User Feedback in the App Store: A Cross-Cultural Study

Emitza Guzman, Luís Oliveira, Yves Steiner, Laura C. Wagner, Martin Glinz
Department of Informatics, University of Zurich, Switzerland
guzman@ifi.uzh.ch,luis.piheiroolivera@uzh.ch,yves.steiner2@uzh.ch,laura.wagner@uzh.ch,glinz@ifi.uzh.ch

ABSTRACT

App stores allow globally distributed users to submit user feedback, in the form of user reviews, about the apps they download. Previous research has found that many of these reviews contain valuable information for software evolution, such as bug reports or feature requests, and has designed approaches for automatically extracting this information. However, the diversity of the feedback submitted by users from diverse cultural backgrounds and the consequences this diversity might imply have not been studied so far.

In this paper, we report on a cross-cultural study where we investigated cultural differences in app store reviews and identified correlations to cultural dimensions taken from a well-established cultural model. We analyzed 2,560 app reviews written by users from eight countries with diverse national culture. We contribute evidence about the influence of cultural factors on characteristics of app reviews. Our results also help developers of automated feedback analysis tools to avoid cultural bias when choosing their algorithms and the data for training and validating them.

CCS CONCEPTS

• **Social and professional topics → Cultural characteristics**; • **Human-centered computing →** *User studies*; • **Software and its engineering →** *Software evolution*;

KEYWORDS

User Feedback, Feedback Analysis, Culture, Algorithm Bias, Software Evolution

ACM Reference Format:
Emitza Guzman, Luís Oliveira, Yves Steiner, Laura C. Wagner, Martin Glinz. 2018. User Feedback in the App Store: A Cross-Cultural Study. In *ICSE-SEIS'18: 40th International Conference on Software Engineering: Software Track, May 27-June 3, 2018, Gothenburg, Sweden.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3183428.3183436

1 INTRODUCTION

App stores are software distribution platforms that allow users to search, buy and download software for mobile devices, as well as to provide feedback about their satisfaction and experience with apps through reviews.

Popular app stores such as the Apple App Store and Google Play host millions of apps [2], [39], have a presence in over 150 countries [1] and are accessed by a global audience of users who download their apps and write reviews on a regular basis.

Previous studies [27], [35] showed that app store reviews include information that is useful to analysts and app designers, such as user requirements and bug reports. This feedback represents a "voice of the users" and can be used to drive the development effort and improve forthcoming releases.

Existing work, e.g., [6], [8], [17] has addressed the large number of reviews received by popular apps, their unstructured nature and varying quality by proposing to automatically analyze the feedback with data mining techniques—reducing developers' and analysts' effort when analyzing the reviews.

Users who provide feedback through app store reviews are distributed all over the world, and it is highly probable that the culture in which users are rooted influences the way they provide feedback. However, this phenomenon has not been studied so far by the software engineering research community.

Studying the influence of cultural differences on feedback provided in app reviews is not just motivated by the general scientific goal of backing believed or probable phenomena with data. There is also a strong practical motivation concerning the automatic analysis of user feedback: So far, possible cultural differences have not been accounted for when designing algorithms for automatic feedback analysis. This omission may result in *algorithm bias*.

Such bias occurs when algorithms for feedback analysis are trained and evaluated on data sets that are not representative of the users' diverse demographic and cultural backgrounds. For example, algorithms for classifying user reviews into different categories (e.g., [14], [30], [38]) could be more apt to correctly categorize reviews from users having the same or a similar cultural context as the reviews with which the algorithms were trained and evaluated. Ranking algorithms prioritizing user feedback (e.g., [6],[16],[43]) could, for instance, favor reviews written by users from cultures that tend to a more verbose or emotional language, or those that tend to faster reactions when giving feedback. This could lead to a *selected user empowerment*—in which feedback by users from particular cultural contexts has a stronger influence on the evolution of an app than feedback by other users.

To our best knowledge, the effects of cultural diversity on the automatic analysis of user reviews have not been studied yet. Moreover, there is little understanding to what extent user reviews differ from one country to another and which differences should be considered when designing algorithms for analyzing user feedback.

In this work we take a step towards this direction and study differences among culturally diverse countries when giving feedback about software applications. We selected a representative sample of 2,560 reviews from a data set of 59,203 reviews that we collected over a period of two months. The sample size of 2,560 reviews is small enough for the required manual labeling of reviews and

ICSE-SEIS'18, May 27-June 3, 2018, Gothenburg, Sweden
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5661-9/18/05…$15.00
https://doi.org/10.1145/3183428.3183436

large enough for drawing statistically valid conclusions. We studied cultural differences and similarities when giving feedback in terms of sentiment expression, gender representation, timing, rating, content and length by using content analysis techniques and statistical tests. We then identified common patterns by using a well established cultural model [20].

The contribution of this work is twofold. First, we provide data and statistical evidence about the influence of cultural factors on characteristics of app reviews—while previously there were only beliefs that, with rather high probability, cultural differences observed in other fields would also somehow manifest in app reviews. By making our data available[1], we enable replication and encourage further research in this direction. Second, our results shed light on potential algorithm bias due to cultural factors in automatic feedback analysis. While an in-depth analysis of algorithm bias is subject to future research, our results provide initial evidence that algorithm bias due to cultural factors is not just a hypothetical threat.

2 RELATED WORK

2.1 User Feedback and Software Evolution.

Previous research [35] found that user feedback is essential for software quality and for identifying ideas for improvement. Pagano and Maalej [36] and Hoon [25] conducted exploratory studies and analyzed characteristics of user feedback from app stores, such as frequency, length, rating and content. In contrast to the study presented in this work, they focused on feedback available in the United States App Store.

User feedback mining has received a considerable amount of attention in the recent years. Among the most studied platforms for obtaining user feedback are app stores. Martin et al. [31] surveyed the most relevant work in the area. Much of the existing work has focused on the classification, summarization and prioritization of the feedback. Supervised machine learning approaches have often been applied for automatically classifying user feedback e.g., [14], [30], [38]. Topic modeling and clustering algorithms have been used for its summarization [8], [17], [26], [43]. Existing approaches for ranking user feedback use weighted functions e.g., [6] or supervised machine learning e.g., [43]. Approaches have also been proposed for detecting spam [5], retrieving reviews with different opinions about specific features [13] and linking the reviews to source code [37].

