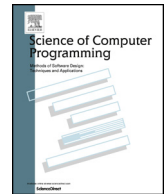




Contents lists available at ScienceDirect

Science of Computer Programming

www.elsevier.com/locate/scico


Original software publication

SATDBailiff-mining and tracking self-admitted technical debt

Eman Abdullah AlOmar^a, Ben Christians^a, Mihal Busho^a, Ahmed Hamad AlKhalid^a, Ali Ouni^b, Christian Newman^a, Mohamed Wiem Mkaouer^{a,*}

^a Rochester Institute of Technology, NY, USA^b ETS Montreal, University of Quebec, Canada

ARTICLE INFO

Article history:

Received 9 June 2020

Received in revised form 29 June 2021

Accepted 29 June 2021

Available online 8 August 2021

Keywords:

Self-admitted technical debt

Mining software repositories

ABSTRACT

Self-Admitted Technical Debt (SATD) is a metaphorical concept to describe the self-documented addition of technical debt to a software project in the form of source code comments. SATD can linger in projects and degrade source-code quality, but it can also be more visible than unintentionally added or undocumented technical debt. Understanding the implications of adding SATD to a software project is important because developers can benefit from a better understanding of the quality trade-offs they are making. However, empirical studies, analyzing the survivability and removal of SATD comments, are challenged by potential code changes or SATD comment updates that may interfere with properly tracking their appearance, existence, and removal. In this paper, we propose SATDBailiff, a tool that uses an existing state-of-the-art SATD detection tool, to identify SATD in method comments, then properly track their *lifespan*. SATDBailiff is given as input links to open source projects, and its output is a list of all identified SATDs, and for each detected SATD, SATDBailiff reports all its associated changes, including any updates to its text, all the way to reporting its removal. The goal of SATDBailiff is to aid researchers and practitioners in better tracking SATDs instances, and providing them with a reliable tool that can be easily extended. SATDBailiff was validated using a dataset of previously detected and manually validated SATD instances. SATDBailiff is publicly available as an open source, along with the manual analysis of SATD instances associated with its validation, on the project website.¹

© 2021 Elsevier B.V. All rights reserved.

* Corresponding author at: Rochester Institute of Technology, NY, USA.

E-mail addresses: eman.alomar@mail.rit.edu (E.A. AlOmar), bbc7909@rit.edu (B. Christians), mb5185@rit.edu (M. Busho), aa5130@rit.edu (A.H. AlKhalid), ali.ouni@etsmtl.ca (A. Ouni), cdnvse@rit.edu (C. Newman), mwmvse@rit.edu (M.W. Mkaouer).

¹ https://smilevo.github.io/self-affirmed-refactoring/SCP20_index.html

Software metadata

(executable) Software metadata description	
Current software version	1.0
Permanent link to executables of this version	https://github.com/smilevo/SATDBailiff
Legal Software License	N/A
Computing platform / Operating System	Windows, Mac, Linux, Docker image is also available
Installation requirements & dependencies	The following versions are required to run the tool: <ul style="list-style-type: none"> • Java 1.8+ * • MySql 5.4+ <p>If building the tool from source:</p> <ul style="list-style-type: none"> • Maven 3
Link to user manual	https://smilevo.github.io/self-affirmed-refactoring/SCP20_index.html
Support email for questions	mwmvse@rit.edu eman.alomar@mail.rit.edu

Code metadata

Code metadata description	
Current Code version	1.0
Permanent link to code / repository used of this code version	https://github.com/ScienceofComputerProgramming/SCICO-D-20-00107
Legal Code License	N/A
Code Versioning system used	Git
Software Code Language used	Java
Compilation requirements, Operating environments & dependencies	The following versions are required to run the tool: <ul style="list-style-type: none"> • Java 1.8+ * • MySql 5.4+ <p>If building the tool from source:</p> <ul style="list-style-type: none"> • Maven 3
Link to developer documentation / manual	https://smilevo.github.io/self-affirmed-refactoring/SCP20_index.html
Support email for questions	mwmvse@rit.edu eman.alomar@mail.rit.edu

1. Introduction

Technical debt (TD) is a metaphor that describes taking shortcuts in software development that will require additional time to fix (or payback) in the future [1]. Technical debt commonly occurs when developers conclude development of a section of code before it is complete and optimized [2]. This is done with an understanding that this creates a *debt* that will need additional time and effort to be managed later on. Developers tend to understand *some* quality-related implications of adding this debt to their projects, which can be seen unmistakably when looking specifically at *Self-Admitted Technical Debt* (SATD) [3].

Self-Admitted Technical Debt is a candid form of technical debt in which the contributor of the debt self-documents the location of the debt. This admission is typically accompanied by a description of a known or potential defect or a statement detailing what remaining work must be done. Well-known and frequently used examples of SATD include comments beginning with *TODO*, *FIXME*, *BUG*, *XXX*, or *HACK*. SATD can also take other forms of more complex language void of any of the previously mentioned keywords. Any comment detailing a *not-quite-right* implementation present in the surrounding code can be classified as SATD.

Modern Integrated Development Environments (IDEs) have begun recognizing the utility of SATD. It is common for them to highlight comments containing the aforementioned keywords, or for them to add SATD to a project when automatically generating unimplemented stubbed functionality to be implemented manually at a later time. Developers and IDEs both contribute SATD with the common assumption that including a self-admission will make their technical debt easier to pay back, or at least reduce the likelihood of it being forgotten. The effectiveness of this strategy needs to be brought into question, as it may have significant impacts on development practices. Understanding the implications of this assumption is vital to assure high-quality performance in software development teams.

Since SATD is a written testament of an existing negative manifestation in the source code, several studies have observed the removal of SATD instances to better understand how developers manage and resolve TD. Detecting the removal of TD is of interest to both researchers and practitioners, as, in addition to indicating the disappearance of a problem, it indicates the changes containing the fix to that TD. Such TD fixes are important to locate, since they can be valuable when taking

Table 1
Details of the studied 5 projects.

Project	# Java files	# SLOC	File versions	# Contributors
Camel ¹	15,091	800,488	254,920	289
Gerrit ²	3,059	222,476	53,298	270
Hadoop ³	8,466	996,877	79,232	160
Log4j ⁴	1,112	30,287	12,609	35
Tomcat ⁵	3,187	297,828	46,716	32

¹ <https://github.com/apache/camel>

² <https://github.com/GerritCodeReview/gerrit>

³ <https://github.com/apache/hadoop>

⁴ <https://github.com/apache/log4j>

⁵ <https://github.com/apache/Tomcat>

corrective actions against similar TDs. Several recent studies have been focused on accurately identifying the removal of SATD. For instance, Bavota and Russo [4] have shown that up to 57% of SATD is addressed, and those instances are typically addressed by the same developer who initially contributed to the SATD. However, the identification of SATD removal is complex when taking into account common occurrences when code changes overshadow SATD removals [5], such as the renaming of code elements containing the instance or the accidental deletion of containing source files. Another challenge that threatens the correctness of empirical studies related to SATD is the changes that may interfere with properly tracking the survivability of SATDs. For example, if the class containing the SATD gets renamed or moved, without properly handling such refactorings, one SATD may look like it was deleted, and another one appeared, while it is the same SATD. Similarly, if the SATD text gets updated, without properly handling such change, this can be detected by a removal of one SATD instance, and the appearance of another one. While not all accidental *disappearances* of SATD comments imply the correction of any associated technical debt, there is a need for a tool that can reliably track the appearance and removal of SATD comments, along with any potential changes associated with the text of the SATD. This effort would provide valuable support for existing studies by clearly capturing these removals without the need to manually validate them.

