Images Don’t Lie: Duplicate Crowdtesting Reports Detection With Screenshot Information

Junjie Wang¹³, Mingyang Li¹³, Song Wang⁴, Tim Menzies⁵, Qing Wang¹²³,*

¹Laboratory for Internet Software Technologies, ²State Key Laboratory of Computer Science, Institute of Software Chinese Academy of Sciences, Beijing, China
³University of Chinese Academy of Sciences, Beijing, China, *Corresponding author
⁴Electrical and Computer Engineering, University of Waterloo, Canada
⁵Department of Computer Science, North Carolina State University, Raleigh, NC, USA

Email: {wangjunjie,limingyang,wq}@itechs.iscas.ac.cn, song.wang@uwaterloo.ca, tim@menzies.us

Abstract

Context: Crowdtesting is effective especially when it comes to the feedback on GUI systems, or subjective opinions about features. Despite of this, we find crowdtesting reports are highly duplicated, i.e., 82% of them are duplicates of others. Most of the existing approaches mainly adopted textual information for duplicate detection, and suffered from low accuracy because of the lexical gap. Our observation on real industrial crowdtesting data found that when dealing with crowdtesting reports of GUI systems, the reports would be accompanied with images, i.e., the screenshots of the tested app. We assume the screenshot to be valuable for duplicate crowdtesting report detection because it reflects the real context of the bug and is not affected by the variety of natural languages.

Objective: We aim at automatically detecting duplicate crowdtesting reports that could help reduce triaging effort.

Method: In this work, we propose SETU which combines information from the ScrEenshots and the TextUal descriptions to detect duplicate crowdtesting reports. We extract four types of features to characterize the screenshots (i.e., image structure feature and image color feature) and the textual descriptions (i.e., TF-IDF feature and word embedding feature), and design a hierarchical algorithm to detect duplicates based on the four similarity scores derived from the four features respectively.

Results: We investigate the effectiveness of SETU on 12 projects with 3,689 reports from one of the Chinese largest crowdtesting platforms. Results show that recall@1 achieved by SETU is 0.44 to 0.79, recall@5 is 0.66 to 0.92, and MAP is 0.21 to 0.58 across all experimental projects. Furthermore, SETU can outperform existing state-of-the-art approaches significantly and substantially.

Conclusion: Through combining the screenshots and textual descriptions, our proposed SETU can improve the duplicate crowdtesting reports detection performance.

Keywords: Crowdtesting, Duplicate report, Similarity detection

1. Introduction

Crowdtesting is an emerging trend in software testing which accelerates testing processes by attracting online crowd workers to accomplish various types of testing tasks [35, 12, 37, 13, 36, 15]. It entrusts testing tasks to crowd workers whose diverse testing environments/platforms, background, and skill sets could significantly contribute to more reliable, cost-effective, and efficient testing results.

The benefit of crowdtesting must be carefully assessed with respect to the cost of the technique. At first place, crowdtesting is a scalable testing method under which large software systems can be tested with appropriate results. This is particular true when the testing is related with the feedback on GUI systems, or subjective opinions about different features.

One aspect of crowdtesting which is not received enough attention in prior work is the confusion factors in crowdtesting results. Our observation on real industrial data shows that an average of 82% crowdtesting reports are duplicate (see Table 3 for details), which suggests much of the crowdtesting work can be optimized. A significant problem with such a large number of duplicate reports is that the subsequent analysis by software testers becomes extremely complicated. For example, we find that merely working through 500 crowdtesting reports to find the duplicate ones takes almost the whole working day of a tester. This paper mostly removes that effort by a novel method for detection of duplicate reports.

The issue of duplicate reports has been studied in terms of textual descriptions [28, 38, 18, 30, 31, 29, 26, 33, 3, 44, 24, 4, 1, 17, 27, 42] (see details in Section 3). However, in practice, it is common that different people might use different terminologies, or write about different phenomena to describe the same issue [38, 24, 33, 42], which makes the descriptions often confusing. Because of this, most existing approaches for duplicate report detection suffer from low accuracy. However, when dealing with crowdtesting reports of GUI systems, besides the textual descriptions, often the feedback is in the form of images. Our observation on real industrial crowdtesting data re-
veals that an average of 94% crowdtesting reports are accompanied with an image, i.e., screenshot of the app. We suppose this is another valuable source of information for detecting duplicate crowdtesting reports. Compared with the textual description, a screenshot can reflect the real context of the bug and is not affected by the variety of natural languages.

In this paper, we propose SETU which combines information from the Screenshots and the Textual descriptions to detect duplicate crowdtesting reports. We first extract two types of features from screenshots (i.e., image structure feature and image color feature), and two types of features from textual descriptions (i.e., TF-IDF feature and word embedding feature). We then obtain the screenshot similarity and textual similarity through computing the similarity scores based on the four types of features. To decide the duplicates of a query report, SETU adopts a hierarchical algorithm. Specifically, if the screenshot similarity between the query report and candidate report is higher than a threshold, we treat the candidate report as the first class and rank all reports in the first class by their textual similarity. Otherwise, we treat it as the second class (follow behind the first class) and rank all reports in this class by their combined textual similarity and screenshot similarity. Finally, we return a list of candidate duplicate reports of the query report, with the ranked reports of the first class followed by the ranked reports of the second class.

We experimentally evaluate the effectiveness of SETU on 12 projects with 3,689 crowdtesting reports from one of the Chinese largest crowdtesting platforms. Results show that the recall@1 achieved by SETU is 0.44 to 0.79, recall@5 is 0.66 to 0.92, and MAP is 0.21 to 0.58 across all experimental projects. This denotes, for a new-coming crowdtesting report, in 44% - 79% of the time, SETU can correctly identify its duplicate reports with only a single recommendation. And in 66% - 92% of the time, SETU can correctly identify its duplicate reports with five recommendations. These results significantly and substantially outperform three state-of-the-art and typical duplicate detection approaches. In addition, we also experimentally evaluate the necessity of screenshots and textual descriptions in duplicate detection, as well as the relative effect of the four types of features.

This paper makes the following contributions:

- We show that the screenshots are valuable in duplicate crowdtesting reports detection and we further demonstrate the need to use both the screenshots and the textual descriptions in detecting duplicate crowdtesting reports.

- We propose a novel approach (SETU) for duplicate crowdtesting report detection, which combines the information from screenshots and textual descriptions hierarchically.

- We evaluate the effectiveness of SETU on 12 projects from one of the Chinese largest crowdtesting platforms, and results are promising.

The rest of this paper is organized as follows: Section 2 describes the background and motivation of this study, while Section 3 surveys related work. Section 4 presents our proposed approach of duplicate detection. Section 5 and 6 show the experimental setup and evaluation results respectively. Section 7 provides a detailed discussion and threats to validity. Finally, we summarize this paper in Section 8.

2. Background and Motivation

2.1. Background

In this section, we present a brief background of crowdtesting to help better understand the challenges we meet in real industrial crowdtesting practice.

![Figure 1: Procedure of crowdtesting](image)

Our experiment is conducted with Baidu CrowdTest crowdtesting platform. As shown in Figure 1, in general, the task requester prepares the software under test and testing tasks, and distributes them on the crowdtesting platform. Then, the crowd workers can sign in to conduct the tasks and are required to submit the crowdtesting reports, i.e., to describe the process and result of the testing task s/he carried on, which include the input, operation steps, result description, and screenshot. Table 1 demonstrates an example of the crowdtesting report.

