Beyond the Stars: Towards a Novel Sentiment Rating to Evaluate Applications in Web Stores of Mobile Apps

Phillipe Rodrigues
CEFET-MG
Brazil
caixeta.phillipe@gmail.com

Ismael Silva
CEFET-MG
Brazil
ismaelsantana@decom.cefetmg.br

Glívia Barbosa
CEFET-MG
Brazil
gliviabarbosa@decom.cefetmg.br

Flávio Coutinho
CEFET-MG
Brazil
coutinho@decom.cefetmg.br

Fernando Mourão
Seek AI Labs
Brazil
fernando.mourao@catho.com

ABSTRACT

This paper proposes an approach to evaluate mobile applications which complements the information provided by the number of stars and downloads in app stores. The goal is to provide novel information to assist users in the decision-making process regarding the choice of applications. In this sense, we conducted experiments to verify the relationship between the number of stars and the content of review comments. Results indicated that there is information in reviews not properly represented by stars. Thus, we present a sentiment rating generated automatically by aggregating opinions reported in the reviews related to each application. We evaluated this new rating using 26,996 reviews related to six applications present on the Google Play Store. The obtained results allow us to demonstrate that: (1) it is possible and feasible to generate a sentiment rating automatically and (2) the rating is useful for web stores of mobile applications to improve their mechanisms of ranking and recommendation as well as to assist users and developers to evaluate the quality and/or acceptance of the offered mobile applications.

Categories and Subject Descriptors

H. Information Systems: H.1 MODELS AND PRINCIPLES: H.1.2 User/Machine Systems: Human information processing.

Keywords

Sentiment Analysis, User Review, Web Stores, Mobile Apps, Decision-making, Machine Learning.

1 INTRODUCTION

The massive use of mobile applications has boosted the supply of applications (apps) in several categories, such as Communication Apps (e.g., WhatsApp, Telegram and Hangouts) and Navigation Apps (e.g., Waze, GoogleMaps and Navigator). Considering this new reality, users are challenged to choose the apps that best meet their needs [22, 13].

WWW 2017 Companion, April 3-7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054139

In order to assist users in this choice, app web stores (e.g., Google Play and Apple Store) allow users to evaluate and compare their products through mechanisms such as the number of stars, amount of downloads and textual reviews that describe the perceptions and/or user experiences with the applications [22, 28].

In practice, these mechanisms are useful tools that assist users in decision-making through easily read information. Indeed, the number of stars and the volume of downloads are straightforward information easily interpretable. However, the information contained in the reviews, despite intuitive, is not displayed in an aggregated manner. Hence, whether a user desires to take into account previous experiences of other users with the app, she/he should read each review individually. This task may be impractical due to the large amount of reviews usually related to each app [7, 6]. Thus, when users decide to use the reviews, they are able to manually evaluate only a small and unrepresentative sample.

According to Hoon et al. [6], due to the impracticability of manual analysis of the reviews, most users consider the number of stars as the main reference for decision-making. However, the number of stars may not reflect the underlying information contained in the reviews, which could assist users in the evaluation/choice of applications (e.g., user opinions that may be useful for choosing the mobile application [6]).

Additionally, Vasa et al. [26] and Hoon et al. [6] showed that the number of stars assigned to an app does not necessarily reflect the sentiment conveyed on the review comments. Such situations surface when two different users express similar perceptions on the comments, but assign a different number of stars each. For instance, while "user 1" comments "*This app is good*" and rates the app with 3 stars, "user 2" comments "*Good app*" but assigns 4 stars to it. This lack of consensus (or multiple individual biases) may result in a mean number of stars that does not represent the actual reviews related to each app [26, 6].

Aware of this issue, several researchers have shown the need to create mechanisms, complementary to the existing ones, that are able to properly summarize the user experiences expressed in review comments [20, 8, 2, 5, 11]. Such information might provide a sentiment rating that reflects opinions, experiences and perceptions reported on the messages and, consequently, assist users in assessing the applications without reading each review [7]. Thus, this sentiment rating can be used as criteria in the decision-making process of choosing applications in web stores, suppressing the impractical task of reading all user comments.

Motivated by this scenario, this work aims to assist users in the decision-making process regarding the choice of mobile applications, so that they can consider not only the number of

stars, but also the sentiment expressed in reviews. In this sense, we propose a sentiment rating generated from the automatic aggregation of user opinion in reviews available in app stores. This rating consists of a real number ranging from 1 to 5 and expresses the collective sentiment of users, wherein the closer to 5, the more positive.

As a practical use of this sentiment rating, we propose an intuitive display of it on the interfaces of mobile application stores, complementing the number of stars and downloads. In this sense, we present a visual metaphor based on emoticons. Furthermore, the information conveyed by a sentiment rating does not present the aforementioned problem of lack of consensus (multiple individual biases) [8, 26, 6, 20]. Whereas the number of stars is generated from the number of stars assigned by different users with different interpretations of the meaning of stars, the sentiment rating being proposed is automatically generated from a single bias learned from a labeled training data.

