2017 IEEE/ACM 39th International Conference on Software Engineering

Recommending and Localizing Change Requests for Mobile Apps based on User Reviews

Fabio Palomba[1], Pasquale Salza[2], Adelina Ciurumelea[3], Sebastiano Panichella[3],
Harald Gall[3], Filomena Ferrucci[2], Andrea De Lucia[2]
[1]Delft University of Technology, The Netherlands — [2]University of Salerno, Italy — [3]University of Zurich, Switzerland

***Abstract*—Researchers have proposed several approaches to extract information from user reviews useful for maintaining and evolving mobile apps. However, most of them just perform automatic classification of user reviews according to specific keywords (e.g., bugs, features). Moreover, they do not provide any support for linking user feedback to the source code components to be changed, thus requiring a manual, time-consuming, and error-prone task. In this paper, we introduce CHANGEADVISOR, a novel approach that analyzes the structure, semantics, and sentiments of sentences contained in user reviews to extract useful (user) feedback from maintenance perspectives and recommend to developers changes to software artifacts. It relies on natural language processing and clustering algorithms to group user reviews around similar user needs and suggestions for change. Then, it involves textual based heuristics to determine the code artifacts that need to be maintained according to the recommended software changes. The quantitative and qualitative studies carried out on $44\,683$ user reviews of 10 open source mobile apps and their original developers showed a high accuracy of CHANGEADVISOR in (i) clustering similar user change requests and (ii) identifying the code components impacted by the suggested changes. Moreover, the obtained results show that CHANGEADVISOR is more accurate than a baseline approach for linking user feedback clusters to the source code in terms of both precision ($+47\%$) and recall ($+38\%$).**

***Index Terms*—Mobile Apps; Mining User Reviews; Natural Language Processing; Impact Analysis**

I. INTRODUCTION

Nowadays, the development and the release planning activities moved from a traditional paradigm, in which software systems are released following a clearly defined road map, towards a paradigm in which continuous releases become available for an upgrade every week [1]–[4]. It is particularly true in the case of mobile apps, where developers manage updates (i.e., new features, enhancements, or bug fixes) through online app stores, such as *Google Play Store*, *Apple Store*, and *Windows Phone App Store* [3], [5], [6].

This kind of distribution is accompanied by mechanisms which allow end users to evaluate releases by using scores, usually expressed as five stars values and user reviews. These reviews are free text that may informally contain relevant information for the development team [7] such as bugs or issues that need to be fixed [5], summaries of the user experience with certain features [8], requests for enhancements [9], ideas for new features [5], [10], and comparison with other apps.

Thus, not only do user reviews represent the simplest and fastest way end users have to express their opinions or report their suggestions, but also a powerful crowd feedback mechanisms that can be used by developers as a backlog for the development process [11], [12], aiming to improve the success/distribution of their apps [5], [13], [14].

The main problem for developers is that existing app distribution platforms provide limited support to systematically filter, classify, and aggregate user feedback to derive requirements [15]. Moreover, manually reading each user review to gather the useful ones is not feasible considering that popular apps (e.g., Facebook) receive hundreds of reviews every day [5], [16]. To address the problem and reducing the manual effort, in the recent past the research community has proposed approaches to select the useful feedback from user reviews [9], [14], [17]–[19], essentially consisting of performing only an automatic classification of the review content according to specific keywords without considering sentence structures and semantics. Moreover, the information that can be gathered is restricted to user reviews [20], and there not exists any systematic way for linking user feedback to the related source code components to change, a task that requires an enormous manual effort and is highly error-prone.

To better illustrate the problem statement, let us consider the following scenario. *John* is a mobile developer of the `FrostWire` app, an open source BitTorrent client for Android, a quite popular app on Google Play Store. *John* frequently checks the numerous user reviews the app receives on a daily basis. Not only do they contain compliments, but also some informative feedback such as bug reports or feature requests. He needs to read reviews carefully and carry out a time-consuming activity of analysis and synthesis in order to identify and collect possible suggestions for improvement. Once *John* detects an informative feedback, he has to locate the source code components related to the requested changes. This is not always easy, even if he is the main developer of the application. Indeed, user reviews are usually informal text written by non-technical users, therefore poorly understandable and/or not containing enough information to match the corresponding source code components correctly. For instance, *John* finds a user review verbatim reporting: *"I can't download most off the songs."*. Besides the spelling mistakes, *John* realizes that the user is complaining about download problems, but he is not sure that this is an actual bug or maybe the user is not able to use his app correctly. Thus, he skips the review and continues to read the other ones. Afterward, he finds other negative user reviews mentioning and detailing the same problem as the one above. As an example, a user claims that he is not able

1558-1225/17 $31.00 © 2017 IEEE
DOI 10.1109/ICSE.2017.18

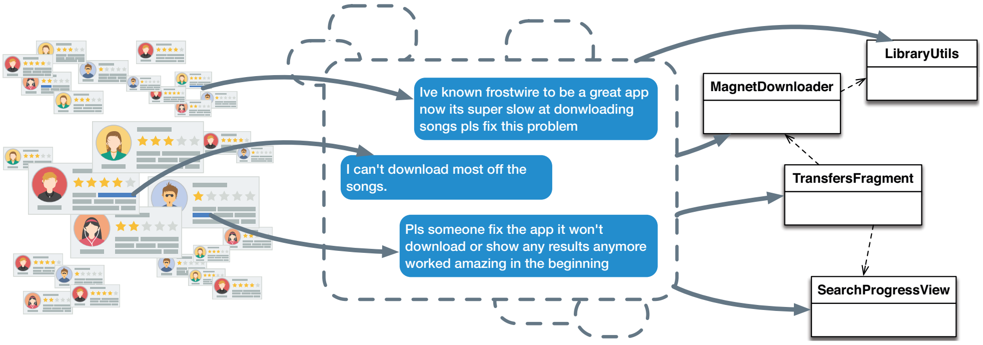
Ive known frostwire to be a great app now its super slow at donwloading songs pls fix this problem
I can't download most off the songs.
Pls someone fix the app it won't download or show any results anymore worked amazing in the beginning
LibraryUtils
MagnetDownloader
TransfersFragment
SearchProgressView

Figure 1: ChangeAdvisor – Motivating example.

to download songs through magnet links. At this point, *John* understands that this is a real bug by reading the collection of these user reviews and identifies a possible solution.

As an alternative, *John* might have used existing tools. For instance, he might have exploited SURF [19], [21] to summarize reviews or URR [22] and CLAP [23] to prioritize groups of reviews. Nevertheless, he would not be able to (i) extract only the useful information hidden behind different user reviews, (ii) group together fine-grained information, (iii) understand the actual impact of each change request.

To overcome these limitations, in this paper we introduce ChangeAdvisor, a novel approach that facilitates the analysis of the enormous amount of user reviews developers receive on a daily basis, by automatically (i) extracting user feedback relevant from a maintenance perspective, (ii) clustering them based on similar user needs, and (iii) identifying the set of source code components that need to be changed to accommodate the user requests.

By using our approach, the work done by *John* in the previous scenario would have been drastically reduced. Indeed, *John* can download the list of user reviews from the developer console of *Google Play Store* and give them as input to ChangeAdvisor, together with the source code of his app. Following the steps described in Section III, our approach extracts only user feedback useful from a maintenance perspective and groups entries together into clusters representing similar user needs. Then, ChangeAdvisor locates the specific source code components that would likely have to be changed. Figure 1 depicts the way ChangeAdvisor supports the specific task of *John*. In particular, the user feedback is extracted and clustered in a group of sentences which describe a (i) slow download performance and (ii) the inability in showing the transfer results.