In order to attract more workers, testing tasks are often financially compensated. Under this context, workers can submit hundreds of reports for a crowdtesting task. This platform delivers approximately 100 projects per month, and receives more than 1,000 test reports per day on average. An observation on our experimental dataset shows that an average of 82% crowdtesting reports are duplicates of other reports (see Table 3 for details).

Currently in this platform, the testers need to manually inspect these crowdtesting reports to identify the duplicate ones. However, inspecting 500 reports manually could take almost the whole working day of a tester. Obviously, such process is time-consuming and low-efficient. Most of the existing approaches mainly adopted textual information for duplicate detection, and suffered from|

---

1 Baidu (baidu.com) is the largest Chinese search service provider. Its crowdtesting platform (test.baidu.com) is also one of the largest crowdtesting platforms in China.

2 The task requesters are usually the testers in practice, so we also call them as the testers in the following paper.

3 Note that, since all the reports are written in Chinese in our experiment projects, we translate them into English to facilitate understanding.
low accuracy because of the lexical gap in natural language. This is especially true for crowdtesting because the reports are written by crowdworkers who are usually unprofessional in software testing. Therefore, we propose to include screenshots in duplicate detection to help address the issue. Section 2.2 will provide two motivating examples to show how textual descriptions lead to confusing understanding, and how screenshots and textual descriptions can complement each other in duplicate detection.

2.2. Motivation

In this section, we present two examples from Baidu CrowdTest crowdtesting platform to motivate the need of using both the screenshots and the textual descriptions in duplicate crowdtesting report detection. These examples draw from a sport application, i.e., JIAJIA Sport. It can automatically record and analyze users’ sport-related information such as running, rope skipping, etc. To better test this application, its testers distribute the testing task on Baidu CrowdTest crowdtesting platform, and 462 crowdtesting reports are submitted by the crowd workers. By analyzing these crowdtesting reports, we find the following two motivating examples.

2.2.1. Motivating Example 1: Descriptions Could Be Confusing

Crowdtesting reports Rope-145 and Rope-270 are about the sharing function. Their descriptions are as follows:

**Rope-145:** I press the sharing button in the bottom and want to share my rope skipping record, but nothing happens.

**Rope-270:** I press the sharing button and want to share my rope skipping record to qzone, but nothing happens except a failure notice.

Both descriptions contain such words as “sharing button”, “rope skipping record”, and “qzone”. Using traditional duplicate detection approaches, these two reports would be identified as duplicates with a high probability. However, if the screenshot information is considered, they can be easily determined as non-duplicates, which is the ground truth. The screenshot of Rope-145 is about the billboard of rope skipping record, and the crowdtesting report reveals a bug about sharing the ranking of rope skipping record. For the screenshot of Rope-270, it demonstrates the detail page of rope skipping record, and the report reveals a bug about sharing the detailed record. In this sense, the screenshot provides the context-related information and can help better detect duplicate reports. One may argue that the above information should be in the operation steps. However, the crowd workers are far from professional testers, and in our datasets only few reports contain the detailed and correct operation steps.

**Finding 1:** The textual descriptions of crowdtesting reports may easily lead to confusing understanding. With the help of context-related information provided by screenshots, the duplicate crowdtesting reports can be detected more accurately.

2.2.2. Motivating Example 2: Screenshots Usually Lack Details

In this example, crowdtesting reports Rope-62 and Rope-217 have similar screenshots (see Figure 3). If the duplicate detection is only based on the screenshot information, these two reports would be determined as duplicates with a high probability. However, the ground truth is just the opposite. Their descriptions are as follows:

**Rope-62:** I walked for 10 minutes, but the steps only increased by 10. An hour later, I just sat on my chair, but the steps increased sharply.

**Rope-217:** In the detail page, there is indeed step record for today, but there is no step record for this week.

From the descriptions, we can easily observe that these two crowdtesting reports involve two different bugs, although under the same function which is denoted by the two highly similar screenshots. In this sense, a screenshot merely demonstrates the context-related information about a crowdtesting report. We still need to refer to the textual descriptions to finally determine whether they are duplicates.
3. Related Work

3.1. Crowdtesting

Crowdtesting has been applied to facilitate many testing tasks. Chen and Kim [7] applied crowdtesting to test case generation. They investigated object mutation and constraint solving issues underlying existing test generation tools, and presented a puzzle-based automatic testing environment. Musson et al. [23] proposed an approach, in which the crowdworker was used to measure real-world performance of software products. Gomide et al. [16] proposed an approach that employed a deterministic automata to help usability testing. Adams et al. [14] proposed MoTIF to detect and reproduce crashes in mobile apps after their deployment in the wild.

These studies leverage crowdtesting to solve the problems in traditional testing activities, while some other approaches focus on the new encountered problem in crowdtesting.

Feng et al. [12, 13] proposed approaches to prioritize test reports in crowdtesting. They designed strategies to dynamically select the most risky and diversified test report for inspection in each iteration. Jiang et al. [19] proposed the test report fuzzy clustering framework by aggregating redundant and multi-bug crowdtesting reports into clusters to reduce the number of in-
spected test reports. Wang et al. [35, 37, 36] proposed approaches to automatically classify crowdtesting reports. Their approaches can overcome the different data distribution among different software domains, and attain good classification results. Liu et al. [20] proposed an automatic approach to generate descriptive words for the screenshots based on the language model and Spatial Pyramid Matching technique. Cui et al. [11, 10] and Xie et al. [40] proposed crowd worker selection approaches to recommend appropriate crowd workers for specific crowdtesting tasks. These approaches considered the worker’s experience, relevance with the task, diversity of testing context, etc., and recommend a set of workers who can detect more bugs.

In this work, we focus on detecting the duplicated crowdtesting reports to facilitate real industrial crowdtesting practice.

3.2. Duplicated Bug Report Detection

Many approaches have been proposed to detect duplicate bug reports [28, 38, 18, 30, 31, 29, 26, 33, 3, 44, 24, 4, 1, 17, 27, 42]. The main focus of duplicate report detection is to obtain the similarity of two reports. Runeson et al. [28] proposed the first duplicate report detection approach which uses natural language information of bug reports to compute the similarity. Later, other sources of information were utilized in similarity measurement, e.g., execution trace [38], fields information as product and component [33, 27, 42]. Meanwhile, techniques to improve the similarity measurement precision were also introduced, e.g., BM25F similarity measurement [29], topic modeling [24], word embedding technique [42], etc.

Table 2 presents a summary of existing duplicate bug report detection researches.

These approaches mainly leverage the textual information and fields information to detect the duplicate reports. In this work, we introduce the screenshot information to facilitate the duplicate crowdtesting report detection. Our proposed approach combines the information from both the screenshots and the textual descriptions, and can detect duplicate crowdtesting reports with high accuracy.

To evaluate the effectiveness of our proposed approach (in Section 5 and 6), we choose three state-of-the-art and typical approaches [24, 27, 42] as the baselines. The reason why we use [27, 42] is because they are the latest two researches which can be treated as the state-of-the-art approaches. The reason why we use [24] is because they can achieve the highest performance across all existing researches. We had planned to include [17], which is also one of the latest researches, as baseline. However, the contextual words utilized in that work are English-oriented; simply translating these words to Chinese or translating our crowdtesting reports to English would bring serious information loss. Therefore, this paper utilize three state-of-the-art and typical approaches [24, 27, 42] as baseline in evaluation (see details in Section 5.3).

4. Approach

Motivated by the two examples in Section 2.2, we propose a duplicate detection approach (SETU), which combines information from both the Screenshots and the Textual descriptions to detect duplicate crowdtesting reports. Figure 4 illustrates the overview of SETU.

