What People Like in Mobile Finance Apps – An Analysis of User Reviews

Johannes Huebner[1], Remo Manuel Frey[1], Christian Ammendola[2], Elgar Fleisch[1], Alexander Ilic[1]
[1]ETH Zurich, Zurich, Switzerland, [2]42matters AG, Zurich, Switzerland
{jhuebner,rfrey,efleisch,ailic}@ethz.ch, christian@42matters.com

ABSTRACT
Even though app store reviews provide highly valuable information on how people use mobile apps and what they expect from them, the systematic, timely analysis of an ever-growing volume of such unstructured reviews across many apps remains a challenge. We analyzed more than 300'000 review sentences belonging to 1'610 finance apps using a machine learning-based approach to investigate the impact that different aspects of finance apps have on their ratings. Additionally, we manually categorized all apps into sub-categories such as payment or trading apps to discuss our findings on an extra level of detail. This work illustrates how different aspects of mobiles apps affect their ratings, how this varies across sub-categories, and discusses the role of privacy, user interfaces, signup experiences, notifications, when the use of location services may be appropriate, and other aspects of mobile finance apps, to provide detailed insights into users' expectations and perception of finance apps.

Author Keywords
App store reviews; mobile finance apps; machine learning

ACM CCS Concepts
• Information systems~Information systems applications • Human-centered computing~Human computer interaction (HCI)

INTRODUCTION
Smartphones and mobile applications have given billions of people unprecedented access to information systems. Many of these users leave reviews with valuable information regarding what they appreciate, and what they dislike in mobile apps. Yet, app stores do not provide an easy way for developers to extract insights from such reviews, other than manually inspecting them, which becomes impracticable for popular apps with hundreds or even thousands of reviews generated each month. Previous work has helped app developers extract app store information for individual apps, such as bugs, or feature requests mentioned in user reviews. Other work has analyzed larger sets of apps, usually without going into a great level of detail. This paper thus introduces a fine-granular, fully automated system to systematically analyze a large volume of app store reviews belonging to many different apps in a more detailed fashion than done previously. Using finance apps as an example, it demonstrates how it can be used to understand what aspects of mobile apps are most important in the perception of their users. This work focused on finance apps because the recent FinTech revolution, which started after the 2008 financial crisis, has given rise to an extraordinary amount of market entrants who challenged almost all aspects of the financial services value chain, and they predominantly used mobile apps as their customer-facing communication channel. We enriched this analysis with domain-specific knowledge to provide more insightful and detailed results than what would be possible when investigating apps over a broader range of categories. This work seeks to answer the following two research questions:

RQ1. *To what extent do different aspects of mobile finance apps impact their app store ratings?*

RQ2. *How does the importance of those aspects vary across different sub-categories of finance apps*?

We thus acquired a large dataset of more than 300'000 app store review sentences, and detected the topics mentioned in each sentence, as well as the corresponding sentiment. In doing so, this paper makes two important contributions: First, it provides an overview of what aspects of financial information systems their users focus on. Second, this paper quantifies the impact that different aspects of finance apps have on their rating, and thus gives meaningful guidelines to app developers seeking to improve their products.
The next chapters present related research, methodology, results and their discussion, followed by a conclusion.

LITERATURE REVIEW

Understanding User Reviews
User reviews in the app stores provide the app developer with valuable information about how the app is perceived by users. The developer learns if something does not work and even receives suggestions for improvement. At the same time, positive comments attract new users and are therefore free advertising, whereas bad comments are unfavorable for business. It is thus crucial to better understand users' feedback and why they are happy or complaining. [9] provided first insights in a systematic

MUM '18, November 25–28, 2018, Cairo, Egypt © 2018 Association for Computing Machinery. ACM ISBN 978-1-4503-6594-9/18/11…$15.00
ACM ISBN 978-1-4503-6594-9/18/11…$15.00
https://doi.org/10.1145/3282894.3282895

investigation of user reviews. The authors found that the most frequent complaints were functional errors, feature requests, and app crashes, though even updates themselves, which should improve the app, led to negative comments. A total of 11 percent of negative reviews were due to a recent app update. Researchers also sought to understand different characteristics of user reviews. For instance, [25] found that users tend to leave longer messages when they rate an application badly, and the depth of feedback in certain categories was significantly higher than for others. [15] found that as the number of downloads and releases increased, the total number of reviews also increased.

Automated Analysis of User Reviews

The systematic acquisition and analysis of user reviews still remains a challenge today. Even though the reviews of all apps in the app stores are available digitally, publicly, and free of charge, they cannot easily be exported across multiple apps, let alone in a machine-readable form. [27] proposed a semi-automated framework to collect and mine user opinions from app stores using a keyword-based approach. However, the sheer amount of feedback for popular apps is difficult to process and requires a fully automated solution, and many computational resources. Especially for app developers, the timely understanding of user problems and integration into release cycles is crucial. [2] offered a solution for better release planning by using machine learning and information retrieval techniques to automatically classify reviews according to a taxonomy. Based on that, their solution gave recommendations for the review of specific source code to handle the issue described in the user review. [13] also described an automated classification approach for user reviews to identify bug reports and feature requests for individual apps. The work of [26] went in a similar direction, where user reviews were categorized as either bug report, feature suggestion, or other, were then clustered, and automatically prioritized for the subsequent app release. A similar approach was also proposed by [1], with the aim of presenting the groups of most informative reviews. The system of [4] was able to automatically discover inconsistencies in reviews; identify reasons why users like or dislike a given app, provide a view of how users' reviews evolve over time; and identify users' concerns and preferences of different types of apps.

Star Rating

This work measured the success of an app by the star rating in the app stores. The selection of this criterion for marketing success is also aligned with previous research by [7] who found a strong correlation between app ratings and the total number of downloads of an app. However, it is important to note that overall star ratings may not be dynamic enough to measure continuous improvements of an app. [23] found in their study with over 10,000 unique mobile apps that while many improved from version to version, their store rating remained relatively stable once an app had gathered a substantial number of raters.

Sometimes, people's rationale for star ratings are difficult to identify. For example, the use of third-party ad libraries was found to have an effect on ratings [22]. Such ad libraries are used by app developers to monetize their app. In their work, [17] tried to predict the star rating based on user reviews. They found out that not all fine-granular opinions were of importance for rating predictions. [20] conducted experiments to check the relationship between the number of stars and the content of the review comments. The results showed that some text contents of reviews were not correctly represented by the star rating.