Then, ChangeAdvisor links the grouped user feedback into four classes, i.e., `TransfersFragment`, `SearchProgressView`, `MagnetDownloader`, and `LibraryUtils`. The `TransfersFragment` class is responsible for the visualization of the downloads that are being transferred, and a progress bar (implemented by the `SearchProgressView` class) shows the status of the searches initialized by the user using a magnet link. *John* discovers a connection delay bug in the `LibraryUtils` class which is used by the `MagnetDownloader` class responsible for the download of torrents related to magnet links. Given the output of ChangeAdvisor, *John* is now able to solve the issue because he immediately identifies the problem experienced by users, and receives suggestions about which source code components need to be modified.

In this paper, we empirically evaluated the performances of ChangeAdvisor by using 44 683 user reviews and the source code of 10 open source mobile apps. As a preliminary analysis, we quantified the ChangeAdvisor ability to cluster user feedback toward similar change requests, finding that our approach is able to identify cohesive user feedback clusters. Then, we evaluated ChangeAdvisor capabilities in identifying the code components impacted by the suggested changes, comparing our technique with a baseline implementation of the tool proposed by Saha et al. [24] in the context of bug localization. We observed that our approach is able to achieve 81 % of *precision* and 70 % of *recall*, being 47 % more precise and 38 % more complete than the baseline technique. Finally, we qualitatively evaluated ChangeAdvisor by surveying the 10 original developers of the apps in our dataset, who confirmed the actual usefulness of the approach in practice. The prototypical implementation of ChangeAdvisor, as well as the material and working data sets used in our study, are publicly available [25].

Structure of the Paper. Section II discusses the related literature, while Section III presents the proposed approach. In Section IV, we describe the case study conducted to evaluate the proposed approach. The qualitative case study involving the original developers is reported in Section V, whereas Section VI concludes the paper.

II. Related Work

In the following, we summarize the main relevant research in the context of mining user reviews of mobile apps and linking informal textual information to the source code. The interested reader can find a comprehensive literature review in the work by Martin et al. [26].

A. Mining User Reviews

Harman et al. [27] introduced the concept of app store mining by identifying correlations between the customer ratings and the download rank of a mobile app. Iacob and Harrison [9] empirically assessed the extent to which users of mobile apps rely on reviews to describe change requests discovering that a noticeable percentage of user reviews (23 %) describe feature requests. Moreover, Pagano and Malej [5] found that 33 % of the user reviews are related to requirements and user experience, and that developers use the feedback provided by the users to gather requirements. These papers motivated our work since they clearly indicate that users actually exploit the review mechanism in order to suggest improvements to apply to the current version of an app.

Chen et al. [28] devised AR-MINER, an approach to filtering and ranking informative reviews using a semi-supervised learning based approach. They demonstrated that, on average, 35 % of reviews contain informative content. Khalid et al. [29] reported a study with 6390 user reviews aimed at qualitatively classifying them into 12 types of complaints. The results suggest that more than 45 % of the complaints are related to problems that developers can solve. Such results highlight the importance to have tools supporting developers in evolutionary tasks, in the same way as the one proposed in this paper.

Guzman et al. [8] proposed an automatic approach to assigning a score to reviews indicating the user sentiment. Di Sorbo et al. [19] devised SURF, a tool to summarize user reviews to gather new requirements. Panichella et al. proposed ARDOC [30], an approach that combines natural language processing, sentiment analysis, and text analysis techniques, through a Machine Learning (ML) algorithm to detect sentences in user reviews. CHANGEADVISOR relies on ARDOC to classify user reviews as change requests (e.g., fix a bug or add a new feature). Villarroel et al. [23] devised CLAP, an approach to classifying, clustering, and then prioritizing user reviews on the basis of the information they contain (i.e., suggestions for new features or bug reporting). Unlike our work, this approach takes into account entire user reviews rather than the more fine-grained information given by user feedback, and has as a final goal of prioritizing user reviews rather than locate change requests. Finally, Gu and Kim [20] defined an approach able to summarize sentiments and opinions and classify them in aspect evaluation, bug reports, feature requests, praise, and others.

The approaches mentioned above perform an automatic classification of user reviews according to predefined topics (e.g., bugs, features) [20]. Moreover, they do not provide any support for linking user feedback to the source code components to be changed.

B. Linking Informal Textual Documentation to the Source Code

Traceability between textual artifacts (e.g., requirements) and the source code was widely studied in the past (see, e.g., [31], [32]). Similarly, several approaches to locating features in the source code [33], and tracing informal textual documentation, such as e-mails [34]–[36], forum discussions [37]–[39], and bug reports [24] to the source code have been proposed.

In this context, three works are closer to the one we proposed in this paper. Firstly, Saha et al. [24] proposed the use of structured Information Retrieval based on code constructs, i.e. class and method names, to improve bug localization. Their approach, named BLUIR, exploits the Vector Space Model [40] to link bug reports to the source code. In Section IV we report the detailed comparison between BLUIR and our approach.

Secondly, Asuncion et al. [41] devised TRASE, an approach that uses LDA-based topic modeling to enhance the information provided by prospective traceability. Since there is no way to know a priori the number of latent topics, the authors configured the parameter α of the LDA algorithm (i.e., the number of topics) using different settings, namely $\alpha = 10$, 20, and 30. However, as previously demonstrated by Panichella et al. [42], the configuration used to set the clustering algorithm is an important component of topic modeling techniques, and an optimal choice of the parameters generally results in better performance. This is especially true in cases where there are no hints about the right number of clusters to create as for user reviews [42]. Moreover, the technique exploited by TRASE to retrieve links toward the source code requires additional information by developers about the part of the project that needs to be modified, which is not required by our approach.

Finally, Palomba et al. [13] proposed CRISTAL, a tool for tracing informative crowdsourced reviews to the source code commits and for monitoring the extent to which developers accommodate user requests and follow-up user reactions as reflected in their ratings. Unlike CRISTAL, the intent of CHANGEADVISOR is to recommend the location of code changes on the current version of an app, rather than monitor the changes already applied during the history of a project.

III. THE CHANGEADVISOR APPROACH

The goal of the proposed approach is to extract from user reviews feedback relevant from a maintenance perspective and suggest the location of such changes in the source code. To this aim, it applies the following steps:

1) user feedback identification and classification (i.e., bug fixing tasks, features enhancement, and new features requests);
2) source code and user feedback preprocessing;
3) user feedback clustering, representing similar user needs (i.e., code change requests);
4) determining the code artifacts related to the suggested software changes.

A. User Feedback Classification

During the first step CHANGEADVISOR extracts and classifies the informative sentences (i.e., the feedback) contained in the user reviews. To achieve this goal it employs ARDOC, a *review classifier* previously defined by Panichella et al. [14], which is able to automatically mine feedback in user reviews. Specifically, ARDOC combines Natural Language Processing (NLP), Sentiment Analysis (SA) and Text Analysis (TA) techniques through a Machine Learning (ML) algorithm to detect the user feedback that belongs to one of the following categories: *Information Giving*, *Information Seeking*, *Feature Request* and *Problem Discovery*. We relied on the original implementation of the tool, publicly available as a Java Library [30]. We used this tool to classify and filter feedback in user reviews categorized as (i) request to fix bugs (sentences of the ARDOC category *Problem Discovery*), and (ii) feature enhancements or feature requests (*Feature Request* category). Note that we filtered out feedback in user reviews categorized as *Information Giving* and *Information Seeking* since they are not strictly related to change requests in the source code of an app. As an example, a review reporting *"It's an awesome app!"* is classified as *Information Giving* and does not provide any useful information from a maintenance perspective.