To address the above-mentioned challenge, in this tool paper, we propose SATDBailiff, a tool based on the existing classification model, called SATD Detector [6]. This tool is able to (i) mine, identify, and track the additions, removal, and changes to SATD comments, while providing an overview of their *lifespan* in the project, (ii) detect all textual changes associated with the SATD, detect all changes associated with the class containing the SATD (moving, renaming) that it underwent throughout the later commits up until the commit in which each instance was removed, if applicable, and (iii) allow the integration of SATD detection and classification tools, and (iv) allow the input of software repository link, and the output of all commits associated with additions, changes, and deletions of identified SATDs.

SATDBailiff was validated using a dataset of previously detected and manually validated SATD instances [7]. SATDBailiff is challenged in identifying and tracking those instances throughout the evolution of five long-lived open-source projects from different application domains, namely *Gerrit*, *Camel*, *Hadoop*, *Log4j*, and *Tomcat*. The detail of the studied projects is summarized in Table 1. We manually analyzed the ability of SATDBailiff in correctly identifying SATD changes and removals (SATD additions were provided by the dataset). Results show that SATDBailiff is efficient by averaging an accuracy score of 0.97 when tracking SATD instances from their appearance in the project until their disappearance.

Tool, documentation, docker and demo video. SATDBailiff is publicly available as an open source tool,² along with a continuous integration feature with a docker and a demo video. The raw data and the manual analysis of SATD instances are also available on the project website.³

The rest of the paper is organized as follows. Section 2 gives an overview of the necessary information related to SATD and summarizes the related work. Section 3 describes our approach, SATDBailiff. Section 4 describes how SATDBailiff can be used in practice while Section 5 shows the applicability of the tool. Section 6 details the results of our experiments to evaluate SATDBailiff. In Section 7, we report the tool's limitation. Section 8 discusses the known threats to validity, while Section 9 draws our conclusions and future investigations.

2. Background & related work

This paper focuses on mining and tracking SATD instances from Git repositories. Thus, in this section, we are interested in discussing related work on SATD. A summary of these state-of-the-art studies is depicted in Table 2.

The investigation of Self-Admitted Technical Debt began to gain traction in 2014 with the study of Potdar and Shihab [3]. Initial approaches to classifying source comments as SATD involved intensive manual efforts. Potdar and Shihab manually classified 101,762 Java code comments and generated a string matching heuristic based on 62 commonly occurring comment patterns. This heuristic inspired Maldonado and Shihab [8] to apply this classification to ten projects by manually analyzing 33 K code comments. They indicated that SATD items point to five types of debt: design, requirement, defect, test, and documentation. Their findings show that design debt is the most common type of debt. Later, Bavota and Russo

² <https://github.com/smilevo/SATDBailiff>.

³ https://smilevo.github.io/self-affirmed-refactoring/SCP20_index.html.

Table 2

A summary of the literature on SATD.

Study	Year	Focus	Detection technique	SATD tool	SATD type	Project size
Potdar & Shihab [3]	2014	SATD identification	Manual analysis	N/A	Not mentioned	4
Maldonado & Shihab [8]	2015	SATD detection	Manual analysis	N/A	design/requirement/ defect/test/ documentation	10
Bavota & Russo [4]	2016	SATD identification	Mining-based technique	N/A	design/requirement/ code/test/architecture/ defect/people/build/ documentation	159
Wehaibi et al. [5]	2016	Impact of SATD on quality	Defect-based measurement	N/A	Not mentioned	5
Maldonado et al. [7]	2017	SATD automatic classification	NLP-based technique	Maximum entropy classifier	design / requirement	10
Maldonado et al. [9]	2017	SATD removal	NLP-based technique	Maldonado et al.'s tool	Not mentioned	5
Zampetti et al. [10]	2018	SATD removal	Mining-based technique	N/A	Not mentioned	5
Huang et al. [12]	2018	SATD automatic classification	Machine learning technique	N/A	design/requirement/ defect/test/ documentation	8
Liu et al. [6]	2018	SATD detection	Mining-based technique	SATD Detector	design/requirement/ defect/test/ documentation	8
Lammarino et al. [11]	2019	SATD removal	Mining-based technique	N/A	Not mentioned	4
Farias et al. [13]	2020	SATD identification	Contextualized vocabulary technique	Not mentioned	design/requirement/ code/test/architecture/ defect/people/build/ documentation	3
Zampetti et al. [14]	2020	SATD removal	Deep learning technique	SARDELE	Not mentioned	5
Xavier et al. [15]	2020	SATD identification	Mining-based technique	Liu et al.'s tool [6]	design/requirement/ code/test/infrastructure/ build/security/UI/ documentation/ performance	5

[4] replicated the work of Potdar and Shihab and carried out an empirical study across 159 projects to explore the diffusion and evolution of SATD and its impact on software quality. Moreover, Wehaibi et al. [5] investigated the relationship between SATD and quality. They show that technical debt increases the system's difficulty and may lead to complex software changes.

Maldonado et al. [7] expanded the classification approach to use Natural Language Processing (NLP) to identify design and requirement debts from ten open source projects. They also reported the top-10 words appearing within design and requirement SATD-based comments. Despite the increased performance of classification models, little work has been done to develop an empirical understanding of SATD in Java projects. One of the empirical studies has been conducted by Maldonado et al. [9] to analyze SATD comments removal by looking at the change history of five large open source projects. Results of the study indicate that there is a high percentage of SATD comments removed and their survivability varies by project. The quality of this dataset has been brought into question by Zampetti et al. [10], who manually altered and improved the dataset in an effort to improve its quality. While this filtered dataset is regarded to be high quality, the filtering process removed a significant number of entries, there does not exist a means to expand its size. In their in-depth investigation of SATD removal, they found that between 20% and 50% of SATD comments are removed when either the whole class or method is removed. Another study examined the relationship between refactoring and technical debt. Lammarino et al. [11] particularly studied the co-occurrence of refactorings and SATD removals. The authors show that refactorings are more likely to co-occur with SATD removals than with other commits. Huang et al. [12] developed a new SATD classification model using text mining. They built a composite classifier that combined multiple classifiers from eight different source projects which improved the F1-score of classification over Maldonado et al. by 27.95%. Liu et al. [6] proposed an Eclipse plugin and Java library tool called SATD detector to automatically detect SATD using text mining and highlight the detected comments in an integrated development environment (IDE).

