first uploaded to Google Play, or when the application is updated with new permissions. This way, users will not have to rely on others to answer a permissions-related question.

Another possibility is to create a separate section for just questions on the application's Google Play page. In the current design, privacy and security questions and explanations are buried within reviews, meaning such information may go unnoticed. A "questions-only" section would separate users' questions from reviews, and would allow knowledgeable users to easily find and address concerns relating to privacy and security, as well as concerns regarding an application's overall functionality. It would also allow potential application downloaders to be able to easily see other's concerns and responses before downloading the application. Moreover, a "questions-only" section could potentially help users who are unaware of or don't pay attention to permissions to become aware of and notice them.

Felt et al. [8] suggested incentivizing reviewers to address privacy and security concerns. There is potential in leveraging the current practices of message boards and other ecommerce sites to reward users for addressing concerns, particularly those related to privacy and security. This includes creating a reputation system. While Google Play allows users to currently "thumbs up" and "thumbs down" a review, it might be helpful to include fun "expert" level titles of reviewers and the ability to access a reviewer's past comments. We imagine that the number of responses and like/dislikes would help establish a user's "expert" level. By including expertise level and history of comments, users can gauge whether or not the reviewer is trustworthy.

If users are not writing about privacy and security risks, then what are they writing about? We found that the majority of reviews commented on an application's quality and functional aspects (e.g., whether it works, how it looks, etc.). 61% (338/556) of reviews discussed an application's feature and functionality. We found a negatively reviewed feature/functionality will have a much lower star rating than an application whose feature/functionality are rated positively (2.95 vs. 4.47). Surprisingly, applications missing a feature/functionality are not negatively rated. This indicates that consumers are willing to overlook this missing feature, so long as the overall application works.

Whether an application worked or didn't work was also important to reviewers. Reviews where the reviewer indicated that the application worked for them were rated much higher than those that reported that the application didn't work—a difference of 2.25 stars, on average. We also observed that when an application did not work, people were more likely to explain "why" it did not work by describing its behavior when it was installed. Conversely, when an application did work, people were less likely to provide any technical description. By providing additional information, reviewers are able to warn other users of a particular application; a working application may not need any additional information.

Our study also indicates that people care if an application's interface is visually unappealing or hard to use. Reviews that negatively describe the application's aesthetics or usability tend to have a lower star rating; but reviews that positively mention these qualities do not seem to be associated with an inverse star rating. Thus, if an application is "ugly," or if it's interface is hard to use, the application's ratings will be negatively impacted. Perhaps users expect applications to be usable and good-looking, and thus do not award extra stars for providing these qualities—but they do lower their ratings if the application falls short.

People also do not like being misled into paying for an application or service within a "free" application. We found the mean for "Money-Negative" was low. This suggests, perhaps, that when people download a free application, they expect the information in the application to be free as well.

Lastly, our research suggests that Google Play is no different from other e-commerce sites. Similar to past research on other sites, the overall ratings in Google Play are overwhelmingly positive, which suggests that purchasing bias also affects the overall ratings. The spike in 1 and 5-star ratings indicate that reviewers are writing reviews primarily when they are extremely satisfied or dissatisfied.

VI. CONCLUSION

Overall, reviews in Google Play contain substantive information about applications, though most information is about the quality of the application and not about privacy and security concerns. However, there were a small number of users who did explicitly question the use of permissions in applications, indicating that there is an opportunity to educate and address such concerns in application markets. More broadly, our study provides some visibility into what people write about in reviews and what qualities of applications are most important to reviewers. As applications continue to grow in popularity and usage, developers and designers should become increasingly aware of what their customers want and expect from an application.

This paper is a first step to understanding what users write about in the Google Play. We realize that much work can still be done. For instance, it would be interesting to study the accuracy of reviews. It would also be interesting to investigate whether it is possible to automatically summarize all reviews of an application. This would allow consumers to quickly see common concerns or strengths of an application without having to individually read reviews. Lastly, since user reviews appear to be informative, it would be interesting to see if we could use reviews to predict whether an application is "good" or not [10].