Our approach is organized as a pipeline comprising three phases: feature extraction, similarity calculation, and duplicate detection.

In the feature extraction phase, for the screenshots, we extract two types of features, i.e., image structure feature and image color feature (details are in Section 4.1). For the textual descriptions, we also extract two types of features, i.e., TF-IDF (Term Frequency and Inverse Document Frequency) feature and word embedding feature (details are in Section 4.2).

In the similarity calculation phase, based on the four types of features, we compute four similarity scores between the query report and each of the pending reports, and obtain the screenshot similarity and textual similarity. Cosine similarity, which is commonly used for measuring distance [35, 28, 27, 39], is employed in our approach.

In the duplicate detection phase (details are in Section 4.3), we design a hierarchical algorithm. In detail, if the screenshot similarity between the query report and pending report is larger than a specific threshold, we treat the pending report as first class and rank all reports in the first class by their textual similarity. Otherwise, we treat it as second class (follow behind the first class) and rank all reports in this class by their combined textual similarity and screenshot similarity. Finally, we return a list of candidate duplicate reports of the query report, with the ranked reports of the first class followed by the ranked reports of the second class. The reason why we separate reports in two classes is to take the advantages of the information provided by screenshots. As described in Section 2.2, the screenshots can provide the context-related information, and reports with different screenshots are very unlikely to be duplicate with each other. We will further discuss the performance of other combinations of the screenshot similarity and textual similarity in Section 7.4.
4.1. Extracting Screenshot Features

We extract screenshot features from the image of screenshot accompanied with each crowdtesting report. Note that, in our experimental projects, all the crowdtesting reports contain zero or one screenshot (with details in Section 5.2). For reports which do not have a screenshot, we use a default blank picture to extract the features. Future work will consider the situation that one report contains several screenshots. We use the following two types of features.

4.1.1. Image Structure Feature

The geometric structure of an image exhibits fundamental information for distinguishing screenshots [8, 34]. Images with highly similar geometric structures (e.g., line segments) would be probably the same screenshot. Gist descriptor [25] can capture the spatial structure of an image through the segmentation and processing of individual regions. Its general idea is as follows: propose a set of perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness) to represent the dominant spatial structure of an image; estimate these dimensions based on the spectral and coarsely localized information; generate a multidimensional space in which scenes sharing membership in semantic categories are projected closed together.

Figure 5 shows the image structure features for the four screenshots in Figure 2 and Figure 3. The x axis is the feature’s dimension (i.e., 128), while the y axis is the value of the feature in each dimension. We can easily find that the image structure feature vectors of Rope-145 and Rope-270 are obviously different with each other, while the structure feature vectors of Rope-62 and Rope-217 are almost the same. This coincides with our visual perception.

4.1.2. Image Color Feature

Color is another basic indicator of visual contents, and is often used to describe and represent an image [8, 34]. Images with highly similar color distributions are probably the same screenshot. MPEG-7 descriptor [21] can capture the representative colors on a grid superimposed of an image based on segmentation and clustering. Its general idea is as follows: use the edgeflow algorithm to segment image; perform color clustering on each segmented region to obtain its representative colors and calculate the percentage of these colors; each representative color and its corresponding percentage form a pair of attributes that describe the color characteristics in an image.

We also use the publicly available software⁶ to extract the image color feature. It results in a 189-dimensional image color feature vector for each image.

![Figure 5: Image structure features](image1.png)
(a) Rope-145 and Rope-270
(b) Rope-62 and Rope-217

![Figure 6: Image color features](image2.png)
(a) Rope-145 and Rope-270
(b) Rope-62 and Rope-217

Figure 5 and Figure 6 demonstrate the image color feature vectors for the four screenshots in Figure 2 and Figure 3.

---

⁵http://people.csail.mit.edu/torralba/code/spatialenvelope/

⁶http://www.semanticmetadata.net/
4.2. Extracting Textual Features

We extract textual features from the textual descriptions of crowdtesting reports.

We first collect different sources of textual descriptions together (input, operation steps, and result description), and then conduct the natural language processing to remove noise and extract core terms. Specifically, because the crowdtesting reports in our experiment are written in Chinese, we adopt ICTCLAS\(^7\) for word segmentation, and segment descriptions into words. We then remove stopwords (i.e., “am”, “on”, “the”, etc.) to reduce noise. Note that, crowd workers often use different words to express the same concept, so we introduce the synonym replacement technique to mitigate this problem. Synonym library of LTP\(^8\) is adopted. The remaining terms are saved and will be used to extract the following two types of features.

4.2.1. TF-IDF Feature

TF-IDF (Term Frequency and Inverse Document Frequency) is one of the most popular feature for representing textual documents in information retrieval. The main idea of TF-IDF is that if a term appears many times in one report and a few times in the other report, the term has a good capability to differentiate the reports, and thus the term has high TF-IDF value. Specifically, given a term \(t\) and a report \(r\), \(TF(t, r)\) is the number of times that term \(t\) occurs in report \(r\), while \(IDF(t)\) is obtained by dividing the total number of reports by the number of reports containing term \(t\). TF-IDF is computed as: \(TF - IDF(t, r) = TF(t, r) \times IDF(t)\).

With the above formula, the textual description of a report \(r\) can be represented as a TF-IDF vector, i.e., \(r = (w_1, w_2, ..., w_n)\), where \(w_i\) denotes the TF-IDF value of the \(i^{th}\) terms in report \(r\).

4.2.2. Word Embedding Feature

Word embedding is a feature learning technique in natural language processing where individual words are no longer treated as unique symbols, but represented as \(d\)-dimensional vector of real numbers that capture their contextual semantic meanings \[22, 5\].

We use the publicly available software\(^9\) to obtain the word embedding of a report. With the trained word embedding model, each word can be transformed into a \(d\)-dimensional vector where \(d\) is set to 100 as suggested in previous studies \[42, 41\]. Meanwhile a crowdtesting report can be transformed into a matrix in which each row represents a term in the report. We then transform the report matrix into a vector by averaging all the word vectors the report contains as previous work did \[42\]. Specifically, given a report matrix that has \(n\) rows in total, we denote the \(i^{th}\) row of the matrix as \(r_i\) and the transformed report vector \(v_d\) is generated as follows:

\[
v_d = \frac{\sum_i r_i}{n}
\]

With the above formula, each crowdtesting report can be represented as a word embedding vector.

The TF-IDF feature focuses on the similarity of reports considering the term matching, while the word embedding feature concerns more on the relationship of terms considering the context they appear. We adopt both of them in our approach and investigate their performance in duplicate report detection in Section 6.

4.3. Conducting Duplicate Report Detection

Following the previous studies \[28, 38, 18, 30, 31, 29, 26, 33, 3, 44, 24, 4, 1, 17, 27, 42\], our duplicate crowdtesting report detection problem is formulated as follows: Given a query report of a crowdtesting project, our approach would recommend a list of duplicate reports from all the pending reports of the project and rank them by their probabilities to be duplicates. We design a hierarchical algorithm (Algorithm 1) to detect the duplicate reports.