Sentiment Analysis in User Reviews

The sentiment analysis tries to uncover the emotional status of reviewers. Are they satisfied with the app or rather annoyed? Research in this area is already well established. Various techniques are used, such as natural language processing, text analysis, and computational linguistics. The two approaches of [6] and [12] identified fine-grained app features in the reviews and extracted the user sentiments about these features, as a tool for individual developers to understand app-specific review contents. [21] provide a list of 19 state-of-the-art approaches for sentiment analysis of user reviews about products and services in general. They compared these approaches with their own machine learning-based approach in a sentiment analysis of Italian reviews. [20] suggested introducing a sentiment rating in addition to the existing star rating, which provides novel information to assist users in the decision-making process regarding the choice of applications. It is generated from the automatic aggregation of opinions reported in the reviews. The results obtained evince that it is possible and useful to generate a sentiment rating automatically. [11] investigated how the sentiments of different topics in online reviews affect app sales. They developed a multifaceted sentiment analysis method for analyzing textual sentiments from the perspective of product quality and service quality. Results indicate that although reviews on product quality occupied a larger portion, comments on service quality had a greater impact on app sales rankings. [3] include Kano's Customer Satisfaction Model in a sentiment analysis in a case study of five apps, which categorizes product and service attributes in five groups: *must-have quality*, *satisfiers*, *delighters*, *indifferent quality*, *reverse quality*. This approach can help app developers find areas for improvements and prioritize improvement tasks.

METHODOLOGY

Data Collection

Finance Apps and Acquisition of App Store Reviews

The strategy for the app selection was to focus on a wide-range of well-adopted finance apps, which were identified using the mobile app explorer provided by 42matters, an app store analytics company, who aggregate data on all mobile apps found in the two largest app stores, Google Play Store and Apple's App Store. Using 42matters' app explorer, we thus selected all English-speaking Android apps that were also available on iOS, have been downloaded at least 10'000 times, and have been assigned

the 'Finance' category by the app developer, leading to a total of 2'616 apps. The snapshot was taken in April 2018.

For acquiring an extensive sample of app store user reviews for the selected finance apps we employed web scraping techniques. Since a single review could mention multiple aspects of an app, each review was split up into individual sentences for this analysis. This led to a dataset including a total of 354'040 review sentences that belonged to the selected finance apps, which were collected from the Play Store between February 2016 and April 2018. It should be noted that the analyzed dataset does not include all reviews. Instead, this work only included reviews that did not merely provide a star rating, but also contained an English review text. In addition, due to the high computational costs of this endeavor, the web scraper was configured to visit each app at least once per day (apps with more reviews were visited more frequently), and obtained the first page of reviews that were sorted in the default order ("most useful first") by Google. Because of this approach, we had a slightly more complete picture of apps with only few reviews, since the scraping process was likelier to catch all of their available reviews. However, especially for very popular apps with many reviews, comments repeat, and thus do not add as much information as reviews for less popular apps.

Definition and Coding of App Categories

This work hypothesizes that the effect of different app aspects might vary across different categories of finance apps, *e.g.* the presence of a tutorial explaining app functionality might be more crucial in very novel or more complex areas such as investing and trading than in traditional general-purpose banking apps (see *RQ2*). Therefore, the acquired finance apps needed to be broken down into functional categories. To define the taxonomy of categories this work followed the proven iterative approach proposed by [18]. The first iteration was conceptual-to-empirical, starting with the six functions of financial

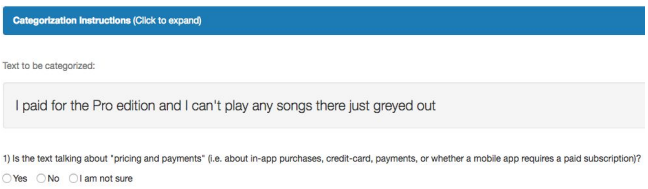

Figure 1. MTurk interface for obtaining a test dataset

services proposed by [16]: Payments, Insurance, Market Provisioning, Deposits & Lending, Investment Management, and Capital Raising. After investigating the first round of 10 apps, general banking apps were identified as a missing category, which was thus added as a seventh category. After three further empirical-to-conceptual iterations based on 25, 50, and 100 further apps, the ending conditions were met. During those three iterations the categories Savings, Distributed Ledger Technologies (DLT, *e.g.* Blockchain), Authentication, and Other were added, and Market Provisioning as well as Capital Raising were removed, because no apps were present for either of those two categories in which customer interactions typically take place through other channels than mobile apps. In addition, the category Deposits & Lending was split up in two, and Deposits was renamed into Saving. Moreover, Investment Management was renamed to also explicitly include the term Trading. The final set of categories is summarized in Table 1. One member of the research team then manually coded the category of all mobile apps with at least 30 review sentences (1'610 apps), resulting in 341'989 (96.6%) of all available review sentences coded by app category. The classification was stopped at this point, since categorizing the last 3.4% of all review sentences would have required a manual classification of the 1'006 apps to which those sentences belonged, which seemed disproportionately expensive. In the rare cases where apps included features of multiple categories (*e.g.* payment and lending), the dominant category was applied. A second

App Category	Description
Banking	General-purpose banking apps, such as banks' main mobile apps, account aggregators, or budget planners. Banking apps often include elements of other categories (such as payments or trading),
Payment (and money transfers)	Apps focused on transferring money, e.g. at the point of sale or between two individuals, or mobile wallets
Trading (and investment management)	Apps focused on helping users to manage their investment portfolios and acquire information on financial markets
Lending	Apps focused around lending money, e.g. providing information about existing loans, allowing pay-offs, or calculating new loans
Distributed Ledger Technologies (DLT)	Apps focused on use cases related to distributed ledger technologies (incl. cryptocurrencies, blockchain), such as wallets, portfolio trackers or exchanges, where cryptocurrencies can be traded
Insurance	Apps focused on insurance-related use cases, such as submitting and tracking claims, or the tracking of existing and purchase of new policies
Saving	Apps focused on helping users save money, often connecting to external data or payment cards in order to automate saving transactions (e.g. towards retirement or consumption goals)
Authentication	Apps providing additional authentication security mechanisms to other financial apps, such as one-time passwords for banking apps
Other	Apps fitting in none of the above categories, such as tax-related apps, or business expense tools.

Table 1. App Categories

coder independently re-coded a randomly selected subset of 200 apps using the same classification scheme. For those apps, the raters agreed in 94.0% of the apps, resulting in a Cohen's kappa of 0.933, indicating almost perfect agreement between the two coders [8], which also demonstrates the robustness of the classification scheme.