B. Input Preprocessing

In this step CHANGEADVISOR preprocesses both the user feedback extracted from the original reviews and the source code components in order to remove noise contained in the data that may hinder the accuracy of the NLP techniques exploited in further steps.

1) Source Code Preprocessing: The source code is first parsed to extract the code components it contains, such as fields, classes, and methods. The extracted code components are then normalized using a typical Information Retrieval (IR) normalization process [40]. In particular, the terms contained in the source code are transformed by applying the following steps: (i) separating composed identifiers using the camel case splitting which separates words on underscores, capital letters, and numerical digits base; (ii) reducing letters of extracted words to lower case; (iii) removing special characters, programming keywords and common English stop words; (iv) stemming words to their original roots via Porter's stemmer [43]. Finally, the normalized words are weighted using the *term frequency - inverse document frequency* (tf-idf) schema [40], which reduces the relevance of too generic words that are contained in several source code components. This step outputs a set of bag-of-words, one for each class of the analyzed app.

2) User Feedback Preprocessing: The challenge of this step is finding a methodology that is able to "parse" correctly all the words contained in a user feedback, because the language used by the end users of an application is generally informal, very noisy [44], and substantially different from the *sectorial language* used in software artifacts [45]–[47]. For this reason, we defined a specialized IR process to transform the end user language into input suitable for textual analysis using the Python libraries NLTK [48] and TEXTBLOB [49], two collections of tools for Natural Language Processing (NLP) and the spell check PYENCHANT library [50]. The involved NLP steps are:

Spelling correction: the words of user feedback are replaced if misspelled, according to the English vocabulary of PYENCHANT library.

Contractions expansion: it substitutes any possible English contractions with the related extended form (e.g., "don't" becomes "do not").

Nouns and verbs filtering: the process is performed by first applying a part of speech (POS) tagging classification, which identifies the logic role of words in user feedback sentences. Then, only the nouns and verbs are selected for the following steps because they are the most representative parts of the meaning of an artifact [47].

Tokenization: user feedback is transformed into lists of words (i.e., tokens), which will be the atomic part of the next steps, excluding numbers and punctuation that usually do not contain information.

Singularization: the singularization function of the TEXTBLOB library is exploited to normalize every word to its related singular form.

Stopword removal: the tokens are intersected with the WordNet English stopword list [51], which is a list of common words that are frequent in written English (e.g., "the", "a", and "an") and only introduce noise to NLP activities.

Stemming: the inflected words are reduced to their stem form (e.g., "pushing" is replaced with "push"). This step reduces the number of tokens and thus the complexity of the NLP work.

Repetitions removal: for each document, only one occurrence per word is preserved. We perform this step since, in this context, it is unlikely to have more occurrences of a word. Therefore, the contribution provided by having repetitions of the same word is not useful for the NLP activities.

Short tokens removal: this step excludes tokens with less than 3 characters (such threshold is a standard applied in NLP [46]) because tokens having a low number of characters are usually conjunctions or generic terms used in informal context and irrelevant for our purposes [52].

Short documents removal: documents with less than 3 tokens are also excluded from the output because short documents (i.e., user feedback) do not have enough terms to explain a change request clearly.

As the final result, the process outputs a bag-of-words for each user feedback that will be the input of the subsequent clustering phase.

C. User Feedback Clustering

The goal of this step is to group together automatically user feedback expressing similar user needs, consisting of similar code change requests (e.g., a problem in the GUI and/or performance and energy consumption problems). It is worth noting that this step is required since linking single user reviews to source code components does not provide high *precision* values, as highlighted by the preliminary analysis reported in our online appendix [25]. Moreover, this phase allows grouping together common user change requests, making them more comprehensible for the developer. Many effective clustering algorithms [53], [54] accept text objects as input. We experimented three different clustering techniques, such as the Latent Dirichlet Allocation (LDA) exploited by Asuncion et al. [41], the application of Genetic Algorithms to LDA (LDA-GA) devised by Panichella et al. [42] and the Hierarchical Dirichlet Process (HDP) algorithm proposed by Teh et al. [55].

The first technique is the classical implementation of the LDA algorithm [56], which requires the specification of the parameter α beforehand, namely the number of clusters (topics) to create starting from the set of sentences to group together (in our case, the user feedback extracted from the user review). The LDA-GA technique [42] is an evolution of the LDA algorithm able to find an optimal configuration of the parameter α during their execution by relying on Genetic Algorithms. Finally, HDP [55] is an extension of the LDA algorithm where the number of topics is not known a priori. In HDP and LDA models, each document in the corpus collection is considered as a mixture of latent topics, with each topic being a multinomial distribution over a known vocabulary of words. To cluster related feedback, HDP implements a nonparametric Bayesian approach [57] which iteratively groups elements based on a probability distribution, i.e., the Dirichlet process [55].

We benchmarked the three techniques concerning the quality of solutions and execution time. We decided to integrate HDP in ChangeAdvisor because it provides a good trade-off between quality and execution time. A detailed analysis of the comparison between the techniques is reported in our online appendix [25].

D. Recommending Source Code Changes

Once the user feedback clusters are formed, grouping together similar change requests, they are subjected to a process of linking to source code components to suggest what is the set of classes to update in order to meet these requests. We linked the user feedback clusters to source code classes by measuring the asymmetric Dice similarity coefficient [40], defined as follows:

$$\text{sim}\left(cluster_j, class_i\right) = \frac{\left|\, W_{cluster_j} \cap W_{class_i} \right|}{\min\left(\left|\, W_{cluster_j}\right|, \left| W_{class_i}\right|\right)}$$

where $W_{cluster_j}$ is the set of words contained in the cluster j, W_{class_i} is the set of words contained in class i and the $\min$ function normalizes the similarity score with respect to the number of words contained in the shortest document (i.e., the one containing fewer words) between the cluster and the class under analysis. The asymmetric Dice similarity ranges between $[0, 1]$. We used the asymmetric Dice coefficient instead of other similarity measures (e.g., the Jaccard coefficient [58]) because often the user feedback clusters are notably shorter than the source code files and therefore considering the minimum cardinality of the sets of words at the denominator of the formula allows to weight the similarity between documents better. The output is represented by a ranked list where the links having the highest similarity values are reported at the top. Pairs of $(cluster, component)$ having a Dice similarity coefficient higher than a threshold are considered by ChangeAdvisor to be a "link". We experimented different values for this threshold, and the best results were achieved when considering the third quartile of the distribution of the Dice similarity coefficients obtained for a given application (detailed results of the calibration are in our online appendix [25]). However, the developers may want to order the list based on other criteria, such as the popularity of a change request expressed by the cluster size. In future, we plan to integrate into ChangeAdvisor a prioritization approach that takes into account both (change) requests popularity and the similarity values with the source code.

Table I: Characteristics of the apps in the dataset.

App	KLOC	Classes	Reviews	Feedback
ACDisplay	47	267	4802	1722
Cool Reader	37	115	4484	1158
FB Reader	104	804	4495	830
Focal	19	81	2642	1285
FrostWire	291	1518	6702	1375
K-9 Mail	108	475	4542	1570
Shortyz Crosswords	22	165	3391	931
SMS Backup +	12	118	4488	1522
Solitaire	10	28	4480	1605
WordPress	110	567	4657	1512
Overall	760	4138	44 683	13 510

IV. Study I: The Accuracy of ChangeAdvisor

The *goal* of the study is to evaluate the effectiveness of ChangeAdvisor in retrieving links between user feedback clusters and source code. Thus, the *quality focus* is on the accuracy of the proposed approach in suggesting source code changes based on user feedback. The *context* of the study consists of 10 Android open source apps, whose source code is hosted on the *F-Droid* [59] open source repository and also published on the *Google Play Store* [60].

