A Measurement Design for the Comparison of Expert Usability Evaluation and Mobile App User Reviews

Necmiye Genc-Nayebi and Alain Abran

Department of Software Engineering and Information Technology,
Ecole de technologie superieure (ETS) – University of Quebec (Montreal, Canada)

necmiye.genc.1@ens.etsmtl.ca,alain.abran@etsmtl.ca

Abstract. Usability and user experience (U&UX) as important components of software quality are now more critical than ever for mobile app store success. Usability experts use different protocols to evaluate the usability of mobile apps while app store user reviews also produce valuable related information. Our research study proposes a measurement design to compare user reviews and expert-based usability evaluation results that includes an exploratory analysis and topic modeling of user reviews. This design is structured to investigate whether mobile app usability features extracted from user reviews align with subject-matter expert usability evaluation results.
Keywords: Usability measurement, Measurement Design, App Store Reviews, Text Mining

1 Introduction

Usability evaluations have been performed traditionally by subject-matter experts and end users, usually in laboratory and field contexts. However, such evaluations can cover only a limited time-span of the applications and re-doing the same usability evaluation for all available app versions would typically lead to large evaluation costs. Furthermore, evaluating some usability dimensions, such as understandability or learnability, requires more complex procedures and indicators [1]. In addition, studies have demonstrated that both experts and end-users are effective in revealing different usability problems [2,3].

With the advent of mobile ecosystems including mobile apps and related meta-data such as ratings and user reviews, app stores now contain a wealth of information about user experience and expectations. However, it is difficult to manually extract this information due to various factors such as the large quantity of reviews, their lack of structure and varying quality.

In this paper, we present a measurement design to compare the findings from subject-matter expert usability evaluations and corresponding app store user reviews. The measurement design proposes two different approaches: one is through an exploratory analysis and the other is through, first, a semi-supervised topic modeling to extract

usability aspects from user reviews and next, comparing these findings with the results from a prior expert-based usability evaluation. To the best of our knowledge, our work is the first to propose topic modeling techniques to automatically extract usability and user experience (U&UX) information from app store user reviews and to compare usability evaluation results of experts and end-users.

2 Related Work

Usability evaluation of mobile apps is an emerging research area that faces a variety of challenges due to the limitations of mobile devices such as processing capacity, screen size, connectivity, and a lack of a consensus on a usability evaluation methodology [4]. Over the years, different methods and techniques have been proposed for usability evaluation. The leading traditional methods fall into two main categories: inspection methods without end users and test methods with end users [5].

App stores are valuable repositories of app and user data where app users can give feedback about different aspects of an app such as its functionality, design or value. A previous study has reported that app store user reviews are valuable to understand user experience and usability aspects [6], while another study reported that 13%-49% of the content of user reviews contains U&UX information that could be used to improve the software quality [7]. However, these reviews permit a limited number of studies on end-user evaluation of usability. For example:

- mining the app store review corpus identified nine different classes of feedback: positive, negative, comparative, price related, missing requirements, issue reporting, usability, customer supports and versioning [8];
- using a support vector machine (SVM) algorithm to classify five main dimensions of usability: memorability, learnability, efficiency, errors/effectiveness and satisfaction [9].

However, we could not identify any related work comparing user reviews and expert-based usability evaluation results.

3 Measurement Design

The first objective of the proposed measurement design is to extract usability related information from user reviews. The second objective is to compare expert-based usability results with user usability evaluation through reviews.

The overview of the proposed measurement design is presented in Figure 1 where Parts 1 and 2 address the first research objective and Part 3 the second research objective. Pre-defined usability keyword frequency analysis of a corpus of reviews was performed in Part 1, while topic modeling was performed in Part 2 to automatically extract usability related topics. The outputs obtained in Parts 1 and 2 and prior subject-matter expert usability evaluation findings are compared in Part 3. The details of the proposed measurement design are presented in the following sub-sections.

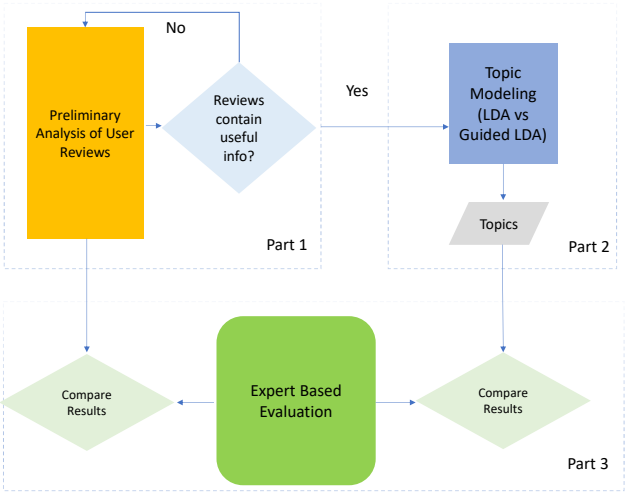

Fig. 1. Overview of the measurement design

3.1 Part 1. Preliminary Analysis of User Reviews for a Set of Apps

In the first part of the research, a preliminary data analysis was performed to discover usability related keyword frequencies in the review corpus through the following four steps:

