Empirical Software Engineering (2022) 27:34
https://doi.org/10.1007/s10664-021-10080-8

# A study of gender in user reviews on the Google Play Store

**Ehsan Noei**[1] · **Kelly Lyons**[1]

Accepted: 8 November 2021 Published online: 20 December 2021


**Abstract**
User reviews that are posted on the Google Play Store provide app developers with important information such as bug reports, feature requests, and user experience. Developers should maintain their apps while taking user feedback into account to succeed in the competitive market of mobile apps. the Google Play Store provides a star-rating mechanism for users to rate apps on a scale of one to five. Apps that are ranked higher and have higher star ratings are more likely to be downloaded. In this paper, we investigate and compare men's and women's participation in user reviews that are posted on the Google Play Store. We analyze 438,707 user reviews of the top 156 Android apps over six months. We find that women give higher star ratings and use more positive sentiment in their reviews than men. Furthermore, women's reviews receive more likes and are ranked higher in the top 10 by the Google Play Store. For the reviews from which user gender can be inferred, we find that men submit more reviews than women, making reviews by men more likely to be visible to app developers and other users. Past research has shown that app developers respond more to negative reviews with fewer stars. We found that developers respond to a greater percentage of men's reviews than women's. The small number of and more positive reviews by women are less likely to be addressed by app developers; thus, the resulting changes to apps will align more with the needs of men users, perhaps causing even less participation by women in the Google Play Store reviews. Our findings suggest that developers should take gender into consideration when responding to reviews to help mitigate a feedback loop of bias. Our observations also suggest a need for future research in this area to understand the motivations of men and women in reviewing apps and how developers respond to reviews.

**Keywords** Empirical study · Gender · Mobile app · App store

Communicated by: Emerson Murphy-Hill

✉ Ehsan Noei
e.noei@utoronto.ca

Kelly Lyons
kelly.lyons@utoronto.ca

[1] University of Toronto, Toronto, Canada

## 1 Introduction

The revenue of mobile applications (apps) has increased in the past few years where app markets, such as the Google Play Store[1], have become immensely competitive for app developers (Statista 2020). The Google Play Store provides users of Android apps with a mechanism for rating apps and leaving feedback (i.e., user reviews). User reviews often include critical information such as bug reports and feature requests that developers can use to improve and maintain their apps, whereas, those developers who do not risk the survival of their apps in the competitive market of mobile apps (Noei et al. 2018). Users have the option to associate their user reviews with ratings from one star (the lowest) to five stars (the highest). Although users' sensitivity to ratings can be different in different countries (Kübler et al. 2018), star ratings can affect the number of downloads (Bavota et al. 2015; Kim et al. 2011a). Developers should keep improving their apps with respect to users' feedback in order to maintain or increase their star ratings and ranks on the Google Play Store (Noei et al. 2018).

Prior research (e.g., May et al. 2019; Merchant et al. 2019; Wachs et al. 2017; Hannák et al. 2017; Terrell et al. 2017; Wagner et al. 2016; Stephens 2013; Ford et al. 2017) has studied gender differences on online platforms, such as Stack Overflow[2] and GitHub[3]. For example, Vasilescu et al. (2014) found that a large proportion of users of Stack Overflow use anonymous identities and the percentage of women who engage in such communities is imbalanced. They asked users of Stack Overflow to provide them with information such as their gender and age; they obtained 127 valid responses where the majority of respondents were men (115 responses from men). May et al. (2019) reported significant differences in the way men and women participate in Stack Overflow discussions, observing that women hold half the reputation of men on average. They suggested that this could be due to the fact that men post more answers than women and men are usually rewarded more for their given answers (May et al. 2019). On the other hand, unlike the findings on Stack Overflow, Terrell et al. (2017) reported that women's contributions are accepted more often than men's on GitHub.

Because of the importance of user reviews for changes and fixes made on Android apps, we are interested in studying gender participation on the Google Play Store. Users of Android apps can *like* a user review that they find useful and developers can *respond* to any user review if desired. We identify genders associated with user reviews on the Google Play Store. There are two limitations in the use of gender. First, users of the Google Play Store (as on many platforms) do not include information about gender in their profile and so it is not possible to know the gender with which people identify. Instead, we infer gender from user names (May et al. 2019; Wachs et al. 2017; Hannák et al. 2017; Wagner et al. 2016; Ford et al. 2017; Vasilescu et al. 2014). Another approach that has been used in the case of GitHub is to link user profiles by email address with social media sites where users self declare gender (Terrell et al. 2017). Since email addresses are not available from the Google Play Store, we inferred gender from names. In cases where gender information is inferred from names, there is the added limitation that the information is typically based only on two genders. We recognize that gender does not exist in opposition or on a binary scale (Scheuerman et al. 2019). To attempt to overcome this limitation, some researchers have used an approach that

[1] http://play.google.com/
[2] http://www.stackoverflow.com/
[3] http://www.github.com/

counts the number of times a name is found in a list of male names in different countries and the number of times it is found in a list of female names in different countries. If the number of times it is counted as male [female] is 2 times greater than the number of times it is counted as female [male], it is labeled male [female], otherwise, it is labeled as what they call "unisex" (Ford et al. 2017). This approach is still limited in its ability to categorize users according to their self-identified gender. We follow the approach of May et al. (2019), Wachs et al. (2017), Hannák et al. (2017), Wagner et al. (2016), and Vasilescu et al. (2014) and use tools that infer gender on a binary scale, recognizing this as a limitation of our study.

We compare the number of user reviews posted by women and men. Then, we compare the star ratings and sentiment scores of user reviews associated with men and women. We identify the topics that are discussed by men and women. Lastly, we study the relationship between gender and star ratings. Women tend to give higher star ratings and post user reviews with higher sentiment scores. For this study, we analyze 438,707 user reviews of 156 Android apps that are among the top-ranked apps on the Google Play Store in Canada for six months from $January$ 1, 2019 to $July$ 12, 2019. We address the following research questions:

**RQ1) Are there the same number of reviews posted by men and women?** Related work suggests lower participation of women on online platforms such as Stack Overflow (May et al. 2019) and a low ratio of women participation in online communities. We investigate men's and women's participation in the Google Play Store. We observe that, for the reviews from which user gender can be inferred, there are two times more user reviews that are posted by men than women. Our findings suggest that remedies may be required to encourage women to participate more in the Google Play Store as users' feedback play an important role in app development processes (Noei 2018).

**RQ2) Do men's and women's user reviews share similar topics?** We apply topic modeling on user reviews and observe statistically significant differences between the proportion of topics by men and women. For example, 28% of women discuss the topic of *Learning* while only 23% of men do. On the other hand, 26% of men discuss *Device & Connectivity* as opposed to 20% of women.

**RQ3) What is the relationship between gender and star rating?** By building a linear regression model with star ratings as the dependent variable, we identify gender as a statistically significant factor for explaining star ratings. Also, star ratings and sentiment scores associated with user reviews posted by women tend to be higher than those posted by men. Higher star ratings and sentiment scores of women can introduce inequality as developers tend to prioritize addressing more negative feedback in the future releases of their apps (Noei et al. 2019a).

**RQ4) How different is the response to men's and women's reviews?** Because we find a statistically significant relationship between gender and star rating, in this research question, we study users' and developers' responses to men's and women's reviews. We observe more likes received for reviews that are posted by women. However, developers tend to respond more to reviews that are posted by men. In addition, by analyzing the top ten (McMillan et al. 2013; Niu et al. 2017) user reviews of each app on the Google Play Store, we observe that women's reviews appear more often than men's among the reviews that users are more likely to see first (i.e., top ten). Our observations suggest that developers should take gender into consideration to mitigate a feedback loop of bias when responding to reviews.

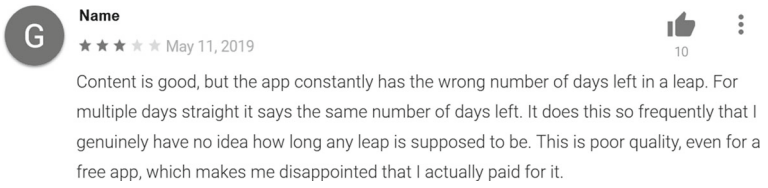


**Fig. 1** An example user review on the Google Play Store

**Paper Organization** Section 2 explains our study setup including data collection, gender inference, and metric measurement. Section 3 describes our research questions and findings. Section 4 introduces the related work. Section 5 discusses the potential threats to the validity of this work. Finally, we conclude the paper in Section 6.

## 2 Study Setup

First, we explain the data collection process in Section 2.1. In Section 2.2, we describe the methods that we use to infer users' genders. Finally, in Section 2.3, we explain the metrics that are measured for this study.

### 2.1 Data Collection

We gathered a list of the top 200 apps that are provided by the Google Play Store in Canada as of $July$ 12, 2019. The Google Play Store limits the total number of user reviews that a user can view (Khalid et al. 2014a), so that we could not access all the user reviews of all the 200 apps. Martin et al. (2015) discuss that using an incomplete set of user reviews in app stores can introduce bias to the findings of empirical studies. To mitigate such a threat, we study 156 of the apps for which we could retrieve all their user reviews for 6 months. In this study, we use 438,707 of the total 739,598 retrieved user reviews (see Section 2.2). Android apps in this study, have a minimum of 11, median of 2,956, average of 4,741, and maximum of 17,840 user reviews.

To retrieve the user reviews, we utilized the same approach as Noei et al. (2019a) and developed a crawler on top of the Selenium automation tool.[4] Selenium provides a set of tools and an application programming interface (API) to facilitate automated web browsing processes. The client API of Selenium lets developers communicate with the backend of Selenium and its web driver sends the Selenium commands to the browser. By employing Selenium, for each app, we retrieved all the available details of user reviews, such as the name of users, reviews, associated stars, number of likes, date, and developers' response (if available).