More recently, Farias et al. [13] improved a set of contextualized patterns or SATD identification vocabulary built to detect SATD by carrying out three empirical studies. With regard to the technical debt items detection, their results show that more than half of the new patterns were considered decisive or very decisive. In a different work, Zampetti et al. [14] developed SATD Removal using DEep LEarning (SARDELE) that highlighted developers' need to cope with SATD removal. Their tool is capable of recommending six SATD removal strategies (e.g., telling that a more complex change is needed). Their evaluation reveals that SARDELE is able to predict the type of change with an average precision of 55% and recall of 57%. In another study, Xavier et al. [15] mined SATD in issue tracker systems. They studied a sample of 286 SATD instances collected from five open source projects. Although SATD instances are not more complex in terms of code churn, their findings show that SATD instances take more time to be closed.

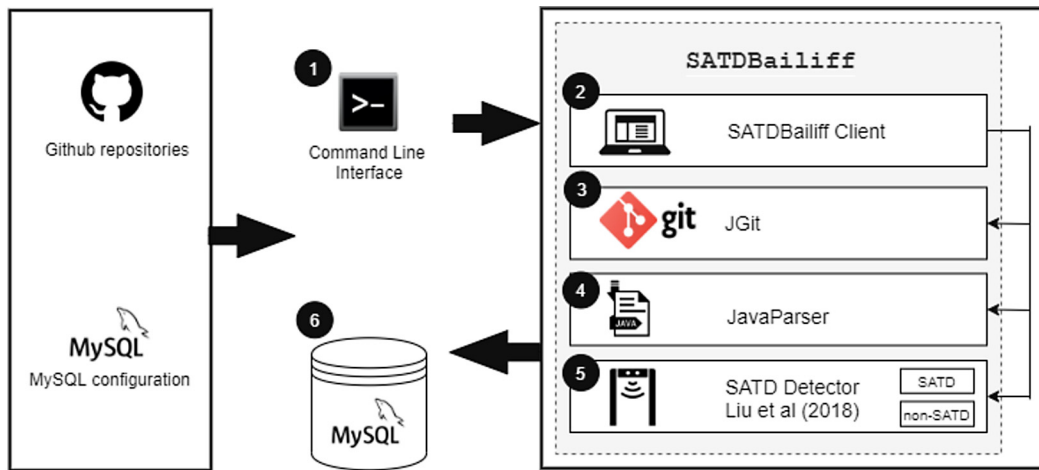


Fig. 1. High-level architecture of SATDBailiff.

There is now an opportunity to take advantage of the improved detection tool by Huang et al. [12] to enhance research efforts with a highly accurate, large scale empirical history of SATD instances in Java projects previously unavailable. This can be accomplished alongside fixing some of the data quality issues noted with Maldonado et al.'s [9] empirical study. This study will aim to package these improvements and model in a tool will allow further efforts to expand past these seven previously available software projects in terms of size and quality. In addition to the publication of this tool, an empirical history of SATD instances in 30 open source software projects will be made available as produced by SATDBailiff.

3. Overview of SATDBailiff

SATDBailiff is a Java tool designed to mine the empirical history of SATD instances from Java project Git repositories on a large scale. This is done with the goal of tracking the additions, removals, and changes to SATD instances that occur during the process of software development. SATDBailiff's output can be used to better understand the prominence of SATD in software projects at different points over the course of those projects' lifetimes, while also offering new ways to interpret and visualize those SATD instances. While SATDBailiff accomplishes its objective using a state-of-the-art classification model [6] and a scalable output format, the tool was also designed with modularity in mind to allow for the use of new classification models and output formats. To collect its data, the tool leverages several existing tools as shown in Fig. 1.

The SATDBailiff command line interface (CLI) (Seen in Fig. 1 as ①) has two main inputs: A CSV file containing a list of GitHub repositories will be mined by the tool and a MySQL configuration file. The CLI has many other optional inputs to optimally configure the tool.

The Eclipse JGit library⁴ (③) is used to collect Java source files from GitHub. This library is also used to collect any commit metadata available, which is output alongside any associated SATD operations found during mining. As explained by Nugroho et al. [16], Git offers a diff utility for users to select various diff algorithms like Myers [17] and Histogram [18], which are utilized to obtain the differences of two identical files located in two different commits. Each algorithm has its own procedures for finding the items presented in the original document but absent in the second one and vice versa. JGit is also used to generate edit scripts between different versions of a project's source code. These edit scripts detail which lines contain removals and additions, and represent all changes to a file between one version of a file to the next. Examples of edit scripts can be seen in Figs. 4, 5, 6, and 7. SATDBailiff is configured to use both the Myers [17] and Histogram [18] difference algorithms to generate edit scripts and the tool provides the ability to change between these two algorithms.

JavaParser⁵ (④) is used to extract source code comments from the Java source files obtained by JGit (③). It also extracts comment metadata, containing method and class, and line numbers. This metadata is output alongside any associated SATD operations found during mining.

The SATD Detector tool presented by Liu et al. [6] (⑤) is used for the binary classification of source comments as SATD. This state-of-the-art tool achieved an average F-score of 0.737 during the classification of comments from 5 major open source projects. This classification interface was designed with modularity in mind, and any future higher-performance models can be used with SATDBailiff as well.

⁴ <https://github.com/eclipse/jgit>.

⁵ <https://github.com/javaparser/javaparser>.

Table 3
Simplified sample data from the Apache Tomcat project.

SATD_id	SATD_instance_id	resolution	commit	comment
13958	652915385	SATD_ADDED	09b640e	TODO: 404
14317	652916048	FILE_PATH_CHANGED	decfe2a	TODO: 404
13665	652915615	FILE_REMOVED	a457153	None

The logic that bridges all of these tools together is located within the SATDBailiff client (2). The client begins by generating parent-child pairs for every single parent commit found under a given head of the git repository. In this pairing, commits with multiple parents (i.e., merge commits) are not handled, as the current version of the client handles only one branch. Then, for each of those pairs, all source code differences (edit scripts) are calculated for each Java file. All SATD instances are recorded from *each file* impacted by a source code modification in the parent commit, as well as the child commit. A mapping approach is taken to identify which SATD instances may have been impacted by these changes. An SATD instance will map between two commits if both commits contain the same comment, under the same method signature (or lack thereof), and the same containing class name (or lack thereof). SATD instances that share all of those identification properties (e.g., two identical SATD comments in the same method), a number is assigned based on the order they occur. All SATD instances that were not mapped between the two commits are then classified as removed or changed. This classification is determined by the edit scripts generated earlier, and the logic is further described in the Section 3. The result of this process is a complete empirical history of all operations to SATD instances between a given point in a project's lifetime and its origination.