Parsing of User Reviews

App Aspects

Aspects are defined as general facets about an app mentioned in app reviews such as the user interface or device compatibility. Those aspects can arise in any category of app, independently of whether it is a finance app, a music player, or a game. Based on prior work [4,14] and a manual inspection of reviews, 17 main aspects were identified in an iterative approach (see Table 2). Other aspects, such as security, were not included since they were not frequently mentioned in reviews, in this example arguably because end users usually cannot judge the security features of mobile apps.

Prediction of App Aspects

To build a system capable of predicting app aspects in an automated manner, supervised learning techniques were applied. In order to cope with reviews mentioning different aspects of an app, we designed a system capable of categorizing each sentence of a review into one aspect separately. For achieving this goal, a training dataset with sentences labeled with one of the app aspects listed in Table 2 was built. This was done by a keyword search over the sentences. Further, topic modeling was used over the datasets generated using keyword search in order to evaluate the quality of the training datasets. The evaluation based on topic modeling was also used to further refine the sets of keywords used for building the training datasets. Several iterations where conducted in order to build the final training dataset. Using the described approach, a training dataset of about 15,000,000 samples (for the 17 aspects) was built; this included review sentences of all apps in the app stores, *i.e.* not just finance apps.

In addition, a test dataset was needed to test the quality of any fitted supervised model, which was done with Amazon Mechanical Turk (MTurk), using the following approach:

1. For each app aspect, a random subset of the training dataset samples was selected (and removed from the training dataset)
2. Those samples were uploaded to MTurk, asking people whether the corresponding app review sentence mentioned a certain app aspect, as shown in Figure 1. For each sample (*i.e.* sentence) three different workers were asked to annotate it
3. To ensure a high-quality test dataset was produced, only such samples were used where all three workers agreed on the annotation. As a result, a test dataset of more than 10,000 samples was created.

To predict the app aspects mentioned in user reviews, a baseline model using a logistic regression with doc2vec

App Aspect	**Definition:** A review sentence mentioning…
Advertising	… ads shown in a mobile app such as ads which pop up, ads notifications, or targeting of ads.
User Interface	… the graphical user interface of a mobile app, such as design, layout, colors, graphics, buttons, etc.
Pricing and Payments	… in-app purchases, credit-card, payments, or whether a mobile app requires a paid subscription.
Resource Usage	… a mobile app's usage of memory, cpu, network, battery, etc.
Device Compatibility	… the compatibility of an app with devices, e.g. does not work on Android 22, works perfectly on iPhone X, etc.
Connectivity	… topics related to the connectivity of a mobile app, e.g. bluetooth, wifi, NFC, etc.
Privacy	… privacy settings of a mobile app, e.g. logs data in the background, ask for too many permissions, etc.
Sign-up Experience	… the signup experience of a mobile app, e.g. log in, can't sign up, can't log out, etc.
Tutorial	… an in-app tutorial, e.g. lack thereof, great tutorial, tutorial too long, etc.
Audio	… the audio characteristics of a mobile app, e.g. audio quality
Video	… the video characteristics of a mobile app, e.g. video streaming, etc.
Notification / Alerts	… the push and in-app notifications of a mobile app, e.g. too many / helpful / non-working notifications etc.
Translation and Internationalization	… the app translations and internationalization issues, e.g. suggestions to improve translations, localize images/graphics for a country, etc.
Location Services	… the location usage of a mobile app, e.g. using location services in the background, forcing users to allow access to location services, an app tracking users' location, etc.
Uninstall*	… whether a user uninstalls a mobile app because of various reasons, e.g. because the app requires too many permissions, is continuously crashing, because too many ads are shown, etc.
Update	… experiences of users when updating a mobile app, e.g. the app improved a lot after the update, some feature does not work anymore in the new version of the app, etc.
Stability	… mobile app failures, e.g. when an app crashes or freezes

Table 2. App Aspects. *Note: The deletion of an app itself is not an app aspect, but rather the consequence resulting from other aspects that have not been further specified by the user. For brevity, we simply label such review sentences with *uninstall* instead of using a more accurate label such as "not otherwise specified app aspects resulting in an app's deletion".

[10] as input features was built, which yielded a macro F1 score of 0.584 and a micro F1 score of 0.722.

The improved, final prediction model used a Convolutional Neural Network (CNN) implemented with TensorFlow and the Python library Keras. We also experimented with different CNN architectures and finally chose one inspired by the one presented in [28]. This final system used GloVe word embeddings [19] as features, which were fitted over English Wikipedia articles. Using this approach, a noticeable performance improvement was achieved compared to the baseline model, as shown in Table 3.

Detection of Review Sentiments

Finally, the rule-based VADER system was used to detect sentiments of each review sentence in the dataset [5]. This approach, which has shown to outperform human raters and other computational systems, detects the direction and intention of the sentiment of a given sentence, and thus classifies the overall sentiment as to be either negative, neutral, or positive. Note that the sentiment detection was based only on the textual content of a review, and not the numeric star rating provided by the users. While the VADER system's impressive performance has been demonstrated by its original authors, we also manually inspected a random sample of the detected sentiments, which appeared to be accurate in the vast majority of all cases. Additionally, we examined the sentiment-rating distribution as an extra quality check of sentiment detection: Sentences with positive sentiments should occur more frequently in reviews with high star ratings than in those with low ratings. As Figure 2 illustrates, this is indeed the case, thus the sentiment detection provides meaningful results, as is also reflected in the results of a simple linear regression, see Model (1) in Table 4.

Data Analysis

The research at hand then used OLS regressions and statistical tests to answer the research questions, the findings of which are presented in the following chapter.

RESULTS AND DISCUSSION

Investigated Apps and Review Sentences

Across the analyzed dataset of 341'989 review sentences, the average star rating was 2.773, whereas the extreme values (one and five stars) occurred much more frequently than the moderate values (two through four stars, see Figure 3), which is in line with previous research [17,20]. The large share of extreme ratings could be explained with herding effects, users' higher propensity to actually leave reviews when they are either very disappointed or pleased by an app, as well as by apps that improve over time, and thus transition from initially receiving mostly 1-star reviews to better ones at a later stage. This last point is supported by the distribution of mean ratings on an app-level, as illustrated in Figure 4: While the center of the app-level rating distribution (M=2.709) is almost identical with that of individual review sentences (M=2.773), the share of apps that managed to receive very high ratings (4.0 stars or better) over their lifetime in the app store is relatively small

Metric	Baseline Model (Logistic regression)	Final Model (CNN)
Accuracy	0.722	0.925
Macro F1 score	0.584	0.921
Micro F1 score	0.722	0.925