Figure 1 shows an example user review posted on the Google Play Store. A user review is an informal piece of text without a predefined structure. User reviews are valuable as they contain useful information for app developers, such as bug reports, feature requests, and user experience (Fu et al. 2013; Palomba et al. 2015). Recent studies have shown that addressing the issues reported in user reviews can significantly help developers to improve their apps and succeed in the competitive market of mobile apps (Noei et al. 2019a; Palomba

[4] http://seleniumhq.org/

et al. 2015). As shown in Fig. 1, each user review is associated with a star rating from 1 (the lowest) to 5 (the highest). Also, for each user review, as in Fig. 1, the name of its author (replaced with the word: *Name* in this figure to protect user privacy), date of the user review, and the number of likes received from other users (indicated by the count under the *thumps up* icon) are accessible.

### 2.2 Gender Inference

Identifying users' genders is typically a complex problem (May et al. 2019). As shown in Fig. 1, each user review is associated with a name; therefore, we infer users' genders with the help of associated names. To this end, we apply and evaluate three of the available solutions in the literature (and combinations of them): Genderize (2019), Genni (Smith et al. 2013), and GenderComputer (Vasilescu et al. 2014).

Genderize (Genderize 2019) is a tool that provides an API to predict the gender of an individual using a database of 114,541,298 names in different languages, such as English, Spanish, and French. We use GenderizeR (Wais 2016) which is an R interface of Genderize to access its API. Genni (Smith et al. 2013) applies a logistic regression classifier to distinguish men's and women's names. It also provides a web API to let users access the implemented classifier. GenderComputer (Vasilescu et al. 2014) infers the gender of an individual by looking up at a table of names that is collected using names associated with users of Stack Overflow.

As discussed earlier, an important limitation of using tools such as Genderize (Genderize 2019) and Genni (Smith et al. 2013) is their inability to identify non-binary genders (Topaz and Sen 2016; Ibrahim et al. 2019), which is, therefore, a limitation of our study.

#### 2.2.1 Evaluation Data

To evaluate and identify the best approach to infer genders on the Google Play Store, we take a random sample of 385 names from our dataset with a 95% confidence level and confidence interval of 5.

#### 2.2.2 Manual Labeling

Three evaluators manually labeled each name in the sample. Each evaluator assigned a label of either *man*, *woman*, or *unknown* (if they could not identify the gender using the associated name). Then, we applied the majority vote rule to solve conflicts among evaluators. Out of 385 names, 205 (54%) names were identified as men, 87 (23%) were women, and 91 (24%) identified as unknown. We could not resolve only two names using the major vote rule; therefore, we excluded these two names from our evaluation set.

#### 2.2.3 Evaluation of Tools

Having our manual labeling as our golden set, we compare the output of each tool against our manual analysis. Table 1 shows the results of our evaluation. As shown in Table 1, Genderize achieves a precision of 85% whereas Genni and GenderComputer achieve a precision of 83% and 81%, respectively.

In order to further improve the gender inference precision, we also tried combinations of the aforementioned solutions and compared the outputs with our manual labeling. In

**Table 1** Gender inference evaluation results

| Solution | Precision | Loss |
|---|---|---|
| Genderize | 0.85 | – |
| Genni | 0.83 | – |
| GenderComputer | 0.81 | – |
| **Genderize and Genni** | **0.94** | **0.20** |
| Genderize and GenderComputer | 0.95 | 0.22 |
| Genni and Gender Computer | 0.87 | **0.14** |
| All Three | **0.96** | 0.27 |

other words, we take a label as man or woman where a combination of solutions agrees on the same label. We evaluate the following combinations: (i) Genderize and Genni, (ii) Genderize and GenderComputer, (iii) Genni and Genderize, and (iv) all three solutions.

The bottom part of Table 1 (last four rows) shows the results of each combination. The last column in Table 1 (i.e., loss) indicates the proportion of outputs where the solutions in each combination do not agree on the same label; therefore, will be omitted. The combination of all three solutions achieves the highest precision (96%) however this combination suffers from the highest loss rate (27%). While the combination of Genni and GenderComputer has the lowest loss rate, it has the lowest precision among the combinations. We selected the combination of Genderize and Genni as our solution by a compromise between precision and loss rate where this combination achieves a precision of 94% and a loss rate of 20%. Therefore, by using both Genderize and Genni, we are able to identify user genders with higher precision (in comparison to applying each method individually) and with a relatively low loss rate.

### 2.3 Measuring Metrics

In addition to the identified gender labels (see Section 2.2), we follow the *Goal/ Question/ Metric* (GQM) paradigm (Basili 1992; Van Solingen et al. 2002) to capture the metrics of user reviews. The GQM paradigm is based on three levels: (i) conceptual, (ii) operational, and (iii) quantitative. The conceptual level defines the goal of measurements with respect to the purpose of a given model. The operational level includes a set of questions in order to describe the goal of the conceptual level. The quantitative level defines a set of metrics that should be measured in order to address the questions of the operational level.

Table 2 shows our GQM model with the goal of quantifying the information provided in user reviews. As shown in Table 2, we measure metrics from six major aspects: (i) star ratings, (ii) sentiment scores, (iii) users' response (likes), (iv) length (number of words and sentences), (v) rank, and (vi) developers' response. For each user review, we capture the overall sentiment score, negative sentiment score, and positive sentiment score. The negative sentiment score is the measurement of negative terms, such as *"terrible"*, and positive sentiment score in the measurement of positive terms, such as *"excellent"*. The overall sentiment score is taking both negative and positive aspects into consideration by summing negative and positive sentiment scores.

**Table 2** GQM model with the goal of quantifying user reviews associated with men and women

| Question | Metric(s) | Description |
|---|---|---|
| How do men and women rate an app? | Star Ratings | Star ratings are one of the most important factors for app developers. To increase the star ratings, developers are more likely to address user reviews with lower stars first (Noei et al. 2018). |
| How do users react to the user review associated with men and women? | Likes | Users can like user reviews and Google Play Store reports the total number of likes received for each user review. |
| How do men and women express their feedback? | Sentiment Scores | The star ratings do not reflect all aspects of the sentiment of user reviews. We apply sentiment analysis on user reviews using the SentiStrength-SE tool (Islam and Zibran 2017) to measure the sentiment scores. |
| What is the length of user reviews associated with men and women? | Number of Words and Sentences | The length of a user review can reflect the helpfulness and the amount of information in a feedback (Kim et al. 2006). We use Stanford parser (De Marneffe et al. 2006) to count the number of words and sentences in user reviews. |
| How does the Google Play Store rank user reviews of men and women? | Rank | The Google Play Store uses its own criteria to rank each user review. We capture the ranks of user reviews (i.e., the order of user reviews sorted and displayed by Google Play Store). |
| To what degree do developers respond to user reviews of men and women? | Developer Response | Developers can respond to each user review on the Google Play Store. We capture whether developers respond to each of the user review in our dataset or not. |

## 3 Research Questions and Results

In this section, for each research question, first, we explain our motivations. Then, we describe our approach and findings.

### 3.1 RQ1) Are there the same number of reviews posted by men and women?

#### 3.1.1 Motivation

Recent studies suggest low participation of women in online communities and platforms such as Stack Overflow (May et al. 2019). The ratio of men and women participation in such communities is different. May et al. (2019) state that while women ask more questions on Stack Overflow, men appear to be more active (e.g., answer more questions and vote for

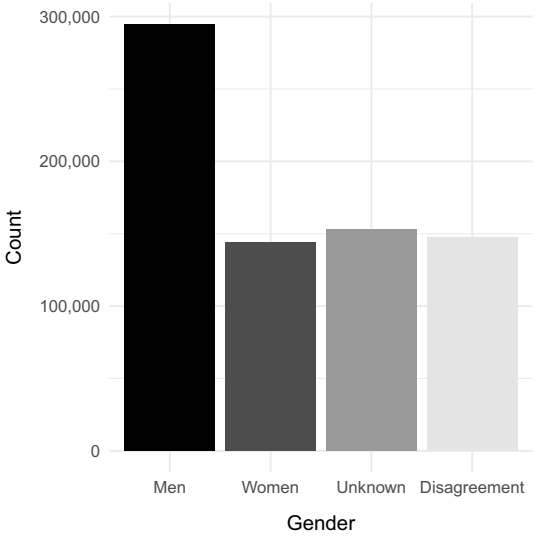


**Fig. 2** Number of user reviews by men and women. The last two columns show the number of user reviews by names that could not be labeled and number of user reviews where Genderize and Genni do not agree on the label

more answers). Burger et al. (2011) reported that 48% of the Tweets[5] in their study belong to women while only 33% belong to men. In contrast, Wachs et al. (2017) find that men post more work on Dribbble[6], an online platform for designers. In this research, we are interested to investigate women's and men's participation in the Google Play Store.

### 3.1.2 Approach

As explained in Section 2.2, we determine the gender of each user with a precision of 94%. Using the identified genders, we compare the number of user reviews associated with women and men to find out which gender generates more user reviews.

### 3.1.3 Findings

Figure 2 shows the number of user reviews by men and women. The first and second columns report the number of user reviews associated with men and women, respectively. We only use the user reviews that are labeled as men or women in this study and eliminate those that could not be labeled or for which agreement was not reached (May et al. 2019; Wachs et al. 2017). The third column reports the number of user reviews that are associated with names that could not be labeled as man nor woman. The last column shows the number of user reviews where Genderize (Genderize 2019) and Genni (Smith et al. 2013) could not agree on the label associated with the name.