The implicit implementation of SATDBailiff outputs to an SQL database (6), but the tool supports a modular implementation allowing for an extension of other output formats. A simplified data-point sample from the Apache Tomcat project is included in Table 3. The data includes some important features:

- **SATD Id and SATD Instance Id.** Each entry has two identifying integers. The SATD Id is a unique identifier for a single operation to an SATD Instance. An SATD Instance ID is an overarching identifier used to group many SATD operations to a single contiguous instance. In Table 3, each entry in this sample would have a different and unique SATD Id.
- **Resolution.** Each SATD operation has a single resolution that impacts the SATD between two commits. These operations include: *SATD_ADDED*, *SATD_REMOVED*, *SATD_CHANGED*, *FILE_REMOVED*, *FILE_PATH_CHANGED*, and *CLASS_OR_METHOD_CHANGED*. The definitions of these operations are described in detail in Section 3.
- **Comment Metadata.** When each SATD operation is recorded, SATDBailiff also records the comment's metadata at the time of the operation. This includes data such as the comment type (Line, Block, or JavaDoc as recorded by JavaParser), start and end line, containing class and method, the file name, and the comment itself.
- **Commit Metadata.** When each SATD operation is recorded, SATDBailiff also records the metadata of both the child and parent commit. This includes author name and timestamp, committer name and timestamp, and SHA1 commit hash.

Operations on SATD

Previously, Maldonado et al. [9] conducted an empirical study on the removal of SATD that pointed changes to SATD instance incorrectly as removals and additions. In addition to mistaking file renames, this would detect instances like the examples in Figs. 6 and 7 as having both resolved the original SATD instance and added the new version to the project, respectively.

SATDBailiff resolves this issue by handling SATD comment and file name changes as operations in-between additions and removals of SATD. In order to observe a more fine-grained change in source code changes, the tool observes edit scripts for changes to specific lines of code made between each commit. Edit scripts detail the addition and removal of specific lines of code within a file. An example edit script can be seen in Fig. 6 detailed by the red and green highlighted source text.

Fig. 2 depicts the SATD operations captured by our tool and Fig. 3 shows the distribution of these operations in our five projects. Our tool captures the following cases of the SATD removal: (1) when the file has been deleted– that is, when SATD disappears because the related code is no longer in the system (i.e., accidental removal), and (2) when SATD comments are removed but the code still exists. It is worth noting that the purpose of the tool is not to evaluate the effectiveness of SATD removal or to explore whether SATD is addressed or not. Our study supports the existing studies that focus on these aspects by providing more details and explanations related to the evolution of SATD (e.g., method changed or not, file removed or not, track when SATD starts and when it is removed, etc). The cases that we plan to capture in the extended version of the tool are: (1) when the SATD comment is removed entirely, and a new SATD comment is added instead, although we have not found any such case based on our manual analysis, (2) when SATD comment has dropped while the code still remains unchanged, and (3) when SATD comments are removed and there are changes in the method.

The next subsections describe the process of identifying each of the operations that SATDBailiff handles. The sections use the variables:

- C_a, C_b , the **parent commit** and the more **recent commit**, respectively.
- S_a, S_b , a specific **SATD instance** in the parent commit and the more recent commit, respectively. It should be assumed that S_a and S_b are intentionally related to each other if not identical.

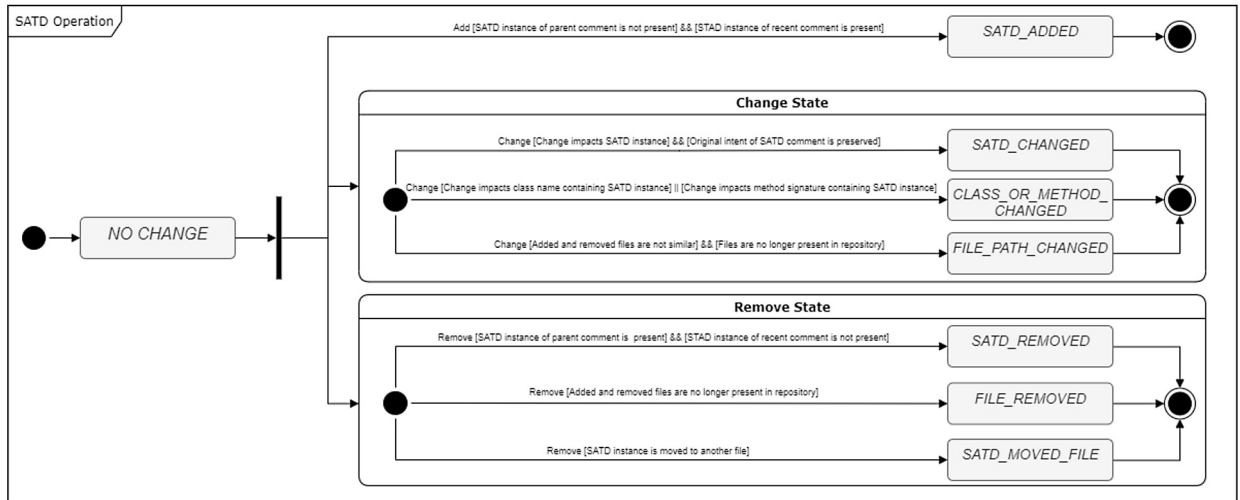


Fig. 2. SATD operations captured by SATDBailiff.

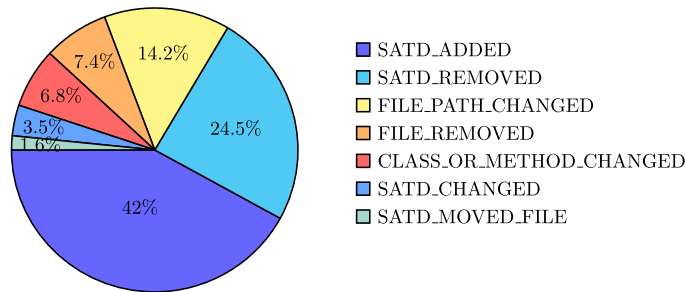


Fig. 3. Distribution of SATD operations in the studied 5 projects. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

```

body = exchange.getOut().getBody();
+ // TODO: what if exchange.isFailed()?
if (body != null) {
  
```

Fig. 4. A basic case **SATD_ADDED** instance.

- SA_a, SA_b , an **arbitrary other SATD instance** in the same file unrelated to S_a or S_b in commits C_a and C_b respectively.
- E_1, E_2, \dots, E_n , the **edit scripts** generated when differencing C_a and C_b that impact the lines of SATD Comment S_1 . Multi-line SATD comments may have multiple edit scripts that impact it where n is used to differentiate these line-based edit scripts.

1. SATD_ADDED & SATD_REMOVED

A naive comment differencing algorithm determines **SATD_ADDED** instances would exist in any C_a where S_a is not present and the associated C_b where S_b is present. This satisfies a basic case in Fig. 4.

However, SATDBailiff needs to account for *changes* in SATD comments. In Fig. 6, these changes would be identified by separate **SATD_REMOVED** and **SATD_ADDED** instances using the naive logic. Instead, SATDBailiff determines that if a single edit script E_n exists such that E_n impacts S_b without impacting SA_a , then S_b was added by C_b .

A naive comment differencing algorithm also determines **SATD_REMOVED** instances would exist in any C_b where S_b is not present and the associate C_a where S_a is present. This satisfies the basic case in Fig. 5.