2.2 Culture

In the words of Hofstede, "culture is the software of the mind" [21], shaping our perceptions, behavior and attitudes. In this paper we expand on Hofstede's work about finding cultural similarities and differences between countries [20], [22]. During a period of 15 years, Hofstede surveyed 116,000 IBM employees based in 67 countries, resulting in a model measuring cultural differences among countries. While there are several culture models (e.g., [10], [19], [28]) we chose this one as it is the most widely used in software engineering contexts [3]. The Hofstede model consists of six dimensions:

- **Power Distance**: refers to the degree to which members of the country accept and expect that power is distributed unequally.
- **Individualism vs. Collectivism**: indicates the extent to which members of a society are integrated into groups.
- **Masculinity vs. Femininity**: evaluates the differentiation between genders in a society. A society with a high Masculinity will have a stronger differentiation than a society with a low Masculinity.
- **Uncertainty Avoidance**: indicates the extent to which people in a society are resistant to unpredictable and ambiguous situations.
- **Long-term vs. Short-term Time Orientation**: refers to people's tendencies to focus on future or present goals. People in societies with a higher long-term index will consider the future more important than those with a short-term orientation.
- **Indulgence vs. Restraint**: indicates the extent to which a society expresses their wants and impulses. More indulgent societies—those with a higher Indulgence index, tend more to gratification than those with a lower Indulgence index.

The indexes of the different dimensions range between 0 and 100, with 50 as an average score. Results from the Hofstede model should be interpreted so that if an index is under 50, the culture scores relatively low on that dimension and if an index is over 50, the culture scores relatively high on the concerned dimension. It is important to note that a country's specific indexes on the dimensions are relative, i.e., the model can only be used purposefully when making comparisons.

Existing work in the software engineering domain has performed cross-cultural studies for understanding the adoption of agile practices across different cultures [3], the impact of agile practices on reducing sociocultural distances [24], as well as for understanding the challenges encountered by globally distributed software teams [9], [29]. Work in human-computer interaction has found that users from diverse cultural backgrounds have different preferences when interacting with software [34], [41], and has proposed approaches for adapting software to these differences [40]. This work motivates our study as users with different interaction preferences will most probably write feedback from different perspectives, leading to culturally diverse feedback.

3 STUDY

3.1 Scope

The goal of our study is to investigate if user feedback submitted in diverse cultural contexts differs in its characteristics and how such differences relate to cultural dimensions in Hofstede's model (cf. Sect. 2.2). We study six characteristics of app reviews:

- **Sentiment:** The affect present in the user feedback text.
- **Content:** The category of the feedback with respect to software evolution (i.e., bug report, feature request or other).
- **Gender:** The gender of the user submitting the feedback.
- **Rating:** Ordinal scale score (from one to five stars) provided by the users together with the textual feedback to express their satisfaction with the concerned app.

[1] The replication package is available at:
https://www.ifi.uzh.ch/en/rerg/people/guzman/culture.html

- **Timing:** Amount of time that has passed between the submission of the feedback and the last release of the concerned app.
- **Length:** The number of characters or words in the feedback text.

We studied sentiment, content, rating, timing and length characteristics as they have been used as input in approaches for automatic classification and prioritization of user feedback [14], [30], [38], [6], [15]. We also analyzed the gender of users submitting feedback across different cultures. Previous work found that there is a hidden gender bias in the product cycle—software included [44] and we were interested in investigating if such bias can also be observed in the reviews of software users across different national cultures.

We limited the scope of our study to user feedback provided as reviews (written in English) in Apple's App Store[2], one of the largest mobile application distribution platforms, which has a global presence, hosts a variety of applications, and has separate stores for most countries. The last feature is the main reason for choosing this platform for our study: it makes it easy to identify the country of origin for each feedback instance. This is not possible in other popular platforms commonly used in research studies, such as Android's Google Play. We chose to study feedback written in English as it is the most widely used language in computing and a language in which all study authors were fluent. Further, the authors would not have been capable of analyzing reviews written in all native languages of the countries included in our study.

As we only consider user feedback in the form of reviews submitted to app stores in our study, we will use the terms user feedback and review interchangeably for the remainder of this paper.

3.2 Research Method

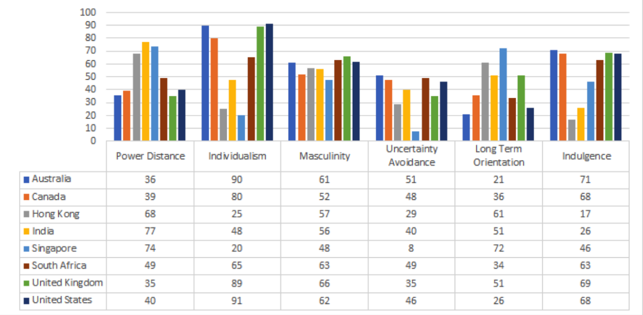
After collecting a raw data set of 59,203 reviews, we selected a representative sample of 2,560 reviews which we then analyzed, partially with automatic processing and partially by manual inspection and labeling. Then we employed statistical tests for identifying significant differences among the feedback given by users from distinct national cultures. The term *national culture* denotes the predominant cultural traits in a country. Finally, we compared the distribution of our observed results to those of Hofstede's culture model [20], [22] (see Section 2.2).

We performed our analysis across all apps, as opposed to per app. The main reason for this is that the per app analysis would have needed a larger manual analysis of data for each of the apps in order to have statistically valid results, and this was not feasible due to limited resources.

The details of data collection and analysis are given in the remainder of this section; the results are reported in Section 4.

3.3 Data Collection

We collected reviews available in Apple's App Store, written in English, from eight countries: Australia, Canada, Hong Kong, India, Singapore, South Africa, United Kingdom and United States. We chose these specific countries because of the differences among them across all six dimensions of national culture present in Hofstede's model [23] (see Figure 1) and because they are representative of different world regions. Further, all of these countries had an App Store presence and use English as an official language, i.e., English is either spoken natively or as a second language in the selected country. This guarantees a certain level of language fluency in the analyzed reviews. Choosing countries from geographically and culturally diverse regions also helps reduce potential bias and improves the validity of our results.



	Power Distance	Individualism	Masculinity	Uncertainty Avoidance	Long Term Orientation	Indulgence
■ Australia	36	90	61	51	21	71
■ Canada	39	80	52	48	36	68
■ Hong Kong	68	25	57	29	61	17
■ India	77	48	56	40	51	26
■ Singapore	74	20	48	8	72	46
■ South Africa	49	65	63	49	34	63
■ United Kingdom	35	89	66	35	51	69
■ United States	40	91	62	46	26	68

Figure 1: Countries in our data set with their Hofstede index in different dimensions [23]. Since each dimension represents specific cultural traits, larger differences in the degree of the dimensions conducts to larger differences in the behaviour of the respective cultures—our selected countries, are thus, culturally diverse under the Hofstede model.