As illustrated in Fig. 2, based on the user reviews for which we are able to infer gender, men have posted 2.05 times more user reviews than women with 294,650 user reviews by men and 144,057 user reviews by women. As we discuss in the third research question, women tend to give higher star ratings when leaving feedback. At the same time, developers

[5] http://www.twitter.com/
[6] http://www.dribbble.com/

**Table 3** List of top ten apps with more user reviews from women, ordered by the ratio of user reviews by women versus men

| App Name | Category | #Reviews | | Ratio |
|---|---|---|---|---|
| | | Men | Women | W/M |
| The Wonder Weeks | Health & Fitness | 65 | 509 | 7.83 |
| Wattpad: Where stories live. | Books & Reference | 114 | 491 | 4.31 |
| Fruits Master | Puzzle | 119 | 389 | 3.27 |
| Global Star Live app V LIVE | Entertainment | 167 | 542 | 3.24 |
| Word Crossy - A crossword game | Word | 995 | 2,718 | 2.73 |
| Wayfair - Shop All Things Home | Shopping | 477 | 1,275 | 2.67 |
| Pinterest | Social | 3,112 | 8,228 | 2.64 |
| Candy Crush Friends Saga | Casual | 1,115 | 2,823 | 2.53 |
| Moodpath | Medical | 296 | 716 | 2.42 |
| Candy Crush Soda Saga | Casual | 355 | 849 | 2.39 |

tend to respond to user reviews with lower star ratings first (Noei et al. 2018). Juxtaposing the two, receiving lower attention from developers may have discouraged women and resulted in a lower number of user reviews from women. Indeed our empirical study cannot claim a reason and, therefore, future studies should shed more light on this matter. However, our findings suggest that more fundamental remedies are required to encourage women to participate more in the Google Play Store. Thus, as users' feedback play a major role in the development and evolution of apps, women can play a stronger role in shaping the next versions of their favorite apps.

Although overall, among the 156 apps that are studied in this paper, we find 39 apps (i.e. 25% of apps) with more user reviews from women. Table 3 lists the top ten apps with more user reviews from women and Table 4 lists the top ten apps with more user reviews from men, with the app category, number of user reviews by men and women, and the ratio of user reviews by women and men. As shown in Table 3, *The Wonder Weeks*[7] is the app with the highest participation ratio from women which is a calendar app that explains the development of a baby. As in Table 4, the top app for men, in terms of participation ratio, is *StbEmu*[8] which belongs to the category of *Video Players & Editors*.

Our study covers 38 categories of Android apps. To find out about each category, please visit Google category guidelines (Google 2020). Despite the fact that we only cover one to two apps in some of the categories (which is not representative enough to draw a conclusion about a category), our observations denote a potential difference in the participation of men and women. Table 5 shows a complete list of categories in this study sorted by the ratio of user reviews by women versus men. As shown in Table 5, the category of *Books & Reference* holds the highest ratio of women participation, followed by the categories of *Puzzle Games*, *Medical*, and *Educational*. On the other hand, the category of *Racing Games* receives its majority of feedback from men, followed by the categories of *Personalization*, *Simulation*, and *Sports*.

[7] https://play.google.com/store/apps/details?id=org.twisevictory.apps
[8] https://play.google.com/store/apps/details?id=com.mvas.stb.emu.pro

**Table 4** List of top ten apps with more user reviews from men, ordered by the ratio of user reviews by men versus women

| App Name | Category | #Reviews | | Ratio |
|---|---|---|---|---|
| | | Men | Women | M/W |
| StbEmu (Pro) | Video Players & Editors | 47 | 2 | 23.50 |
| Iron Marines | Strategy | 338 | 15 | 22.53 |
| 365Scores | Sports | 2,033 | 100 | 20.33 |
| Grindr - Gay chat | Social | 2,275 | 116 | 19.61 |
| Torque Pro (OBD 2 & Car) | Communication | 428 | 23 | 18.61 |
| UFC | Sports | 207 | 16 | 12.94 |
| Tasker | Tools | 233 | 19 | 12.26 |
| Navigation Pro | Maps & Navigation | 354 | 29 | 12.21 |
| Nova Launcher Prime | Personalization | 2,324 | 212 | 10.96 |
| Nitro Nation Drag & Drift | Racing | 609 | 56 | 10.88 |

Similar to our findings, a report on the types of apps and app categories more frequented by men and women concluded that there is a 50-50 split in gender ratio for the popular apps (Bonnington 2013). They also found, similar to our findings, that men download sports- and automotive-related apps more than women do and women are more likely to download catalog apps, health & fitness-related apps, and trivia, puzzle, and education or family-friendly games (Bonnington 2013).

*After excluding user reviews with names that are labeled as unknown or disagreement, in total, there are more user reviews posted by men than women on the Google Play Store. The only exception is for the categories of Books & Reference, Puzzle, Medical, Learning, Word, Music, and Casual where the number of user reviews posted by women is greater than those by men.*

### 3.2 RQ2) Do men's and women's user reviews share similar topics?

#### 3.2.1 Motivation

In this research question, we investigate the topics of reviews by men and women so that we can further understand the differences between user reviews, e.g., do women report more bugs than men?

#### 3.2.2 Approach

We use topic modeling to identify and quantify the topics associated with reviews by men and women. To this end, we apply the Latent Dirichlet Allocation (LDA) technique (Blei et al. 2003). LDA allows sets of observations to be described by unobserved groups of data, and, therefore, it determines the similar parts accordingly (Blei et al. 2003). LDA has been applied in previous studies on user reviews (Iacob and Harrison 2013; Guzman and Maalej 2014).

**Table 5** List of categories with number of apps in each category and number of user reviews by men and women, sorted by the ratio of user reviews from women versus men

| Category | #Apps | #Reviews | | Ratio |
|---|---|---|---|---|
| | | Men | Women | W/M |
| Books & Reference | 1 | 114 | 491 | 4.31 |
| Puzzle | 1 | 119 | 389 | 3.27 |
| Medical | 1 | 296 | 716 | 2.42 |
| Educational | 1 | 81 | 179 | 2.21 |
| Word | 3 | 2,591 | 4,823 | 1.86 |
| Music | 1 | 400 | 636 | 1.59 |
| Casual | 8 | 9,911 | 10,145 | 1.02 |
| Card | 2 | 5,639 | 5,386 | 0.96 |
| Shopping | 4 | 8,321 | 7,066 | 0.85 |
| Social | 6 | 12,842 | 10,326 | 0.80 |
| Education | 4 | 1,008 | 807 | 0.80 |
| Photography | 9 | 9,117 | 7,119 | 0.78 |
| Food & Drink | 5 | 12,714 | 9,811 | 0.77 |
| Entertainment | 11 | 22,360 | 16,804 | 0.75 |
| Health & Fitness | 9 | 9,348 | 6,476 | 0.69 |
| Travel & Local | 3 | 7,962 | 4,846 | 0.61 |
| Arcade | 8 | 5,017 | 2,853 | 0.57 |
| Finance | 2 | 2,049 | 1,095 | 0.53 |
| Board | 1 | 1,125 | 599 | 0.53 |
| Weather | 2 | 948 | 487 | 0.51 |
| Communication | 13 | 45,686 | 18,542 | 0.41 |
| Productivity | 5 | 12,861 | 5,089 | 0.40 |
| Adventure | 2 | 1,787 | 587 | 0.33 |
| Lifestyle | 4 | 15,535 | 4,706 | 0.30 |
| Maps & Navigation | 2 | 4,190 | 1,253 | 0.30 |
| Action | 5 | 6,887 | 1,924 | 0.28 |
| Tools | 5 | 5,831 | 1,598 | 0.27 |
| Music & Audio | 5 | 16,904 | 4,505 | 0.27 |
| Dating | 1 | 1,119 | 298 | 0.27 |
| Strategy | 6 | 8,876 | 2,173 | 0.24 |
| News & Magazines | 5 | 14,695 | 3,554 | 0.24 |
| Auto & Vehicles | 1 | 11,060 | 2,599 | 0.24 |
| Video Players & Editors | 5 | 15,917 | 3,406 | 0.21 |
| Role Playing | 1 | 1,194 | 182 | 0.15 |
| Sports | 10 | 15,816 | 2,146 | 0.14 |
| Simulation | 2 | 894 | 110 | 0.12 |
| Personalization | 2 | 2,828 | 275 | 0.10 |
| Racing | 1 | 609 | 56 | 0.09 |

In order to build a corpus of user reviews, we take a similar approach as (Noei et al. 2019b) when preprocessing user reviews. First, we correct typos and informal terms in user reviews as user reviews usually suffer from typos and typos can impact the results of text analysis techniques (Nord 2005). Similar to Noei et al. (2018, 2019b), we fix typos using Jazzy Spell Checker (Jazzy 2017) with a dictionary of 645,289 English words. We then resolve synonyms in user reviews. Third, we fix negations in user reviews as negations can disturb text processing techniques (Villarroel et al. 2016). For example, we replace *"not good"* with *"bad"* in our user reviews using the Stanford natural language processing toolkit (Manning et al. 2014). Fourth, we remove the punctuation and stop words (Rajaraman and Ullman JD 2012), such as *will* and *that*. Finally, we stem the words (Lovins 1968). For instance, *commented* and *commenting* have the same word stem which is *comment*.

We configure LDA parameters (see (Blei et al. 2003)) with respect to the distribution of data (Panichella et al. 2013). More details about the LDA and its parameters can be found at Blei et al. (2003). To identify the optimum number of topics, we apply the approach proposed by Deveaud et al. (2014) that estimates the number of latent concepts by maximizing the information divergence between all pairs of topics of LDA. The optimum number of topics for our corpus of user reviews is 10 where overall dissimilarity between topics achieves its maximum value. We run LDA with 2,000 Gibbs sampling iterations (Raftery et al. 1992).