Table 3. Accuracy levels of the app aspect prediction models

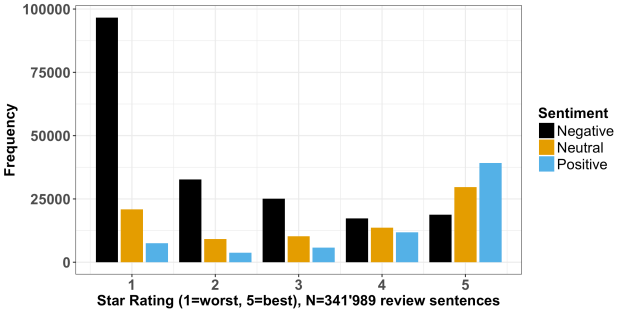

Figure 2. Distribution of sentiments across star ratings

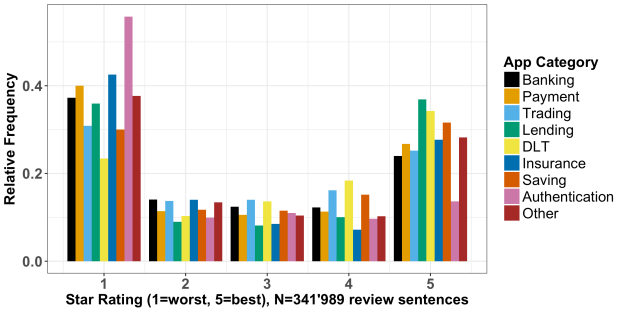

Figure 3. Distribution of star ratings by app category

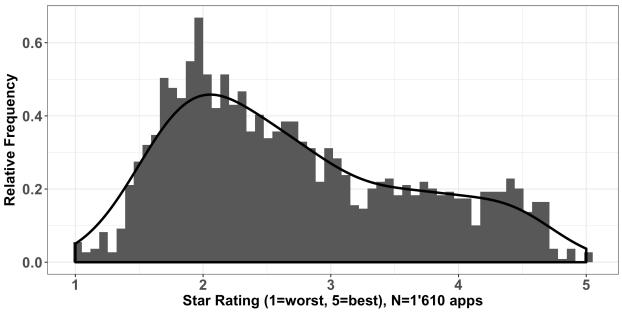

Figure 4. Distribution of mean app-level star ratings

(13.4% or 216 apps), even though 38.1% of the individual rating data points had a 4-star or 5-star rating. It would not be surprising to also find similar distributions of individual reviews and average ratings in other mobile app categories. As can be seen in Figure 3, the share of 2- and 3-star reviews is very similar across all finance sub-categories, whereas authentication apps seem to receive particularly many 1-star reviews, while apps in the categories DLT, savings and lending receive more favorable reviews in general, which reflects in discrepancies of mean ratings across categories (cf. Table 5).

The Impact of Individual App Aspects on Ratings

To study what aspects of apps influence app ratings in the financial sector to what extent, we used an OLS analysis.

For example, do app users in a domain with access to such sensitive data put an emphasis on privacy, what role do user interfaces play, and do consumers appreciate push notifications and alerts? To answer these questions, all five-star ratings were centered by the grand mean in the dataset (2.773 stars) to facilitate the interpretation of our findings; lower ratings than the grand mean can be regarded as penalties relative to this default rating, higher ones as bonuses.

Figure 5 plots the mean ratings associated with each app aspect across the three levels of sentiment (x-axis). For many app aspects, the slopes on the left half of the figure (between negative and neutral sentiments), and on the right half (neutral to positive sentiments), are very similar, *i.e.* the two partial lines form an almost straight overall line. However, the increase in ratings that app developers can expect is slightly higher when improving an app aspect from *negative* to *neutral*, than when moving from there to *positive*. This observation is likely to apply also to mobile apps from other categories such as productivity or health apps, but would have to be investigated separately in future work. Therefore, for almost all included aspects of mobile apps, fixing annoyances before optimizing already-good aspects appears to be a sensible strategy, which makes sense intuitively. It is also noticeable that while many of the lines run almost perfectly parallel, their intercept varies quite a bit: The mean ratings per app aspect span a bandwidth of 1.2 to 2.3 stars for any given sentiment.

Must-Have App Aspects

Compared to the grand mean rating, there exists a cluster of app aspects, like connectivity, that never (or barely) cross the x-axis, *i.e.* even with a positive sentiment they still don't (or barely do), on average, produce an above-average star rating (see Figure 5). Thus, there seem to be aspects of mobile finance apps, which, when they surface to the users' consciousness, they usually do so in a negative, disappointing context; and even if the users perceive these particular aspects as positive in any given app, they still tend to give the app, at best, an average rating. The data thus suggests that when such aspects surface, app developers can only lose in the favor of the users, therefore it is advisable to make sure these aspects never get to the users' minds. In the KANO Customer Satisfaction Model [24], such qualities are described as *must-have qualities*. The aspects included in this group are: *connectivity*, *stability*, *device compatibility, update, uninstall*, *signup experience*, and *privacy*.

Of course, basic technical requirements such as connectivity and app stability are thus expected by consumers, and even when mentioned in a review positively, developers should not expect more than an average rating. Updates to apps unsurprisingly fall in the same category, and of course when disgruntled users uninstall apps and leave a review, they usually won't give stellar ratings. However, a small share (4.5%) of review sentences mentioning the deletion of an app, actually do so with a positive sentiment, *e.g.* "I uninstalled, then

	Star Rating					
	(1)		(2)		(3)	
	ß	SE	ß	SE	ß	SE
Constant	2.102***	.003	2.212***	.028	2.526***	.006
Neutral sentiment	1.163***	.006	.887***	.006	.893***	.006
Positive sentiment	1.949***	.006	1.519***	.007	1.526***	.007
App Aspect						
User Interface			.640***	.028		
Signup Experience			-.453***	.028		
Pricing / Payments			.313***	.028		
Update			-.292***	.029		
Stability			-.391***	.029		
Notifications / Alerts			.340***	.030		
Uninstall			-.701***	.031		
Resource Usage			.055	.033		
Connectivity			-.414***	.034		
Device Compatibility			-.426***	.039		
Privacy			-.472***	.046		
Video			.082	.056		
Location Services			.065	.060		
Audio			.155**	.064		
Tutorial			.517***	.087		
Translation and I18n			.241***	.093		
App Aspect Group						
Must-Have					-.728***	.006
Satisfiers					-.268***	.015
User Interface					.321***	.007
N	341,989		341,989		341,989	
R^2	0.236		0.304		0.302	
F	52,717.060***		8,289.389***		29,599.110***	

Note: ***p<0.001. Negative sentiments are the reference in Model (1); in Model (2), negative sentiments and Advertising are the reference; in Model (3), negative sentiments and Delighters are the reference.