We collected reviews for seven apps from the eight chosen countries. We selected popular apps that were among the most downloaded in the App Store, since their high number of reviews helps ensure the statistical validity of our results. To enable direct comparisons among the chosen countries, we chose apps that were available in the App Stores of all countries. To prevent bias we chose apps belonging to different domains according to the App Store taxonomy (e.g., social media, health and fitness).

In total we collected 59,203 reviews in the period between May 1st and June 30th, 2017 (see Table 1). For this purpose, we modified an openly available scraper[3].

3.4 Sampling

We used the 95% confidence interval by country to compute the minimal sample size to perform a valid statistical analysis. For this purpose, we took the total number of reviews in our data set for each country, and then inferred the respective number of necessary reviews that would need to be present in our sample set to ensure statistical significance of 95%. Furthermore, we used a stratified sampling, assuring that all selected apps were equally represented in the sample.

The resulting sample has a size of 2,560 reviews. We did not choose a larger sample because we needed a size which is still manageable for manual analysis and labeling. Table 2 shows the number of reviews per app and country present in our sample set.

[2] https://www.apple.com/lae/ios/app-store/

[3] https://github.com/grych/App StoreReviews

Table 1: Data set overview per country and app.

App	Domain	Country								
		Australia	Canada	Hong Kong	India	Singapore	South Africa	UK	US	Total
Calorie Counter	Health & Fitness	131	305	6	13	7	12	437	2,576	3,487
Facebook	Social media	846	946	312	848	138	57	1,421	7,724	12,292
Fitbit	Health & Fitness	70	234	10	19	9	16	313	1,499	2,170
Instagram	Photo & Video	970	1,609	270	744	178	118	1,886	13,959	19,734
Pinterest	Social media	211	395	16	149	14	50	474	4,426	5,735
Waze	Navigation	264	617	6	13	57	101	829	10,381	12,268
Whatsapp	Social media	72	109	185	1,069	114	115	532	1,321	3,517
Total		2,564	4,215	805	2,855	517	469	5,892	41,886	59,203

Table 2: Sample and study set overview per country and app.

App	Australia	Canada	Hong Kong	India	Singapore	South Africa	UK	US	Total
Calorie Counter	18	26	3	3	4	6	28	24	112
Facebook	124	90	108	103	63	26	92	73	679
Fitbit	10	20	4	3	5	7	20	15	84
Instagram	126	132	94	91	81	54	116	127	821
Pinterest	28	33	6	19	7	23	30	41	187
Waze	35	51	3	3	27	47	51	94	311
Whatsapp	10	10	65	130	52	53	33	13	366
Total	351	362	283	352	239	216	370	387	2,560

3.5 Automated Analysis

We extracted the rating, timing and length characteristics of the analyzed user feedback by using some of the data available through the collection processing, and in some cases performing additional calculations. We extracted the *rating* score of each review by directly using the rating score obtained during the collection of each individual review. We computed the *timing* as the difference between the date in which the review was written and the release date of the version of the app concerning the review. The review *length* was acquired by counting the number of words and the number of characters of the comment in each review.

3.6 Manual Content Analysis

While the rating, length and timing of reviews can be determined easily with fully automated procedures, the accuracy achievable with existing algorithms for automatic classification of sentiment and content in app store reviews is still not good enough in comparison to what human annotators achieve [14], [17], [30], [38]. In order to achieve truly accurate results, we therefore decided to label the sentiment and content of the reviews manually. As we needed human annotators anyway, and there are no existing approaches for classifying the gender of app store review writers that we are aware of, we decided to also do the gender labeling manually.

Three of the authors used the content analysis techniques described by Neuendorf [33] to identify the gender, content and sentiment in the reviews of our sample and annotated the reviews accordingly. We detail the main steps as follows.

Annotation guide. To systematize our manual analysis, we created an annotation guide with definitions and examples of the content categories and sentiment scales, as well as instructions for the gender assessment. To avoid strong disagreements, we conducted three annotation trials of 50 reviews each. After each trial, the guide was refined to avoid further disagreements.

For the analysis of the *sentiment* expressed in the reviews, the annotators were instructed to use a five-level Likert scale from very positive to very negative for coding the sentiment.

For the analysis of the feedback *content*, we used a simple taxonomy consisting of three categories: "bug report", "feature request" and "other". This taxonomy is inspired by the results of previous work [14], [12], [15],[35], [43] which found that user feedback from app stores and social media contains the aforementioned categories. Annotators could label the reviews as belonging to more than one category (e.g., a review containing both a bug report and a feature request).

Although we had filtered for reviews written in English when creating our data set, it turned out that our sample contained reviews that were partially or fully written in another language. Also, annotators found reviews with illegible characters (e.g., "???!!#$!??!") and reviews that had no clear meaning with respect to the software (e.g., "It's not me!"). All these reviews were marked as "noise" by the annotators.

For the analysis of the *gender* of the users writing the review, annotators chose between "male", "female", "unisex" and "unclear". For this step, annotators were instructed to identify the firstname from the username and then check it with the database genderize.io[4] which contains approximately 216,000 first names used in different countries. This consultation helped reduce annotators' own cultural bias, as gender anticipation based on first names is a subjective task. For example, the name "Andrea" usually has a male connotation in an Italian context and a female connotation in a German context. For each name entered, genderize.io returns the probability of a name being associated to a female or male gender, except when there is no occurrence in the database. Since we wanted to rule out too many false positives, we instructed annotators to only assign "male" or "female" in case the genderize.io database suggested a probability higher than 95%. In case of a 95% probability or lower annotators

[4] https://genderize.io/

assigned the “unisex” label and for names not occurring in the database they assigned the “unclear” label. Further, annotators also considered the inclusion of words indicating specific genders, such as “boy”, “lady”, “woman” or “Mr”; these identifiers were considered conclusive for identifying the gender.

Annotation process. Three co-authors acted as annotators. Each of them independently labeled two thirds of the total 2,560 reviews contained in the sample, so that each review was labeled by two annotators. The annotation was done through a specialized web tool that was developed for the task. The tool displayed the name of the app that the review referred to, the name of the user who wrote the review, the review title and the review text. During this task annotators followed the instructions and recommendations detailed in the annotation guide. Each annotator reported an average of 18 hours for the completion of this step.

Disagreement handling. For all reviews where the annotators did not agree about their labeling, the third annotator (who until this point had not been in involved in the concerned review) solved the disagreement by choosing those labels from the two original annotations that he or she considered most appropriate. In over 79% percent of all reviews, there was no disagreement.