### 3.2.3 Findings

Table 6 shows the results of the topic modeling. The second column in this table shows the most frequent words of each topic and the third and fourth columns denote the ratio of men and women that discuss each topic, respectively. The last column indicates the adjusted p-values comparing the proportion of user reviews by men and women for each topic. We identified ten major topics as follows:

– *Bug Reports.* Users report bugs and glitches they encounter while using an app. For example: *"Trying to access messages is very glitchy. I can browse ads fine but pulling up old messages is filled with bugs... constantly says network problem when there is actually no issues."*

**Table 6** Topics of user reviews with the proportion of men and women participated in each topic

| Topic | Key Terms | Proportion of | | Adjusted |
|---|---|---|---|---|
| | | Men | Women | p-value |
| Bug Report | error, save, crash, time, long | 0.22 | 0.21 | 4.05e-21 |
| Device & Connectivity | phone, google, call, connect, android | 0.26 | 0.20 | 0.00e-00 |
| Game | game, play, fun, level, addict | 0.27 | **0.29** | 8.17e-21 |
| Learning | help, interest, feel, idea, learn | 0.23 | **0.28** | 1.38e-249 |
| Multimedia | add, download, movie, song, quality | 0.23 | 0.22 | 1.81e-14 |
| Notifications & Messages | email, text, open, load, notification | 0.23 | 0.20 | 3.43e-107 |
| Purchase & Order | buy, service, time, money, place | 0.23 | **0.24** | 3.67e-19 |
| Review | review, star, rate, full, people | 0.21 | 0.17 | 3.66e-131 |
| Social | day, account, friend, user, number | 0.22 | 0.20 | 2.36e-32 |
| Speed | better, browser, experience, fast, super | 0.24 | 0.19 | 7.70e-288 |
| Video & Music | video, watch, music, screen, player | 0.23 | 0.20 | 3.02e-134 |

- *Device & Connectivity.* Users share their experience when using an app on different devices. Their major issue seems to be calls and connections on their devices. For example: *"Using ver. 7.1.2 on my Google Pixel 2 XL, Android 8.1. The app forced on me several times. I uninstalled and reinstalled the app twice, but the result was the same. The most disappointing issue is that it absolutely does not work with Android Auto."*
- *Game.* Users share their experience running a game. They usually talk about how addictive a game or discuss different levels of a game. For example: *"I love this game!!! Best and most addictive and addicting game I've ever played!!! I would so recommend this game!!! This is a well designed game and fun for sure!!!"*
- *Learning.* Users talk about learning aspects of an app, such as how helpful it is. For example: *"So far it has been very helpful and I am learning more about simple meditation no matter where I am."*
- *Multimedia.* Users write about their experience of media consumption on an app, such as its quality and user friendliness. For example: *"...buffers or pauses every 10 seconds.both through Chromecast and the Android TV app. i am on a 300mbps internet service... please help!"*
- *Notifications & Messages.* User reviews within this topic are more about text messages, emails, and notifications received by the apps and the issues around messages and notifications. For example: *"Takes an excessively long time to load anything. ...and notifications are on even if you turn them off unless you try off all notifications."*
- *Purchase & Order.* User reviews in this topic include user complaints or reports when purchasing a service or paid version of an app. It also includes user reviews about the apps that helps them in buying and selling online. For example: *"Don't buy. There are too many new launchers run by developers who are excited to hear from the users about issues and improvements."*
- *Review.* Users speak about the reviews of an app. A very common issue within this topic is asking for a feature request or bug fix so that they improve their current given star rating and feedback. For example: *"Won't modify the pics on google photos. 'Modify Original' option does not help. Earlier it used to create another copy now it won't. Kindly fix it and I will give 5 stars (which is actually this apps worth)."*
- *Social.* Users explain social aspects of an apps, their profiles, and how it connects them with other people and friends. For example: *"Good app to find new friends with similar interests."*
- *Speed.* Users share their experience on how speedy an app is for fulfilling their requirements. For example: *"super slow and bad quality even when you have a very fast internet connection."*
- *Video & Music.* User reviews within this topic include user experiences and reports about watching video or listening to music. For example: *"I used to listen all my music with ok app. Now it doesn't play anymore in background. You need a subscription for that."*.

The third and fourth columns of Table 6 show the proportion of men participating in each topic versus the proportion of women. For three topics of *Learning*, *Game*, and *Purchase & Order* (reported with a bold font in the table), the ratio of women is more than men. Although both men and women share similar topics of user reviews, we investigate their proportion of contribution to each topic using $\chi^2$ test (Bolboacă et al. 2011; Ugoni and Walker 1995). First, we form a contingency table for each topic. Each contingency table contains the number of user reviews that discuss a topic versus other topics for men and women. Then, we apply $\chi^2$ test to each contingency table. We adjust the p-values that

are derived from each test (i.e., one test per topic) using Benjamini-Yekutieli method (Benjamini and Yekutieli 2001). All the adjusted p-values (as listed in the last column of Table 6) indicate significant differences between the proportion of topics that are covered by men versus women.

Table 5 shows the number of user reviews in each category of apps in our dataset and Table 6 shows the topics of user reviews that are shared between men and women. By crosschecking the two tables, we observe that in the categories of *Books & References*, *Puzzle*, *Medical*, *Educational*, *Word*, and *Music*, the number of user reviews that are posted by women are more than men for all the identified topics of user reviews. Having more user reviews by women in the aforementioned categories is not an unusual observation as it is in line with our observations in Table 5. In addition to the six aforementioned categories, we observe more user reviews by women in the categories of *Educational* and *Social* discussing the topic of *Learning*, in the category of *Shopping* discussing *Multimedia*, and in the categories of *Card* and *Casual* talking about *Game*.

> *Despite similar topics of user reviews between men and women, there are significant differences in their proportion of contribution to each topic. For example, 28% of user reviews by women are about learning while 23% of user reviews by men are about learning.*

### 3.3 RQ3) What is the relationship between gender and star rating?

#### 3.3.1 Motivation

Users usually rely on star ratings for choosing an app to download (Bavota et al. 2015). (Harman et al. 2012) reported a statistically significant relationship between the number of downloads and star ratings. In this research question, we are interested to check whether gender relates to star ratings, and if it does, whether men or women provide higher star ratings and sentiment scores. This is important as developers tend to address reviews with a low number of stars and more negative user reviews first (Noei et al. 2018).

#### 3.3.2 Approach

Recent studies (Martin et al. 2016; Noei and Lyons 2019) identified important metrics that are statistically significantly related to star rating such as sentiment scores and length of user reviews. We build a linear regression model (Weisberg 2005) using star ratings of each user review and the aforementioned metrics as control variables. We also include the number of likes received for each user review, ranks of user reviews by the Google Play Store, and developers' response to user reviews (see Section 2.3).

After building a linear regression model using the metrics listed in Table 2, we add gender as a new metric to the model. Then, by applying the ANOVA test (Pinheiro et al. 2007), we compare the two models: (i) the one with gender as a metric, and (ii) the one without gender as a metric. A p-value $\leq 0.05$ denotes a statistically significant difference between the two models where gender contributes statistically significantly to the explanatory power of the model.

We identify the correlated metrics because the inclusion of correlated metrics negatively affects the stability of the linear regression model and, therefore, prevents us from understanding the full impact of each metric (Harrell 2001). We apply variable clustering analysis (Hmisc 2020) to build a hierarchy of the correlated explanatory metrics. For each metric sub-hierarchy with correlation $|\rho| > 0.7$ (Nguyen et al. 2010), we select only one metric for inclusion in the model. Figure 3 shows the hierarchical clustering of metrics according to Spearman's $|\rho|$. In Fig. 3, the horizontal line shows the threshold at which the metrics are considered highly correlated. As shown in Fig. 3, the numbers of words and sentences are highly correlated in addition to positive sentiment scores and overall sentiment scores. We pick the number of words and overall sentiment scores for inclusion in our model.

### 3.3.3 Findings

By applying the ANOVA test (Pinheiro et al. 2007), we compare the two models: (i) *Base Model*: all the metrics except gender included with 438,701 degrees of freedom and (ii) *Full Model*: all the metrics plus gender are included as explanatory metrics (with star rating as the target variable) with 438,700 degrees of freedom and $R^2 = 0.39$. The p-value associated with the ANOVA test is $< 2.2e^{-16}$ which denotes a statistically significant difference between the two models. In other words, the addition of gender to the model significantly increases its explanatory power.

Table 7 shows the base regression model and full model. In this table, the significant metrics have $P(> |t|)$ less than 0.05. $P(> |t|)$ is the p-value associated with the t-statistic (McClave and Sincich 2006). The third column in Table 7 shows the coefficients of each metric and the fourth column shows the standard error.

Gender is a categorical metric indicating *man* or *woman*. As in Table 7, gender as a man shares a statistically significantly *negative* relationship with star ratings while gender as a woman share a statistically significantly *positive* relationship. Women tend to give higher stars to apps with an average star rating of 3.94 in comparison with men who give an average star rating of 3.70. We plug median values for all the variables into our model while flipping the gender. Switching the gender from *man* to *woman* increases the star rating (i.e., dependent variable) by 0.10. Moreover, given that we study the top apps on the Google Play Store, 49% of star ratings given by women users are associated with full five stars while only 39% of men have given five stars to the studied apps.