- Step 1. Selection of apps with the information available to build a review corpus. A reference review corpus was identified and selected, which contained for the same apps both the results of expert-based usability evaluation and user reviews.
- Step 2. Converting usability attributes to aspect words. After the review corpus was populated for the selected versions of the apps, the usability attributes or heuristics were converted into a bag-of-words (BOW). Step 3. Stemming. The Porter stemming algorithm [10] was run to remove affixes from the words and then stemmed versions of the aspect words were searched in the review corpus.
- Step 4. Querying the review corpus. The stemmed words were next queried within individual review corpuses per app and their term frequencies recorded. Query results (e.g., usability aspect term frequencies) were analyzed to understand if (i) user reviews convey good information about usability aspects and (ii) user reviews align with expert-based usability evaluation results.

3.2 Part 2. Usability Topic Modeling

In part 2 of the measurement design, a topic modeling technique was used to help identify individual topics in the document and understand the document corpora in an automated manner. However, unsupervised topic models often lead to topics that are not completely meaningful and/or topics discovered in an unsupervised way that may not

match the true topics in the data. To address this limitation, we leveraged the guided latent Dirichlet allocation (LDA) topic model [11] given in Eqs. 1 to 3 that use Gibbs sampling as an inference method and usability related seed words to improve topic-word distribution.

Step 1: For $k = 1 \ldots \ldots K$: (1)

(a) Choose regular topic $\phi^k{}_r \sim Dirichlet(\beta\_r)$

(b) Choose seed topic $\phi^k{}_s \sim Dirichlet(\beta_s)$

Step 2: For each seed set $s = 1 \ldots \ldots S$, (2)

(a) Choose group-topic distribution $\psi_d \sim Dirichlet(\alpha)$

Step 3: For each document d: (3)

(a) Choose a binary vector $\vec{b}$ of length S

(b) Choose a document-group distribution $\zeta^d \sim Dirichlet(\tau\vec{b})$

(c) Choose a group variable $g \sim Multinominal(\zeta^d)$

(d) Choose $\theta_d \sim Dirichlet(\psi_g)$

(e) For each token $i = 1 \ldots \ldots N_d$,

(1) Select a topic $z_i \sim Multinominal(\theta_d)$
(2) Select an indicator $x_i \sim Binomial(\pi_{zi})$
(3) If x_i is 0
Select a word $w_i \sim Multinominal(\theta_{zi}{}^r)$ //choose from LDA style topic
(4) If x_i is 1
Select a word $w_i \sim Multinominal(\theta_{zi}{}^s)$

The generative process for a document collection D under the guided LDA model is as follows – see Figure 2:

1. First, the T topic-word distribution ϕ^k and group-topic distribution ψ_{ds} were generated.
2. Then for each document, a list of seed sets allowed for the document, represented as a binary vector $\vec{b}$, was generated, and then
3. $\vec{b}$ was populated based on the document words, and hence treated as an observed variable.

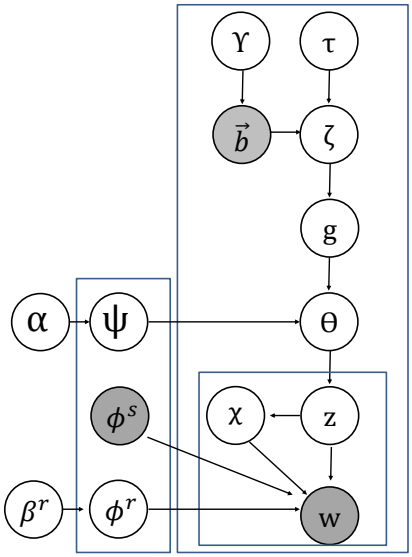

Fig. 2. Graphical model representation of a guided LDA [12]

Step 1. Pre-Processing on Review Corpus.

The review corpus generated in Part 1 was used in the topic modeling. To reduce the dimensionality of the document term matrix, certain data pre-processing and cleaning steps were carried out before proceeding with topic modeling. This pre-processing consisted of:

I. tokenization to segment the review corpus into its atomic elements using the Natural Language Toolkit (NLTK) *tokenize.regexp* module;
II. lower case conversion; and
III. stop-word (e.g. 'the', 'and', 'or', 'a'), punctuation and non-alphabetic phrase removal.

Step 2. Guided LDA Modeling

The guided LDA library1 developed in Python was used in this study. The document term matrix that was generated in the pre-processing step was given as an input to the LDA model. The training step required the input parameters, such as seed topics, seed confidence, be set at 0.15 to bias the seeded words by 15% towards the seeded topic, the number of topics be set at 5, 10 and 20, chunksize at 2000, refresh at 20 and iterations at 100.