**Algorithm 1 Duplicate report detection algorithm**

**Input:**
- Pending crowdtesting report set \(R\);
- Query report \(q\); Threshold \(\text{thres}\)

**Output:**
- A list of duplicate reports \(D\);  

1. for each report \(r\) in \(R\) and \(q\) do 
2. Extract the image structure feature vector and image color feature vector from its screenshot;  
3. Extract the TF-IDF feature vector and word embedding feature vector from its textual description;  
4. end for  
5. for each report \(r\) in \(R\) do 
6. for \(fType\) in \{structure, color, tfidf, embedding\} do 
7. Compute the cosine similarity between the \(fType\) vector of \(q\) and \(r\), and denote as \(s_{fType}\);  
8. end for  
9. \(s_{screenshot} = (s_{structure} + s_{color})/2\);  
10. \(s_{textual} = (s_{tfidf} + s_{embedding})/2\);  
11. \(s_{total} = (s_{screenshot} + s_{textual})/2\);  
12. if \(s_{screenshot} > \text{thres}\) then 
13. put \(r\) in \(D_{first}\);  
14. else  
15. put \(r\) in \(D_{second}\);  
16. end if  
17. end for
18. Rank \(D_{first}\) based on reports’ \(s_{textual}\) and put them in \(D\) sequentially  
19. Rank \(D_{second}\) based on reports’ \(s_{total}\) and put them in \(D\) sequentially (follow behind the reports of \(D_{first}\))

---

\(^7\)ICTCLAS (http://ictclas.nlp.rzt.org) is widely used Chinese NLP platform.

\(^8\)LTP (http://www.ltp-cloud.com/) is considered as one of the best cloud-based Chinese NLP platforms.

\(^9\)https://code.google.com/archive/p/word2vec/
In the algorithm, the screenshot similarity can be seen as a filter because the screenshot provides the context-related information (see Section 2.2 for details). If two crowdtesting reports have different screenshots, they are unlikely to be duplicate even if the textual similarity between them is high. In addition, if two crowdtesting reports are accompanied with the same screenshot, whether they are duplicate reports mainly depends on the similarity of their textual descriptions.

The threshold to determine whether two screenshots are the same one is an input parameter. We explore the influence of this parameter on the detection performance in Section 5.6.

We have also experimented with different weights for $s_{structure}$ and $s_{color}$ when combining them to obtain $s_{screenshot}$ (Line 9 in Algorithm 1), the weights for $s_{tfidf}$ and $s_{embedding}$ when combining them to get $s_{textual}$ (Line 10 in Algorithm 1), as well as the weights for $s_{screenshot}$ and $s_{textual}$ when combining them to obtain $s_{total}$ (Line 11 in Algorithm 1). Results turned out that when the two similarities have an equal weight, SETU can achieve a relative good and stable performance. Due to space limit, we do not present the detailed results. Note that, this does not imply the screenshot and textual descriptions are equally important, because the weights for $s_{screenshot}$ and $s_{textual}$ are only used in the second class of reports (Line 10 and 19 in Algorithm 1). Another note is that, for the reports in the second class, we have experimented with other ranking manners, i.e., by $s_{screenshot}$, by $s_{textual}$, and results turned out that with the ranking manner shown above, the detection performance is relative good and stable.

5. Experiment Design

5.1. Research Questions

Our evaluation addresses the following research questions:

- **RQ1 Effectiveness**: How effective is SETU in detecting duplicate crowdtesting reports?

RQ1 aims at evaluating the effectiveness of SETU in duplicate reports detection. We also compare SETU with the state-of-the-art approaches (see details in Section 5.3) to investigate whether and to what extent it improves over prior work.

- **RQ2 Necessity**: Are both screenshots and textual descriptions necessary in detecting duplicate crowdtesting reports?

This paper proposes to utilize both screenshots and textual descriptions in duplicate detection. RQ2 is to investigate whether both of them are necessary. We employ two additional experiments to investigate it. See details in Section 5.4.

- **RQ3 Replaceability**: What is the relative effect of the four types of features (i.e., TF-IDF, word embedding, image color, and image structure) in detecting duplicate crowdtesting reports?

SETU employs four types of features to characterize the screenshots (i.e., image color and structure) and the textual descriptions (i.e., TF-IDF and word embedding). RQ3 is to investigate the relative effect of these features. We use another four experiments for investigating this RQ, with details in Section 5.4.

5.2. Experimental Dataset

We mentioned that our experiment is based on crowdtesting reports from the repositories of Baidu CrowdTest crowdtesting platform. We collect all crowdtesting projects closed between June 1st 2017 and June 10th 2017. There are totally 12 crowdtesting projects.

Table 3 presents the detailed information of the projects with the application domain, the number of reports (i.e., # report), the number and percentage of reports which have screenshots (i.e., # screenshots and % screenshots).

There is a label accompanied with each report. It signifies a specific type of bug assigned by the tester in the company. Reports with the same label denote they are duplicates of each other. In this sense, we treat the pair of reports with the same label as duplicates, while the pair of reports with different labels as non-duplicates.

Table 3 also presents the percentage of duplicates (i.e., % duplicates) and percentage of duplicate pairs (i.e., % duplicate pairs).

\[
\%\text{duplicates} = \frac{\#\text{duplicates}}{\#\text{reports}} = \frac{\#\text{reports} - \#\text{unique labels for reports}}{\#\text{reports}} \tag{2}
\]

\[
\%\text{duplicate pairs} = \frac{\#\text{duplicate pairs}}{\#\text{total pairs}} = \frac{\sum \#\text{pairs of reports with same label}}{\#\text{pairs of all reports}} \tag{3}
\]

\[
\text{pairs of reports} = \frac{\#\text{reports} \times (\#\text{reports} - 1)}{2} \tag{4}
\]

To verify the validity of these stored labels, we additionally conduct the random sampling and relabeling. In detail, we randomly select 4 projects, and sample 30% of crowdtesting reports from each selected project. A tester from the company is asked to relabel the duplicate reports, without knowing the stored labels. We then compare the difference between the stored duplicate results and the new labeled duplicate results. The percentage of difference for each project is all below 4%. Therefore, we believe the ground truth labels are relatively trustworthy.

For **training the word embedding model**, we use another textual dataset. In detail, we crawl the textual description of crowdtesting reports and task requirements of 500 crowdtesting projects from the experimental platform. The reason why we use this dataset is that previous studies have revealed that to train an effective word embedding model, a domain-specific dataset with large size is preferred [42, 41]. The size of our training dataset is 520M.
5.3. Baselines

To explore the performance of our proposed SETU, we compare it with three state-of-the-art and typical baseline approaches. Note that, since there is no approach designed for duplicate crowdtesting report detection, we choose the approaches for duplicate bug report detection as our baselines. Section 3.2 has presented why we choose these three baselines.

**Information retrieval with word embedding (IR-EM)** [42]: It is the state-of-the-art technique for duplicate bug report detection. This approach first builds TF-IDF vector and word embedding vector and calculates two similarity scores based on them respectively. Meanwhile, it calculates a third similarity score based on bug product field and component field. Finally, it combines the three similarity scores into one final score and makes similar bug recommendation with it.

**Similarity based on bug components and descriptions (NextBug)** [27]: It is another state-of-the-art similarity-based approach for duplicate bug report detection. This approach first checks whether two reports have the same bug component field, if yes, processes the reports with standard information retrieval technique, calculates the cosine similarity of the reports, and ranks the reports with the similarity value.

Note that, for these two baselines, because the crowdtesting reports do not have the product or component fields, we use the most similar field, i.e., test task id for substitution.

**Information retrieval with topic modeling (DBTM)** [24]: It is the commonly-used technique for detecting duplicate bug reports. DBTM supposes a report as a textual document describing one or more technical issues, and duplicate reports as the documents describing the same technical issues. It then utilizes term-based and topic-based features to detect duplicates.

### 5.4. Experimental Setup

This section illustrates the experimental setup for answering each research question.