However, in the case of Fig. 6, it is seen that a more robust algorithm must be used to detect SATD removals. SATDBailiff handles this case such that if a single edit script E_n exists such that E_n impacts S_a without impacting SA_b , then S_a was removed by C_b .

```
protected void connectIfNecessary() {
- // can we avoid copy-pasting?
  if (!client.isConnected()) {
```

Fig. 5. A basic case **SATD_REMOVED** instance.

```
logger.log('`Init successful`');
- // Moved this config to the bottom
+ // Moved this config
+ // to the bottom
super.init();
```

Fig. 6. A case of a would-be false **SATD_ADDED** and **SATD_REMOVED** instances.

```
try {
- // Maybe this already existst
+ // Maybe this already exists
  success = client.changeDir(dirName);
```

Fig. 7. A valid **SATD_CHANGED** instance.

```
c = endpoint.createChannel(session);
- // TODO: what if creation fails?
+ // Bug 1402
c.connect();
```

Fig. 8. A possibly valid but false **SATD_CHANGED** instance.

It can also be the case that the previous logic used by SATDBailiff to classify additions and removals to be false, and for the tool to still classify an operation as an **SATD_ADDED** or **SATD_REMOVED**. This case is better identified by the logic in the following **SATD_CHANGED** section.

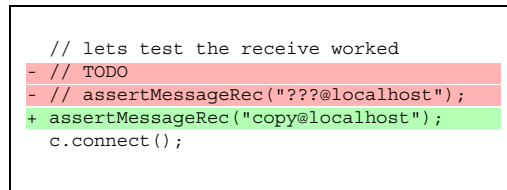
2. **SATD_CHANGED**

Changes in an SATD comment can be difficult to determine because they can remove the SATD comment entirely or replace it with a new non-SATD, or irrelevant, comment. Ideally, only SATD comments which preserve the original intent of the SATD comment should be recorded as a **SATD_CHANGED**. However, this requires the comprehension of the comment and its relationship with the TD described. Any updates on the comment text may be related to the addition or deletion of more details, as seen later in the manual validation. Since we cannot assess whether an update to an SATD means partially or totally addressing it, our tool will report any form of textual change to the SATD, as **SATD_CHANGED**. To do so, the edit scripts of the file are first observed. If a single edit script E_n exists such that E_n impacts both S_a and SA_b , then it can be determined that a change may have occurred. Our logic currently flags as **SATD_CHANGED** additions of newlines (see Fig. 6), spelling corrections (see Fig. 7), addition or removal of adjacent related and unrelated comments, and URL updates.

SATDBailiff also checks the updated comments to determine if they still can be classified as SATD instances (see Figs. 8 and 9). Fig. 9 shows an example of an SATD instance which is recorded in C_a as "lets test the receive worked\nTODO" due to how the tool groups adjacent comments. In C_b , the removal of the \nTODO substring of the instance results in the instance no longer being classified as SATD, and thus SATDBailiff reports this instance as **SATD_REMOVED**. Without this additional verification, this SATD instance would be incorrectly reported as having only been changed. However, this approach has its own limitation: if a developer removes one SATD and adds another entirely unrelated SATD, for the same method, in the same commit, like in Fig. 8, our tool would flag this as **SATD_CHANGED**. Fortunately, this case does not seem to be frequent, as shown later in the manual validation, the precision of our model is up to **96%**. Yet, we report the limitation of our tool in Section 7, and we discuss our current explorations to address it.

3. **CLASS_OR_METHOD_CHANGED**

The final edit script source code change detected by SATDBailiff is modification to an SATD instance's containing class or method. These cases are detected if any E_n impacts the class or method containing S_a such that the method signature or



```
// lets test the receive worked
- // TODO
- // assertMessageRec("???@localhost");
+ assertMessageRec("copy@localhost");
c.connect();
```

Fig. 9. SATD_REMOVED instance removing only part of a comment.

the class name are changed. SATDBailiff does not currently identify when SATD is moved throughout a file by multiple separate edit scripts.

4. FILE_REMOVED & FILE_PATH_CHANGED

File removals are detected implicitly by Git, where a similarity between added and removed files determines whether a file is removed or renamed when it is no longer present in the repository when committed. This detection method was available as part of the JGit library, and was utilized for identifying **FILE_REMOVED** and **FILE_PATH_CHANGED** instances.

4. SATDBailiff usage

This section describes the usage of SATDBailiff and its features.

4.1. Installation

The most up-to-date precompiled binaries can be found on the project's GitHub⁶ repository and on the tool's website.⁷ The project can be run using a Java version 8+ and is otherwise OS independent as it could be run through Docker.

4.2. Usage

SATDBailiff can be used either through its Command Line Interface (CLI) or Application Program Interface (API), and can be easily modified to support different output types and classification models.

The SATDBailiff Command Line Interface (CLI) (Seen in Fig. 1 as ①) has two main inputs: A CSV file containing a list of GitHub repositories and a MySQL configuration file. The CLI has many other optional inputs to optimally configure the tool.

The CSV file containing repository information details which repositories will be mined by the tool, and where the mining will terminate. A terminal commit value can be added next to the repository URI to add terminal point in time to which the tool can mine. This is done primarily to assure reproducibility between datasets mined at different times. If absent, the terminal commit value will default to the most recently available commit in the repository. The output format of the data also allows for a manual filtering of SATD operations by date. However, it should be noted that un-merged branches of the repository at certain timestamps are likely to cause difference between a pre- and post-execution commit filtering. Git credentials for private repositories can be added as a separate program argument.

By default, SATDBailiff outputs to a MySQL database. The tool intends for a MySQL database to be set up to receive the system's output. Configuration fields for this database must be supplied to the program to connect to the database. A description for the required fields, as well as the required schema for the database can be found in the GitHub repository.

Other runtime variables available within SATDBailiff include:

- File differencing algorithms available through JGit - currently only Histogram and Myers;
- The Normalized Levenshtein distance threshold (between 0 and 1) described in Section 3;
- A toggle for error display;
- A help menu display;

When run, SATDBailiff will display using the interface in Fig. 10. This output includes a detailed description of the runtime duration of the tool, the number of commits differences, and a description of each entry that caused any sort of detectable error in the system, if errors display is toggled. If error display is disabled, then only the number of errors encountered will be displayed.

⁶ <https://github.com/smilevo/SATDBailiff>.

⁷ https://smilevo.github.io/self-affirmed-refactoring/SCP20_index.html.

```
$ java -jar SATDBailiff.jar -d mySQL.properties -r repos.csv
Completed analyzing 78 diffs in 26,103ms (334.65ms/diff, 0 errors) - analogweb/core
3wks/thundr -- Mining SATD (13.3%, 46/346, 0 errors) - e048736
```

Fig. 10. Sample runtime snapshot.