3.7 Statistical Analysis

Test selection. We chose the statistical tests of our study by following the recommendations by McCrum-Gardner [32] which take into account the number of groups being compared (eight in our case), the scale in which the data is presented and its distribution.

Noise removal. In total, the annotators found 152 reviews that they had to label as “noise” for various reasons (see above). To avoid making conclusions about such reviews, we removed them from the sample. So we eventually used 2,560 - 152 = 2,408 reviews for our statistical analysis.

4 RESULTS

Sentiment. Many of the reviews in our studied data have a neutral sentiment, leading to a 0 sentiment median (neutral sentiment) of the the whole analyzed data. Reviews from Australia, Hong Kong, India and Singapore have a median of 0 (neutral sentiment), whereas Canada, United Kingdom and United States have a median of 1 (positive sentiment) and South Africa has a median of 2 (very positive sentiment). Figure 2 shows an overview of the results. The differences among the countries are significant (Kruskal-Wallis, p-value<2.2e-16). In particular, Hong Kong and every other country (p-value<0.01) have a significant difference according to a Tukey-Kramer test[5], as well as India and United States (p-value=0.0005), India and South Africa (p-value=3.9e-5), South Africa and Singapore (p-value=0.0005), South Africa and Australia (p-value=2.1e-5), Australia and United States (p-value=0.0003), and United States and Singapore (p-value=0.006). The sentiment scores in Figure 2 show a slightly similar distribution across countries as the Indulgence and Individualism dimensions, and an inversely similar distribution than the Power Distance dimension (see Figure 1). A lower Power Distance index together with a higher index in Indulgence—the case of the countries in the Anglo group, could suggest that users write more liberally and in a less restrained manner, resulting in reviews with a higher sentiment. The opposite could be true for the Confucian-Asian countries, which tend to be more restrained and with higher Power Distance and lower Indulgence and Individualism.

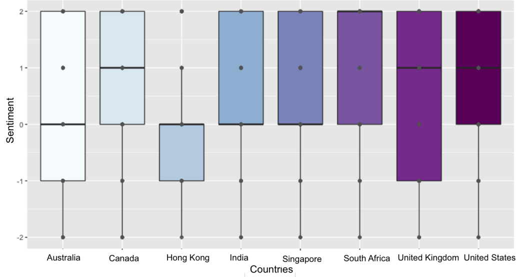


Figure 2: Sentiment scores across the different countries.

Thus, we hypothesize the following:

[H1] *The reviews' sentiment of a specific country positively correlates with the Indulgence of that country and negatively correlates with the Power Distance of the country.*

Using Pearson coefficients, we correlate the sentiment score of each country and its respective Indulgence and Power Distance indexes. We find that users from countries with higher Indulgence (r=0.57, p-value=0.14) tend to give more positive feedback and that users from countries with lower Power Distance tend to give more positive feedback (r=-0.53, p-value=0.18). The correlations, however, are not statistically significant.

Content. To simplify the content analysis process, we assign user reviews with more than one content category to a single one. For this we use the following prioritization: bug report > feature request > other. This results in the following distribution: 24.83% of the reviews contain bug reports, 32.48% include feature requests (but no bug reports) and 42.69% have other type of content only. The country with the highest proportion of reviews that are useful for software evolution (i.e., bug reports and feature requests) is Hong Kong with 77.47%, whereas the country with the lowest proportion is the United States with 45.12%. Figure 3 shows a bar chart of the content categories by country. The proportion of content categories in the analyzed reviews is different among the studied countries. This difference is statistically significant (Chi-square test, p-value<2.2e-16). There are significant differences between Hong Kong and all other countries except Singapore and India (Chi-square test[6], p-value<0.0001), Singapore and South Africa, United Kingdom, United States, Canada (Chi-square test, p-value<0.0001), India and United States, South Africa, United Kingdom (Chi-square test, p-value<=0.0001), as well as Australia and South Africa, United States (Chi-square test, p-value<=0.0021). When comparing the distribution of reviews that are relevant for software evolution (i.e., bug reports and feature requests) among the different countries (see Figure 1) we found that the distribution is similar to the distribution of the Power Distance index among the different countries. Additionally, we also found a slight inverse similarity to the Indulgence and Individualism index distribution. Therefore, we hypothesize the following:

[5] all Tukey-Kramer tests in our work include a Tukey-Dist approximation.

[6] all pairwise comparisons in this step were executed with a Bonferroni correction.

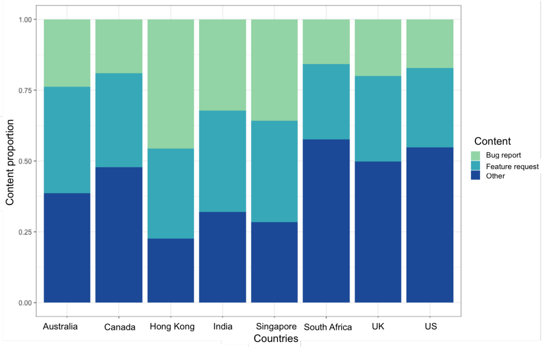


Figure 3: Content categories by country.

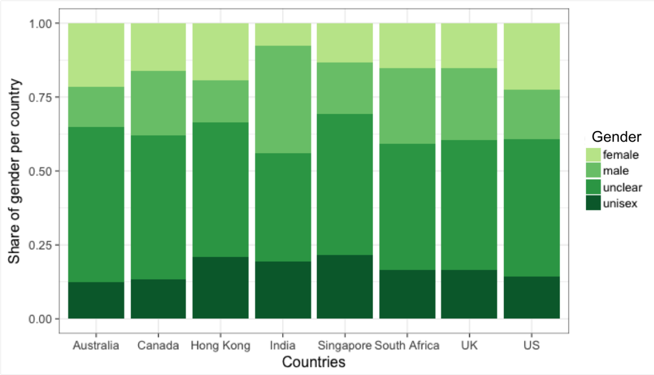


Figure 4: Gender share by country.

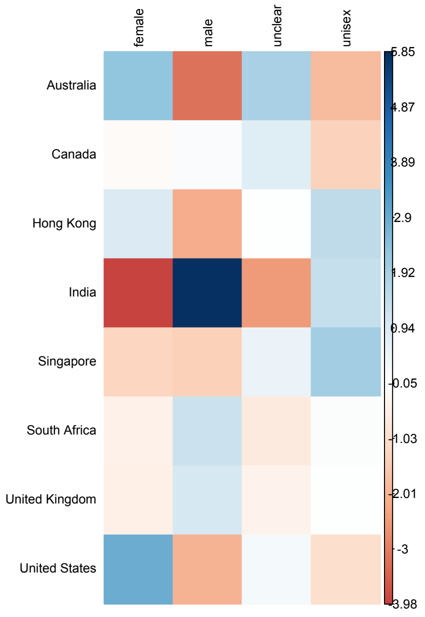


Figure 5: Gender Pearson residuals. Positive associations are shown in blue and negative in red. The intensity of the color is proportional to the strength of the association.