Table 4. Regression analysis on detected sentiments

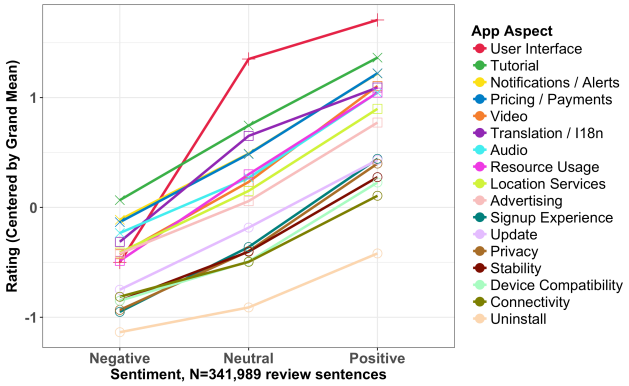

Figure 5. Mean ratings across app aspects

	Apps	Review sentences (whole sample)			Negative sentences			Neutral sentences			Positive sentences		
Category	N	N	M	SD	N	M	SD	N	M	SD	N	M	SD
Banking	868	199172	2.716	1.622	112387	2.067	1.337	48700	3.210	1.622	38085	3.997	1.391
Payment	202	47407	2.733	1.681	26153	2.057	1.408	11567	3.186	1.670	9687	4.017	1.424
Trading	191	40636	2.911	1.592	22602	2.262	1.390	9781	3.389	1.511	8253	4.121	1.262
Lending	68	9942	3.029	1.762	4815	2.061	1.476	2493	3.563	1.644	2634	4.291	1.265
DLT	65	11030	3.297	1.583	5452	2.612	1.517	2884	3.659	1.457	2694	4.295	1.115
Insurance	60	8767	2.634	1.700	5228	1.884	1.295	1976	3.379	1.698	1563	4.203	1.381
Saving	14	3407	3.066	1.652	1718	2.288	1.456	845	3.465	1.579	844	4.251	1.187
Authentication	11	857	2.155	1.504	573	1.682	1.156	180	2.833	1.612	104	3.587	1.658
Other	131	20771	2.779	1.680	11611	2.092	1.400	5063	3.294	1.663	4097	4.091	1.398
All categories	**1610**	**341989**	**2.773**	**1.641**	**190539**	**2.102**	**1.371**	**83489**	**3.264**	**1.619**	**67961**	**4.051**	**1.367**

Table 5. Overview of app store review sentence data set

reinstalled and so far so good". One important aspect that deserves mentioning, is privacy: An inspection of privacy-related reviews reveals that while most users certainly cannot proficiently judge the technical, organizational, and legal measures that the app developer put into place to protect their users' privacy, they i) do check what information is displayed where (*e.g.* account balances very prominently visible in the app, which might expose them to third parties in the vicinity of the user, and thus make use of the app in public uncomfortable). Users also ii) refer to proxies such as the amount of permissions an app is requesting (*e.g.* "Why does this app, my bank, need access to my photos and contacts, have the ability to take photos? This seems to be quite evasive (...) As of tomorrow I shall close my account", and iii) they do, sometimes, read and even compare different versions of privacy policies (*e.g.* "their privacy policies have been getting more and more vague and relaxed as far as what they disclose to 3rd parties"). We thus conclude that privacy is a *must-have* quality of finance apps, which is on users' minds, and likely will become more prominent in the future. While app developers often have legitimate reasons to ask for device permissions, they should explain precisely why these requests are made. Additionally, mobile finance apps are more frequently used on the go, or when family members and friends are around, which stands in contrast to the more private use of online-banking websites. It may thus be worth considering to design features around this fact, *e.g.* adding an option to only reveal account balances based on an additional interaction by the user, such as a swipe or tap.

Satisfier App Aspects

Further, there is a group of app aspects, which would be classified as *satisfiers* in the Kano model, which have a noticeable negative impact on ratings when not fulfilled, and a positive impact when fulfilled. The aspects included in this group are: *resource usage, advertising, video, translation / internationalization,* and *location services*. For resource usage, it is not surprising that users would penalize bad experiences in the reviews, but apparently great performance is rewarded (*e.g.* "Very very nice and smooth performance" or "This is so convenient and perfect performance, for all banking needs."). The same goes for translation / internationalization (*e.g.* "This is the only bank and app that is user friendly, has adequate English translation and is accurate.") as well as location services (*e.g.* "Works great and has lots of good features, I like the GPS based ATM lookup when you are out and about."), although the sample size for these two app aspects is rather small, therefore we would caution the reader with interpretations. With advertising, the picture seems to be a bit more one-sided: By and large, users do not appreciate ads, thus positive advertising-related review sentences often mention the absence of advertising (*e.g.* "An absolute gem in the world of apps; no ads, absolutely free, great advice."). However, some users explicitly show understanding for developers' needs to finance apps through advertising (*e.g.* one five-star review says "It's totally free, which in this day and age of course means it's LOADED with advertising, but considering what you get, and with how they present the ads, it's completely worth it"). Finally, regarding the use of videos in finance apps, positive reviews mostly express joy over an easy-to-use video chat, or a great tutorial video explaining how the app works. More negative reviews mention technical failures (*e.g.* non-functional video chat or video playback) or over-use of the video format (*e.g.* too many videos in a stock trading news feed), or wasteful use of data-intensive videos where other formats could have just as easily been used. Especially for the use of location services and videos, it is thus recommended to carefully consider the use cases where users might appreciate these aspects.

Delighter App Aspects

Third, an additional set of app aspects is unlikely to yield in rating penalties even when mentioned with a negative connotation, but they are rewarded highly when mentioned in a positive context. The Kano model labels such qualities

delighters. The aspects in this group are: *notifications / alerts, tutorial, audio,* and *pricing / payments*.

Regarding notifications and alerts, developers should obviously employ common sense and not overuse them, i.e. they should bear in mind that most users carry their mobile phones at all times, and thus be even more sensitive to disrupting the user's current activity, or else they might receive negative reviews, as evidenced by the dataset at hand. However, in the financial domain, people appear to appreciate meaningful alerts *e.g.* when payments are made or incoming transactions are detected (*e.g.* "I LOVE getting a notification when payments are received"), but also for other use cases such as alerts when a trading portfolio value crosses a threshold, or tips on how to optimize one's portfolio.