We validated our proposal through experimental analyses with 26,996 reviews, gathered from the Google Play Store, related to three categories of mobile applications. These analyses aimed to demonstrate that: (1) it is possible to extract, quantify and aggregate the opinion expressed in reviews; (2) the information contained in reviews differ from the information represented by the number of stars; (3) the extracted opinions may reflect in a sentiment rating useful to complement the qualifications of mobile apps; and (4) it is viable to automatically generate this sentiment rating using Sentiment Analysis techniques and Machine Learning [19].

Thus, we believe this work presents significant contributions, both practical and scientific. In practical terms, the work contributes not only to the web stores of mobile applications but also to other e-commerce systems. A proper collective sentiment allows these systems to enhance their ranking and recommendation methods, as well as to assist users and developers in the evaluation of the offered products/services. That is made possible because the sentiment rating automatically extracts and represents information from the comments, which aids user in the evaluation of the mobile applications. Without the approach being proposed here, this task could only be achieved through the reading of all the reviews.

In scientific terms, this work reinforces the applicability of Sentiment Analysis techniques and Machine Learning [19] to extract, quantify and summarize experiences and perceptions of users. In addition, this study demonstrates differences in data aggregation (e.g., number of stars) when presenting multiple biases (e.g., mean stars) and a single one (e.g., sentiment rating). Thus, developers and researchers should reflect on what type of information they wish to represent from these aggregations.

Finally, we highlight that the dataset generated during the experiments of this study (available at: *https://github.com/ismasantana/datasets*) can be exploited by other researchers as input for evaluations of Sentiment Analysis techniques.

2 RELATED WORKS

Reviews related to products and services on the Web have been explored under different perspectives [26, 6, 4, 20, 11, 2, 5]. Among the possible threads of investigation, we highlight works that investigate the relationship between the information contained in reviews and the number of assigned stars, in order to examine whether such stars represent properly what is expressed in the text. Recent studies have concluded that, despite reviews contain important information describing user experiences, the number of assigned stars does not reflect the sentiment expressed in reviews [26, 6]. Thus, users may express similar sentiments in reviews (e.g., enjoying the app) but assign a different number of stars for the same product and/or service.

As reviews on the Web related to products and/or services may express experiences and perceptions of users, another research direction widely explored refers to identifying and measuring quality attributes (e.g., accessibility, usability, user experience). In this direction, a study conducted by Korhonen et al. [11] argues that although relevant, most existing methods for evaluating user experience with products do not report relevant information about these experiences on a daily basis. According to the authors, a promising source of user experience reports is the list of reviews associated with the products. Motivated by this hypothesis, Korhonen et al. [11] investigated user experiences with products using solely the reviews. The authors manually analyzed reviews of products such as smartphones and MP3 players. The results support the hypothesis that reports contained in reviews comprise a rich source of information about user experience.

In turn, the work performed by Anam and Yeasin [2] infers the accessibility of mobile apps through user review comments. The authors collected reviews from 25 applications and proposed a system to detect automatically reviews related to accessibility, as well as the polarity of these reviews (i.e., positive or negative). The experiment was conducted considering accessibility from the perspective of users having low vision or blindness. The results indicated that the proposed system can be used to improve the ranking of apps and, therefore, it contributes to enhancing the user experience.

Hedegaard and Simonsen [5] conducted a study to investigate whether reviews related to software and video game contain information describing comments of two categories: usability and/or the user experience. The authors analyzed 5,198 reviews of 3,492 distinct products in order to verify whether the reviews matched one of those categories. Further, a vocabulary was generated for each category using the reviews assigned to each one. The results indicated that from 13% to 49% of the reviews contained information about usability and/or user experience. Hence, Hedegaard and Simonsen [5] came to conclude that reviews can be used as inputs to measure quality attributes.

Although the studies presented in this section indicate that user reviews contain relevant information that complements the evaluation of products and/or services on the Web, Korhonen et al. [11] and Anam and Yeasin [2] point out the need for further studies in this direction, which includes techniques to automate the review analysis for the extraction of attributes that may assist users evaluating products and/or services. In this sense, the present study differs from the others by proposing and demonstrating the usefulness and feasibility of generating a sentiment rating from reviews. We derive this rating by aggregating opinions reported in the reviews available in app stores. This sentiment rating aims to assist users in the decision-making process regarding the choice of mobile applications so that they can consider not only the number of stars but also the collective sentiment expressed in the reviews.

3 STUDY PREPARATION

In this study, we conducted experiments considering reviews from Google App Store (i.e., Google Play) [22]. We chose this platform because it hosts mobile applications for one of the most popular operating systems, Android [25].

Given the variety of mobile applications available, we defined a subset of target categories for our analysis. To that end, as performed by Platzer [21], we established popularity as the selection criteria for these categories and applications. We measure popularity by the number of downloads in Google Play Store, ignoring the category Games. We excluded Games from this study since in this category comments may be associated with personal experiences of the player (e.g., frustration due to game fails) rather than expressing opinion about the app quality (e.g., usability issues) [21].