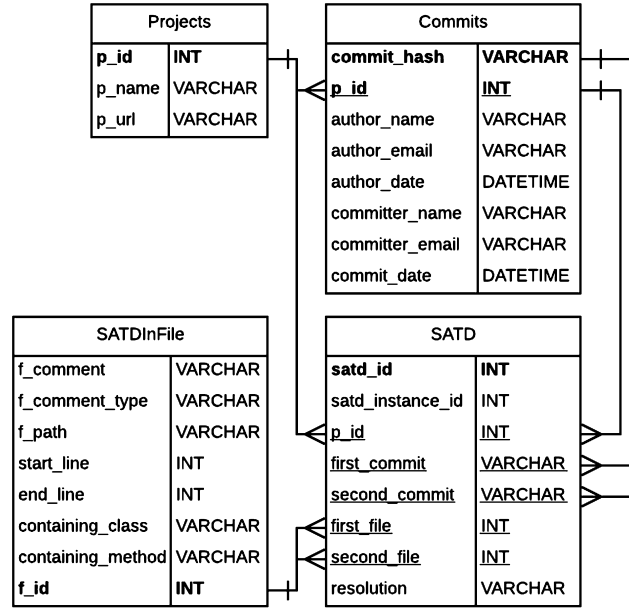


Fig. 11. SATDBailiff output schema.

4.3. Interpreting output

The released implementation of SATDBailiff outputs to a SQL database. The schema for the output is detailed in Fig. 11. Project table contains the list of open-source Java projects hosted on GitHub that are utilized in the study. Commits table stores commit level metadata by looping through the commit log for each studied project. The metadata includes author name and timestamp, committer name and timestamp, and SHA1 commit hash. A high level information about SATD is stored in the SATD table, whereas SATD comment's metadata including the comment type, start and end line, containing class and method, the file name, and the comment itself is recorded in SATDInFile table.

SATDBailiff also generates a CSV file and an HTML file that contain the same information saved in the database. The CSV file is helpful for results parsing, while the HTML helps with the visualization of the results.

5. SATDBailiff applicability

Tracking SATDs are vital for maintaining healthy software systems and reducing maintenance cost. As reported by Wehaibi et al. [5], SATD has negative implications on the software development process since it makes it hard to perform any future changes. Our tool could serve as a guide; helping developers pay technical debt by recommending the changes they should perform. Tracking the additions, removals, and changes to SATD instances that occur during the process of software development can be used to better understand the prominence of SATD in software projects at different points over the course of those projects' lifetimes. Additionally, tracking SATD instances would provide valuable support for existing studies by capturing the removals without the need to manually validate them and will help future research to explore how SATD is addressed and how SATD removal should be measured (e.g., evaluate the effectiveness of the SATD removal when a method has changed along with the removal). Additionally, our tool can assist in exploring ways to reduce/manage technical debt. For instance, a recent study [11] shows that there is a higher chance for refactoring actions to occur together with SATD removals than with other changes. Another recent study on refactoring documentation [19–21] shows that developers reported "Fix Technical Debt" in their commit messages which shows that refactoring actions could be used to cope with technical debt.

Martini et al. [22] conducted an industrial case study with practitioners and their results show that 25.9% of the average development time and effort is spent by 215 practitioners to manage technical debt. In their survey, they found that some respondents report spending more than 40% of their time managing technical debt. However, only 26% of the participants

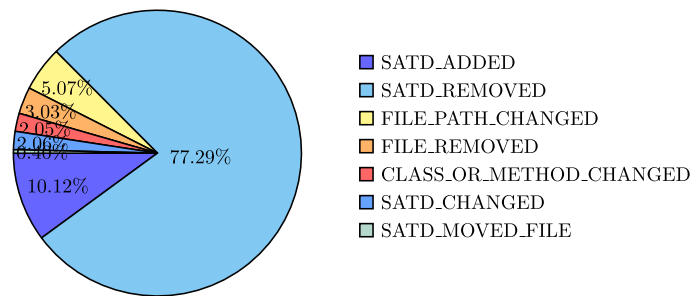


Fig. 12. Distribution of SATD operations in the studied 30 projects.

used a tool to track technical debt. The respondents mentioned the tools used to track technical debt are: comments, documentation, issues, backlog, static analyzer, lint and test coverage. Backlogs (i.e., Jira, Hansoft, and Excel) are the most used tool among the participants. According to their survey with the practitioners, respondents show awareness of indicators of technical debt like comments or documentation. This sheds light on the need of developing tools to help with automatically tracking technical debt early in the development stages. Here are some examples that showcase the usefulness of the tool:

- **Promoting the adoption of tracking SATD in practice.** Fixing bugs or adding new features would be challenging and a time consuming task as the system evolves, and software engineers might experience a negative impact on the quality. Adopting technical debt tracking processes by managers or experienced developers, who understand the importance of technical debt, reduces maintenance cost. Our tool is one of the methods that helps promote the adoption of tracking SATD practice. This feature would be fully achieved with future refinement of the tool, once we gather more feedback from researchers and practitioners who are using the tool.
- **Indicating refactoring technique that removed the SATD.** Since the management of technical debt has been heavily correlated with refactoring, the tool shows the history of SATDs and the refactoring operations that helped remove the debt. This assists in analyzing the necessary changes needed to remove technical debt.

6. SATDBailiff validation

6.1. Internal validation:

To verify the accuracy of SATDBailiff, a manual analysis was performed on a sample of 1,882 entries mined from 5 large open source Java projects described in Table 1, each as their own strata. The number of samples taken from each project is determined by the total number of SATD instances mined from the projects. Each of the SATD instances selected includes SATD operations (removed or changed) performed on a single instance of SATD. Each instance in the sample will represent an entirely unique instance of SATD. A simplified example of a single SATD instance can be seen in Table 3. The results of this analysis can be seen in Tables 5 and 6.

For the tool validation, we selected five projects that are widely used by the state-of-the-art research in the context of SATD [9–11,14], as these projects are claimed to have an adequate number of SATD instances. We build on top of findings of Maldonado et al. [9] and start from their SATD removal dataset. In addition to the publication of this tool, an empirical history of SATD instances in 30 open source software projects, produced by SATDBailiff, is available on our website.

The detail of the studied 30 projects is summarized in Table 4. Fig. 12 shows the distribution of SATD operations captured by our tool in the 30 projects. We observe that the dominant category is the SATD_REMOVED with 77.29%. Our future work involves exploring the technique developers used to remove SATDs from their projects.

SATDBailiff was configured to only mine SATD instances between the original commit to each project, and the most recent commit reported by the Maldonado et al. study.

To perform the analysis, one of the authors was given the set of SATD instances and asked to locate the exact location of each of the SATD operations using the GitHub website. A “correct” entry was identified as an entry in which every operation made to the SATD instance could be located using the GitHub interface. Any unnecessary additional, missing, or inaccurate operations found on GitHub would result in the entire entry being incorrect. For entries that were not removed from the project, their existence in the terminal commit supplied to SATDBailiff was confirmed. For transparency of this analysis, a GitHub link to the exact source modification was recorded in each of the projects where available. These results are available on the project’s website.⁸ During validation, it was assumed that all binary classifications of source code comments as SATD were correct.