For answering RQ1, we compare SETU with three state-of-the-art approaches (see details in Section 5.3) to investigate whether and to what extent it improves over prior work.

For answering RQ2, we employ two additional experiments, i.e., onlyText and onlyImage, to investigate whether both screenshots and textual descriptions are necessary in duplicate reports detection. In detail, onlyText denotes ranking the reports only based on the textual similarity, while onlyImage denotes ranking the reports only based on the screenshot similarity.

For answering RQ3, we use another four experiments, i.e., noTF, noEmb, noClr, and noStrc, to investigate the relative effect of the four types of features utilized in duplicate detection. Each experiment denotes conducting the duplicate detection by removing one specific type of feature. For example, noTF denotes applying other three features except TF-IDF (i.e., only use word embedding, image color and image structure feature) for duplicate detection.

For all these experiments, we employ the commonly-used leave-one-out cross validation [6]. In detail, we use one crowdtesting project as the testing dataset to evaluate the performance of duplicate detection, and use the remaining crowdtesting projects as the training dataset to determine the optimal parameter value (see detail in Section 5.6).

We ran our experiments on a 3.40 GHz Intel Core i7-3770 machine with 8 GB of RAM. For the time-cost of feature extraction, extracting the image structure features of 188 screenshots consumes 132 seconds and extracting the image color features of 188 screenshots consumes 214 seconds. For training the word embedding model, it consumes 540 seconds. Extracting the TF-IDF features and word embedding features of 213 reports (i.e., P1 in Table 3) all consume less than 1 second.

### 5.5. Evaluation Metrics

We use three evaluation metrics, i.e., recall@k, mean average precision (MAP), and mean reciprocal rank (MRR), to evaluate the performance of duplicate detection. These metrics are commonly-used to evaluate the duplicate detection approaches (see Table 2 for details).

Given a query report q, its ground truth duplicate reports set G(q), and the top-k recommended duplicate reports list produced by duplicate detection approach R(q), we use another four experiments, i.e., noTF, noEmb, noClr, and noStrc, to investigate the relative effect of the four types of features utilized in duplicate detection.

For answering RQ1, we compare SETU with three state-of-the-art approaches (see details in Section 5.3) to investigate whether and to what extent it improves over prior work.

For answering RQ2, we employ two additional experiments, i.e., onlyText and onlyImage, to investigate whether both screenshots and textual descriptions are necessary in duplicate reports detection. In detail, onlyText denotes ranking the reports only based on the textual similarity, while onlyImage denotes ranking the reports only based on the screenshot similarity.

For answering RQ3, we use another four experiments, i.e., noTF, noEmb, noClr, and noStrc, to investigate the relative effect of the four types of features utilized in duplicate detection. Each experiment denotes conducting the duplicate detection by removing one specific type of feature. For example, noTF denotes applying other three features except TF-IDF (i.e., only use word embedding, image color and image structure feature) for duplicate detection.

For all these experiments, we employ the commonly-used leave-one-out cross validation [6]. In detail, we use one crowdtesting project as the testing dataset to evaluate the performance of duplicate detection, and use the remaining crowdtesting projects as the training dataset to determine the optimal parameter value (see detail in Section 5.6).

We ran our experiments on a 3.40 GHz Intel Core i7-3770 machine with 8 GB of RAM. For the time-cost of feature extraction, extracting the image structure features of 188 screenshots consumes 132 seconds and extracting the image color features of 188 screenshots consumes 214 seconds. For training the word embedding model, it consumes 540 seconds. Extracting the TF-IDF features and word embedding features of 213 reports (i.e., P1 in Table 3) all consume less than 1 second.

### 5.5. Evaluation Metrics

We use three evaluation metrics, i.e., recall@k, mean average precision (MAP), and mean reciprocal rank (MRR), to evaluate the performance of duplicate detection. These metrics are commonly-used to evaluate the duplicate detection approaches (see Table 2 for details).

Given a query report q, its ground truth duplicate reports set G(q), and the top-k recommended duplicate reports list produced by duplicate detection approach R(q), we define \( recall@k \) for a query report q as follows:

\[
recall@k = \begin{cases} 
1, & \text{if } G(q) \cap R(q) \neq \emptyset \\
0, & \text{Otherwise}
\end{cases}
\]  

(5)

According to the formula, if there is at least one ground truth duplicate report in the top-k recommendation, the top-k recommendation is useful for the query report q. Given a set of query reports, we compute the proportion of useful top-k recommendations by averaging the \( recall@k \) of all query reports to get an
overall recall@k. As previous approaches, we set k as 1, 5, and 10 to obtain the performance.

MAP (Mean Average Precision) is defined as the mean of the Average Precision (AP) values obtained for all the evaluation queries. The AP of a single query q is calculated as follows:

$$AP(q) = \frac{\sum_{n=1}^{\|G(q)\|} \text{Precision} @ k(n)}{\|G(q)\|}$$

In the above formula, Precision@k(q) is the retrieval precision over the top-k reports in the ranked list, i.e., the ratio of ground truth duplicate reports of the query report q in the top-k recommendation:

$$\text{Precision} @ k(q) = \frac{\# \text{ground truth in top } k}{n}$$

MRR (Mean Reciprocal Rank) is defined as the mean of the Reciprocal Rank (RR) values obtained for all the evaluation queries. RR of a single query q is the multiplicative inverse of the rank of first correct recommendation first_q (i.e., first ground truth duplicate report in the recommendation list):

$$RR(q) = \frac{1}{\text{first}_q}$$

In addition, for each evaluation metric, we obtain the Improvement of SETU compared with other approaches (e.g., the baseline). Taken metric MRR and approach DBTM as an example, Improvement is calculated as follows:

$$\text{Improvement} = \frac{\text{MRR of SETU} - \text{MRR of DBTM}}{\text{MRR of DBTM}}$$

To further demonstrate the superiority of our proposed approach, we perform the MannWhitney U test between our proposed SETU and other approaches (i.e., the baseline). We obtain the p-value to demonstrate the significance of the test, and the Cliff’s delta to demonstrate the effect size of the test. MannWhitney U test is often employed to check whether the difference in two data groups is statistically significant (which corresponds to a p-value of less than 0.05) or not. We use one-tailed MannWhitney U test with the following hypotheses:

H0: Performance produced by other approach (e.g., the baseline) is no smaller than the performance produced by SETU.

H1: Performance produced by other approach (e.g., the baseline) is smaller than the performance produced by SETU.

We include the Bonferroni correction to counteract the impact of multiple hypothesis tests.

Cliff’s delta is often used to check if the difference in two data groups are substantial. We include the Bonferroni correction to counteract the impact of multiple hypothesis tests. The range of Cliff’s delta is [-1, 1], where -1 or 1 means all values in one group are smaller or larger than those of the other group, and 0 means the data in the two groups is similar. The mappings between Cliff’s delta and effectiveness levels are shown below.

<table>
<thead>
<tr>
<th>Cliff’s delta</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; -1</td>
<td>Negative</td>
</tr>
<tr>
<td>-1 &lt; Cliff’s delta &lt; 0.147</td>
<td>Negligible</td>
</tr>
<tr>
<td>0.147 &lt;= Cliff’s delta &lt; 0.33</td>
<td>Small</td>
</tr>
<tr>
<td>0.33 &lt;= Cliff’s delta &lt; 0.474</td>
<td>Medium</td>
</tr>
<tr>
<td>0.474 &lt;= Cliff’s delta &lt; 1</td>
<td>Large</td>
</tr>
</tbody>
</table>

Note that, we use all results from all query reports to compute p-value and Cliff’s delta. For each query report, we have one value for other approach (i.e., the baseline) and another value for our proposed approach SETU. By computing the p-value and Cliff’s delta, the extent of which our approach improves over other method can be more rigorously assessed.