[H2] *The proportion of reviews that is relevant for software evolution for a specific country is positively correlated to the Power Index of that country and negatively correlated to the Individualism and Indulgence.*

A correlation analysis shows that the proportion of reviews that is relevant for software evolution is positively correlated to the Power Distance of the concerned country (r=0.75, p-value<0.03), and negatively correlated to its Indulgence (r=-0.80, p-value<0.02) and Individualism (r=-0.78, p-value<0.03). All three correlations are statistically significant.

Gender. Most of the studied reviews (45.51%) are written by users with an unidentifiable gender (i.e., falling into the "unclear" category in the annotation, see Section 3.6). 21.68% were written by users with male usernames, 16.45% by users with female usernames, and 16.36% by users with unisex names. The gender distribution of the review writers' usernames is different among the countries, with statistical significance (Chi-square test, p-value=4.583e-15). Figure 4 shows the gender distribution of review writers by country. We see two potential reasons for the extraordinarily high number of review writers with an unclear or unisex gender. (1) The tool we used for deciding the gender of the usernames (see Section 3.6) lacks cultural context. (2) Many users do not use genderized names as usernames but rather generic names, such as "Enamul", "chubby_duck", "VeryVeryVeryAnnoyedC" and "Hunnibunni bear".

Due to the abundance of this ambiguous data we did not perform any additional statistical tests, as they would not have any explanatory power. However, we visualized the Pearson residuals of the Chi-square test, shown in Figure 5. The visualization shows that India has a strong positive association to male usernames and a strong disassociation to females. South Africa and United Kingdom are in a similar situation, but with a much weaker strength. Australia and the United States have a strong association to females, while the disassociation to males is strong too. Hong Kong shows a similar tendency, but with a weaker strength. Singapore has a rather strong disassociation to male and female usernames, but a quite high association to unisex usernames. Canada does not have clear associations or disassociations towards any specific gender, but only slightly to unclear and unisex. In short, the abundance of unclear and unisex labels in our data does not allow for any conclusive results with respect to the gender distribution of the usernames. However, our data do show some differing tendencies—with India tending more towards male usernames and Australia and United States tending more towards female names than the other countries.

Rating. Ratings in our data are overall positive, leading to a rating median of 4 ($\bar{x}$=3.25, s=1.73). Reviews from South Africa and United States have a median of 5, while Canada and United Kingdom have a 4 median rating. Australia, India and Singapore have a 3 median rating; Hong Kong has a rating median of 1. Figure 6 shows an overview of the rating variance in our data. The rating differences between reviews from the studied countries is statistically significant (Kruskal-Wallis, p-value<2.2e-16). We found

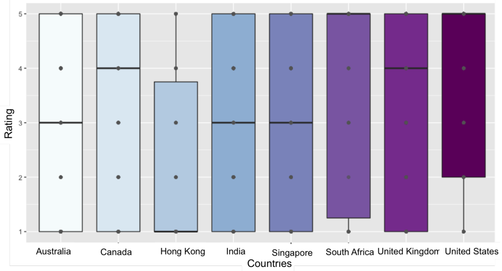


Figure 6: Rating variance across countries.

that the ratings between Hong Kong and all other countries (p-value<0.0007), India and United States (p-value=0.0004), India and South Africa (p-value=0.0007), Singapore and United Kingdom (p-value=0.041), Singapore and South Africa (p-value=0.00017), Singapore and United States (p-value=0.001), Australia and South Africa (p-value=0.00049), and Australia and United States (p-value=0.0002) are different with a statistical significance, as per a Tukey-Kramer test.

The rating scores in Figure 6 show a slightly similar distribution per country as the Indulgence vs. Restraint dimension, and, similarly to the sentiment scores, an inverse similarity to the Power Distance dimension (see Figure 1). A higher indulgent society with a lower power distance could be more willing to rate software in a more positive manner, than a more restrained one. Additionally a society with a lower Power Distance might be more willing to express their dissatisfaction with the reviewed apps. Therefore, we hypothesize that:

[H3] *The rating of the reviews of a specific country positively correlates with the Indulgence of the country and negatively correlates with its Power Distance.*

Through a correlation analysis we find that review rating has a significant positive correlation to the countries’ Indulgence (*r*=0.77, p-value<0.03), while negatively correlating to the countries’ Power Distance, albeit without statistical significance (*r*=-.57, p-value=.14).

Timing. The overall review timing median in our data is 6 days, with users taking a minimum of 0 days after the software has been released to submit feedback and a maximum of 709 days. As Figure 7 suggests, users tend to give less feedback over time. Canada, Hong Kong, Singapore and the United Kingdom are the countries with the most responsive users, with a median of 5 days after the software has been released, whereas South African users are the least quick with a median of 8 days. The review timing among the studied countries is different, and this difference is statistically significant (Kruskal-Wallis, p-value<7.409e-05). However, it is only one country that drastically differs from all others, with the exception of India: South Africa (Tukey-Kramer, p-value<0.03). When comparing to Hofstede’s cultural model along the different dimensions (see Figure 1), we could not find any pattern that could help explain this difference.

Length. The length of the reviews in our data set ranges between 1 and 1,898 characters, and 1 and 335 words, with a median of 17 words and 89 characters. Reviewers from Australia and United Kingdom wrote the longest reviews, whereas Hong Kong has the shortest reviews. The character length difference between the reviews of the studied countries is statistically significant (Kruskal-Wallis, p-value<2.2e-16). Figure 8 shows the length variance across countries. In particular, the character length of reviews from Hong Kong is significantly different to all other countries (p-value<3.7e-8), whereas there are significant differences in review length between India and United States (p-value=0.009), India and Australia (p-value=8.8e-7), India and United Kingdom (p-value=1.6e-5), India and Canada (p-value=0.01), South Africa and Canada (p-value=0.01), South Africa and Australia (p-value=4.5e-16), South Africa and United Kingdom (p-value=5.3e-5), and South Africa and United States (p-value=0.01) according to a Tukey-Kramer test.