For each app considered in our dataset, Table I shows: (i) the application name, (ii) the size of the app in terms of KLOC and number of classes, (iii) the number of user reviews, and (iv) the number of user feedback entries extracted for the considered user reviews. The selected apps belong to different categories, have different sizes and reviews written by users with different requirements and expectations.

A. Empirical Study Definition and Design

The study aims at answering the following research questions:

- **RQ$_1$:** *Does* ChangeAdvisor *identify cohesive user feedback clusters representing related change requests?*
- **RQ$_2$:** *Does* ChangeAdvisor *correctly link change requests represented by user feedback clusters to code artifacts that need to be modified and how well it works compared to a state-of-the-art technique relating informal textual documentation to source code?*

To address our research questions, we firstly ran ChangeAdvisor against the apps in our dataset. In a normal scenario of using ChangeAdvisor, the developer owns both the user reviews (by downloading them as CSV file from the app distribution platform) and the source code needed as input. In our case, we first had to retrieve the reviews related to the chosen apps. To this aim, we developed a web scraper that extracts the user reviews directly from the *Google Play Store*, where they are publicly available. For the source code, we downloaded the last version of the considered apps from the corresponding repositories (e.g., *GitHub*) and we used a Java parser to extract single components (see Section III-B).

Before focusing on the performance of the proposed approach in suggesting source code changes (addressed by

RQ$_2$), with **RQ$_1$** we wanted to ensure that the clusters created by ChangeAdvisor using HDP are *cohesive* and thus meaningful. To this aim, we involved as inspectors 2 external mobile developers having 5 years of experience. The participants were asked to evaluate the *cohesiveness* of the clusters generated by ChangeAdvisor for each app in our study. To avoid bias, the inspectors were not aware of the experimental goals and of the specific algorithm used by the approach to cluster user feedback. To express their opinion the participants used a Likert scale intensity from *very low* to *very high* values [61], i.e., giving a value between 1 and 5 (where 1 means *very low*, while 5 *very-high*). This process required approximately 3 weeks of work.

In **RQ$_2$** our goal was to evaluate ChangeAdvisor accuracy in linking user feedback clusters to the source code. However, as pointed out in Section II, several approaches able to link informal textual documentation to the source code have been proposed in the literature to solve other software engineering tasks (e.g., bug localization). Therefore, rather than simply evaluating the performance of ChangeAdvisor, we are also interested in understanding whether the definition of our approach is actually needed or similar performance can be achieved by using existing state-of-the-art techniques when used to link user feedback to the source code. To this aim, we used a baseline approach inspired to BLUiR, the technique proposed by Saha et al. [24] for linking bug reports to the source code using an Information Retrieval infrastructure. Note that bug reports may be affected by the same noise (e.g., errors, different language compared to the source code programming language) of user reviews. We chose this approach as the baseline because it tries to solve a similar problem as ChangeAdvisor. As discussed in Section III-C, linking single user reviews to source code components does not provide good results. Hence, to conduct a fair comparison, BLUiR links each of the user feedback cluster produced using HDP to the source code by computing the cosine similarity between the vectors of terms built through the application of the Vector Space Model [40]. The baseline reports a link if the similarity between a cluster and a class is ≥ 0.6.

To compare the two approaches, we needed a set reporting the actual links between user feedback contained in the clusters and the source code. Due to the absence of this data set, we had to build our own oracle manually. We asked the 2 mobile developers previously involved in the evaluation of the *cohesiveness* of user feedback clusters to analyze the change requests contained in the clusters of reviews and determine the software artifacts that need to be maintained for performing such changes. In this case as well, the participants were not aware of the details of the study. We provided the participants with both the source code of the considered apps and the clusters generated by HDP [55]. The task consisted of analyzing each user feedback cluster in order to find the set of classes that need to be modified according to the requests reported in the cluster. Each inspector performed the task independently. Once completed the task, the two different sets of links were compared and the inspectors discussed the differences they found (e.g., links marked as correct by one inspector, but not by the other) in order to resolve the disagreement and reach a common decision. Finally, we considered as *golden set* all the links declared as correct by both the two inspectors after the discussion. To measure the level of agreement between the inspectors, we computed the Jaccard similarity coefficient [58], i.e. the number of traceability links identified by both the inspectors over the union of all the links identified by them. The overall agreement between the inspectors was of 76 %. A spreadsheet reporting the data about the agreement computation is available in our replication package [25]. This process required approximately 5 weeks of work.

Table II: Evaluation provided by the inspectors about the cohesiveness of the clusters generated by ChangeAdvisor.

App	Clusters	Cohesiveness		
		Min	Median	Max
ACDisplay	6	3	5	5
Cool Reader	8	3	4.5	5
FB Reader	9	3	4	5
Focal	6	4	4	5
FrostWire	7	3	4	5
K-9 Mail	7	4	4.5	5
Shortyz Crosswords	6	3	4	4
SMS Backup +	12	2	3.5	5
Solitaire	8	3	4	5
WordPress	13	3	4	5
Overall	82	2	4	5

Once defined the oracle, we answered **RQ$_2$** by reporting *precision* and *recall* [40] achieved by ChangeAdvisor and by the baseline.

Replication Package. ChangeAdvisor is publicly available in the online appendix [25], together with the dataset used in the study. The prototype is a runnable Docker container which allows not only the replication of the experiment, but also the application of the approach to any given app data (i.e., reviews and source code).

B. Analysis of the Results

In the following, we discuss the results of the study.

1) Evaluating the Cohesiveness of User Feedback Clusters: Table II reports the minimum, median, and maximum of the scores assigned by the inspector during the evaluation of the clusters generated by ChangeAdvisor.

As we can see, our approach identified a minimum of 5 and a maximum of 13 user feedback clusters, for a total of 82 clusters on the 10 apps. All the data related to clusters and their validation are available in our online appendix [25]. From a qualitative perspective, the *cohesiveness* of clusters produced was generally evaluated by the inspectors as *high* or *very high* (i.e., 4 and 5 values of the Likert scale, respectively). Indeed, the median of the distribution is 4, while the values are mainly placed between 4 and 5. This result highlights the ability of ChangeAdvisor to group together correctly similar changes. Nevertheless, there is one specific case for the SMS Backup + app where the clustering phase did not provide the expected results. This app offers the possibility to backup SMS, MMS, and calls log entries by providing space in the cloud. Users

Table III: Comparison between ChangeAdvisor and the baseline approach in linking change request feedback.

App	ChangeAdvisor		Baseline	
	precision (%)	*recall* (%)	*precision* (%)	*recall* (%)
ACDisplay	82	68	41	31
Cool Reader	79	68	48	68
FB Reader	83	73	43	17
Focal	83	63	27	50
FrostWire	81	71	29	20
K-9 Mail	84	66	38	69
Shortyz Crosswords	71	63	13	38
SMS Backup +	67	59	38	38
Solitaire	75	60	12	40
WordPress	79	74	38	40
Overall	81	70	34	32

often complain of energy consumption problems or the lack of compatibility with previous versions of the Android operating system (i.e., "Android KitKat 4.4"). Manually analyzing the clusters created by ChangeAdvisor for this application, we discovered that the cluster which obtained the lowest score included user requests related to both these problems. Thus, the presence of co-occurrence of terms in the two different change requests generated some noise that did not allow the correct clustering of the corresponding user feedback.

In Summary. ChangeAdvisor correctly clusters user feedback expressing similar change requests. An exception occurred because of the co-occurrence of terms expressing different change requests reported in the same user feedback.