The results of the manual analysis (Tables 5 and 6) find SATDBailiff to have a precision of 0.99 and 0.96 for SATD removal and SATD changes, respectively. It is infeasible to always achieve a perfect precision of the results extracted from the tool

⁸ https://smilevo.github.io/self-affirmed-refactoring/SCP20_index.html.

Table 4
Details of the studied 30 projects.

Project	# Commits	# Contributors	# detected SATD
gitools/gitools ¹	2,051	4	1,704
twilio/twilio-java ²	118	3	89
hbutani/sqlwindowing ³	424	2	26
pulse00/twig-eclipse-plugin ⁴	499	4	359
eclipse/vert.x ⁵	4,950	196	452
ngdata/lilyproject ⁶	2,717	7	3,908
hibernate/hibernate-validator ⁷	4,198	78	885
wocommunity/wonder ⁸	13,885	70	2,215
apache/maven-plugins ⁹	15,063	36	2,300
davemckain/qtiworks ¹⁰	2,065	2	635
icy-imaging/icy-kernel ¹¹	665	5	978
jfxtras/jfxtras-labs ¹²	1,834	26	3,136
romanchyla/montysolr ¹³	1,628	6	927
apache/httpclient ¹⁴	3,225	51	690
socialsoftware/blended-workflow ¹⁵	849	6	824
crosswire/jsword ¹⁶	1,865	8	1,275
motech/motech-whp ¹⁷	2,190	14	97
eclipse/bpmn2-modeler ¹⁸	1,442	6	1,072
eclipse/ecf ¹⁹	11,793	11	1,559
gettrai/railo ²⁰	3,990	15	862
msbarry/xtest ²¹	341	1	367
projectdanube/xdi2 ²²	2,504	3	254
scribble/scribble-java ²³	2,139	4	6,803
tinkerpops/gremlin ²⁴	1,227	15	159
adangel/pmd ²⁵	17,691	99	3,997
qcadoo/mes ²⁶	14,906	30	1,440
belaban/jgroups ²⁷	19,213	72	1,193
merks/xcore ²⁸	118	3	89,175
rvonmassow/xdoc ²⁹	151	548	6
nasa/certware ³⁰	237	4	1,091

¹ <https://github.com/gitools/gitools>

² <https://github.com/twilio/twilio-java>

³ <https://github.com/hbutani/sqlwindowing>

⁴ <https://github.com/pulse00/twig-eclipse-plugin>

⁵ <https://github.com/eclipse/vert.x>

⁶ <https://github.com/ngdata/lilyproject>

⁷ <https://github.com/hibernate/hibernate-validator>

⁸ <https://github.com/wocommunity/wonder>

⁹ <https://github.com/apache/maven-plugins>

¹⁰ <https://github.com/davemckain/qtiworks>

¹¹ <https://github.com/icy-imaging/icy-kernel>

¹² <https://github.com/jfxtras/jfxtras-labs>

¹³ <https://github.com/romanchyla/montysolr>

¹⁴ <https://github.com/apache/httpclient>

¹⁵ <https://github.com/socialsoftware/blended-workflow>

¹⁶ <https://github.com/crosswire/jsword>

¹⁷ <https://github.com/motech/motech-whp>

¹⁸ <https://github.com/eclipse/bpmn2-modeler>

¹⁹ <https://github.com/eclipse/ecf>

²⁰ <https://github.com/gettrai/railo>

²¹ <https://github.com/msbarry/xtest>

²² <https://github.com/projectdanube/xdi2>

²³ <https://github.com/scribble/scribble-java>

²⁴ <https://github.com/tinkerpops/gremlin>

²⁵ <https://github.com/adangel/pmd>

²⁶ <https://github.com/qcadoo/mes>

²⁷ <https://github.com/belaban/jgroups>

²⁸ <https://github.com/merks/xcore>

²⁹ <https://github.com/rvonmassow/xdoc>

³⁰ <https://github.com/nasa/certware>

due to the inconsistent nature of open source projects and their development practices. Open source projects are varied in size, contributors, number of comments, and SATDs. Thus, the performance of the tool might differ based on development practices of the selected projects. While a higher level of accuracy could have been achieved, it should be noted that many of the incorrect instances were partially correct. For example, instances frequently were found to be incorrect because they became dissociated with one another, where a connection between an SATD instance's addition to the project and its deletion from the project was not made by the tool. In cases where only the additions or removals are observed from the dataset, the accuracy of the data provided by the tool is much more reliable.

Table 5

SATDBailiff manual analysis results for SATD removal.

Project	# Entries	# False Positive	Precision
Camel	20	1	0.99
Gerrit	284	6	0.99
Hadoop	608	3	0.99
Log4j	7	0	1.00
Tomcat	432	8	0.99
Total	1351	18	0.99

Table 6

SATDBailiff manual analysis results for SATD changed.

Project	# Entries	# False Positive	Precision
Camel	19	2	0.99
Gerrit	223	25	0.92
Hadoop	112	13	0.96
Log4j	9	0	1.00
Tomcat	168	24	0.93
Total	531	64	0.96

```
protected void processSoapConsumerOut
(Exchange exchange) throws Exception {
LOG.info("processSoapConsumerOut:" + exchange);
- // TODO
+ // TODO check if the message is oneway mes-
+ // Get the method name form the soap end-
+ point
...
}
```

Fig. 13. An expansion case of **SATD_CHANGED** instance.

Difficulties in solving many of the tool's issues came from the imperfect nature of working with edit scripts produced by Git differencing tools. Edit scripts are used to show an algorithm's best guess of changes in files inside of a Git repository, and do not always reflect the true intentions of the developer who made them [23]. An example of an edit script can be seen in Fig. 7 depicted as the red and green highlight used to represent a source code change. JGit uses the Myers [17] and the Histogram [18] diff algorithm to produce edit scripts. SATDBailiff provides the ability to change between these two algorithms, and the Myers algorithm was used during performed manual validation. Both of these algorithms maintain a manually validated accuracy of less than 0.9 [23]. While, in many cases, an invalid edit script will not directly invalidate SATDBailiff's ability to identify operations to SATD instances, this inaccuracy still serves as a significant limitation in the upper bound of accuracy achievable by this tool.

To assure that these errors are not able to silently pollute the dataset, the tool reports any known errors that are encountered during the mining process. This workaround was taken as an optimistic precaution for an issue that may not have a perfect alternative solution. For example, SATD that is added during a merge commit which was not present in either of the merge branches is not detected with an SATD_ADDED entry. If that SATD is modified or removed later, then the entry would be added to the project before the SATD_ADDED entry was found. Because the search occurs chronologically starting with the oldest commit in the project, the system can detect this as an issue and will output an error to the terminal during runtime.