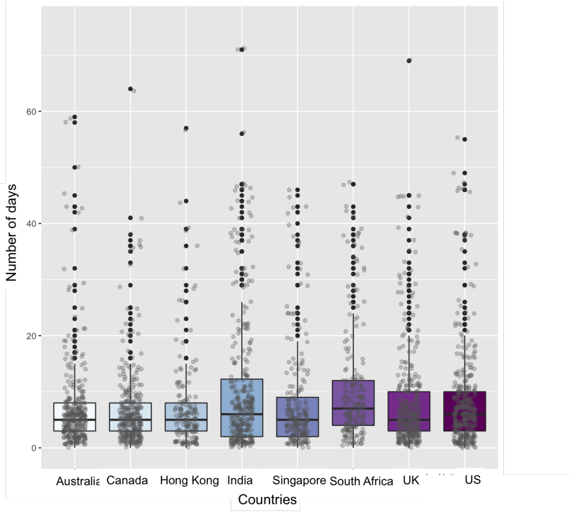


Figure 7: Number of days variance across countries (y-max: 75).

The same differences were observed when analyzing the length in terms of word counts, with a slight difference: Australia and Singapore (p-value=0.03) also differed with a statistically significant difference.

Overall, there is a strong distinction between countries that have English as a native language and those that do not. While comparing the length distribution to the index distribution among the different Hofstede’s dimensions (see Figure 1), we found a slight similarity to the Individualism and Indulgence distribution indexes, as well as an inverse similarity to the Power Distance index. Thus, we made the following hypothesis:

[H4] *The length of the reviews of a specific country is positively correlated to the Individualism and Indulgence of that country, and is negatively correlated to its Power Distance.*

A correlation analysis indicates that users from countries with higher Individualism (*r*=0.71, p-value<0.05), higher Indulgence (*r*=0.80, p-value<0.02) and less Power Distance (*r*=-0.71,p-value<0.04) tend to write lengthier reviews .

5 DISCUSSION

Is user feedback about software applications different across national cultures? Our study shows that the answer is yes. We found

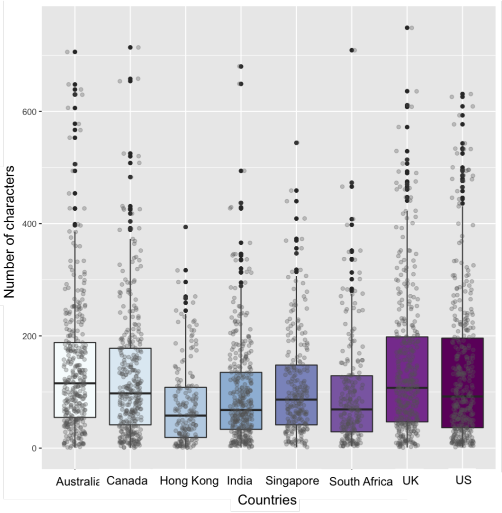


Figure 8: Character length variance across countries (y-max: 750).

that user feedback characteristics such as sentiment, content, rating and length significantly differ at the country level. Further, we also found that these differences follow cultural patterns. When comparing the analyzed data to Hofstede's culture model we found statistically significant correlations between (1) review content and Power Distance, Indulgence and Individualism, (2) review rating and Indulgence, and (3) review length and Power Distance, Individualism and Indulgence.

The significant diversity in terms of sentiment, content, rating and length shown in our results could have implications for both practitioners and researchers developing tools for automatically analyzing user feedback. Failure to account for the diversity in the feedback could result in algorithm bias, leading to the devaluation of feedback that is given by groups of people from regions, countries or cultural contexts that are typically not considered when training and evaluating the algorithms or models. This is a major threat when we recall that most of the existing work in automatic analysis of user feedback has used data for training and validating their approaches exclusively from the United States or from English speaking countries without taking cultural differences into account. We discuss this problem in more detail in the subsequent paragraphs. For our discussion we focus on three analysis problems frequently studied in the literature.

When using supervised machine learning approaches for the *classification* of user feedback e.g., [14], [30], [38], the lack of consideration of diversity in the data used for training and evaluating the algorithms could lead to more frequent misclassifications of reviews from national cultures or groups that were not considered. As a consequence, user feedback from misrepresented groups containing valuable information for software evolution, such as usage scenarios, bug reports or feature requests, could be misclassified as irrelevant and discarded due to using a language with a different cultural context than the one used in the training and evaluation of the model.

A similar problem could occur when *ranking* user feedback using machine learning approaches e.g., [43]. Ranking approaches using weighted functions e.g., [6], [15] are also not exempt from possible bias. These approaches employ user feedback characteristics such as content, length, timing and sentiment to determine the relevance of user feedback for software evolution. If applied indiscriminately on user feedback written by users from different cultures, those that use a more restrained language or that are less emotionally expressive could be disfavored.

The lack of attention in *summarization* approaches could also lead to unwanted biases. For example, using homogeneous data sets and word-embedding—a common approach for summarization [42] could lead to word associations that are not representative of the heterogeneous user audience, a problem already found to be present when summarizing data from other domains [4].

Finally, our findings on significant and potential correlations between user feedback characteristics and different dimensions from Hofstede's model could aid researchers and practitioners when choosing the data with which to train, validate and test their algorithms. Though our results are not conclusive, they hint that there could be similarities between feedback from countries with a similar cultural background. Thus, our results suggest that it might not be necessary to train, validate and test user feedback analysis algorithms on data from all available countries, but only from those that are culturally different.

Algorithm bias has only been recently recognized as a problem in computer science and there is little understanding of the extent of its effects and countermeasures for avoiding it [7]. From our work we can conclude that there are significant differences in user feedback from national cultures. However, our study only provides initial pointers. For determining the actual extent to which these variables affect and bias the current algorithms used for automatically analyzing user feedback, further, in-depth studies will have to be conducted.

6 THREATS TO VALIDITY

There are numerous potential threats to the validity of our study. We present and discuss them as follows.

The notion of culture. A major limitation in our study is the risk of stereotyping individual users based on their country of origin or residence. This study should be considered as a "general behavior" analysis and one should regard the high variability among individual users in the same country.

Furthermore, while Hofstede's model is the most widely used in software engineering contexts [3], Hofstede's definition of culture in terms of nations is rather simplistic [3]—as it assumes a shared and stable national culture that does not change over time or national regions. Future work should explore additional views of culture, in which temporal aspects and conflicting perceptions are also considered.

Causality of correlations. In our four hypotheses (cf. Section 4), we correlated our observations to different dimensions in Hofstede's cultural model. Although some of these observations are statistically significant, they could in fact be caused by other variables, such as geographical, social [11] or economic indicators [11]. Therefore, future work should perform a regression analysis in these cases to better assess the actual impact of the cultural dimensions. Further studies should also analyze other possible affecting variables.