2) Evaluating the Linking Accuracy: Table III reports the *precision* and the *recall* achieved by ChangeAdvisor and by the baseline approach for each of the considered apps when retrieving links between user feedback clusters and source code elements. The last row (i.e., *overall*) shows the results achieved when considering all the links suggested by the two approaches as a single dataset.

As we can see, the results achieved by ChangeAdvisor are quite positive, since it is able to discover correctly, overall, 81 % of the links between user feedback clusters and classes (i.e., 308 traceability links out of the total 381). At the same time, the suggestions provided by ChangeAdvisor are quite close to the actual classes that need to be modified to implement a change request. Indeed, the *recall* ranges between 59 % and 74 % (overall is 70 %).

The lowest *precision* of ChangeAdvisor is achieved on the `SMS Backup +` app, where the proposed approach outputs 5 false positive links. A possible cause behind this result can be related to an incorrect cluster of user requests for this app, as observed before. Besides the mentioned clustering issue, by manually investigating the links suggested by ChangeAdvisor we discovered two other reasons for this result: (i) poor cohesion of the classes composing the application, and (ii) the poor vocabulary of the identifiers. In particular, we found several cases in which a single class of the project is responsible for multiple functionalities, leading to the co-occurrence of different topics in the same class. As an example, the class `com.zegoggles.smssync.service.state.State` is responsible for both the management of error messages and additional services (e.g., user notifications and the information about the state of the execution of the app). Moreover, the same attributes are used in different methods of the class: for instance, the variable `resource` that is used to update the user interface layout and images contained in the project is referenced in all the methods. As a consequence of these observations, the similarity between user feedback clusters and the source code is blurred by the presence of different topics.

Thus, the main limitation of ChangeAdvisor is that it is not always possible to generate reliable links from the user feedback when the vocabulary of the source code is poor, or the source code elements implement functionalities related to multiple topics. This result is expected since our approach relies on topic analysis techniques which are themselves affected by the same problem.

On the other hand, it is interesting to observe that ChangeAdvisor results are highly accurate when the source code is well modularized and has a good vocabulary. For instance, let us consider one of the correct links found on the `Frostwire` project. This app is an open source BitTorrent client and the main problem experienced by users is related to the storage of the downloaded contents. In this example, one of the user feedback cluster identified by ChangeAdvisor contains two reviews, both of them reporting the inability of users in finding songs they have previously downloaded. ChangeAdvisor suggests modifying the class `AlbumSongFragment` contained in the package `com.andrew.apollo.ui.fragments.profile` and the classes `GUIUtils` and `GUIMediator` of the package `com.limegroup.gnutella.gui` in order to deal with the problems reported in the user feedback. By analyzing the source code of the suggested classes, we found that our approach correctly suggests the set of classes to change to address the user feedback. Indeed, the class `AlbumSongFragment` implements the methods `onCreateLoader` and `onLoadFinished` that are responsible for loading the multimedia content stored on the mobile phone, whereas the classes `GUIUtils` and `GUIMediator` manage the user interface related to the music library. In this case, the only class missed by ChangeAdvisor is `MusicUtils` of `com.andrew.apollo.utils`, which contains utility methods used by the class `AlbumSongFragment` to perform the loading of the multimedia contents. This example practically shows the potential usefulness of ChangeAdvisor from the developers perspective.

As for the baseline approach, it always has low *precision* in retrieving links (i.e., 34 % overall). The poor performances are also highlighted by the value of *recall* which reaches only 32 % overall. We manually investigated some of the links provided by the baseline approach to understanding the reasons behind these results. The major issue is related to the fact that a linking done using the Vector Space Model [40] does not provide optimal performances when relating user feedback clusters and source code. User feedback clusters are almost shorter than the set of words belonging to the source code, and most of the times the words used in the two sets of terms are similar because the users tend to explain problems using the same vocabulary present in the user interface of the app [5]. For this reason, the use of the asymmetric Dice coefficient instead of the cosine similarity results in a winning choice for CHANGEADVISOR because it is able to match directly the strings belonging to the two sets of terms, rather than computing their similarity using vectors.

We also perform an additional analysis aimed at verifying the overlap between the approaches, i.e. the number of links correctly retrieved by one approach and missed by the other, as well as the number of links correctly found by both the techniques. A complete report is available in our online appendix [25]. This analysis confirmed the findings discussed above since we observed that 72 % of correct links between user feedback clusters and the source code are retrieved by CHANGEADVISOR only, while both the approaches found 23 % of the correct links. Therefore, only a small percentage of the links (5 %) are retrieved by BLUIR and missed by our approach, highlighting the small contribution given by the alternative approach in finding links between clusters and source code. This result is significant because it shows how existing tools locating informal textual documentation to the source code are not particularly suitable in the context of reviews of mobile applications.

We repeated the experiment by distinguishing the different types of user feedback considered by the approaches (i.e., *Problem discovery* and *Feature request*) and did not find any meaningful differences. For the sake of space limitation, we report these results in the online appendix [25].

In Summary. Despite few cases discussed above, CHANGEADVISOR exhibits high *precision* (81 %) and *recall* (70 %) in linking user feedback clusters to source code components. When compared to the baseline, our approach results are 47 % more precise and 38 % more complete.

C. Threats to Validity

Threats to *construct validity* concern the relationship between the theory and the observation. For the evaluation of the experimented approaches we relied on error-prone human judgment, because of the subjectivity in deciding if a cluster of user feedbacks is cohesive or not and whether a link suggested by an approach is correct or not. To alleviate this issue we built a golden set based on the judgment of more professional inspectors. Moreover, the two inspectors firstly performed the task separately and then all disagreements were discussed and resolved.

The proposed approach itself could have been influenced from intrinsic imprecision of other approaches on which it relies, e.g., ARDOC [14], HDP [55]. In future, we plan to investigate better the relationship between the outcome of CHANGEADVISOR and the use of the techniques discussed above with the goal of increasing the already high performance of our approach.

On the one hand, if the clusters have been qualitatively evaluated by two professional developers in terms of cohesiveness, we cannot speculate about their completeness. Indeed, we cannot ensure that all the user feedback of the apps in our dataset have been taken into account and correctly clustered. However, as future work, we plan to build an oracle reporting the actual clustering of user feedback of the apps in the dataset and compare such an oracle with the results provided by CHANGEADVISOR.

Finally, another threat in this category is related to our implementation of the baseline approach, which was needed to compare our approach with the state-of-the-art. However, we applied the same algorithms and the same steps than the BLUIR approach described by Saha et al. [24].

Threats to *internal validity* concern any confounding factors that could influence our results. A first threat is related to the choice of the versions of the apps we considered in our study. Specifically, the analyzed source code refers to the latest version of an app, while the reviews can be related to previous versions. Even though in a normal context a developer is always able to retrieve the reviews for specific versions of his or her app through the *Google Developer Console*, we were not able to distinguish them during the user reviews scraping phase. Nevertheless, this is not an issue of this study since the goal was to quantify the accuracy of the proposed approach in linking user feedback to the source code, independently from the version of the app to which the feedback refers. At the same time, user feedback provided by users could have already been implemented by developers. Although there may be additional opportunities for increasing the practical applicability of CHANGEADVISOR, by combining it with a monitoring system such as CRISTAL [13], this is outside the scope of this paper. Hence, we leave this exploration for future work. Another threat that could affect the validity of the results of $\mathbf{RQ_2}$ is the presence of potential missing links. While the inspectors followed a clearly defined evaluation procedure, we cannot ensure that the set of correct links identified in this phase actually reflect the total set of correct links present in a given application. Finally, the threshold adopted for the Dice similarity coefficients may have influenced our results. However, we chose its value by performing a tuning process, detailed in the online appendix [25].