Thus, we selected three categories of apps for analysis: communication, finance and social network. In turn, for each category, we selected the two most popular apps in the Google Play Store ranking of July 2015. Specifically, the selected apps were: (1) Facebook Messenger and WhatsApp as communication apps; (2) Bradesco and Caixa as financial apps, and (3) Facebook and Instagram as social network applications.

Upon completing these settings, we identified the data Google Play Store makes public about the selected applications and their respective reviews. We found that the store provides name, ID, average stars, description and developer of each application. Regarding the reviews, it provides content, rating (i.e., number of stars) and application code. We gathered the aforementioned data through a collector written in Java and using the Application Programming Interface (API) Android-Market-API. We collected a total of 26,996 comments, between 07/04/15 and 07/07/2015, distributed among the six applications, since the Google Play Store limits the amount of comments obtained via the API in 4,500 messages per application.

3.1 Dataset Characterization

Among the 26,996 collected comments, 4,499 are related to Facebook, 4,500 to Instagram, other 4,499 to WhatsApp, 4,499 to Facebook Messenger, 4,500 to Caixa and 4,499 to Bradesco.

The dataset contains 11,911 unique terms and Figure 1 depicts the frequency ranking of these terms. The most frequent term is "good", appearing 7,931 times, as showed in Table 1, which presents the top 10 most popular terms. There are 7,489 (62.87%) distinct words appearing only once in the collection, furthermore, nine of the ten most frequent terms express feelings about the experience and perception of the user with regard to using the apps.

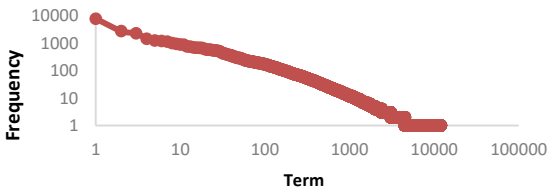

Figure 1. Ranking per term frequency.

In the next sections, we present the experiments performed with the data that was characterized in this section to demonstrate that: (1) it is possible to extract, quantify and aggregate the opinion expressed in reviews; (2) the extracted opinions may reflect in a sentiment rating useful to complement the evaluation of mobile apps; and (3) it is viable to automatically generate this sentiment rating using Sentiment Analysis and Machine Learning techniques.

Table 1. Most frequent terms in the collected data

Terms	Frequency
good	7,931
great	2,811
more	2,313
works	1,475
great	1,265
liked	1,219
better	1,158
cool	1,022
bad	962
loved	911

4 EXTRACTING SENTIMENT RATING FROM REVIEWS

We divide the first step of this study into two phases. The first one aims to demonstrate that there is underlying information in the reviews, not represented by the star rating, potentially useful to describe/evaluate mobile applications. In the second phase, we intend to extract, quantify and aggregate opinions, experiences and perceptions of users, expressed in the reviews, generating a sentiment rating to appraise applications.

4.1 Phase 1: Reviews vs. Assigned Stars

The main motivation to generate complementary mechanisms to evaluate mobile applications based on user reviews is that the existing mechanisms may not reflect the underlying information in these reviews [6, 20, 2, 5, 11]. Indeed, this gap arises when different users express similar feelings about a particular application but assign different numbers of stars. For instance, consider that two users have downloaded the WhatsApp application. While "*user 1*" has commented: *"This app is exceptional"* and assigned 4 stars to WhatsApp, "*user 2*" has commented: *"whatsapp is exceptional for me"* but assigned 5 stars to it. Thus, this lack of consensus (or existence of multiple individual biases) may result in a mean number of stars that does not represent the actual reviews related to each app [26, 6].

Thus, the first phase aims to assess the existence of a strong relationship between the content of the reviews and the number of stars. When this relationship is weak or does not exist, the content expressed in the reviews provides additional information to the star rating [19,1].

In this sense, we performed two experiments. In the first one, we used Machine Learning techniques to predict the number of stars representing the content expressed in the reviews. We evaluated the prediction quality by contrasting the predicted value against the actual number of stars assigned by the users. According to Pang et al. [19] and Alpaydin [1], the higher the prediction quality, the stronger the relationship between reviews and the number of stars.

In the second experiment, we calculate the entropy related to the vocabulary of terms occurring in reviews associated with each

number of stars [23]. The entropy calculation, according to Shannon [23], is obtained using Equation 1.

$$H = -\sum [P(a)] * \log_2 P(a) \quad (1)$$

In this equation, *H* denotes the entropy and *P(a)* represents the probability of a review to contain the term *a*. The lower the entropy, the stronger the relationship between the vocabulary of terms and the number of stars [23].

Before performing both experiments, we applied traditional text pre-processing steps to the collected data. Specifically, we removed accentuation, punctuation, stopwords and words with occurrence frequency lower than two. In addition, we converted all letters to lowercase and removed consecutive repeated ones (e.g., 'GOOOD' became 'good'). We addressed the star prediction as a multi-class text classification task and used the bag-of-words model [12]. Further, the number of stars assigned to reviews was used as the class of the instances [18].