5.6. Parameter Setting

SETU has a parameter, i.e., thres, to determine the reports of first class (Section 4.3). Figure 7 presents how the duplicate detection performance is influenced by different parameter values. Note that, we only use MAP, which is obtained considering the whole recommendation list of duplicates, to investigate the parameter’s influence. Another note is that, we have experimented with thres from 0.1 to 1.0, and due to space limit, we only present the results with better performance. We can easily observe that, almost for all our experimental projects, with the increase of thres, the duplicate detection performance would first increase, reach a peak, and then decrease.

In our evaluation, we tune the optimal parameter value based on the training dataset (see Section 5.4) and apply it in the testing dataset to evaluate the performance of duplicate detection. In detail, for each parameter value, we first compute the average MAP across all the crowdtesting projects in the training dataset and treat the parameter value under which the largest average MAP is achieved as the optimal thres. In this way, the tuned optimal parameter value is 0.94 for projects P1 - P10, and 0.92 for projects P11 and P12. For other experiments in our evaluation (i.e., noClr), we use the same method to tune the optimal parameter value.

6. Results and Analysis

This section presents the results and analysis of the evaluation.
6.1. Answering RQ1: Effectiveness

Table 5 presents the recall@1, recall@5, recall@10, MAP, and MRR for each experimental project for SETU and three baselines. To facilitate reading, Figure 8a presents the box plot of these results.

We can see that, our approach SETU can achieve the highest performance in all experimental projects for all five evaluation metrics. recall@1 is 0.44 to 0.79 across all experimental projects, denoting in 44% to 79% circumstances, our approach can find the duplicate report in the first recommendation. recall@5 is 0.66 to 0.92 across all experimental projects, denoting in 66% to 92% circumstances, our first five recommendations contain the duplicate report. MAP is 0.21 to 0.58 across all experimental projects. The reported recall@1 in existing duplicate bug report detection approach is 0.16 to 0.67, and the reported MAP is 0.26 to 0.53 (see Table 2). Since our experiment is conducted on crowdtesting reports which is different from bug reports in open source projects, these figures are not comparable. However, the fact that these figures being the same order of magnitude proves the effectiveness of our approach.

Compared with the three baselines, SETU brings great improvement in all five evaluation metrics for all experimental projects. The improvement of recall@1 is 20% to 211% compared with the three baselines, while the improvement of MAP is 28% to 241% in all experimental projects. Other evaluation metrics also undergo similar improvements.

We further conduct Mann-Whitney U Test for the five metrics between SETU and each baseline. compared with the three baselines, SETU statistically significantly (i.e., p-value for all tests is less than 0.05) and substantially (i.e., Cliff’s delta for all tests is large ) achieves a better performance in terms of all the evaluation metrics for all experimental projects. This further indicates the effectiveness and advantage of our approach. To make our presentation more catchy, we do not provide the detailed results.

Among the three baselines, the IR-EM and NextBug employ the product field or component field to help detect duplicate reports. We use a different field test task id in this experiment. The reason is that the crowdtesting reports do not have the field product or component, and test task id is the most similar field. Moreover, we also experiment with four other fields, i.e., phone type, operation system, ROM information, and network environment (as shown in Table 1) and the performance is even worse.

The performance of NextBug is almost the worst. This might because the reports from different test tasks are already distinguishable in their textual descriptions. Therefore, the utilization of the test task id field could not provide extra information in detecting duplicate reports. The low performance of IR-EM in our crowdtesting reports dataset might due to the similar reason. As the test task id almost could not contribute to duplicate detection, the baseline IR-EM degenerates to, to some extent, the approach of only using textual descriptions (i.e., the TF-IDF and word embedding features, see results in Section 6.2).

The DBTM utilizes term-based and topic-based features for duplicate detection. The low performance of this baseline might because, unlike the large-scale open source projects, the reports of one crowdtesting project only have very few topics. The optimal topic number is about 100 to 300 for Eclipse, OpenOffice, and Mozilla [24]. However, the optimal topic number for our experimental projects is about 5 to 10. The tiny number of topics cannot effectively help distinguish duplicate reports.

Compared with three state-of-the-art and typical baselines, SETU significantly and substantially achieves a better performance in terms of all the evaluation metrics for all experimental projects. recall@1 is 0.44 to 0.79, recall@5 is 0.66 to 0.92, and MAP is 0.21 to 0.58 across all experimental projects.

6.2. Answering RQ2: Necessity

Figure 8b and Table 6 presents the recall@1, recall@5, recall@10, MAP, and MRR for each experimental project for SETU and onlyText, onlyImage (see Section 5.4 for detail).

We can observe that the performance obtained by onlyText or onlyImage is worse than SETU. The improvement in recall@1 of SETU is 23% to 211% compared with onlyText, and 17% to 319% compared with onlyImage. The improvement in MAP of SETU is 31% to 241% compared with onlyText, and 40% to 552% compared with onlyImage.

We further conduct Mann-Whitney U Test for the five metrics between SETU and onlyText, onlyImage. Compared with only using screenshots or textual descriptions (i.e., onlyText, onlyImage), SETU significantly (i.e., p-value for all tests is less than 0.05) and substantially (i.e., Cliff’s delta for all tests is large ) achieves a better performance in terms of all the evaluation metrics for all experimental projects. This further indicates that only using screenshots or textual descriptions is not effective enough, and combining these two sources of information is a sensible choice for duplicate crowdtesting report detection. To make our presentation more catchy, we do not provide the detailed results either.

The performance of onlyImage is a little lower than the performance of onlyText. This might because the screenshot mainly provides the context-related information. Without the assistance of textual descriptions, the screenshot can not distinguish the duplicate reports in many circumstances. Moreover, duplicate detection with only textual descriptions can neither achieve equivalent performance with SETU, denoting the screenshot plays an indispensable role in detecting duplicate crowdtesting reports.

Duplicate detection with only screenshots or textual descriptions achieves significantly and substantially worse performance than SETU in all the evaluation metrics for all experimental projects. This indicates the necessity of using both screenshots and textual descriptions in detecting duplicate crowdtesting reports.

6.3. Answering RQ3: Replaceability

Figure 8b and Table 7 presents the recall@1, recall@5, recall@10, MAP, and MRR for each experimental project for
Table 5: Performance comparison between SETU and baselines (RQ1)

<table>
<thead>
<tr>
<th>SETU</th>
<th>IR-EM</th>
<th>Next Bug</th>
<th>DB TM</th>
<th>Improvement</th>
<th>SETU</th>
<th>IR-EM</th>
<th>Next Bug</th>
<th>DB TM</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>recall@1</td>
<td>0.792</td>
<td>0.640</td>
<td>0.440</td>
<td>0.424</td>
<td>23% - 38%</td>
<td>MAP</td>
<td>0.188</td>
<td>0.158</td>
<td>0.151</td>
</tr>
<tr>
<td>MAP</td>
<td>0.307</td>
<td>0.549</td>
<td>0.922</td>
<td>0.815</td>
<td>0.280</td>
<td>0.922</td>
<td>0.191</td>
<td>0.119</td>
<td>0.141</td>
</tr>
<tr>
<td>MRR</td>
<td>0.384</td>
<td>0.431</td>
<td>0.500</td>
<td>0.335</td>
<td>95% - 122%</td>
<td>MRR</td>
<td>0.584</td>
<td>0.527</td>
<td>0.417</td>
</tr>
<tr>
<td>recall@5</td>
<td>0.972</td>
<td>0.720</td>
<td>0.640</td>
<td>0.500</td>
<td>23% - 38%</td>
<td>MAP</td>
<td>0.259</td>
<td>0.090</td>
<td>0.080</td>
</tr>
<tr>
<td>MAP</td>
<td>0.562</td>
<td>0.235</td>
<td>0.215</td>
<td>0.218</td>
<td>135% - 162%</td>
<td>MRR</td>
<td>0.552</td>
<td>0.400</td>
<td>0.370</td>
</tr>
</tbody>
</table>