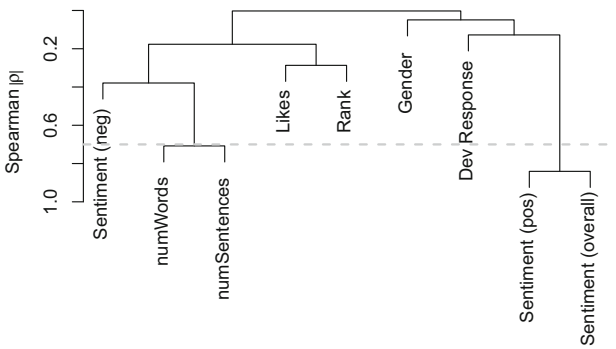


**Fig. 3** Hierarchical clustering of metrics according to Spearman's $|\rho|$

**Table 7** Linear regression model with star ratings as the target variable

| | Metric | Estimate | Std. Error | **P**(> \|**t**\|) |
|---|---|---|---|---|
| Base Model | Developer Response (True) | −8.58e-01 | 5.72e-03 | < 2e-16 |
| | Number of Likes | 6.24e-05 | 8.77e-06 | 1.16e-12 |
| | Number of Words | −2.47e-02 | 1.25e-04 | < 2e-16 |
| | Rank | 5.10e-06 | 4.36e-07 | < 2e-16 |
| | Sentiment Score (Negative) | −1.51e-01 | 3.61e-03 | < 2e-16 |
| | Sentiment Score (Overall) | 6.23e-01 | 2.10e-03 | < 2e-16 |
| Full Model | Developer Response (True) | −8.52e-01 | 5.72e-03 | < 2e-16 |
| | **Gender (Man)** | **−9.47e-02** | **3.99e-03** | **< 2e-16** |
| | Number of Likes | 6.18e-05 | 8.76e-06 | 1.84e-12 |
| | Number of Words | −2.47e-02 | 1.25e-04 | < 2e-16 |
| | Rank | 5.58e-06 | 4.37e-07 | < 2e-16 |
| | Sentiment Score (Negative) | −1.48e-01 | 3.61e-03 | < 2e-16 |
| | Sentiment Score (Overall) | 6.18e-01 | 2.11e-03 | < 2e-16 |

Furthermore, sentiment scores of user reviews by women are associated with higher scores (0.9 on average) than user reviews by men (0.6 on average). The negative sentiment scores of both men and women are similar (1.43 and 1.40 on average for men and women, respectively); however, user reviews by women hold higher positive sentiments, 2.37, compared to 2.08 for men. This indicates that women tend to use positive terms, such as *great* and *excellent*, more often.

*Gender shares a statistically significant relationship with star ratings. Despite a lower ratio of participation by women, women tend to rate apps with more stars and leave reviews with more positive sentiment scores in comparison to men.*

### 3.4 RQ4) How different is the response to men's and women's reviews?

#### 3.4.1 Motivation

On most online platforms and communities, such as GitHub, users are explicitly associated with their user profiles which typically include a name, profile picture, links to their websites, and other information about them. However, on the Google Play Store, users do not own a page or profile of their own and the only information visible to other users is their profile picture and name. In the previous research question, we observed that women tend to rate apps with higher stars and post user reviews with higher sentiment scores than men. In this research question, we study users' and developers' responses to reviews posted by men and women. The outcome of this research question can help the software development community understand gender in the mobile app development community and on app stores. For example, the Google Play Store can use gender as an additional metric in their rankings and developers can consider gender in responding to reviews.

### 3.4.2 Approach

In the last research question, we study stars and sentiment scores and their relationship with gender. In this research question, we study the remaining metrics (see Table 2) and analyze the differences between reviews by men and women.

### 3.4.3 Findings

**Developer Response** Developers' response rate is different for men and women. Overall, developers have responded to 40,227 user reviews by men (14% of men's reviews) and only to 14,669 user reviews by women (10% of women's reviews). We apply $\chi^2$ test (Bolboacà et al. 2011; Ugoni and Walker 1995) on a contingency table given gender on one side and developers' response on the other side. The $\chi^2$ test confirms a statistically significant difference in developers' response with a p-value $< 2.2e-16$. Our study does not reveal the reason behind this difference in response rates; however, this can be due to the statistically significant negative relationship of developers' response with star ratings (see Table 7) which is in line with findings revealed by recent studies (Noei et al. 2018) that developers tend to address more negative user reviews first.

**Number of Likes** As shown in Table 7, the number of likes shares a statistically significant relationship with star ratings. Users of the Google Play Store can like user reviews of other users. In our dataset, user reviews by women received more likes (11.1 likes on average) while user reviews by men received 8.8 likes on average. There might be different reasons for this observation. For example, we find that women tend to give higher star ratings to apps and star ratings correlate with the number of likes (see Table 7). Also, in a study by Hong et al. (2017), they found that women receive more likes than men on Facebook[9] while men use comment and like functionalities more often on Facebook. Also, Ford et al. (2017) observed that women are more likely to engage with a post on Stack Overflow if they see other women involved in the conversation. However, further research is required to fully understand the reason behind the higher number of likes for reviews by women.

**Rank** The Google Play Store ranks user reviews and presents them based on their rank. As users rarely check beyond the first ten results of search engines (McMillan et al. 2013), we study the top ten user reviews of each app and compare ranks of women versus men user reviews. Figure 4 shows kernel density estimates of the appearance of men and women user reviews in the top ten ranked user reviews for each app. As shown in Fig. 4, user reviews posted by women tend to get higher ranks from the Google Play Store than men. This observation is in line with our earlier finding that user reviews by women tend to receive more likes from users.

**Number of Words and Sentences** The length of user reviews in terms of the number of words and sentences may impact users' likes and developers' responses. For example, Asiri and Chang (2018) report that participants of their study do not read all user reviews especially the longer ones. However, longer user reviews might provide more useful information (Vasa et al. 2012). We observe that the length of user reviews in terms of the number of words and sentences is fairly similar for men and women. User reviews by both men and

[9] http://www.facebook.com/

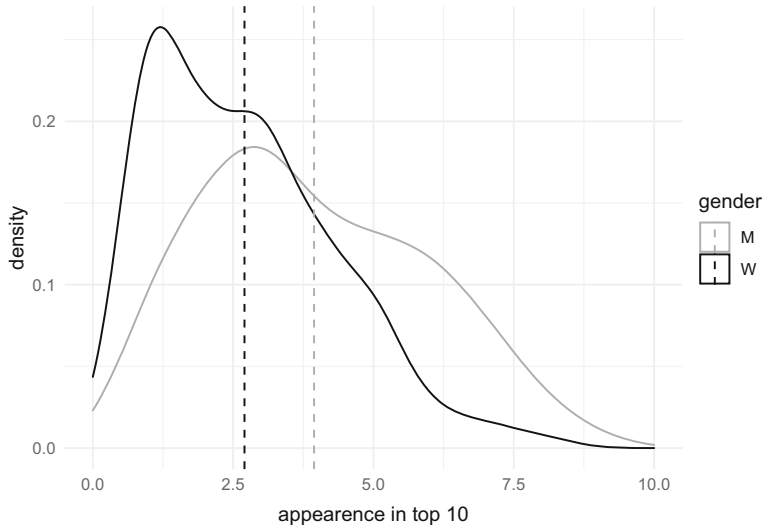


**Fig. 4** Kernel density estimates of the appearance of men and women user reviews among top ten ranked user reviews for each app

women are around two sentences and 18 words long on average. In contrast, the lengths of men's and women's feedback are different in some other communities. For example, Otterbacher (2010) reported shorter movie reviews by women yet richer reviews in vocabulary (Hemphill and Otterbacher 2012).

Future research should shed more light on such similarity and also into other characteristics of user reviews by men and women, such as linguistic patterns (Karimi et al. 2016; Fink et al. 2012; Otterbacher 2010), that can be related to the differences in the number of likes and responses for user reviews of men and women. Lakoff (1973) and Lakoff (1975) states that women use hedges, such as *kind of* and *sort of*, more than men to soften their arguments. However, in our dataset, we observed similar use of hedging by both men and women on the Google Play Store with 22% and 23% for men and women, respectively. Argamon et al. (2003) found that women use a higher rate of pronouns. Similar to Otterbacher (2010), we count the ratio of pronoun use in user reviews by men and women. As shown in Table 8, both men and women in our study use pronouns similarly; however, women tend to use 1$^{st}$ person singular pronouns, such as *I* and *me*, more frequently than men. Foltz et al. (1999) provides a basic measure of writing complexity that has been used for automatic essay scoring. Otterbacher (2010) reported that men's reviews have more complex words and sentences than women's. We also investigated user review complexity on our dataset. We observe that men and women tend to produce user reviews with similar complexity: (i) 8.98 and 9.08 words per sentence for men and women, respectively, and (ii) 5.35 and 5.27 characters per word for men and women, respectively. The similarity of linguistic patterns for both men and women on the Google Play Store can occur because of various reasons such as (i) the unstructured format of user reviews, (ii) using a small mobile device to post user reviews, or (iii) the hasty nature of using such platforms so that users can post feedback on the go, such as on a train or bus.

**Table 8** Proportion of usage of pronouns in user reviews of men and women

| Pronoun | Proportion of | |
|---|---|---|
| | Men | Women |
| 1st person singular | 0.369 | 0.484 |
| 2nd person singular | 0.149 | 0.148 |
| 3rd person singular, men | 0.004 | 0.005 |
| 3rd person singular, women | 0.002 | 0.004 |
| 3rd person singular, neutral | 0.284 | 0.331 |
| 1st person plural | 0.025 | 0.028 |
| 2nd person plural | 0.148 | 0.148 |
| 3rd person plural | 0.058 | 0.065 |

*User reviews by women tend to be liked more by other users and they are among the higher-ranked user reviews by the Google. However, user reviews by women receive a lower number of responses from developers.*

### 3.4.4 Discussion

In this section, we explore two additional scenarios in which we assume that *unknowns* and *disagreements* (see Section 2.2) are either posted only by men or women. The goal of such additional experiments is to test whether our observations still hold even with such strict and unlikely assumptions. The two scenarios are as followed:

*Scenario$_1$:* We assume that user reviews that are either labeled as unknown or disagreement are posted by *men*.