In the first experiment, we used the WEKA implementations of Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Naïve Bayes, which are among the most effective and popular algorithms used for a range of classification domains. Indeed, SVM is deemed as the state-of-the-art classifier in several textual domains [3, 27, 14]. All results were found using 10-fold cross-validation [10] and Table 2 summarizes the mean accuracy found for all algorithms.

Table 2. Analysis of the accuracy in the star prediction task by different traditional text classifiers

App	# Reviews	# Terms	Naive Bayes	KNN	SVM
Bradesco	4500	1288	66.088%	63.222%	66.733%
Caixa	4500	1501	71.844%	70.222%	71.822%
Facebook	4499	1309	62.191%	59.479%	63.125%
Instagram	4500	942	78.222%	76.866%	79.266%
Messenger	4499	1187	63.880%	60.858%	64.614%
WhatsApp	4499	1187	74.883%	74.038%	75.816%

As shown in Table 2, the accuracy values for the star prediction task ranged from 60% to 79%. Compared to other tasks of text classification in the literature (e.g., spam filtering, automatically language detection, e-mail classification), these values are considered low to evince a high regularity of the data [19]. In this case, it means a weak relationship between the vocabulary used in the reviews and the number stars.

To confirm this conclusion, the second experiment was carried out by calculating the vocabulary entropy of reviews related to each number of stars. Table 3 presents the entropy values.

According to Shannon [23] and Islam et al. [9], the minimum value for entropy is zero, whereas its maximum value is given by $\log_2 |A|$ (i.e., log of the number of attributes). In our case, the maximum value is 13.54, since the evaluated collection presents 11,911 distinct terms. As the entropy values presented in Table 1 are closer to the maximum value than to the minimum one, we consider them high.

This result reinforces the conclusion of the first experiment. There is a weak relationship between reviews and number of stars. Hence, users who assign a certain amount of stars for a mobile application do not use a similar set of words to describe their experiences and opinions.

Table 3. Entropy of the term vocabularies used in reviews with the same number of stars.

	Entropy
5 stars	7.7682
4 stars	8.1823
3 stars	8.6318
2 stars	8.7750
1 star	8.7385

Our observations support the arguments presented by Hoon et al. [6], Hoon et al. [20], Hu et al. [8] Anam and Yeasin [2], Hedegaard and Simonsen [5] and Korhonen et al. [11]. These works argue that the number of stars may not represent properly the content of reviews, reinforcing the need for complementary mechanisms to evaluate user experiences and opinions about mobile applications.

4.2 Phase 2: Deriving a novel sentiment rating based on reviews

Motivated by the foregoing discussion presented in the last section, we propose a novel strategy to extract the information related to user experience and opinion hidden in reviews, consolidating a sentiment rating for mobile apps.

Towards this goal, first, we manually evaluated each of the 26,996 collected reviews and assigned a score, from 1 to 5, that represents the sentiment polarity of each review. The higher the score, the more positive the review. This manual classification was performed by two distinct users of mobile applications for Android (henceforth named readers). In a first round, each reader classified all reviews separately. In this process, each reader was asked to classify each review considering the following criteria:

- Sentiment Level 1 (I hated it): Reviews conveying comments of users who hated the application, considering it bad or reporting failures that prevent its usage (e.g. *"This application is very bad. It's been days since I cannot post anything! Too slow! Hated using it").*
- Sentiment Level 2 (I didn't like it): Reviews from users who did not like the application, considering it bad or describing problems that hinder its use (e.g. *"I do not recommend it because I cannot change my pictures").*
- Sentiment Level 3 (It's ok): Reviews about apps that fulfill their purpose without exceeding the user expectations (e.g., *"It helps me to organize my tasks"*).
- Sentiment Level 4 (I liked it): Reviews about apps that exhibit quality of use, encouraging its use (e.g. *"I like it. It is easy to use."*).
- Sentiment Level 5 (I loved it): Reviews about apps that exceed the user expectation (e.g. *"A great app to share pictures with family and friends!"*).

In a second round, to fix possible errors in the individual annotations arising from the painstaking and exhausting nature of the task and also from the possibility of different interpretations, the readers re-classified together all reviews whose classification in the first round had diverged. Among the 26,996 reviews analyzed, 24,625 (91.22%) were classified according to the sentiment level perceived by the readers. In turn, 2,371 (8.78%) were not categorized due to not exhibiting any information about

opinion or user with regard to the application. For instance, comments like *"I'm downloading and, if I like it, I'll give more stars to it."* were not associated with any sentiment level. The whole annotation process lasted five and a half weeks.

This result shows that the majority of the reviews contains information about opinion or related to the user experience. Further, by contrasting the manually classified reviews according to the sentiment against the number of stars, we observe that this information differs from the information summarized by the number of stars, such as shown in Figure 2.