ACKNOWLEDGEMENT

This work was supported by Intel through the ISTC for Secure Computing. We thank Chris Hoofnagle for inspiring this research project. We also thank Adrienne Porter Felt, Nathan Good, Jennifer King, and anonymous reviewers for reading previous drafts and providing insightful feedback and suggestions. Huge thanks to Ben Dong and George Li for developing the scraper, and Leah Dorner for validating our topics and sub-topics.

REFERENCES

[1] Appbrain. Distribution of free vs. paid Android apps. http://www.appbrain.com/stats/free-and-paid-android-applications. Accessed May 22, 2012.

[2] N. Archak, A. Ghose, and P. G. Ipeirotis. 2007. "Show me the money!: deriving the pricing power of product features by mining consumer reviews." In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '07).

[3] S. Afroz, M. Brennan, and R. Greenstadt. "Detecting hoaxes, frauds, and deception in writing style online," In *Proceedings of the 2012 IEEE Symposium on Security and Privacy* (SP '12).

[4] J.A. Chevalier and D. Mayzlin. 2006. "The effect of word of mouth on sales: online book reviews." Journal of Marketing Research, Vol. 43, No. 3, pp. 345-354

[5] P. H. Chia, Y. Yamamoto, and N. Asokan. 2012. "Is this app safe?: a large scale study on application permissions and risk signals." In *Proceedings of the 21st international conference on World Wide Web* (WWW '12).

[6] A. P. Felt, M. Finifter, E. Chin, S. Hanna, and D. Wagner. "A survey of mobile malware in the Wild." In *Proceedings of the ACM Workshop on Security and Privacy in Mobile Devices* (SPSM' 11).

[7] A. P. Felt, K. Greenwood, and D. Wagner. 2011. "The effectiveness of application permissions." In *Proceedings of the 2nd USENIX conference on Web application development* (WebApps'11).

[8] A. P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D.Wagner. "Android permissions: user attention, comprehension, and behavior." In *Proceedings of the Eighth Symposium on Usable Privacy and Security* (SOUPS '12).

[9] N. Hu, J. Zhang, and P. A. Pavlou. 2009. "Overcoming the J-shaped distribution of product reviews." *Commun. ACM* 52, 10 (October 2009), 144-147.

[10] R.Lee. "Symantec introduces new Norton mobile solutions." http://asia.cnet.com/symantec-introduces-new-norton-mobile-solutions-62216875.htm. Accessed August 18, 2012.

[11] J. Mackiewicz. "Assertions of expertise in online product reviews." Journal of Business and Technical Communication" January 2010 24: 3-28, first published on September 10, 2009

[12] NielsenWire. State of the Appnation – "A year of change and growth in U.S. smartphones." http://blog.nielsen.com/nielsenwire/?p=31891. Accessed July 2, 2012.

[13] NielsenWire. "Games dominate America's growing appetite for mobile apps." http://blog.nielsen.com/nielsenwire/online_mobile/games-dominate-americas-growing-appetite-for-mobile-apps/. Accessed July 2, 2012.

[14] NielsenWire. "Consumers and mobile apps in the U.S.: all about Android and Apple iOS." http://blog.nielsen.com/nielsenwire/online_mobile/consumers-and-mobile-apps-in-the-u-s-all-about-android-and-apple-ios/. Accessed July 2, 2012.

[15] S. Sen, and D. Lerma. 2007. "Why are you telling me this? An examination into negative consumer reviews on the Web." Journal of Interactive Marketing, 21:4, 76–94.

[16] G. Wang, S. Xie, B. Liu, and P. S. Yu. 2011. "Review graph based online store review spammer detection." In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining* (ICDM '11).

[17] F. Zhu, X. (Michael) Zhang. 2010. "Impact of online consumer reviews on sales: the moderating role of product and consumer characteristics." Journal of Marketing 74:2, 133-148.