Threats to *external validity* concern the generalizability of our findings. In this study, we selected 10 different apps belonging to different app categories, having different sizes and user reviews written by different audiences. To confirm the generality of our findings as future work we would have to replicate the study exploiting the data from more apps belonging to different app stores.

V. STUDY II: OPINIONS OF THE ORIGINAL DEVELOPERS

Even though our technique achieved good results in the previous study, software engineering techniques have to address the real needs of developers. For this reason, the *goal* of this second study is to investigate the opinions of original app developers about CHANGEADVISOR, with the *purpose* of analyzing its usefulness from a developer's perspective. The *context* of the study consists of the 10 apps already used in the first study.

A. Empirical Study Definition and Design

In this study, we address the following research question:

- **RQ$_3$:** *Are the suggestions provided by* ChangeAdvisor *actually useful for developers?*

To answer our research question, we sent a direct invitation to the developers of the apps taken into account. To identify them, we retrieved the e-mail addresses of the developers from the source code repositories of the considered apps, filtering the developers that have contributed with at least one commit on the main branch. We chose to involve original developers rather than relying on external developers because we wanted to collect the opinions of the developers that actually worked on the systems under analysis and, therefore, have a sound knowledge about the structure of the app. We contacted 102 developers in total, receiving 10 responses, one for each app involved in the study. Note that even though the number of respondents appears to be low (9.8 % response rate), our results are close to the suggested minimum response rate for the survey studies, which is defined around 10 % [62].

The idea behind the study design was to show to each developer a subset of the links found by ChangeAdvisor for the app he/she developed. For this reason, we selected the top 10 links identified by our approach on each app of the study, those with the higher Dice coefficient resulting from the first study (see Section IV). This was done to avoid having a long questionnaire that might have discouraged developers to take part in our study. For each link composed of a user feedback cluster and the related classes, the participants had to answer the following questions:

1) *Could you please summarize what the users wanted to say in the shown user reviews?*
2) *How well the user reviews are grouped together according to the number of source code components that need to be modified? Please, rate your opinion from 1 = very poorly to 5 = very well.*
3) *How well the proposed set of classes matches the actual set of classes that need to be changed in order to satisfy the user requests? Please, rate your opinion from 1 = very poorly to 5 = very well.*
4) *Evaluate the completeness of the set of classes suggested above compared to the actual set of classes that need to be modified in order to satisfy the user requests. Please, rate your opinion from 1 = unacceptable to 5 = excellent.*

Besides the questions reported above, the developers also filled in a pre-questionnaire that allowed us to assess their background. They also answered a brief post-questionnaire where we asked to estimate how much time they would save using an approach like the one proposed in this paper instead of manually analyzing user reviews to identify change requests.

The survey was designed to be completed within approximately 15 minutes. To automatically collect the answers, the survey was hosted using a web-based survey software, i.e., *Google Forms*. Developers were given 20 days to respond to the survey. At the end of the response period, we collected the answers of the 10 complete questionnaires in a spreadsheet to perform data analysis.

To answer **RQ$_3$**, we computed:

1) the distribution of values assigned by developers when evaluating the *cohesiveness* of user feedback clusters (question #2 of the survey);
2) the distribution of values assigned by developers when assessing the *precision* of ChangeAdvisor (question #3 of the survey);
3) the distribution of values assigned by developers when evaluating the *completeness* of the suggestions provided by ChangeAdvisor (question #4 of the survey).

Moreover, when answering to question #1, we collected the descriptions provided by the developers as free opinions about the clusters produced by ChangeAdvisor.

B. Analysis of the Results

The 10 respondents declared a mobile development experience ranging from 3 to 5 years with 2 to 3 applications each and a *Google Play Store* score of 4 stars. The developers also claimed they look at user reviews "most of the time" to gather requirements for new releases of their apps, spending about 2 hours per week.

Figure 2 shows the boxplots of the distributions of the responses, divided by app, provided by the developers involved in the study. We reported the Likert scale values of the 3 aspects analyzed in **RQ$_3$**, i.e., cluster *cohesiveness*, linking *precision*, and linking *completeness*.

As we can see, the level of *cohesiveness* assigned by the developers is always high (Likert value ≥ 4) except for the `WordPress` app where, on average, the cohesiveness is equal to 3.5. However, by further analyzing his answers we observed that he was always able to summarize the requests contained in the feedback clusters indicating specific problems in the app, therefore explaining well-focused issues (e.g., "*Options in the Dashboard were deactivate for a while and users are just reporting such issues*"). This observation makes us conclude that the lowest values assigned by the developer do not indicate problems in the clustering phase of ChangeAdvisor, but rather the judgment of the developer has been just more conservative than that of other developers. The results obtained by this analysis confirmed again the ability of ChangeAdvisor in grouping together similar user needs.

The *precision* and *completeness* of the approach were instead evaluated in general as "well" (4) and "very well" (5), respectively. While the values assigned by developers to assess the *precision* of ChangeAdvisor somehow reflect the quantitative analysis conducted in Section IV (precision = 81 %), the most interesting thing is related to the *completeness*. Indeed, the developers perceived the links provided by our approach closer to the actual set of code components to modify with respect to what we estimated in our quantitative evaluation (*recall* = 70 %). This highlights the potential of the proposed approach as a tool useful for developers in a

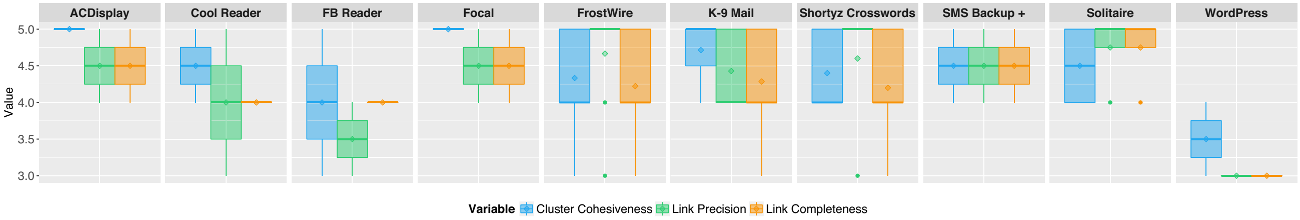

Figure 2: Survey results.

real-world environment. The claim is also supported by some of the answers provided by the developers when they filled in the post-questionnaire. In particular, we asked to estimate how much time they would save using ChangeAdvisor instead of only manually analyzing user reviews. In general, they claimed that a tool like ChangeAdvisor would be "very useful" and that they would be able to save "almost all the necessary time compared to the manual task".

One of the developers pointed out that:

> *"Trivially, I will not need to read each single review to know what requirements I should implement."*

This represents exactly the first goal of ChangeAdvisor, i.e. trying to reduce a huge number of user reviews to a smaller set of useful contents through the clustering process.

On the other hand, ChangeAdvisor links user feedback directly to source code components. Although a developer usually is aware of the source code components responsibilities, the linking process would reveal itself to be very useful, as pointed out by two developers:

> *"I know what changes I have to make in my app when implementing a change. However, a tool like this may help in quantifying the number of classes to be modified."*
>
> *"It would be useful for letting me immediately know to what extent a change is complicated to apply."*

Thus, the original developers of the apps in our dataset highlighted that ChangeAdvisor not only provides a toolkit able to support them during the daily activities of user reviews reading and understanding, but also a tool to estimate the possible cost of some change.