This comparison evinces the difference in categorizing reviews using these two perspectives. Moreover, there was no equivalence between sentiment and number of stars to 6,516 (26.46%) of the reviews (e.g., a review with 3 stars received a sentiment score 5). Thus, the sentiment expressed in the reviews, do not always converge to the number of stars assigned to them.

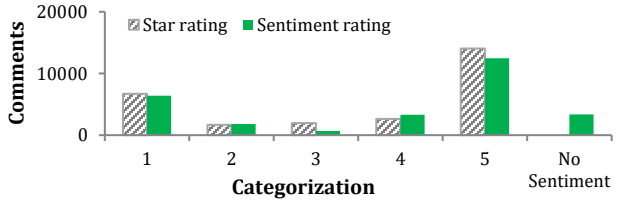

Figure 2. Distribution of comments by number of stars vs. sentiment rating

Aiming to address this issue, this paper proposes a sentiment rating, based on the aggregation of sentiments contained in the reviews of mobile apps. The sentiment score of each app rating range from 1 to 5 and we calculated it by averaging the sentiment attributed to its reviews, as shown in Equation 2.

$$s = \frac{\sum_c^C G(c)}{|C|} \quad (2)$$

where s denotes the aggregated sentiment score of the app; C is the set of input reviews related to the app; c is a review belonging to C; and $G(c)$ represents the sentiment score of c. The aggregated sentiment score is then mapped to a sentiment classification, as shown in Table 4.

Table 4. Correspondence between sentiment score and class

Aggregated Sentiment Score	Class
1 = *Aggregated Sentiment Score*	Hating it
1 < *Aggregated Sentiment Score* ≤ 2	Disliking it
2 < *Aggregated Sentiment Score* ≤ 3	It's ok
3 < *Aggregated Sentiment Score* ≤ 4	Liking it
4 < *Aggregated Sentiment Score* ≤ 5	Loving it

We calculated the proposed sentiment score for WhatsApp, Instagram and Bradesco (each one belonging to a distinct category defined in Section "Study Preparation") and the results are presented in Table 5.

Table 5. Sentiment Rating of the apps

App	Sentiment Rating	
	Aggregated	Sentiment Class
WhatsApp	4.5	Loving it
Instagram	4.7	Loving it
Bradesco	3.4	Liking it

In the next section, we conduct experiments to evince the usefulness of the proposed sentiment rating in practice.

5 USEFULNESS OF THE SENTIMENT RATING

We contrasted the proposed sentiment rating against two mechanisms for classifying/evaluating mobile apps. The goal was to identify similarities and complementarities with the results provided by a user experience (UX) evaluation, the number of stars and our sentiment rating.

User experience (UX) refers to how a person feels about using a system or service. It includes the practical, experiential, affective, meaningful and valuable aspects of human–computer interaction and product ownership. Additionally, it includes a person's perceptions of system aspects such as utility, ease of use and efficiency [16, 17]. User experience is dynamic as it is constantly modified over time due to changing usage circumstances and changes to individual systems as well as the wider usage context in which they can be found [16, 17]. True user experience goes far beyond giving customers what they say they want, or providing checklist features. In order to achieve high-quality user experience in a company's offerings there must be a seamless merging of the services of multiple disciplines, including engineering, marketing, graphical and industrial design, and interface design [16, 17].

Considering this definition, the comparison of sentiment rating with regard to the UX evaluation is relevant since the perceptions, experiences and feelings of users during the use of technology are related to UX principles adopted in interface design and interaction of products and services [16, 17]. Thus, the similarity between the UX evaluation and the sentiment rating proposed in this paper reinforces the consistency and usefulness of the sentiment rating.

It is important to emphasize that this work does not treat the UX evaluation and the sentiment rating as identical concepts. However, as demonstrated earlier, UX can influence user sentiment. That's because a bad UX can reflect on negative feelings during interaction as a product or service, as well as a good UX can reflect on positive feelings during the interaction. Thus, if the sentiment rating follows the same UX indicator pattern, found during the evaluation of the application, it is possible to demonstrate its consistency with UX and, consequently, its usefulness in the context of the qualification of applications in web stores [11, 5].

We conducted the UX evaluation using the Heuristic Evaluation method [15] following the UX guidelines proposed by Nielsen and Budiu [17]. We did this adaptation because, although this method was originally proposed to assess the usability of systems, its author argues that the heuristics set that guide its assessments can be adapted according to the evaluation goals [15].

In this study, two evaluators with experience in applying the Heuristic Evaluation conducted the UX evaluation during a period of 7 consecutive days, from 09/28/2015 to 10/04/2015. Further, the results were validated by an expert in Human Computer Interaction area (HCI) with over seven years of experience in performing this type of research.

In order to enable a quantitative comparison of the UX evaluation results with the other app evaluation mechanisms, we proposed a numerical UX score ranging from 0 to 5 which is derived from the total of UX problems identified, the amount of UX guidelines violated (i.e., the number of principles not found in the interface) and the incidence (i.e., frequency) of each guideline violated in relation to the total of problems.