*Scenario$_2$:* We assume that user reviews that are either labeled as unknown or disagreement are posted by *women*.

In terms of the number of reviews posted by men and women, given the two scenarios and the results that have been demonstrated in Fig. 2, applying Scenario$_1$ results in more reviews by women, and vice versa for Scenario$_2$. This observation was expected since 41% of the initial data were labeled as either unknown or disagreement.

However, when we repeat the remaining experiments, we observe results that are in line with the results of the original experiments for each research question. Table 9 shows the proportion of the topics of user reviews that are posted by men or are labeled as unknown or disagreement (i.e., Scenario$_1$) versus the proportion of the topics of user reviews that are posted by women or are labeled as unknown or disagreement (i.e., Scenario$_2$). For example, for the topic of *Bug Reports*, Scenario$_1$ shows the proportion of user reviews among all the user reviews that are marked as man, unknown, or disagreement is 0.22. By comparing Tables 6 and 9, we find that even by applying each of the scenarios, the results of the second research question remain similar. Moreover, similar to the results of the second research question, applying $\chi^2$ test implies significant differences between the proportion of men and women covering each topic.

Finally, we build two linear models after applying each scenario. Table 10 shows both linear models. The relation of each variable to star ratings and their significance is similar to the original model (see Table 7) with a slightly lower $R^2$ value for both models (0.38). The metrics of user reviews that are labeled as unknown or disagreement come right in between

**Table 9** Topics of user reviews with the proportion of men and women participated in each topic given $Scenario_1$ (unknown and disagreements considered as men) and $Scenario_2$ (unknown and disagreements considered as women)

| Topic | $Scenario_1$ | $Scenario_2$ |
|---|---|---|
| Bug Report | 0.22 | 0.22 |
| Device & Connectivity | 0.24 | 0.21 |
| Game | 0.28 | **0.29** |
| Learning | 0.25 | **0.27** |
| Multimedia | 0.23 | 0.23 |
| Notifications & Messages | 0.23 | 0.22 |
| Purchase & Order | 0.23 | **0.23** |
| Review | 0.20 | 0.19 |
| Social | 0.22 | 0.22 |
| Speed | 0.23 | 0.21 |
| Video & Music | 0.23 | 0.22 |

men and women; for example, women give 3.9 stars on average and men 3.7 stars. The average of stars associated with unknown and disagreement user reviews is 3.8.

### 3.5 Implications for Research and App Development

Related work suggests different rates of participation by women in online platforms; for example, lower participation on Stack Overflow (May et al. 2019) but higher on Twitter (Burger et al. 2011). In this study, we observe that there are some differences and similarities between user reviews posted by men and women. One of the similarities is the topics of user reviews despite the differences in the ratio of contributions between men and women. We identified ten major topics, such as *bug reports* and *device & connectivity*.

**Table 10** Linear regression models for $Scenario_1$ and $Scenario_2$ with star ratings as the target variable

| | Metric | Estimate | Std. Error | $P(>\|t\|)$ |
|---|---|---|---|---|
| $Scenario_1$ | Developer Response (True) | −8.24e-01 | 4.33e-03 | < 2e-16 |
| | **Gender (Man)** | **−6.80e-02** | **3.65e-03** | **< 2e-16** |
| | Number of Likes | 6.24e-05 | 8.77e-06 | < 2e-16 |
| | Number of Words | −2.42e-02 | 9.84e-05 | < 2e-16 |
| | Rank | 6.18e-06 | 3.38e-07 | < 2e-16 |
| | Sentiment Score (Negative) | −1.40e-01 | 2.83e-03 | < 2e-16 |
| | Sentiment Score (Overall) | 6.21e-01 | 1.64e-03 | < 2e-16 |
| $Scenario_2$ | Developer Response (True) | −8.27e-01 | 4.33e-03 | < 2e-16 |
| | **Gender (Man)** | **−6.67e-02** | **2.94e-03** | **< 2e-16** |
| | Number of Likes | 4.93e-05 | 5.89e-06 | < 2e-16 |
| | Number of Words | −2.41e-02 | 9.84e-05 | < 2e-16 |
| | Rank | 6.04e-06 | 3.38e-07 | < 2e-16 |
| | Sentiment Score (Negative) | −1.40e-01 | 2.83e-03 | < 2e-16 |
| | Sentiment Score (Overall) | 6.21e-01 | 1.64e-03 | < 2e-16 |

Topics of user reviews can provide app developers with valuable information when maintaining their apps (Noei et al. 2019b). When studying topics of user reviews, developers and researchers should keep in mind that, despite lower star ratings from men, both men and women contribute to the topics. Therefore, user reviews should not be prioritized solely based on the associated star ratings.

The second similarity is the length of user reviews. Both men and women, on average, post user reviews with two sentences and 18 words. Therefore, gender classification based on the length of user reviews seems to be less relevant than findings from other studies such as Mukherjee and Bala (2017) and Otterbacher (2010) that use text length to identify genders in other communities.

In spite of the similarities, we observed various differences as well. Besides the number of user reviews, women tend to associate their user reviews with higher star ratings and with more positive content. This can lead to attracting less attention from developers as developers are more likely to address problems in user reviews that are associated with lower star ratings first (Noei et al. 2018). This could also result in a feedback loop where women's reviews are not responded to as often as men's, further decreasing their participation in the app review process. Our findings indicate that developers should take gender into consideration when responding to reviews and the Google should consider including gender as an additional factor when considering user feedback in their app rankings.

Last but not least, user reviews by women are more likely to be liked by other users. Also, user reviews by women are usually ranked higher in comparison to those by men. This is important for app developers as users usually read the first few user reviews when deciding to download and install an app on their devices. Women's reviews are potentially having a greater influence on users but attracting less attention from developers. Developers will want to make sure that reviews from both women and men are responded to equitably.

## 4 Related Work

In this section, we summarize the related work along with two research directions: (i) user reviews on app stores and (ii) gender studies in different communities.

### 4.1 User Reviews on App Stores

Star ratings, users' feedback, and app ranks are important elements of success on the Google Play Store (Martin et al. 2016). Harman et al. (2012) observed that there is a statistically significant relationship between the number of downloads and star ratings of apps and Kim et al. (2011b) discussed how star ratings can impact users' decisions when downloading an app. Therefore, many studies investigate various metrics that are related to star ratings (Tian et al. 2015; Ruiz et al. 2016; Bakar and Mahmud 2013; Linares-Vásquez et al. 2013; Bavota et al. 2015); however, Kübler et al. (2018) discussed that app popularity and users' sensitivity to factors like ratings and price can be different in different countries. Linares-Vásquez et al. (2013) found that low-quality APIs can affect star ratings of mobile apps. Bavota et al. (2015) showed that the user-perceived quality of an app shares a relationship with fault- and change-proneness of the APIs that are used by the app. To the best of our knowledge, none of the recent work has empirically investigated the relationship between star ratings and gender using a large dataset from the Google Play Store. In 2013, Baek (2013) surveyed 191 participants (57% male) in Korea. They reported no significant relationship between their participants' gender and their attitudes (positive or negative) towards apps. However, unlike

their study, based on the empirical study presented in this paper, we observe that women tend to leave feedback with higher stars and more positive sentiment scores in comparison to men.

Given the importance of user reviews, some studies propose solutions to summarize user reviews (Galvis Carreño and Winbladh 2013; Guzman and Maalej 2014; Iacob and Harrison 2013). For example, (Panichella et al. 2016) applied natural language processing, text analysis, and sentiment analysis to classify user reviews into five different intentions of information giving, information seeking, feature requests, problem discovery, and others. Villarroel et al. (2016) proposed an approach to classify user reviews into bug reports and feature requests. Khalid et al. (2014b) manually identified a set of user complaints such as app crashing, compatibility, and feature requests and showed their relationship with star ratings. In this paper, we observe that user reviews by men are addressed more by developers while user reviews by women receive more likes from users. Understanding such differences can help developers in identifying important user reviews rather than focusing on, for example, more negative ones.

### 4.2 Gender Studies

We discuss gender studies in different communities in this section.

#### 4.2.1 Software Development Communities

Terrell et al. (2017) studied GitHub to understand gender differences in this software development platform. They observed that women's contributions, such as pull requests, are more likely to be accepted than men's. However, for contributors who are from outside a project and their gender is identifiable by developers, men's contributions tend to be accepted more often. Ford et al. (2017) studied Stack Overflow and reported that despite the lower participation of women, women are more likely to engage with a post that is contributed to by other women. May et al. (2019) also studied gender differences in Stack Overflow. They reported that men post more answers than women to the questions, and, also, men tend to be more rewarded with an increased reputation for their answers (May et al. 2019). Vasilescu et al. (2014) reported that 55% of users in their study present themselves as men, while only 7% as women on Stack Overflow. They also investigated gender participation in two web content management systems, including Drupal[10] and WordPress[11]. They reported that active participation for women in Drupal and WordPress mailing lists is only 10% and both content management systems have more men users than women users. Unlike earlier work, we reveal gender participation and differences in users of the Google Play Store which have interesting implications for both app developers and researchers.

#### 4.2.2 Other Communities

Some of the earlier research targets gender studies in other types of online communities, such as communities for reviewing movies. Wachs et al. (2017) studied the Dribbble online platform to answer and explore the factors behind success in an online design community. They reported that men post more of their work on Dribbble, and, therefore, can demonstrate their work to a larger audience. Wachs et al. (2017) also found that men receive

[10] http://www.drupal.org/
[11] http://www.wordpress.com/

more likes on Dribbble and skills listed by men on their profiles (such as product management), share a statistically significant relationship with the number of received likes, while, women's skills, such as calligraphy and social media, share a negative relationship. Hannák et al. (2017) studied bias in online freelance marketplaces, including TaskRabbit[12] and Fiverr[13]. They reported that being a woman shares a statistically significant negative relationship with the number of reviews received for an individual. Wikipedia, which is an online platform that allows everyone to participate, suffers from a gender gap with fewer than 15% women contributors (Collier and Bear 2012). Wagner et al. (2016) studied gender in the content posted on Wikipedia showing that gender-related topics are more present in biographies about women, while abstract terms tend to be used to describe positive aspects in the biographies of men and negative aspects in the biographies of women.