Time frame. It could be argued that our findings are only applicable to the time frame where we collected our data. We mitigated this threat by collecting data over a time span of eight weeks. Considering the short release cycles for most apps (software applications in the app store have a median of 7.2 days between releases [18]), this is long enough to capture several releases for most apps and the various reactions of its users.

Country congruence. In our study we assume that the reviews available in the app store of a specific country are actually written by users residing in that country or having strong links to the concerned country, which is not necessarily the case. However, users submitting feedback to a specific country app store need to provide billing information concerning the chosen country. Thus, we believe that the probability that most users providing reviews on specific country app stores have a link to the concerned country is high.

Relying on human judgment. For the content analysis we rely on human judgment, which introduces subjectivity when determining the sentiment of the text expressed in a review, the type of content or category of the review text and the gender of the user writing the review. To address this issue we performed our analysis based on the judgment of two annotators. Furthermore, to assure that the annotation task was understood in a uniform manner, we created an annotation guide which contained detailed descriptions of the task and its process, as well as relevant definitions and examples. Also, we conducted three annotation trials in which common misunderstandings were clarified. To increase the confidence of the manual analysis results, disagreements were solved by a third annotator.

Using names as gender indicators. Another threat to validity in our work is the assumption that usernames can be indicative of the gender of their users—as well as the assumption that users can be classified into only three gender groups. Further, due to the limitations of the gender identification tool we could not link the username to specific cultural contexts. Hence, there is an over-representation of reviews identified as unisex in our analyzed set. Due to these limitations we did not draw any hypothesis from the gender distribution of users submitting user reviews across different cultures.

Sample size. We relied on manual content analysis to study the content of the reviews, their sentiment and the gender of the users writing the reviews. A manual analysis on our whole data set of 59,203 reviews was not feasible. For this reason, we used a sample of 2,560 reviews. To mitigate generalizability threats, we selected the size of the sample so that it allowed a generalization to our full data set at a 95% confidence level, accepting an error margin of 3%. Furthermore, we used a stratified sampling, assuring that all selected apps were equally represented in the sample. We had to discard 152 reviews from our sample of 2,560 reviews because they turned out to be noisy in the manual analysis. Therefore we might not have achieved the 95% statistical confidence that we had calculated when creating the sample from our raw data. With hindsight, we should have selected a slightly larger sample to compensate for the noisy reviews that we had to discard.

External validity. We mitigated external validity threats by considering software applications from four different domains and from eight geographically and culturally diverse countries. Analyzing applications with diversity in these three characteristics allows us to obtain insight about national cultural differences concerning different software applications and countries. Nevertheless, we only analyzed user reviews written in English, as this was a language that all annotators were fluent and in which sufficient reviews from culturally and geographically diverse countries could be found in the App Store. Furthermore, we analyzed reviews written in English for countries that use English as a second language, for example, Hong Kong and India. One could argue that in these countries, the community of users who write reviews in English is not representative of the national culture. Further studies should be conducted on reviews from a wider range of countries and written in different languages.

Apple devices are high-end products, which, in many countries, are not affordable for a considerable percentage of the population. Therefore, users buying apps from Apple's App Store and submitting reviews to it may not be representative of the national culture of a country. The reason why we chose the Apple App Store for our study is that this platform allows to collect reviews related to specific countries. This is not possible in other platforms such as Google Play where such distinction is not available in their API. Further studies analyzing user feedback about software submitted through a variety of channels (e.g., various app stores or social media) should be carried to see if the results of this study hold. Moreover, this study did not include applications with low popularity and further research under such conditions is necessary.

7 CONCLUSIONS

We report on a cross-cultural study analyzing user feedback across different countries with diverse cultural backgrounds. We find that feedback characteristics such as sentiment, content, rating and length significantly differ at the country level. Moreover, we also show that these differences follow cultural patterns—and that these patterns can be mapped to a recognized cultural model.

The observations found in our study should be seen as preliminary and further factors affecting the differentiation should be studied. We did not consider user feedback given in other application distribution platforms or social media, such as Facebook or Twitter. We believe that these are interesting directions for future work, as is the analysis of feedback provided by additional national cultures beyond the ones analyzed in this work, and written in other languages. Moreover, the actual impact of the cultural differences of user feedback on automatic analysis algorithms also needs investigation.

We hope that our work inspires software engineering researchers and practitioners to take diversity into account when designing, validating and testing algorithms that use human data or that make decisions for them; and that it helps start a discussion of the consequences that algorithm bias can bring to software users and society in general.

ACKNOWLEDGEMENTS

We thank our reviewers for their valuable feedback. This work was partially supported by the European Commission and the Swiss State Secretariat for Education, Research and Innovation (SERI) within the SUPERSEDE project (ID 644018).

REFERENCES

[1] Apple Inc. 2015. About Internationalization and Localization. (2015). https://developer.apple.com/library/content/documentation/MacOSX/Conceptual/BPInternational/Introduction/Introduction.html

[2] Apple Inc. 2017. App Store shatters records on New Year's Day. (2017). https://www.apple.com/newsroom/2017/01/app-store-shatters-records-on-new-years-day/

[3] Hajer Ayed, Benoît Vanderose, and Naji Habra. 2017. Agile cultural challenges in Europe and Asia: Insights from practitioners. In *39th International Conference on Software Engineering (ICSE 2017)*. 153–162.

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*. 4349–4357.

[5] Rishi Chandy and Haijie Gu. 2012. Identifying spam in the iOS app store. In *2nd Joint WICOW/AIRWeb Workshop on Web Quality*. ACM, 56–59.

[6] Ning Chen, Jialiu Lin, Steven CH Hoi, Xiaokui Xiao, and Boshen Zhang. 2014. AR-Miner: Mining informative reviews for developers from mobile app marketplace. In *36th International Conference on Software Engineering (ICSE 2014)*. 767–778.

[7] Federal Trade Commission et al. 2016. Big data: A tool for inclusion or exclusion? Understanding the issues. *FTC Report* (2016).

[8] Andrea Di Sorbo, Sebastiano Panichella, Carol V Alexandru, Junji Shimagaki, Corrado A Visaggio, Gerardo Canfora, and Harald C Gall. 2016. What would users change in my app? Summarizing app reviews for recommending software changes. In *24th ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE 2016)*. 499–510.

[9] Martin Fowler. 2006. Using an agile software process with offshore development. (2006). http://martinfowler.com/articles/agileOffshore.html

[10] Francis Fukuyama. 1995. *Trust: The Social Virtues and the Creation of Prosperity*. New York: Free Press.

[11] Ruth García-Gavilanes, Yelena Mejova, and Daniele Quercia. 2014. Twitter ain't without frontiers: Economic, social, and cultural boundaries in international communication. In *17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW 2014)*. ACM, 1511–1522.