Tutorials appear to be another feature with which developers can easily climb in the users' favor, as they almost exclusively seem to be received positively, unless the tutorial is forced upon users too frequently (*e.g.* whenever the app opens), or when a tutorial explains overly basic functions, in which case users might reveal their annoyance in a negative review.

When it comes to the use of audio, it should be noted that the sample size of reviews mentioning audio features is small in itself, and the misclassification rate seems to be particularly high (especially due to the use of the word 'sound' as an adjective, *e.g.* "keeps my money safe and sound"). From a manual inspection of reviews, it appears that this aspect would be more appropriately classified as *satisfier*. That said, some reviewers mention innovative features such as paying to nearby individuals or two-factor authentication using ambient sound (*e.g.* "Pay by Nearby by using sound is great funtion."), while negative reviews often complain about apps interrupting audio playback or inexplicable requests for audio recording permissions.

Lastly, the pricing and payments aspect is a peculiar one for finance apps, since i) most finance apps are offered for free, do not implement any in-app purchases or require subscriptions for the apps themselves, and ii) making monetary transfers is a core function of many finance apps, which differs a lot from other app categories with a much higher share of subscription-based or freemium apps. Thus, most of the review sentences the CNN identified as mentioning this aspect, actually seem to refer to topics such as the fee schedule of the underlying financial service (regardless of the mobile app), making payments to other accounts or using credit cards, and so on. What can be concluded from app reviews mentioning this aspect is that the predominant practice is to offer finance apps for free, and monetize using other mechanisms like transaction fees. In some cases, though, individual features are unlocked upon payment (such as the ability to scan bills and use OCR to recognize their content), whereas other vendors offer an ad-free but paid-for premium version of their apps. One exception to this rule appear to be money management tools for individuals (budget trackers) and small businesses (*e.g.* tax-related apps, mileage trackers for company vehicles).

User Interface

Finally, one aspect stands out from the rest quite distinctly– *user interfaces*. With 27.1% of all detected aspects, it is the most frequently-mentioned one, which in itself justifies a separate examination of this aspect. While users seem to penalize particularly bad user interfaces with half a star, just providing an acceptable user interface already results in a significant boost of 1.35 stars compared to the zero point, or 1.85 stars compared to bad user interfaces, which represents the single most-rewarded improvement step an app developer can take. However, enhancing a neutrally-perceived user interface even further appears to be rewarded only marginally (+0.36 stars). Therefore, if user reviews are the metric that an app developer wishes to optimize, it is advisable to provide an acceptable user interface, but then focus on other app aspects before refining the interface even further.

Summary of App Aspect Groups

In addition to the detailed consideration of each app aspect in the previous section, the 17 app aspects were grouped in the four clusters introduced before: i) *must-have aspects*, ii) *satisfiers*, iii) *delighters*, as well as iv) *user interface*, and repeated the same OLS analysis (see Model 3 in Table 4**).** The grouped model explains almost the same amount of variability of the dataset compared to Model 2, but it is arguably easier to interpret (see Figure 6).

Again, an ANOVA test ($F(3)=10844$, $p<.001$, controlling for sentiment) and post-hoc Tukey tests reveal statistically highly significant differences in rating both globally and between any pair of groups.

Differences Between App Categories

The mobile finance app space is a quite broad one in terms of use cases and interaction models, which suggests that the importance of certain app aspects will vary across different types of finance apps. For example, a payment or authentication may be used in a very transactional manner, *i.e.* after completing a specific task (such as making a payment, or logging in to the main banking app), the apps are likely dismissed again. Therefore, user interfaces might play a lesser role than for banking or trading apps, which users might access more frequently to acquire information on their financial situation. Furthermore, when using a mobile app for trading activities, users might appreciate notifications alerting them of market movements much more than they do in other categories, where information is more static, such as in insurance. This question (*RQ2*) is addressed twofold: First, we analyzed how frequently each app aspect is mentioned per category. Second, ANOVA tests and post-hoc Tukey tests were used to determine whether certain aspects were not only more prominent in the reviews, but also had varying impacts on app ratings. Note that the dependent variable for these tests was the category-centered rating (as opposed to the grand mean centered rating in the previous section), since apps in some categories appear to generally receive higher ratings than in others (*e.g.* DLT apps are usually better rated than insurance apps, cf. Table 5).

Figure 7 shows the relative distribution of app aspect mentions per app category. The user interface, for instance, is more of a focal topic for insurance, trading, and savings app reviews than for other categories. The signup experience was mentioned disproportionately often for banking apps. This may be due to the regulation-heavy nature of the financial domain: Signing up to banking apps often involves opening a new bank account, or establishing a connection to an existing bank account, hence the signup experience is arguably more important than in other finance sub-categories and other app categories. Notifications and alerts are mentioned most frequently for Trading, DLT, and Payment apps. Privacy is a feature more likely to be mentioned in DLT and Banking apps than in all other categories. Tutorials are almost exclusively mentioned for Trading apps; Advertising is most commonly mentioned for Trading and Lending apps. These discrepancies in relative frequencies support the hypothesis that the importance of app aspects varies quite dramatically across app categories. However, so far, we have only learned how often certain topics are mentioned across categories, but we have yet to understand whether users mention aspects more frequently