## 5 Threats to Validity

In this section, we discuss potential threats to the validity of this study (Shull et al. 2007).

### 5.1 Construct Validity

Martin et al. (2015) states that using an incomplete set of user reviews can introduce bias to the findings of empirical studies on app markets. To mitigate this threat, we retrieved all the user reviews of our apps over six months and excluded the apps with an incomplete set of user reviews. In the manual evaluation of the gender inference approach, we employ three evaluators from North America. However, for example, a name that is mostly used for men in North America may commonly be used for women in another part of the world. Nonetheless, we rely on Genderize and Genni to infer genders for this study.

### 5.2 Conclusion Validity

In order to increase the accuracy of identifying genders, we use both Genni and Genderize tools. Although this results in the exclusion of some of the data from this study, we could apply our analyses on a large number of user reviews (438,707 user reviews) with a precision of 94%. It is also possible that women hide their gender by using gender-neutral names in online environments or that men adopt female personas online so that other users will behave less aggressively to them (Vasilescu et al. 2014). Those reviews by users who use gender-neutral names or non-name pseudonyms would fall into the unknown category or possibly would result in disagreement among the tools that we used to infer gender. As other research has done (May et al. 2019; Wachs et al. 2017), we focused on those users for whom the tools agreed on the gender category of the name so that we could be more sure of the gender of the users in our study.

### 5.3 External Validity

For app developers who are not interested in star ratings and users' feedback, the result of our study may not be useful. However, it is still an opportunity for them to understand similarities and differences between men's and women's participation in the Google Play

[12] https://www.taskrabbit.com/
[13] https://www.fiverr.com/

Store. The apps and reviews that are used in this paper are based on the data provided by the Google to their Canadian audience. Findings may vary in some locations and countries around the globe. Future studies should investigate the differences and similarities in different geographical locations.

### 5.4 Internal Validity

We study nine metrics in addition to gender to explain star ratings. We do not claim an exhaustive list of explanatory metrics as the goal of our study is to understand the relationship between gender and user feedback.

## 6 Conclusion

User reviews are an important information source that developers can use in order to improve their star ratings. Given, various backgrounds and genders of Android users, we study men's and women's participation in user reviews that are posted on the Google Play Store. We analyze 438,707 user reviews of the top 156 Android apps over six months. We achieve a precision of 94% when inferring genders associated with user reviews. We find that, despite similar topics of user reviews between men and women, there are significant differences in the ratio of men's and women's contributions to each topic. Our findings suggest that remedies and solutions are required to encourage women to participate more in the Google Play Store as user reviews shape developers' direction in the changes they make in future releases. We also build a linear regression model with star ratings as the target variable. We observe that gender shares a statistically significant relationship with star ratings, where being a man shares a negative relationship while being a woman shares a positive relationship. Women's user reviews are more likely to be liked by other users and they appear higher in user reviews ranks, i.e., are among the user reviews that users see first. Our findings suggest that developers should keep gender in mind when addressing user reviews as considering only star ratings or sentiment scores may cause them to miss user reviews that are posted by women. By responding less to women's reviews than men's, developers risk further discouraging women's participation in the app review process. Women's user reviews are important because they are more popular and are ranked higher by the Google which means that they are more likely to influence users. Finally, we highlight similarities, such as similarity in length, and differences, such as different star ratings and ranks, between men's and women's user reviews. Given the statistically significant relationship between gender and star ratings, future research and solutions should consider gender when studying user reviews. Furthermore, future research should investigate men's and women's motivations when posting reviews.

## References

Argamon S, Koppel M, Fine J, Shimoni AR (2003) Gender, genre, and writing style in formal written texts. Text-The Hague Then Amsterdam Then Berlin 23(3):321–346

Asiri O, Chang CK (2018) Investigating users' experiences and attitudes towards mobile apps' reviews. In: International Conference on Human-Computer Interaction. Springer, pp 481–499

Baek Y (2013) Analysis of user's attitude toward apps, intention to use and continual consuming intention-focused on mobile commerce. Int J Content 9(4):35–44

Bakar NSAA, Mahmud I (2013) Empirical analysis of android apps permissions. In: Proceedings of the 2013 International Conference on Advanced Computer Science Applications and Technologies (ACSAT). IEEE, pp 406–411
Basili VR (1992) Software modeling and measurement: the goal/question/metric paradigm. Technical report, Institute for advanced computer studies
Bavota G, Linares-Vasquez M, Bernal-Cardenas CE, Penta MD, Oliveto R, Poshyvanyk D (2015) The impact of api change-and fault-proneness on the user ratings of android apps. IEEE Trans Softw Eng 41(4):384–407
Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann stat:1165–1188
Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
Bolboacă SD, Jäntschi L, Sestraś AF, Sestraś RE, Pamfil DC (2011) Pearson-fisher chi-square statistic revisited. Information 2(3):528–545
Bonnington C (2013) Are men and women using mobile apps differently? do men and women have appreciably different tastes in apps? [Online]. Available: https://www.wired.com/2013/04/men-women-app-usage/
Burger JD, Henderson J, Kim G, Zarrella G (2011) Discriminating gender on twitter. In: Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics, pp 1301–1309
Collier B, Bear J (2012) Conflict, criticism, or confidence: an empirical examination of the gender gap in wikipedia contributions. In: Proceedings of the ACM 2012 conference on computer supported cooperative work, pp 383–392
De Marneffe MC, MacCartney B, Manning CD et al (2006) Generating typed dependency parses from phrase structure parses. In: 5Th international conference on language resources and evaluation, vol 6, pp 449–454
Deveaud R, SanJuan E, Bellot P (2014) Accurate and effective latent concept modeling for ad hoc information retrieval. Doc Numér 17(1):61–84
Fink C, Kopecky J, Morawski M (2012) Inferring gender from the content of tweets: a region specific example. In: Sixth International AAAI Conference on Weblogs and Social Media
Foltz PW, Laham D, Landauer TK (1999) Automated essay scoring: Applications to educational technology. In: EdMedia+ innovate learning, Association for the Advancement of Computing in Education (AACE), pp 939–944
Ford D, Harkins A, Parnin C (2017) Someone Like me: How does peer parity influence participation of women on stack overflow? In: 2017 IEEE Symposium on visual languages and human-centric computing (VL/HCC). IEEE, pp 239–243
Fu B, Lin J, Li L, Faloutsos C, Hong J, Sadeh N (2013) Why people hate your app: Making sense of user feedback in a mobile app store. In: 19th International Conference on Knowledge Discovery and Data Mining. ACM, pp 1276–1284
Galvis Carreño LV, Winbladh K (2013) Analysis of user comments: an approach for software requirements evolution. In: 35th International Conference on Software Engineering. IEEE, pp 582–591
Genderize (2019) Genderize. [Online]. Available: http://www.genderize.io/
Google (2020) Google play store categories. [Online]. Available: https://support.google.com/googleplay/android-developer/answer/113475
Guzman E, Maalej W (2014) How do users like this feature? a fine grained sentiment analysis of app reviews. In: 22nd International Conference on Requirements Engineering. IEEE, pp 153–162
Hannák A, Wagner C, Garcia D, Mislove A, Strohmaier M, Wilson C (2017) Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, pp 1914–1933
Harman M, Jia Y, Zhang Y (2012) App store mining and analysis: Msr for app stores. In: 9th International Conference on Mining Software Repositories, MSR '12. IEEE, Piscataway, pp 108–111
Harrell FE (2001) Regression modeling strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer
Hemphill L, Otterbacher J (2012) Learning the lingo? gender, prestige and linguistic adaptation in review communities. In: Proceedings of the ACM 2012 conference on computer supported cooperative work, pp 305–314
Hmisc (2020) Harrell miscellaneous. [Online]. Available: http://cran.r-project.org/web/packages/Hmisc/index.html
Hong C, Chen ZF, Li C (2017) "liking" and being "liked": How are personality traits and demographics associated with giving and receiving "likes" on facebook? Comput Human Behav 68:292–299