In Summary. The developers confirm the usefulness of ChangeAdvisor in practice. In particular, not only do the developers think that our approach would be useful in grouping together reviews with the same meaning and linking to source components, but also in quantifying the extent of changes.

C. Threats to Validity

Threats to *construct validity* are mainly related to how we measured developers' estimation of cluster *cohesiveness*, linking *precision*, and *completeness*. We used a Likert scale [61] that allows comparing responses from multiple respondents. However, we are aware that questionnaires may only reflect a subjective perception of the problem.

Threats to *internal validity* are related to the selection of the top 10 links based on the higher Dice coefficient values resulting from the first study. We made it to avoid having a long questionnaire that might have discouraged developers to take part in our study. Moreover, a factor that may have affected the results of the survey study is the response rate of 9.8 % although it covers each app in the dataset.

As for the threats to *external validity*, they are related to how we generalized the results. A replication of the study is part of our future agenda.

VI. Conclusions and Future Work

In this paper, we introduced ChangeAdvisor, a novel approach developed with the goal of supporting developers in accommodating user change requests of mobile apps and planning a new release of an application.

We believe that ChangeAdvisor represents a significant step forward to the current state-of-art of app reviews mining since, to the best of our knowledge, it is the first approach able to (i) extract user feedback relevant from a maintenance perspective, (ii) cluster it according to similar user needs, and (iii) determine the code artifacts that need to be maintained to accommodate the change requests.

The first evaluation of ChangeAdvisor demonstrated that the approach is able to (i) identify cohesive user feedback clusters representing similar change requests (**RQ$_1$**), (ii) identify the source code components impacted by the suggested changes with 81 % of *precision* and 70 % of *recall* (**RQ$_2$**), and (iii) it is more accurate than a state-of-the-art technique developed in the context of bug localization [24]. The study conducted with the original developers of the apps in our dataset confirmed the usefulness of our approach in practice (**RQ$_3$**).

Our future research agenda includes (i) the extension of the empirical study to include a larger number and variety of apps from different stores, (ii) the comparison of ChangeAdvisor with other baseline techniques in the context of software traceability (e.g., the technique proposed by Bachelli et al. [63]), and (iii) the integration of a mechanism for prioritizing the suggested changes. Specifically, the prioritization mechanism will take into account both user requests popularity and the interdependencies between the impacted code components to help developers focus on requests important for users and avoid working on conflicting software changes.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. Finally, we acknowledge the Swiss National Science foundation's support for the project SURF-MobileAppsData (SNF Project No. 200021-166275).

REFERENCES

[1] D. A. da Costa, S. McIntosh, U. Kulesza, and A. E. Hassan, "The Impact of Switching to a Rapid Release Cycle on the Integration Delay of Addressed Issues: An Empirical Study of the Mozilla Firefox Project," in *IEEE Working Conference on Mining Software Repositories (MSR)*, 2016, pp. 374–385.

[2] F. Khomh, T. Dhaliwal, Y. Zou, and B. Adams, "Do Faster Releases Improve Software Quality?: An Empirical Case Study of Mozilla Firefox," in *IEEE Working Conference on Mining Software Repositories (MSR)*, 2012, pp. 179–188.

[3] M. Nayebi, B. Adams, and G. Ruhe, "Release Practices for Mobile Apps - What do Users and Developers Think?" in *IEEE International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, 2016, pp. 552–562.

[4] B. Adams and S. McIntosh, "Modern Release Engineering in a Nutshell - Why Researchers Should Care," in *IEEE International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, 2016, pp. 78–90.

[5] D. Pagano and W. Maalej, "User Feedback in the AppStore: An Empirical Study." in *IEEE International Requirements Engineering Conference (RE)*, 2013, pp. 125–134.

[6] W. Martin, F. Sarro, and M. Harman, "Causal Impact Analysis for App Releases in Google Play," in *Proceedings of the 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering, FSE 2016*, 2016, pp. 435–446.

[7] V. N. Inukollu, D. D. Keshamoni, T. Kang, and M. Inukollu, "Factors Influencing Quality of Mobile Apps:Role of Mobile App Development Life Cycle," *Computing Research Repository (CoRR)*, vol. abs/1410.4537, Oct. 2014. [Online]. Available: https://arxiv.org/abs/1410.4537

[8] E. Guzman and W. Maalej, "How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews," in *IEEE International Requirements Engineering Conference (RE)*, 2014, pp. 153–162.

[9] C. Iacob and R. Harrison, "Retrieving and Analyzing Mobile Apps Feature Requests from Online Reviews," in *IEEE Working Conference on Mining Software Repositories (MSR)*, 2013, pp. 41–44.

[10] L. V. Galvis Carreño and K. Winbladh, "Analysis of User Comments: An Approach for Software Requirements Evolution," in *ACM/IEEE International Conference on Software Engineering (ICSE)*, 2013, pp. 582–591.

[11] S. Krusche and B. Bruegge, "User Feedback in Mobile Development," in *International Workshop on Mobile Development Lifecycle*, 2014, pp. 25–26.

[12] T. Vithani, "Modeling the Mobile Application Development Lifecycle," in *International MultiConference of Engineers and Computer Scientists (IMECS)*, 2014, pp. 596–600.

[13] F. Palomba, M. Linares-Vasquez, G. Bavota, R. Oliveto, M. Di Penta, D. Poshyvanyk, and A. De Lucia, "User Reviews Matter! Tracking Crowdsourced Reviews to Support Evolution of Successful Apps," in *IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 2015, pp. 291–300.

[14] S. Panichella, A. Di Sorbo, E. Guzman, C. Visaggio, G. Canfora, and H. Gall, "How Can I Improve My App? Classifying User Reviews for Software Maintenance and Evolution," in *IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 2015, pp. 281–290.

[15] U. Abelein, H. Sharp, and B. Paech, "Does Involving Users in Software Development Really Influence System Success?" *IEEE Software*, vol. 30, no. 6, pp. 17–23, 2013.

[16] S. A. Licorish, A. Tehir, M. F. Bosu, and S. G. MacDonell, "On Satisfying the Android OS Community: User Feedback Still Central to Developers' Portfolios," in *Australasian Software Engineering Conference (ASWEC)*, 2015, pp. 78–87.

[17] E. Ha and D. Wagner, "Do Android Users Write About Electric Sheep? Examining Consumer Reviews in Google Play," in *IEEE Consumer Communications and Networking Conference (CCNC)*, 2013, pp. 149–157.

[18] J. Oh, D. Kim, U. Lee, J.-G. Lee, and J. Song, "Facilitating Developer-User Interactions with Mobile App Review Digests," in *Extended Abstracts on Human Factors in Computing Systems*, 2013, pp. 1809–1814.

[19] A. Di Sorbo, S. Panichella, C. Alexandru, J. Shimagaki, C. Visaggio, G. Canfora, and H. Gall, "What Would Users Change in My App? Summarizing App Reviews for Recommending Software Changes," in *ACM SIGSOFT International Symposium on the Foundations of Software Engineering (FSE)*, 2016, pp. 499–510.

[20] X. Gu and S. Kim, "What Parts of Your Apps Are Loved by Users?" in *IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2015, pp. 760–770.

[21] A. Di Sorbo, S. Panichella, C. Alexandru, C. Visaggio, G. Canfora, and H. Gall, "SURF: Summarizer of User Reviews Feedback," in *ACM/IEEE International Conference on Software Engineering (ICSE)*, To appear in 2017.

[22] A. Ciurumelea, A. Schaufelbühl, S. Panichella, and H. Gall, "Analyzing Reviews and Code of Mobile Apps for better Release Planning." in *IEEE International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, To appear in 2017.