Initially, we calculated the violation incidence, denoted by $V(D_x)$, for each UX guideline that had been violated (D_x) by at least one UX problem detected. It was obtained as the ratio of how frequently D_x was violated in the set of all the detected UX problems, such as shown in Equation 3:

$$V(D_x) = \frac{(n_{dx})}{P} \quad (3)$$

where n_{dx} denotes the frequency of D_x violations; and P represents the total number of UX problems identified in the app during the inspection.

Later, we calculated the percentage of violated guidelines (Δ), by averaging the violation incidences of the guidelines $V(D_x)$ as shown in Equation 4:

$$\Delta = \frac{\sum V(D_x)}{|D|} \quad (4)$$

where $|D|$ is the total number of guidelines used in the UX evaluation.

Finally, we assigned a score ranging from 0 to 5 (the higher the score, the better the user experience) to each app according to the UX evaluation results. The score N for each app was calculated according to Equation 5:

$$N = 5 - (5 * \Delta) \quad (5)$$

For the sake of discussion, Table 6 maps ranges of the score value to five different UX class labels.

Table 6. UX Classes

Score (N)	UX Class
$0 < N \leq 1$	Awful
$1 < N \leq 2$	Bad
$2 < N \leq 3$	Regular
$3 < N \leq 4$	Good
$4 < N \leq 5$	Great

5.1 Analysis of Results

We discuss the usefulness of the proposed sentiment rating based on the analysis of its correlation to UX evaluation and the number of stars related to apps in Google Play Store. Table 7, Table 8 and Table 9 summarize the evaluation of the selected apps considering these three perspectives.

Table 7. Evaluation of WhatsApp

Rating to WhatsApp	Score	Class
Sentiment	4.5	Loving it
Star	4.4	Great
UX evaluation	3.2	Good

Table 8. Evaluation of Instagram

Rating to Instagram	Score	Class
Sentiment	4.7	Loving it
Star	4.5	Great
UX evaluation	3.5	Good

Table 9. Evaluation of Bradesco.

Rating to Bradesco	Score	Class
Sentiment	3.4	Liking it
Star	4.3	Great
UX evaluation	3.5	Good

By contrasting the results, we observe that Bradesco presented a sentiment class equivalent to the UX class. In turn, although WhatsApp and Instagram apps do not present equivalent classes for the sentiment rating and the UX evaluation, both ratings indicate a positive feedback. Such difference may be related to the fact that the UX evaluation considers interface design details that impact the interaction but normally are not made explicit in comments (e.g., color and alignment of buttons on the interface) [15, 16, 17].

Thus, by contrasting sentiment rating against UX evaluation we found a correlation between the results. This observation is relevant since it reinforces that the aggregate sentiment reported in reviews may reflect the experiences and perceptions of users, such as argued by Hedegaard and Simonsen [5].

Regarding the app classification on Google Play Store, the sentiment rating was equivalent for WhatsApp and Instagram, whereas we identified a less positive feedback than the number of stars for Bradesco. Despite this discrepancy, sentiment rating and number of stars do not diverge about the polarity of the feedback, since both identified a positive feedback. Further, the sentiment rating presents additional information since it is equivalent to the UX evaluation, a manual, relevant and expert-generated evaluation method not existing in Google Play. This fact reinforces the usefulness of a sentiment rating as supplementary information for the evaluation of mobile applications since the combination of these two perspectives may reflect both quality and sentiment polarity of users.

Thus, the sentiment rating is useful, since the decision about whether or not use an application (in the view of users) and implementation enhancements (in the view of developers) would be supported not only by the number of downloads and stars, but also by the polarity of the contents expressed in reviews, enriching the selection criteria.

6 FEASIBILITY OF AUTOMATIC SENTIMENT RATING ESTIMATION

Given the impossibility of manually analyzing all reviews related to each application to generate the sentiment rating, the last step of this study aims to verify the effectiveness of Sentiment Analysis and Machine Learning techniques to infer, from reviews, the sentiment rating of each app [19]. The premise is that the higher the success rate of automatic classifiers, the higher the feasibility of automatic sentiment rating estimation.

To conduct this step, we applied to the reviews classified by the two readers the same pre-processing steps described in Section "Reviews vs. Assigned Stars". Further, we used the algorithms Naïve Bayes, KNN and SVM to infer the sentiment rating and evaluated their effectiveness following a 10-fold cross validation process [10]. Again, we used accuracy as the measure of success rate. The classification results are shown in Table 10. We note that SVM presented the highest success rate among all classifiers, especially for Instagram exhibiting 83.3% of accuracy.

To demonstrate the efficiency of this approach, we contrasted the success rate related of predicting the sentiment class against the hit rate to infer the number of stars, presented in Section "Extracting sentiment rating from reviews". Through this analysis, we verify whether the relationship between reviews and the categorization of sentiment is stronger than the relation between reviews and the number of stars assigned by the reviewers. Whether the former is

stronger, it means that the sentiment rating better reflects the content expressed in the reviews. The results are shown in Figures Figure 3, Figure 4 and Figure 5.