[1] https://github.com/vi3k6i5/GuidedLDA

Step 3. Model Accuracy

The guided LDA model performance was tested on a complete review corpus where (i) the number of topics K = 5, 10 and 15 and (ii) LDA model was taken as the baseline. The topic model was run on individual corpuses per app.

3.3 Part 3. Evaluation of Results

In this part of the research, usability aspects extracted from user reviews were compared with expert-based usability evaluation results for the most frequent and the less frequent usability attributes:

- If the user reviews had more positive associations for the usability term, its user review evaluation rating were accepted as *positive*, corresponding to 4-5 stars given by usability experts in prior evaluations.
- If the user reviews had more negative associations for the usability term, its user review evaluation rating was accepted as *negative*, corresponding to 1-2 stars given by usability experts in prior evaluations.
- Equal numbers of positive and negative reviews were given a *neutral* evaluation rating, corresponding to 3 stars given by usability experts in prior evaluations.

4 Work in Progress and Future Work

In this proposed measurement design, usability aspects for expert-based usability evaluation questionnaires were first extracted and then converted into a BOWs in order to trace them back in user reviews. In addition, a guided LDA topic modeling was developed to automatically capture usability aspects intrinsic to the review texts. Expert-based usability evaluation results were compared with user evaluations expressed through reviews and the identified alignment and differences reported. We believe that this proposed measurement design is useful for supporting developers, U&UX designers and researchers to better understand user experience and opinion on mobile application usability aspects, which, finally, can lead to improved software quality.

In future work, we will explore the performance of our guided LDA topic model vs LDA topic model as baseline. Our topic model will also be run on individual review corpuses per app to find the percentage of clustered words that are directly related to usability and usability aspects such as efficiency, errors/effectiveness, etc. Next, for the top 10 and 10 lowest frequency terms identified with preliminary analysis in Part 1 and topic modeling in Part 2, user and expert evaluation ratings will be compared to determine possible alignments or differences between two different usability evaluation methods. A reference review corpus has already been selected which contains for the same apps both the results of expert-based usability evaluation as well as user reviews. This is the app dataset from [13] that includes a set 99 mobile apps evaluated by three usability experts. Since 19 out of the 99 apps from the study are no longer available in the Apple app store and there is no review available for five (5) other mobile apps

within this reference set, our review corpus will be populated with 75 mobile app selected versions of the app dataset.

References

1. T. Grossman, G. Fitzmaurice, and R. Attar, "A survey of software learnability: metrics, methodologies and guidelines," *SIGCHI Conference on Human Factors in Computing Systems*, 2009, pp. 649–658.
2. P.-Y. Yen and S. Bakken, "A Comparison of Usability Evaluation Methods: Heuristic Evaluation versus End-User Think-Aloud Protocol - An Example from a Web-based Communication Tool for Nurse Scheduling," *AMIA Annual Symposium Proceedings,* 2009, pp. 714–8.
3. L. Hasan, A. Morris, and S. Probets, "A Comparison of Usability Evaluation Methods for Evaluating E-Commerce Websites," *Behaviour & Information Technology*, July 2012, vol. 31, no. 7, pp. 707–737.
4. D. Zhang and B. Adipat, "Challenges, Methodologies, and Issues in the Usability Testing of Mobile Applications," *International Journal of Human–Computer Interaction*, 2005, vol. 18, no. 3, pp. 293–308.
5. A. Holzinger, "Usability Engineering Methods for Software Developers," *Communications of the ACM*, Jan. 2005, vol. 48, no. 1, pp. 71–74.
6. N. Genc-Nayebi and A. Abran, "A systematic literature review: Opinion mining studies from mobile app store user reviews", *Journal of System & Software*, 2017, vol. 125, no. Supplement C, pp. 207–219.
7. S. Hedegaard and J. G. Simonsen, "Extracting Usability and User Experience Information from Online User Reviews," *SIGCHI Conference on Human Factors in Computing Systems*, 2013, New York, NY, pp. 2089–2098.
8. C. Iacob and R. Harrison, "Retrieving and Analyzing Mobile Apps Feature Requests from Online Reviews," *10th Working Conference on Mining Software Repositories*, Piscataway, NJ, 2013, pp. 41–44.
9. E. Bakiu and E. Guzman, "Which Feature is Unusable? Detecting Usability and User Experience Issues from User Reviews," *IEEE 25th International Requirements Engineering Conference Workshops (REW)*, 2017, pp. 182–187.
10. M. F. Porter, "Readings in Information Retrieval," K. Sparck Jones and P. Willett, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, pp. 313–316.
11. J. Jagarlamudi, H. Daumé III, and R. Udupa, "Incorporating Lexical Priors into Topic Models," *13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, Stroudsburg, PA, pp. 204–213.
12. D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent dirichlet allocation," *Journal of Machine Learning Research*, 2003, vol. 3, p. 2003.
13. F. Nayebi, "iOS application user rating prediction using usability evaluation and machine learning," PhD Thesis, École de technologie supérieure, University of Quebec, Montreal, Canada, 2015, p. 203.