[23] L. Villarroel, G. Bavota, B. Russo, R. Oliveto, and M. Di Penta, "Release Planning of Mobile Apps Based on User Reviews," in *ACM/IEEE International Conference on Software Engineering (ICSE)*, 2016, pp. 14–24.

[24] R. K. Saha, M. Lease, S. Khurshid, and D. E. Perry, "Improving Bug Localization Using Structured Information Retrieval," in *IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2013, pp. 345–355.

[25] F. Palomba, P. Salza, S. Panichella, A. Ciurumelea, H. Gall, F. Ferrucci, and A. De Lucia, "Recommending and Localizing Code Changes for Mobile Apps based on User Reviews: Online Appendix," 2017. [Online]. Available: https://sites.google.com/site/changeadvisormobile/

[26] W. Martin, F. Sarro, Y. Jia, Y. Zhang, and M. Harman, "A Survey of App Store Analysis for Software Engineering," *IEEE Transactions on Software Engineering*, no. 99, 2016.

[27] M. Harman, Y. Jia, and Y. Zhang, "App Store Mining and Analysis: MSR for App Stores," in *IEEE Working Conference on Mining Software Repositories (MSR)*, 2012, pp. 108–111.

[28] N. Chen, J. Lin, S. Hoi, X. Xiao, and B. Zhang, "AR-Miner: Mining Informative Reviews for Developers from Mobile App Marketplace," in *ACM/IEEE International Conference on Software Engineering (ICSE)*, 2014, pp. 767–778.

[29] H. Khalid, E. Shihab, M. Nagappan, and A. E. Hassan, "What Do Mobile App Users Complain About? A Study on Free iOS Apps," *IEEE Software*, no. 2-3, pp. 103–134, 2014.

[30] S. Panichella, A. Di Sorbo, E. Guzman, C. Visaggio, G. Canfora, and H. C. Gall, "ARdoc: App Reviews Development Oriented Classifier," in *ACM SIGSOFT International Symposium on the Foundations of Software Engineering (FSE)*, 2016, pp. 1023–1027.

[31] G. Antoniol, G. Canfora, G. Casazza, A. De Lucia, and E. Merlo, "Recovering Traceability Links Between Code and Documentation," *IEEE Transactions on Software Engineering*, vol. 28, no. 10, pp. 970–983, 2002.

[32] A. De Lucia, A. Marcus, R. Oliveto, and D. Poshyvanyk, "Information Retrieval Methods for Automated Traceability Recovery," in *Software and Systems Traceability*, J. Cleland-Huang, O. Gotel, and A. Zisman, Eds. Springer London, 2012, pp. 71–98.

[33] B. Dit, M. Revelle, M. Gethers, and D. Poshyvanyk, "Feature Location in Source Code: A Taxonomy and Survey," *Journal of Software: Evolution and Process*, vol. 25, no. 1, pp. 53–95, 2013.

[34] A. Bacchelli, M. Lanza, and R. Robbes, "Linking E-Mails and Source Code Artifacts," in *ACM/IEEE International Conference on Software Engineering (ICSE)*, 2010, pp. 375–384.

[35] A. Di Sorbo, S. Panichella, C. A. Visaggio, M. Di Penta, G. Canfora, and H. C. Gall, "Development Emails Content Analyzer: Intention Mining in Developer Discussions," in *IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2015, pp. 12–23.

[36] A. Di Sorbo, S. Panichella, C. A. Visaggio, M. D. Penta, G. Canfora, and H. C. Gall, "DECA: Development Emails Content Analyzer," in *ACM/IEEE International Conference on Software Engineering (ICSE)*, 2016, pp. 641–644.

[37] C. Parnin, C. Treude, L. Grammel, and M.-A. Storey, "Crowd Documentation: Exploring the Coverage and Dynamics of API Discussions on Stack Overflow," Georgia Institute of Technology, Tech. Rep., 2012.

[38] S. Panichella, J. Aponte, M. D. Penta, A. Marcus, and G. Canfora, "Mining Source Code Descriptions from Developer Communications," in *IEEE International Conference on Program Comprehension (ICPC)*, 2012, pp. 63–72.

[39] C. Vassallo, S. Panichella, M. Di Penta, and G. Canfora, "CODES: Mining Source Code Descriptions from Developers Discussions," in *IEEE International Conference on Program Comprehension (ICPC)*, 2014, pp. 106–109.

[40] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., 1999.

[41] H. U. Asuncion, A. U. Asuncion, and R. N. Taylor, "Software Traceability with Topic Modeling," in *ACM/IEEE International Conference on Software Engineering (ICSE)*, 2010, pp. 95–104.

[42] A. Panichella, B. Dit, R. Oliveto, M. Di Penta, D. Poshyvanyk, and A. De Lucia, "How to Effectively Use Topic Models for Software Engineering Tasks? An Approach Based on Genetic Algorithms," in *ACM/IEEE International Conference on Software Engineering (ICSE)*, 2013, pp. 522–531.

[43] M. F. Porter, "An Algorithm for Suffix Stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[44] M. Mernik, J. Heering, and A. M. Sloane, "When and How to Develop Domain-specific Languages," *ACM Comput. Surv.*, vol. 37, no. 4, pp. 316–344, Dec. 2005.

[45] S. Gupta, S. Malik, L. L. Pollock, and K. Vijay-Shanker, "Part-of-Speech Tagging of Program Identifiers for Improved Text-Based Software Engineering Tools," in *IEEE International Conference on Program Comprehension (ICPC)*, 2013, pp. 3–12.

[46] A. Mahmoud and N. Niu, "On the Role of Semantics in Automated Requirements Tracing," *Requirements Engineering*, vol. 20, no. 3, pp. 281–300, Sep. 2015.

[47] G. Capobianco, A. D. Lucia, R. Oliveto, A. Panichella, and S. Panichella, "Improving Ir-Based Traceability Recovery Via Noun-Based Indexing of Software Artifacts," *Journal of Software: Evolution and Process*, vol. 25, no. 7, pp. 743–762, 2013.

[48] "NLTK, Natural Language Toolkit," http://www.nltk.org.

[49] "TextBlob, Simplified Text Processing," https://textblob.readthedocs.org/en/dev/.

[50] "PyEnchant, a spellchecking library for python," http://pythonhosted.org/pyenchant/.

[51] "WordNet English Stopword List," http://www.d.umn.edu/~tpederse/Group01/WordNet/wordnet-stoplist.html.

[52] P. M. Nadkarni, L. Ohno-machado, and W. W. Chapman, "Natural Language Processing: An Introduction," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 544–551, 2011.

[53] J. B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *Berkeley Symp. Mathematical Statistics and Probability*, 1967, pp. 281–297.

[54] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data*. John Wiley & Sons, Inc., 2008.

[55] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, vol. 101, pp. 1566–1581, 2006.

[56] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.

[57] A. Gelman, C. Robert, N. Chopin, and J. Rousseau, *Bayesian Data Analysis*, 3rd ed. CRC Press, 2014.

[58] P. Jaccard, "Etude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bulletin de la Société Vaudoise des Sciences Naturelles*, 1901.

[59] "F-Droid, Free and Open Source Software App Repository," https://f-droid.org.

[60] "Google Play Store," https://play.google.com.

[61] R. Likert, "A Technique for the Measurement of Attitudes," *Archives of Psychology*, vol. 22, no. 140, 1932.

[62] R. M. Groves, *Survey Methodology*, 2nd ed. Wiley, 2009.

[63] A. Bacchelli, T. D. Sasso, M. D'Ambros, and M. Lanza, "Content Classification of Development Emails," in *ACM/IEEE International Conference on Software Engineering (ICSE)*, 2012, pp. 375–385.