Table 6: Performance comparison between SETU and onlyText, onlyImage (RQ2)

<table>
<thead>
<tr>
<th>SETU</th>
<th>only Text</th>
<th>only Image</th>
<th>Improvement</th>
<th>SETU</th>
<th>only Text</th>
<th>only Image</th>
<th>Improvement</th>
<th>SETU</th>
<th>only Text</th>
<th>only Image</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>recall@1</td>
<td>0.792</td>
<td>0.640</td>
<td>0.500</td>
<td>23% - 38%</td>
<td>MAP</td>
<td>0.188</td>
<td>0.165</td>
<td>0.48% - 60%</td>
<td>MAP</td>
<td>0.570</td>
<td>0.326</td>
</tr>
<tr>
<td>MAP</td>
<td>0.307</td>
<td>0.549</td>
<td>0.922</td>
<td>0.815</td>
<td>0.280</td>
<td>0.922</td>
<td>0.191</td>
<td>0.119</td>
<td>0.141</td>
<td>75% - 124%</td>
<td></td>
</tr>
<tr>
<td>MRR</td>
<td>0.384</td>
<td>0.431</td>
<td>0.500</td>
<td>0.335</td>
<td>95% - 122%</td>
<td>MRR</td>
<td>0.584</td>
<td>0.527</td>
<td>0.417</td>
<td>0.536</td>
<td>52% - 95%</td>
</tr>
<tr>
<td>recall@5</td>
<td>0.972</td>
<td>0.720</td>
<td>0.640</td>
<td>0.500</td>
<td>23% - 38%</td>
<td>MAP</td>
<td>0.259</td>
<td>0.090</td>
<td>0.080</td>
<td>0.100</td>
<td>75% - 222%</td>
</tr>
<tr>
<td>MAP</td>
<td>0.562</td>
<td>0.235</td>
<td>0.215</td>
<td>0.218</td>
<td>135% - 162%</td>
<td>MRR</td>
<td>0.552</td>
<td>0.400</td>
<td>0.370</td>
<td>0.421</td>
<td>23% - 49%</td>
</tr>
</tbody>
</table>

We can observe that, in most circumstances, performance obtained by SETU is better than or equal with the performance of SETU and noClr, noStrc, noTF, noEmb (see Section 5.4 for detail).
obtained by using three features (i.e., noCir, noStrc, noTF, or noEmb). In rare cases, the performance obtained by using three features is a little better than the performance of SETU.

We further conduct Mann-Whitney U Test for the five metrics between SETU and noCir, noStrc, noTF, noEmb. Table 8 shows the p-value, the Cliff's delta, and the interpretation of these tests (see Section 5.5 for details). There are totally 240 tests (12 projects, 4 types of features, 5 evaluation metrics). Among them, in 35% (83/240) tests, the performance obtained by SETU is significantly (i.e., p-value is less than 0.05) and substantially (i.e., Cliff’s delta is not negligible, i.e., small in 21% tests, median in 5% tests, or large in 9% tests) better than the performance of SETU. This further indicates our proposed approach of combining these four types of features is effective.

Among the four experiments with three types of features, we can see that noEmb achieves relatively worse results. This might because the word embedding feature focuses on the relationships of terms by considering the context they appear. Without this feature (i.e., noEmb), simply matching the occurrence of terms by TF-IDF feature, cannot effectively detecting the duplicate reports. This implies word embedding feature is the least replaceable feature, i.e., the performance would undergo a relatively large decrease if removing this feature.

We can also see that noCir achieves relatively better results among the four experiments with three types of features. This indicates image color feature is the most replaceable feature, i.e., the performance would undergo a relatively small decrease if removing this feature. This also implies that image color feature would sometimes bring noise in the duplicate detection. We will present further discussion in Section 7.3.

### Table 7: Performance comparison between SETU and noCir, noStrc, noTF, noEmb (RQ3)

<table>
<thead>
<tr>
<th>Feature</th>
<th>SETU</th>
<th>noCir</th>
<th>noStrc</th>
<th>noTF</th>
<th>noEmb</th>
<th>Improv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRR</td>
<td>0.54</td>
<td>0.53</td>
<td>0.52</td>
<td>0.51</td>
<td>0.55</td>
<td>+9%</td>
</tr>
<tr>
<td>MAP</td>
<td>0.34</td>
<td>0.33</td>
<td>0.32</td>
<td>0.31</td>
<td>0.35</td>
<td>+12%</td>
</tr>
<tr>
<td>Recall@5</td>
<td>0.78</td>
<td>0.75</td>
<td>0.74</td>
<td>0.73</td>
<td>0.80</td>
<td>+7%</td>
</tr>
<tr>
<td>Recall@1</td>
<td>0.55</td>
<td>0.52</td>
<td>0.51</td>
<td>0.50</td>
<td>0.58</td>
<td>+16%</td>
</tr>
</tbody>
</table>

### Table 8: Results of Mann-Whitney U Test between SETU and noCir, noStrc, noTF, noEmb (RQ3)

<table>
<thead>
<tr>
<th>Feature</th>
<th>SETU</th>
<th>noCir</th>
<th>noStrc</th>
<th>noTF</th>
<th>noEmb</th>
<th>Setu vs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRR</td>
<td>0.54</td>
<td>0.53</td>
<td>0.52</td>
<td>0.51</td>
<td>0.55</td>
<td>+9%</td>
</tr>
<tr>
<td>MAP</td>
<td>0.34</td>
<td>0.33</td>
<td>0.32</td>
<td>0.31</td>
<td>0.35</td>
<td>+12%</td>
</tr>
<tr>
<td>Recall@5</td>
<td>0.78</td>
<td>0.75</td>
<td>0.74</td>
<td>0.73</td>
<td>0.80</td>
<td>+7%</td>
</tr>
<tr>
<td>Recall@1</td>
<td>0.55</td>
<td>0.52</td>
<td>0.51</td>
<td>0.50</td>
<td>0.58</td>
<td>+16%</td>
</tr>
</tbody>
</table>

Note: The figures (X/Y/Z) respectively denotes p-value, Cliff’s delta, and interpretation of Cliff’s delta (i.e., Large (L), Medium (M), Small (S), and Negligible (N))
7. Discussion

7.1. Further Exploration of Results

We have noticed that the performance of duplicate detection differ in a large range for the projects. We conducted detailed analysis and summarized two typical reasons for the low performance in some projects.

Firstly, the low quality of screenshots could not effectively provide context-related information. We find that in some projects (e.g., P7, P9), the screenshots of many crowdtesting reports do not match with the textual descriptions. For example, a report describes the short message firewall in its textual description. But its screenshot is about the message sending page (i.e., an intermediate step for the textual description). SETU can be misled by these inaccurate screenshots. This is mainly because the crowdworkers are unprofessional in software testing, and proper training and guidance are needed in crowdtesting [43].