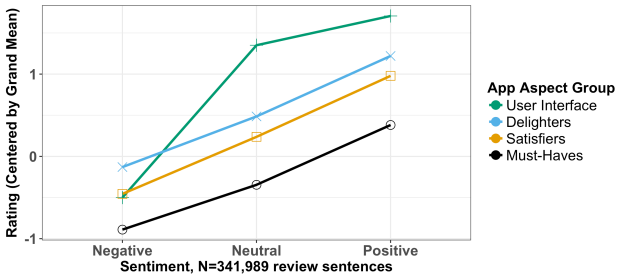

Figure 6. Mean ratings across app aspect groups

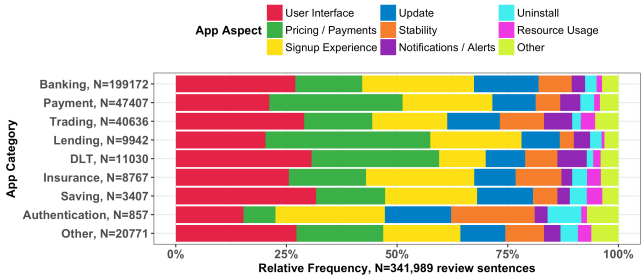

Figure 7. Relative frequency of app aspects across categories

App Aspect	N	ANOVA results (F) Sentiment	Category	Significant pairwise category differences (Tukey)
User Interface	90073	25397.48***	63.38***	A>D***, A>L*, A>S***, A>T**, B>D***, I>B***, P>B***, B>S***, B>T***, I>D***, L>D***, O>D***, P>D***, S>D**, T>D***, I>L***, I>O***, I>P**, I>S***, I>T***, P>L*, L>S*, L>T*, P>O**, O>S***, O>T***, P>S***, P>T***
Signup Experience	76771	4116.45***	92.74***	A>B***, A>D***, A>I***, A>L***, A>O***, A>P***, A>S***, A>T***, B>D***, B>I***, B>L***, B>O***, B>P***, B>S***, B>T***, I>D***, D>L*, O>D*, P>D***, T>D***, I>L***, O>L***, P>L***, S>L***, T>L***, P>O***, T>O***, P>S**, T>S*
Pricing and Payments	63664	4263.58***	82.16***	B>D***, I>B***, B>L***, B>O***, B>P***, B>T***, I>D***, L>D**, O>D***, P>D***, S>D*, T>D***, I>L***, I>O***, I>P***, I>S***, I>T***, O>L***, O>P**
Update	43800	1804.95***	5.611***	O>B**, T>B***
Stability	25850	708.49***	16.41***	A>B***, A>D***, A>I***, A>L***, A>O***, A>P***, A>S*, A>T***, B>I***, B>L***, B>O***, B>P***, B>T***, D>L**, T>I*, O>L**, P>L*, T>L***
Notifications and Alerts	12880	824.831***	8.561***	A>I*, B>I***, L>B***, D>I***, L>D*, L>I***, O>I***, P>I***, T>I***, L>O**, L>P*, L>S***, L>T**
Uninstall	9211	93.3***	24.2***	A>B***, A>D***, A>I***, A>L***, A>O***, A>P***, A>S***, A>T***, B>D***, B>L***, B>O***, B>P***, B>T***, I>L***, O>L**, P>L***, T>L***
Resource Usage	5636	449***	9.25***	S>B**, T>B***, T>I**, T>O*, T>P**
Connectivity	5184	104.063***	3.376***	A>B***, A>I**, A>L*, A>P*, O>P*
Device Compatibility	2519	60.194***	2.787**	T>P**
Advertising	2411	92.57***	4.111***	B>P**, L>P**
Privacy	1436	75.774***	3.013**	O>L**
Video	814	64.144***	4.088***	P>B***, P>D*
Location Services	652	34.039***	0.814	–
Audio	571	33.446***	2.054*	(No significant pairs)
Tutorial	279	24.462***	3.998***	T>B**, T>P*
Translation and I18n	238	22.19***	0.5	–

Table 6. Category-varying effect of app aspects on star rating. Pairwise differences are reported using the initial letter of each app category, e.g. B=Banking, P=Payment, and so on (see Table 1 for a complete list of categories), and using the direction of the effect (e.g. B>P would indicate better ratings for Banking apps than for Payment apps). ***p<0.001, **p<0.01, *p<0.05

simply because they are implemented more often in any given category, or whether they appreciate or dislike them. To learn more about the category-varying impact of different aspects on ratings, this research thus proceeded with an array of ANOVA and post-hoc Tukey tests. For each of the 17 app aspects, the following test was conducted: *rating ~ sentiment + category*.

As can be drawn from Table 6, the differences in sentiment levels were non-surprisingly a significant factor in ratings in all cases, but the primary interest of this section are category-differences, hence post-hoc Tukey tests were conducted only for this variable. Due to the limited space, this paper only discusses findings where both the ANOVA test as well as the post-hoc Tukey test revealed significant results. In addition, statistically significant but practically irrelevant findings (*e.g.* differences between two groups of less than 0.2 stars) were omitted.

User Interface

Regarding the user interface ratings, it is noticeable that DLT receive significantly worse ratings than apps in all other categories, suggesting that blockchain-related app developers may want to improve their interfaces to receive even better reviews. Insurance apps, on the other hand, are already rated significantly better than most other categories, and even the largest category, Banking, is rated comparatively well. It appears that banks and insurers have taken the threat by Fintech startups seriously and are providing well-perceived user experiences. For market entrants, this has an important consequence: Merely offering a better user interface on top of otherwise unchanged value propositions is likely not enough to sustain in the market (anymore).

Notifications and Alerts

The data suggests that push notifications implemented by insurance apps are perceived much more negatively than in all other categories, which appears to be due to the specific nature of such alerts: A manual inspection of reviews revealed that insurance apps often use notifications to promote their products, to remind clients of premium payments, or to otherwise trigger clients to move certain processes forward (such as providing additional details about a claim). Note that this work cannot make any statements about the efficacy of such notifications from the insurers' points of view. However, from the clients' perspective, such use of notifications is penalized in app reviews rather strongly, and it may be advisable to take the user's current context into consideration when promoting products. To the contrary, lending app users appreciate the use of alerts to notify them of changes regarding their loans or credit situation.

Signup Experience

Authentication apps form a positive outlier in terms of the signup experience, arguably because the app developers focus on getting this one key process right, since the functional scope of authentication apps is usually quite minimal. Banking apps were also rated slightly better than the other categories, thus we can only repeat our warning towards market entrants in this category: To select a seamless signup experience as well as a great user interface as sole differentiators is likely not a wise strategy. The signup experience of lending and DLT apps, on the other side, is not perceived to be equally good as is the case in the other categories; considering the *must-have* quality of the signup experience, developers in those two categories would be well-advised to improve their signup experience in order to avoid negative reviews.

Other Aspects

For the remaining app aspects, the data revealed a number of statistically significant differences across app categories, as is summarized in Table 6. However, in some cases, the differences between groups are rather small, thus the practical relevance may be limited; in addition, the analyzed reviews sometimes stem from only a few dozen apps per category, or there are only significant differences between one or two pairs of categories, therefore the generalizability of conclusions drawn from the sample at hand regarding less frequently mentioned app aspects may be limited, which is why we forgo a discussion thereof.

CONCLUSION

On a large-scale dataset of user-generated app store reviews, this work first used machine learning algorithms to detect mentioned aspects as well as sentiments on a sentence level in a fully automated fashion, and manually classified 1'610 apps into sub-categories within the financial space for an extra level of detail. Then, the data was analyzed in-depth to contribute to existing research on how different aspects of mobile finance apps influence their user rating. In addition, we associated individual app aspects with quality categories according to the Kano customer satisfaction model to provide hands-on guidelines for practitioners and insights for researchers about users' expectations towards finance apps. This paper discussed the role of privacy, user interfaces, signup experiences, when the use of location services may be appropriate, and other aspects of mobile finance apps in the previous chapter, and presented findings relevant for researchers and app developers alike. However, this work is not free of limitations. The independent variables describe individual sentences, whereas the star ratings are potentially provided for multiple sentences at a time and may therefore not perfectly reflect all aspects of the review text. However, due to the large number of analyzed data such inaccuracies balance out, such that it is possible to draw valid, unbiased conclusions. Future research will extend this work to other fields such as health apps, productivity apps, or games, and thus discuss in greater detail how the findings presented here generalize to other domains.