[12] Emitza Guzman, Rana Alkadhi, and Norbert Seyff. 2016. A needle in a haystack: What do Twitter users say about software?. In *24th IEEE International Requirements Engineering Conference (RE'16)*. 96–105.

[13] Emitza Guzman, Omar Aly, and Bernd Bruegge. 2015. Retrieving diverse opinions from app reviews. In *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM 2015)*. 1–10.

[14] Emitza Guzman, Muhammad El-Halaby, and Bernd Bruegge. 2015. Ensemble methods for app review classification: An approach for software evolution. In *30th IEEE/ACM International Conference on Automated Software Engineering (ASE 2015)*. 771–776.

[15] Emitza Guzman, Mohamed Ibrahim, and Martin Glinz. 2017. A Little Bird Told Me: Mining Tweets for Requirements and Software Evolution. In *25th IEEE International Requirements Engineering Conference (RE'17)*. 11–20.

[16] Emitza Guzman, Mohamed Ibrahim, and Martin Glinz. 2017. Prioritizing user feedback from Twitter: A survey report. In *International Workshop on Crowd Sourcing in Software Engineering*. 21–24.

[17] Emitza Guzman and Walid Maalej. 2014. How do users like this feature? A fine grained sentiment analysis of app reviews. In *22nd IEEE International Requirements Engineering Conference (RE'14)*. 153–162.

[18] Stuart Hall. 2017. How Often Should You Update Your App? (2017). https://stories.appbot.co/how-often-should-you-update-your-app-9405b85a967c

[19] Charles Hampden-Turner and Fons Trompenaars. 1993. *The Seven Cultures of Capitalism*. New York: Doubleday.

[20] Geert Hofstede. 1984. *Culture's Consequences: International Differences in Work-Related Values*. Sage Publications.

[21] Geert Hofstede. 1991. *Cultures and Organizations: Software of the Mind*. New York: McGrawHill.

[22] Geert Hofstede. 2011. Dimensionalizing cultures: The Hofstede model in context. *Online Readings in Psychology and Culture* 2, 1 (2011). https://doi.org/10.9707/2307-0919.1014

[23] Geert Hofstede. 2018. Compare Countries. (2018). https://geert-hofstede.com/countries.html

[24] Helena Holmström, Brian Fitzgerald, Pär J Ågerfalk, and Eoin Ó Conchúir. 2006. Agile practices reduce distance in global software development. *Information Systems Management* 23, 3 (2006), 7–18.

[25] Leonard Hoon, Rajesh Vasa, Jean-Guy Schneider, and John Grundy. 2013. An analysis of the mobile app review landscape: Trends and implications. *Swinburne University of Technology, Tech. Rep* (2013). http://hdl.handle.net/1959.3/352848

[26] Claudia Iacob and Rachel Harrison. 2013. Retrieving and analyzing mobile apps feature requests from online reviews. In *10th Working Conference on Mining Software Repositories (MSR 2013)*. 41–44.

[27] Hammad Khalid, Emad Shihab, Meiyappan Nagappan, and Ahmed E Hassan. 2015. What do mobile app users complain about? *IEEE Software* 32, 3 (2015), 70–77.

[28] Florence R Kluckhohn and Fred L Strodtbeck. 1961. *Variations in Value Orientations*. Row, Peterson.

[29] Seiyoung Lee and Hwan-Seung Yong. 2010. Distributed agile: project management in a global environment. *Empirical Software Engineering* 15, 2 (2010), 204–217.

[30] Walid Maalej and Hadeer Nabil. 2015. Bug report, feature request, or simply praise? On automatically classifying app reviews. In *23rd IEEE International Requirements Engineering Conference(RE'15)*. 116–125.

[31] William Martin, Federica Sarro, Yue Jia, Yuanyuan Zhang, and Mark Harman. 2017. A Survey of App Store Analysis for Software Engineering. *IEEE Transactions on Software Engineering* 43, 9 (2017), 817 – 847.

[32] Evie McCrum-Gardner. 2008. Which is the correct statistical test to use? *British Journal of Oral and Maxillofacial Surgery* 46, 1 (2008), 38 – 41.

[33] Kimberly Neuendorf. 2002. *The Content Analysis Guidebook*. Thousand Oaks, CA: Sage Publications.

[34] Nigini Oliveira, Nazareno Andrade, and Katharina Reinecke. 2016. Participation differences in Q&A sites across countries: Opportunities for cultural adaptation. In *Nordic Conference on Human-Computer Interaction (NordiCHI'16)*.

[35] Dennis Pagano and Bernd Bruegge. 2013. User involvement in software evolution practice: A case study. In *Proc. of the International Conference on Software Engineering*. 953–962.

[36] Dennis Pagano and Walid Maalej. 2013. User feedback in the appstore: An empirical study. In *21st IEEE International Requirements Engineering Conference (RE'13)*. 125–134.

[37] Fabio Palomba, Pasquale Salza, Adelina Ciurumelea, Sebastiano Panichella, Harald Gall, Filomena Ferrucci, and Andrea De Lucia. 2017. Recommending and localizing change requests for mobile apps based on user reviews. In *39th International Conference on Software Engineering*. 106–117.

[38] Sebastiano Panichella, Andrea Di Sorbo, Emitza Guzman, Corrado Visaggio, Gerardo Canfora, and Harald Gall. 2015. How can I improve my app? Classifying user reviews for software maintenance and evolution. In *31st International Conference on Software Maintenance and Evolution (ICSME 2015)*. 281 – 290.

[39] Statista. The Statistics Portal. 2017. Number of available applications in the Google Play Store from December 2009 to September 2017. (2017). https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/

[40] Katharina Reinecke and Abraham Bernstein. 2011. Improving performance, perceived usability, and aesthetics with culturally adaptive user interfaces. *ACM Transactions on Computer-Human Interaction (TOCHI)* 18, 2 (2011).

[41] Katharina Reinecke and Krzysztof Z Gajos. 2014. Quantifying visual preferences around the world. In *2014 SIGCHI Conference on Human Factors in Computing Systems (CHI'14)*. 11–20.

[42] Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685* (2015).

[43] Lorenzo Villarroel, Gabriele Bavota, Barbara Russo, Rocco Oliveto, and Massimiliano Di Penta. 2016. Release planning of mobile apps based on user reviews. In *38th International Conference on Software Engineering (ICSE 2016)*. 14–24.

[44] Gayna Williams. 2014. Are you sure your software is gender-neutral? *Interactions* 21, 1 (2014), 36–39.