Table 10. Automatic classifiers success rate to infer the sentiment rating.

App	# Reviews	# Terms	Naive Bayes	KNN	SVM
Bradesco	4180	1182	73.038%	72.272%	**75.861%**
Caixa	4167	1398	78.617%	79.793%	**81.209%**
Facebook	4037	1140	67.525%	67.872%	**72.058%**
Instagram	4259	856	81.263%	82.249%	**83.305%**
Messenger	3978	1001	68.175%	69.381%	**71.694%**
WhatsApp	4004	1018	75.999%	76.673%	**78.721%**

Figure 3. Comparison of the success rate of Naive Bayes to infer the number of stars and sentiment rating.

Figure 4. Comparison of the success rate of KNM to infer the number of stars and sentiment rating.

Figure 5. Comparison of the success rate of SVM to infer the number of stars and sentiment rating.

The charts in Figure 3, Figure 4 and Figure 5 demonstrate that the success rate of all classifiers to infer the sentiment rating was higher than to infer the number of stars. Table 11 shows the improvement in the success rate of the sentiment rating prediction over the star rating prediction.

These results evince a stronger relationship between the content of reviews and the sentiment rating than between the reviews and the number of stars. When inferring the sentiment class, the same classifiers presented up to 14% of enhancements on the success rate.

Therefore, by replacing multi-bias signals of feedback by a single-bias, defined by the readers, we decrease the subjectivity in categorizing reviews. Hence, the approach proposed in this work better discriminates the experiences and perceptions of users reported reviews. Also, the success rates exhibited by the classifiers in our collection reinforce the feasibility of generating the sentiment rating automatically for mobile apps.

Table 11. Improvement of inferring the sentiment rating over the inference of the number of stars

App	Naive Bayes	KNN	SVM
Bradesco	10.52%	14.32%	13.68%
Caixa	9.43%	13.63%	13.07%
Facebook	8.58%	14.11%	14.15%
Instagram	3.89%	7.00%	5.10%
Messenger	6.72%	14.01%	10.96%
WhatsApp	1.49%	3.56%	3.83%

7 USAGE SCENARIOS FOR SENTIMENT RATING

The content expressed in reviews can be explored and summarized in different manners at the interfaces of web stores of mobile app. For instance, this information may be presented through tag clouds of the most common terms, summary of reviews or highlighting the most relevant comments [8, 11, 5]. To ease the interpretation of the sentiment rating by users, we propose to present it on the interfaces through visual metaphors familiar to users. Specifically, we decided to explore the emoticons for this purpose.

Figure 6 presents (a) examples of possible graphical representations for the sentiment classes related to our sentiment rating; and (b) examples of how the sentiment rating can be displayed on the Google Play Store interface to help users and developers in the evaluation of the quality/acceptance of apps.

Hating it **Disliking it** **It's ok** **Liking it** **Loving it**

Figure 6 (a). Possible representations for the sentiment rating.

Figure 6 (b). Prototype of utilization on Google Play Store's interface.

Figure 6. Graphical representations for the sentiment rating

8 CONCLUSIONS AND FUTURE WORKS

This paper presents a novel approach to complement existing mechanisms that evaluate mobile apps. In practice, such mechanisms are useful to assist users in the decision-making process regarding the choice of apps. First, we conducted experiments to verify the relationship between the number of stars and the content of reviews. By observing that there is information in the reviews not properly represented by the number of stars, we proposed a sentiment rating 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. This novel rating can be incorporated into the app web store's interface as a parameter, complementing the number of stars and downloads. Thus, users can acquire the collective sentiment of each app without having to read each review individually.

Thus, this work stands out due to its practical and scientific contributions. In scientific terms, we reinforce the applicability of Sentiment Analysis and Machine Learning techniques to extract and quantify user experiences contained in reviews. In this sense, this research supports initiatives to explore the use of these techniques, outlining their advantages and disadvantages to evaluate distinct technologies with regard to quality attributes (e.g., usability, accessibility, user experience). Also, this work emphasizes the differences between aggregating information with multi-bias (e.g., number of stars) and a single bias (e.g., sentiment rating). Thus, developers and researchers should reflect on what type of information they wish to aggregate.

In practical terms, the work contributes to both web stores of mobile application and other e-commerce systems with a complementary quality indicator of their products/services. This novel information is also useful for these stores/systems to enhance their ranking and recommendation systems, as well as to assist users and developers in the evaluation of products/services offered. As the sentiment rating automatically extracts and represents the information present on the comments, it aids users in their evaluation of the mobile applications, which would not be feasible otherwise.

Finally, we highlight that the dataset generated during the experiments (available at: *https://github.com/ismasantana/datasets*) can be exploited by other researchers as input for evaluation of algorithms for Sentiment Analysis.

As future work, the proposed sentiment rating can be contrasted against other review summarization proposals. Another relevant direction is to investigate the feasibility of reducing the cost of review labeling for composing the training set through active and semi-supervised Machine Learning methods [24].