ACKNOWLEDGEMENTS

We thank our project partner, 42matters AG, for the unique opportunity to conduct research with a large volume of app review data, and we gratefully acknowledge the grant from the Swiss CTI (26989.1 PFES-ES).

REFERENCES

1. Ning Chen, Jialiu Lin, Steven C H Hoi, Xiaokui Xiao, and Boshen Zhang. 2014. AR-miner: mining informative reviews for developers from mobile app marketplace. In *Proceedings of the 36th International Conference on Software Engineering*, 767–778.
2. Adelina Ciurumelea, Andreas Schaufelbühl, Sebastiano Panichella, and Harald C Gall. 2017. Analyzing reviews and code of mobile apps for better release planning. In *Software Analysis, Evolution and Reengineering (SANER), 2017 IEEE 24th International Conference on*, 91–102.
3. Sharmistha Dey. 2017. Aspect Extraction and Sentiment Classification of Mobile Apps using App-Store Reviews. *arXiv preprint arXiv:1712.03430*.
4. Bin Fu, Jialiu Lin, Lei Li, Christos Faloutsos, Jason Hong, and Norman Sadeh. 2013. Why people hate your app: Making sense of user feedback in a mobile app store. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1276–1284.
5. C J Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*.
6. Emitza Guzman and Walid Maalej. 2014. How do users like this feature? a fine grained sentiment analysis of app reviews. In *Requirements Engineering Conference (RE), 2014 IEEE 22nd International*, 153–162.
7. Mark Harman, Yue Jia, and Yuanyuan Zhang. 2012. App store mining and analysis: MSR for app stores. In *Mining Software Repositories (MSR), 2012 9th IEEE Working Conference on*, 108–111.
8. Gary G Koch J. Richard Landis. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1: 159–174. Retrieved from http://www.jstor.org/stable/2529310
9. Hammad Khalid, Emad Shihab, Meiyappan Nagappan, and Ahmed E Hassan. 2015. What do mobile app users complain about? *IEEE Software* 32, 3: 70–77.
10. Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, 1188–1196.
11. Ting-Peng Liang, Xin Li, Chin-Tsung Yang, and Mengyue Wang. 2015. What in consumer reviews affects the sales of mobile apps: A multifacet sentiment analysis approach. *International Journal of Electronic Commerce* 20, 2: 236–260.
12. Washington Luiz, Felipe Viegas, Rafael Alencar, Fernando Mourão, Thiago Salles, Dárlinton Carvalho, Marcos Andre Gonçalves, and Leonardo Rocha. 2018. A Feature-Oriented Sentiment Rating for Mobile App Reviews. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, 1909–1918.
13. Walid Maalej, Zijad Kurtanović, Hadeer Nabil, and Christoph Stanik. 2016. On the automatic classification of app reviews. *Requirements Engineering* 21, 3: 311–331.
14. Stuart McIlroy, Nasir Ali, Hammad Khalid, and Ahmed E Hassan. 2016. Analyzing and automatically labelling the types of user issues that are raised in mobile app reviews. *Empirical Software Engineering* 21, 3: 1067–1106.
15. Stuart Mcilroy, Weiyi Shang, Nasir Ali, and Ahmed E Hassan. 2017. User reviews of top mobile apps in Apple and Google app stores. *Communications of the ACM* 60, 11: 62–67.
16. Jesse McWaters, Giancarlo Bruno, Abel Lee, and Matthew Blake. 2015. *The Future of Financial Services - How disruptive innovations are reshaping the way financial services are structured, provisioned and consumed*. Retrieved from http://www3.weforum.org/docs/WEF_The_future__of_financial_services.pdf
17. Dagmar Monett and Hermann Stolte. 2016. Predicting star ratings based on annotated reviews of mobile apps. In *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*, 421–428.
18. Robert C Nickerson, Upkar Varshney, and Jan Muntermann. 2013. A method for taxonomy development and its application in information systems. *European Journal of Information Systems* 22, 3: 336–359.
19. Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.
20. Phillipe Rodrigues, Ismael Santana Silva, Gl\'\ivia Angélica Rodigues Barbosa, Flávio Roberto dos Santos Coutinho, and Fernando Mourão. 2017. Beyond the Stars: Towards a Novel Sentiment Rating to Evaluate Applications in Web Stores of Mobile Apps. In *Proceedings of the 26th International Conference on World Wide Web Companion*, 109–117.
21. Emanuele Di Rosa and Alberto Durante. 2016. App2Check: a Machine Learning-based System for Sentiment Analysis of App Reviews in Italian Language. In *SIDEWAYS@ LREC*, 8–13.
22. Israel J Mojica Ruiz, Meiyappan Nagappan, Bram Adams, Thorsten Berger, Steffen Dienst, and Ahmed E Hassan. 2014. Impact of ad libraries on ratings of

android mobile apps. *IEEE Software* 31, 6: 86–92.

23. Israel Mojica Ruiz, Meiyappan Nagappan, Bram Adams, Thorsten Berger, Steffen Dienst, and Ahmed Hassan. 2017. An examination of the current rating system used in mobile app stores. *IEEE Software*.
24. Elmar Sauerwein, Franz Bailom, Kurt Matzler, and Hans H Hinterhuber. 1996. The Kano model: How to delight your customers. In *International Working Seminar on Production Economics*, 313–327.
25. Rajesh Vasa, Leonard Hoon, Kon Mouzakis, and Akihiro Noguchi. 2012. A preliminary analysis of mobile app user reviews. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*, 241–244.
26. Lorenzo Villarroel, Gabriele Bavota, Barbara Russo, Rocco Oliveto, and Massimiliano Di Penta. 2016. Release planning of mobile apps based on user reviews. In *Proceedings of the 38th International Conference on Software Engineering*, 14–24.
27. Phong Minh Vu, Tam The Nguyen, Hung Viet Pham, and Tung Thanh Nguyen. 2015. Mining user opinions in mobile app reviews: A keyword-based approach (t). In *Automated Software Engineering (ASE), 2015 30th IEEE/ACM International Conference on*, 749–759.
28. Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.