Secondly, the short textual descriptions can only provide limited information and make it hard for duplicate detection. Analytical tasks (such as classification, clustering, etc.) involving short texts have long been recognized as challenging due to the lack of context and owing to their sparseness [2, 32]. We find that in project P1, the median term length of crowdtesting reports is 16, and the median term length of reports of project P10 is only 9. This is another important reason for the low duplicate detection performance of P10.

However note that, existing duplicate detection approaches also suffer from the performance variation in different experimental projects. For example, the recall@1 is 42% - 57% in [24], and the MAP is 29% - 51% in [17].

7.2. Further Exploration of Image Features

We have mentioned that noClr (i.e., duplicate detection without image color feature) can sometimes achieve a slightly better performance than SETU, e.g., in P4, P6 (see details in Table 7). This indicates that image color feature would bring noise to the duplicate detection in some cases.

We have examined the screenshots of the experimental projects, and found that, in these projects (i.e., P4, P6), screenshots of different functionalities (i.e., denoting non-duplicate reports) can sometimes have very similar color distribution (as Figure 6b shows). This is somehow coincident with our common sense, because many apps pursue a unified interface design to make it more user-friendly. Under this case, the non-duplicate reports might share very similar image color features, and removing this feature (i.e., noClr) can instead increase the detection performance.

More than that, we also noticed that, in some projects as P9, P10, the performance achieved by noStrc (i.e., duplicate detection without image structure feature) is a little better than SETU.

We further examined the screenshots of these two projects (i.e., P9, P10), and found that, the images’ structure for different functionalities (i.e., denoting non-duplicate reports) can sometimes exert no much difference, just like Figure 5b shows. In this case, the non-duplicate reports would share very similar image structure features, and removing the structure feature (i.e., noStrc) can instead increase the detection performance.

The above analysis indicates that image color feature can occasionally bring noise and removing it can increase the performance for duplicate detection, and so does the image structure feature. We have also analyzed the reasons and found that if non-duplicate reports tend to share a similar color distribution, removing image color feature can achieve a better performance; this also holds true for image structure feature when non-duplicate reports usually share a similar image structure.

This motivates us to conduct feature selection for specific crowdtesting projects to further improve the duplicate detection performance. In practice, this can be done based on the received reports of the project during the crowdtesting process. One can examine whether a large number of non-duplicate reports share the similar color distribution or structure, so as to suggest which features to be used for detecting the new-coming reports for this specific project.

7.3. Advantage of Hierarchical Approach

Our proposed SETU employs a hierarchical algorithm to detect duplicate reports, i.e., first classify the reports into two classes, and rank them separately. Under SETU, the class information indicates different levels of certainty of being duplicates. In detail, the first class contains the reports whose screenshot similarity is large enough, thus they can be more likely to be the duplicates of the query report. On the contrary, the second class contains the reports whose screenshots are quite different with the query report, thus they are less likely to be the duplicates of the query report.

Therefore, when we provide the users with the duplicate detection results, these class information can give more insights of the detection. For example, we encourage the users put more focus on the first class since they have large probability to be the true duplicates. For the second class, the users only need to glance off the reports to examine whether they are the true duplicates, especially when there have been a large number of recommended reports in the first class. We believe our two-classes hierarchical approach can provide more practical assistance for duplicate detection in real-world crowdtesting practice. Meanwhile, existing approaches treated all the reports as a whole, thus could not provide this kinds of guidelines.

7.4. Alternative Combination Manner

The motivating examples in Section 2.2 indicate that the role of screenshots is mainly to demonstrate the context-related information, while the role of textual descriptions is to provide
detailed illustration of the reported problem. This is why our proposed approach SETU first uses screenshot similarity to filter the reports to first class and conducts the ranking separately. However, one may still argue that other combination manners could achieve better performance. We did conduct experiments to explore whether other combination manners could outperform SETU in duplicate detection. Due to word limit, we provide the detailed results and analysis, as well as the experimental setup in the external link. Here, we briefly summarize the main findings.

We had designed three new combination manners, i.e., addCmb, multiplyCmb, and textFirst. Specifically, addCmb denotes adding screenshot similarity and textual similarity as one similarity value and ranking the reports based on it, which is a straight-forward manner. multiplyCmb denotes multiplying the screenshot similarity with textual similarity as one similarity value and ranking the reports based on it, which is borrowed from [42]. textFirst denotes treating the reports with high textual similarity as the first class and ranking them with the screenshot similarity (the second class is treated as SETU does).

We further conduct Mann-Whitney U Test for the five metrics between SETU and addCmb, multiplyCmb, textFirst. There are totally 180 tests (12 projects, 3 alternative combination manners, 5 evaluation metrics). Among them, in 73% (131/180) tests, the performance obtained by SETU is significantly (i.e., p-value is less than 0.05) and substantially (i.e., Cliffs delta is not negligible, i.e., small in 30% tests, median in 15% tests, or large in 28% tests) better than the performance of other combination manners. In other 27% (49/180) tests, the performance obtained by SETU demonstrates negligible difference (i.e., Cliff’s delta is negligible although some p-value is less than 0.05) with the performance of other combination manners. In none of the 180 tests, the performance of alternative combination manners is significantly and substantially better than the performance of SETU. To summarize, the combination manner proposed in SETU can achieve relatively highest performance than other alternative combination manners.

7.5. Threats to Validity

The external threats concern the generality of this study. First, our experiment data consists of 12 projects collected from one of the Chinese largest crowdtesting platforms. We can not ensure that the results of our study could generalize beyond this environment in which it was conducted. However, the various domains of projects and size of data relatively reduce this risk. Second, all crowdtesting reports investigated in this study are written in Chinese, and we cannot assure that similar results can be observed on crowdtesting projects in other languages. But this is alleviated due to the fact that we did not conduct semantic comprehension, but rather simply tokenize sentence and use word as token for experiment.

Internal validity of this study mainly questions the selection and implementation of baselines. Because there is no approach for duplicate crowdtesting report detection, we can only employ the approaches for duplicate bug report detection. Moreover, as the crowdtesting reports do not have the product and component fields as the original approaches, we employ the most similar field (i.e., test task id) for substitution. In addition, as these three baseline approaches are not publicly available, we implement our own versions rigorously following the steps described in their papers.

Construct validity of this study mainly questions the data processing method. We rely on the stored duplicate labels of crowdtesting reports to construct the ground truth. However, this is addressed to some extent due to the fact that testers in the company have no knowledge that this study will be performed for them to artificially modify their labeling. Besides, we have verified its validity through random sampling and relabeling.

8. Conclusion

In this work, we propose SETU, which combines the information from both the screenshots and the textual descriptions to detect duplicate crowdtesting reports. We evaluate the effectiveness of SETU on 12 commercial projects with 3,689 reports from one of the Chinese largest crowdtesting platforms, and the results are promising.

Note that, the application scope of SETU is not limited to crowdtesting reports. It can also be applied in traditional software testing, especially testing of GUI system, where the screenshots can be easily obtained. The adoption of SETU in other scenarios and further demonstration of its feasibility will be conducted in future.

This paper is just the starting point of the work in progress. We are closely collaborating with Baidu CrowdTest crowdtesting platform and planning to deploy the proposed duplicate crowdtesting report detection approach online. The feedback from real-world case studies will further validate the effectiveness, as well as guide us in improving SETU. For the future work, we would like to explore other features, as well as conduct intelligent selection of optimal features to further improve the duplicate detection performance.

9. Acknowledgments

This work is supported by the National Natural Science Foundation of China under grant No.61602450, No.6143200, and China Scholarship Council. We would like to thank the testers in Baidu for their great efforts in supporting this work.

10. Reference