9 REFERENCES

[1]. Ethem Alpaydin. 2004. Introduction to Machine Learning (Adaptive Computation and Machine Learning). The MIT Press.

[2]. Asm Iftekhar Anam and Mohammed Yeasin. 2013. Accessibility in smartphone applications: what do we learn from reviews?. In *Proc. of the 15th International ACM SIGACCESS (ASSETS '13)*, 2 pages.

[3]. Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory (COLT '92)*. 144-152.

[4]. Kavita Ganesan and Chengxiang Zhai. 2012. Opinion-based entity ranking. Inf. Retr. 15, 2 (April 2012), 116-150.

[5]. Steffen Hedegaard and Jakob Grue Simonsen. 2013. Extracting usability and user experience information from online user reviews. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*, 2089-2098.

[6]. Leonard Hoon, Rajesh Vasa, Gloria Yoanita Martino, Jean-Guy Schneider, and Kon Mouzakis. 2013. Awesome!: conveying satisfaction on the app store. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration (OzCHI '13)*, 229-232.

[7]. Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04)*, 168-177.

[8]. Nan Hu, Paul A. Pavlou, and Jennifer Zhang. 2006. Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of Online word-of-mouth communication. In *Proceedings of the 7th ACM conference on Electronic commerce (EC '06)*, 324-330.

[9]. Zahurul Islam, Md. Rashedur Rahman, and Alexander Mehler. 2014. Readability Classification of Bangla Texts. In *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing* (CICLing 2014), Alexander Gelbukh (Ed.), Vol. 8404. Springer-Verlag New York, Inc., New York, NY, USA, 507-518.

[10]. Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on Artificial intelligence (IJCAI'95)*, Vol. 2. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1137-1143.

[11]. Hannu Korhonen, Juha Arrasvuori, and Kaisa Väänänen-Vainio-Mattila. 2010. Let users tell the story: evaluating user experience with experience reports. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems (CHI EA '10)*, 4051-4056.

[12]. Andrew McCallum and Kamal Nigam. 1998. A comparison of event models for Naive Bayes text classification. In *Proc. of the AAAI-98 Workshop on Learning for Text Categorization*, pages 41--48

[13]. Aliaksei Miniukovich and Antonella De Angeli. 2016. Pick me!: Getting Noticed on Google Play. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*, 4622-4633.

[14]. Renata F. P. Neves, Cleber Zanchettin, and Alberto N. G. Lopes Filho. 2012. An efficient way of combining SVMs for handwritten digit recognition. In *Proceedings of the 22nd international conference on Artificial Neural Networks and Machine Learning - Volume Part II (ICANN'12)*, 229-237.

[15]. Jakob Nielsen. 1993. Usability Engineering. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[16]. Jakob Nielsen and Don Norman. 2010. The Definition of User Experience. Retrieved August 22, 2016 from http://www.nngroup.com/articles/definition-user-experience.

[17]. Jakob Nielsen and Raluca Budiu. 2015. User Experience for Mobile Applications and Websites. In Design Guidelines. 3a Ed., 506 pages.

[18]. Bo Pang and Lillian Lee. 2005. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05), PA, USA, 115-124.

[19]. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10 (EMNLP '02),* Vol. 10. Association for Computational Linguistics, Stroudsburg, PA, USA, 79-86. http://dx.doi.org/10.3115/1118693.1118704

[20]. Dae Hoon Park, Mengwen Liu, ChengXiang Zhai, and Haohong Wang. 2015. Leveraging User Reviews to Improve Accuracy for Mobile App Retrieval. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*, 533-542.

[21]. Elisabeth Platzer. 2011. Opportunities of automated motive-based user review analysis in the context of mobile app acceptance. In *Proceedings of the 22nd Central European Conference on Information and Intelligent Systems (CECIIS'11)*, 309-316.

[22]. Google Play. Retrieved August 13, 2016 from https://play.google.com/store/apps

[23]. Claude Elwood Shannon. 1948. A mathematical theory of communication. Bell System Technical Journal, 27, 379-423.

[24]. Ismael Santana Silva, Janaína Gomide, Adriano Veloso, Wagner Meira, Jr., and Renato Ferreira. 2011. Effective sentiment stream analysis with self-augmenting training and demand-driven projection. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR '11)*, 475-484.

[25]. Mobile/Tablet Top Operating System Share Trend. Retrieved August 16, 2016 from https://goo.gl/Ne1S1N

[26]. 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 (OzCHI '12)*, Vivienne Farrell, Graham Farrell, Caslon Chua, Weidong Huang, Raj Vasa, and Clinton Woodward (Eds.), 241-244. http://dx.doi.org/10.1145/2414536.2414577

[27]. Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. 2007. Top 10 algorithms in data mining. Knowl. Inf. Syst. 14, 1 (Dec 2007), 1-37.

[28]. Peifeng Yin, Ping Luo, Wang-Chien Lee, and Min Wang. 2013. App recommendation: a contest between satisfaction and temptation. In *Proceedings of the sixth ACM international conference on Web search and data mining (WSDM '13),* 395-404.