Iacob C, Harrison R (2013) Retrieving and analyzing mobile apps feature requests from online reviews. In: 10th Working Conference on Mining Software Repositories, MSR '13. IEEE, pp 41–44
Ibrahim H, Abdel-Razig S, Stadler DJ, Cofrancesco J, Archuleta S (2019) Assessment of gender equity among invited speakers and award recipients at us annual medical education conferences. JAMA Netw Open 2(11):e1916222–e1916222
Islam MR, Zibran MF (2017) Leveraging automated sentiment analysis in software engineering. In: 14th International Conference on Mining Software Repositories. IEEE Press, pp 203–214
Jazzy (2017) Jazzy spell checker. [Online]. Available: http://jazzy.sourceforge.net/
Karimi F, Wagner C, Lemmerich F, Jadidi M, Strohmaier M (2016) Inferring gender from names on the web: A comparative evaluation of gender detection methods. In: Proceedings of the 25th International Conference Companion on World Wide Web, pp 53–54
Khalid H, Nagappan M, Shihab E, Hassan AE (2014a) Prioritizing the devices to test your app on: A case study of android game apps. In: 22nd International Symposium on the Foundations of Software Engineering, pp 370–379
Khalid H, Shihab E, Nagappan M, Hassan AE (2014b) What do mobile app users complain about? IEEE Softw 32(3):70–77
Kim HW, Lee H, Son J (2011a) An exploratory study on the determinants of smartphone app purchase. In: 11th International DSI and the 16th APDSI Joint Meeting
Kim HW, Lee H, Son J (2011b) An exploratory study on the determinants of smartphone app purchase. In: Proceedings of the 11th International Decision Science Institute and the 16th Asia Pacific Decision Sciences Institute Joint Meeting
Kim SM, Pantel P, Chklovski T, Pennacchiotti M (2006) Automatically assessing review helpfulness. In: 2006 Conference on empirical methods in natural language processing, Association for Computational Linguistics, pp 423–430
Kübler R, Pauwels K, Yildirim G, Fandrich T (2018) App popularity: Where in the world are consumers most sensitive to price and user ratings? J Mark 82(5):20–44
Lakoff G (1975) Hedges: A study in meaning criteria and the logic of fuzzy concepts. In: Contemporary research in philosophical logic and linguistic semantics. Springer, pp 221–271
Lakoff R (1973) Language and woman's place. Lang Soc 2(1):45–79
Linares-Vásquez M, Bavota G, Bernal-Cárdenas C, Di Penta M, Oliveto R, Poshyvanyk D (2013) Api change and fault proneness: A threat to the success of android apps. In: 9th Joint Meeting on Foundations of Software Engineering. ACM, pp 477–487
Lovins JB (1968) Development of a stemming algorithm. MIT Information Processing Group, Electronic Systems Laboratory
Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D (2014) The stanford corenlp natural language processing toolkit. In: 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp 55–60
Martin W, Harman M, Jia Y, Sarro F, Zhang Y (2015) The app sampling problem for app store mining. In: 12th Working Conference on Mining Software Repositories. IEEE, pp 123–133
Martin W, Sarro F, Jia Y, Zhang Y, Harman M (2016) A survey of app store analysis for software engineering. IEEE Trans Softw Eng PP(99)
May A, Wachs J, Hannák A (2019) Gender differences in participation and reward on stack overflow. Empir Softw Eng 24(4):1997–2019
McClave JT, Sincich T (2006) Statistics: Technology manual [and CD-ROM]. Pearson, Prentice Hall
McMillan C, Poshyvanyk D, Grechanik M, Xie Q, Fu C (2013) Portfolio: Searching for relevant functions and their usages in millions of lines of code. ACM Trans Softw Eng Methodol (TOSEM) 22(4):1–30
Merchant A, Shah D, Bhatia GS, Ghosh A, Kumaraguru P (2019) Signals matter: understanding popularity and impact of users on stack overflow. In: The World Wide Web Conference, pp 3086–3092
Mukherjee S, Bala PK (2017) Gender classification of microblog text based on authorial style. IseB 15(1):117–138
Nguyen TH, Adams B, Hassan AE (2010) Studying the impact of dependency network measures on software quality. In: Proceedings of the 26th International Conference on Software Maintenance. IEEE, pp 1–10
Niu H, Keivanloo I, Zou Y (2017) Learning to rank code examples for code search engines. Empir Softw Eng 22(1):259–291
Noei E (2018) Succeeding in mobile application markets (from development point of view). PhD thesis, Queen's University, Canada
Noei E, Lyons K (2019) A survey of utilizing user-reviews posted on Google Play Store. In: Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering, pp 54–63

Noei E, Da Costa DA, Zou Y (2018) Winning the app production rally. In: 26Th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering, ESEC/FSE. ACM, New York, pp 283–294

Noei E, Zhang F, Wang S, Zou Y (2019a) Towards prioritizing user-related issue reports of mobile applications. Empir Softw Eng 24(4):1964–1996

Noei E, Zhang F, Zou Y (2019b) Too many user-reviews, what should app developers look at first? IEEE Transactions on Software Engineering

Nord C (2005) Text analysis in translation: Theory, methodology, and didactic application of a model for translation-oriented text analysis, pp 94. Rodopi

Otterbacher J (2010) Inferring gender of movie reviewers: exploiting writing style, content and metadata. In: Proceedings of the 19th ACM international conference on Information and knowledge management, pp 369–378

Palomba F, Linares-Vásquez M, Bavota G, Oliveto R, Di Penta M, Poshyvanyk D, De Lucia A (2015) User reviews matter! tracking crowdsourced reviews to support evolution of successful apps. In: 31st International Conference on Software Maintenance and Evolution. IEEE, pp 291–300

Panichella A, Dit B, Oliveto R, Di Penta M, Poshynanyk D, De Lucia A (2013) How to effectively use topic models for software engineering tasks? an approach based on genetic algorithms. In: 2013 35th International Conference on Software Engineering (ICSE). IEEE, pp 522–531

Panichella S, Di Sorbo A, Guzman E, Visaggio CA, Canfora G, Gall HC (2016) Ardoc: app reviews development oriented classifier. In: Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering. ACM, pp 1023–1027

Pinheiro J, Bates D, DebRoy S, Sarkar D et al (2007) Linear and nonlinear mixed effects models. R package version 3:57

Raftery AE, Lewis S et al (1992) How many iterations in the gibbs sampler. Bayesian Stat 4(2):763–773

Rajaraman A, Ullman JD (2012) Mining of massive datasets, vol 77. Cambridge University Press, Cambridge

Ruiz IJM, Nagappan M, Adams B, Berger T, Dienst S, Hassan AE (2016) Examining the rating system used in mobile-app stores. IEEE Softw 33(6):86–92

Scheuerman MK, Spiel K, Haimson OL, Hamidi F, Branham SM (2019) Hci guidelines for gender equity and inclusivity

Shull F, Singer J, Sjøberg DI (2007) Guide to Advanced Empirical Software Engineering. Springer, New York

Smith BN, Singh M, Torvik VI (2013) A search engine approach to estimating temporal changes in gender orientation of first names. In: Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries, pp 199–208

Statista (2020) Number of apps available in leading app stores as of march 2017. [Online]. Available: https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/

Stephens M (2013) Gender and the geoweb: divisions in the production of user-generated cartographic information. GeoJournal 78(6):981–996

Terrell J, Kofink A, Middleton J, Rainear C, Murphy-Hill E, Parnin C, Stallings J (2017) Gender differences and bias in open source: Pull request acceptance of women versus men. PeerJ Comput Sci 3:e111

Tian Y, Nagappan M, Lo D, Hassan AE (2015) What are the characteristics of high-rated apps? a case study on free android applications. In: Proceedings of the 2015 IEEE International Conference on Software Maintenance and Evolution (ICSME). IEEE, pp 301–310

Topaz CM, Sen S (2016) Gender representation on journal editorial boards in the mathematical sciences. PLoS One 11(8):e0161357

Ugoni A, Walker BF (1995) The chi square test: an introduction. COMSIG Rev 4(3):61

Van Solingen R, Basili V, Caldiera G, Rombach HD (2002) Goal question metric (gqm) approach. Encyclopedia of software engineering

Vasa R, Hoon L, Mouzakis K, Noguchi A (2012) A preliminary analysis of mobile app user reviews. In: Proceedings of the 24th Australian Computer-Human Interaction Conference, pp 241–244

Vasilescu B, Capiluppi A, Serebrenik A (2014) Gender, representation and online participation: a quantitative study. Interact Comput 26(5):488–511

Villarroel L, Bavota G, Russo B, Oliveto R, Di Penta M (2016) Release planning of mobile apps based on user reviews. In: 38th International Conference on Software Engineering. ACM, pp 14–24

Wachs J, Hannak A, Vörös A, Daróczy B (2017) Why do men get more attention? exploring factors behind success in an online design community. In: Eleventh International AAAI Conference on Web and Social Media

Wagner C, Graells-Garrido E, Garcia D, Menczer F (2016) Women through the glass ceiling: gender asymmetries in wikipedia. EPJ Data Sci 5(1):5
Wais K (2016) Gender Prediction Methods Based on First Names with genderizeR. R J 8(1):17–37
Weisberg S (2005) Applied linear regression, vol 528. Wiley



**Dr. Ehsan Noei** has a PhD degree from Queen's University. He is currently an Instructor and a Postdoctoral Fellow at the University of Toronto. His primary research interests include software engineering, data science and machine learning, software project management, and release engineering.

**Dr. Kelly Lyons** is a Professor in the Faculty of Information at the University of Toronto with a cross appointment to the Department of Computer Science. Prior to joining the Faculty of Information, she was the Program Director of the IBM Toronto Lab Centre for Advanced Studies (CAS). Her current research interests include service science, knowledge mobilization, data science, social media, collaborative work, and software engineering. From 2015 to 2020, Kelly served as Associate Dean, Academic in the Faculty of Information. From 2020 to 2021, Kelly is serving as the Dean's Advisor on Pandemic Planning and Response. Kelly has co-authored several papers, served on program committees for conferences, given many keynote and invited presentations, and co-chaired several workshops. She has been the recipient of an NSERC Strategic Partnership Grant, NSERC Discovery Grants, an NSERC Collaborative Research and Development Grant with SAP, two NSERC Engage Grants (with Sciencescape and Dell), MITACS Accelerate Grants (with CA, IBM, and Cerebri AI), a SSHRC Knowledge Synthesis Grant, two University of Toronto/UCL research grants, an IBM Smarter Planet Faculty Innovation Grant, an IBM Advanced Studies grant, and has received funding through the GRAND Networks of Centres of Excellence (NCE). Kelly is an IBM Faculty Fellow and a Faculty Affiliate of the Schwartz Reisman Institute for Technology and Society. She is currently on the Board of CS-Can/Info-Can and on the Board of the Informs Service Science Section. From 2008 to 2012, she was a Member- at-Large of the ACM Council and a member of the Executive Council of ACM-W.