When performing the manual validation, we noticed that developers changed SATD comments as follows:

- An expansion form of SATD comment: a comment that has more explanation of the SATD case. For instance, in Fig. 13, the comment contains more than just the TODO tag, as the developer added a couple of functionalities to implement, as part of the TODO.
- An abbreviation form of SATD comment: the opposite of the expansion in which a shortened form of SATD comment is provided. For instance, in Fig. 14, the text explaining the task, tagged with the TODO, has been removed, however, the TODO has been kept, and therefore, the comment is still flagged as SATD.
- A generalization form of SATD comment: a comment has broad SATD context. In Fig. 15, the details of the TODO have been removed, and the reference to the active issue has been kept.
- A specialization form of SATD comment: the opposite of the generalization in which a comment has focused SATD context. For example, in Fig. 16, details on how to fix the test and make it support the Object message, have been added.

```

public void onExchange(HttpExchange exchange)
{
- // we need an external HTTP client such as
commons-httpclient// TODO
+ // TODO
}

```

Fig. 14. An abbreviation case of SATD_CHANGED instance.

```

- //TODO: YARN-3284//The containerLocality met-
rics will be exposed from AttemptReport
+ // TODO:YARN-3284
private void createContainerLocalityTable
(Block html) {
...
}

```

Fig. 15. A generalization case of SATD_CHANGED instance.

```

- // TODO fix this test
+ // TODO fix this test, it looks like AMQP
don't support Object message
public void xtestJmsRouteWithObjectMessage()
throws Exception {
...
}

```

Fig. 16. A specialization case of SATD_CHANGED instance.

```

public void afterPropertiesSet() {
- // TODO: is needed when we add support for
when predicate
if (getOutputs().size() == 0) {
+ // no outputs
return;
}
}

```

Fig. 17. A case of SATD_REMOVED instance when there is no overlap.

As for the SATD removal case, we noticed the case in which there is entirely no overlap between the SATD comments as shown in Fig. 17.

6.2. External validation:

To further assess the usefulness of the tool, we perform an external validation by involving 15 graduate students from the Department of Software Engineering and Data Science at Rochester Institute of Technology. All participants volunteered to participate in the experiment. For each participant, the experiment consists of (1) randomly choosing Java-based open-source project and fork it, (2) detecting 10 issues in the open-source project, using PMD and/or SpotBugs, (3) writing a comment, describing the issue detected inside the method, while making sure the comment is written in the form of SATD, then committing it (4) addressing the issues one by one by making the necessary code changes recommended by PMD and/or SpotBugs, while making sure to remove the comment describing it, then committing the changes (5) running SATDBailiff tool to identify these added and removed comments, and finally (6) filling out the survey to give feedback about the tool. The idea behind this experiment is to identify potential TD instances, and document their existence, then addressing them and removing their corresponding comments. We used PMD and SpotBugs because they are known to be good indicators for technical debt in the source code [24,25]. Upon the completion of the experiment, we calculated the average recall, and

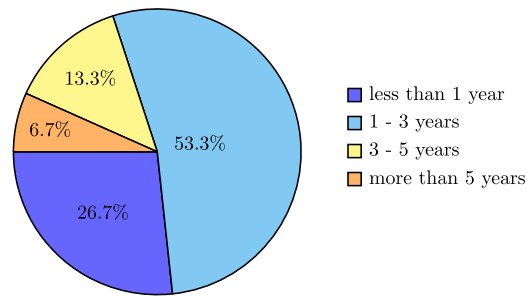


Fig. 18. Participant programming experience in years.

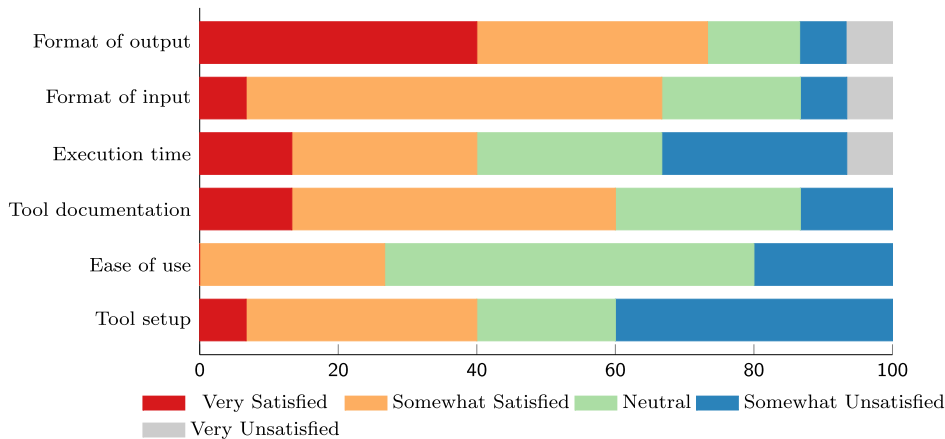


Fig. 19. Level of satisfaction with the aspects of SATDBailiff tool.

we found the average recall is 0.90, which is considered acceptable. The replication package for the whole experiment is available in the project website.⁹

The survey consisted of 6 questions that are divided into 2 parts. The first part of the survey includes demographics questions about the participants. In the second part, we asked about the (1) satisfaction of the tool aspects, (2) preferred features of the tool, (3) suggested features of the tool to be added in the future, and (4) any comments about the tool. We constructed the survey to use 5-point ordered response scale (“Likert scale”) question on the aspects of the tool, and 3 open-ended questions on the preferred features, suggested features, and general comments about the tool.

As shown in Fig. 18, the experience of these participants with programming ranged from less than a year to more than 5 years. As for the familiarity with the concept of technical debt, 66.7% of the participants are familiar with the technical debt, whereas 33.3% are not familiar with the concept of technical debt. Prior to the execution process, the participants were provided with a 75-minute tutorial on technical debt along with reference materials.

Fig. 19 presents an overview of the aspects of the tool and the satisfaction of the participants. With respect to tool setup, some of the respondents reported that they are satisfied with the tool. However, there are a few participants who are usually not happy with how the tool is set up and we are planning on improving the tool setup in the future. For the easy to use aspect, a larger group (10 participants) selected neutral. Some of the participants found that the tool is not easy to use, so we will work on improving the usability of the tool. The main feedback from the participants is to add GUI features, and we think that GUI might be an option to be considered. Adding GUI is not one of our priorities because we have more features that we would like to improve, but we will consider adding it in our future extension of the tool. Regarding the tool documentation, the majority of the respondents agreed that the documentation is useful; only 2 participants are somewhat unsatisfied. Concerning the execution time, the participants are happy with it, although we don’t have a consensus between participants and we believe that execution time depends on the selected project size. For the format of the input, the vast majority of respondents agree that the format is acceptable. Likewise, the participants are happy with the format of the output.

With respect to the preferred features of the tools, the participants listed a variety of features which are centered around six main topics: (1) tool correctness, (2) tool accuracy, (3) tool documentation, (4) execution time, (5) data storage, and (6) Docker integration. We demonstrated some of the responses:

⁹ https://smilevo.github.io/self-affirmed-refactoring/SCP20_index.html.

"The tool performed more smartly and accurately than I had anticipated, since one of my SATD resolving commits was actually in 2 parts because its was long, I assumed that would cause some inaccuracy and confusion in the tool. But surprising and delightfully the tool identified and clubbed the 2 commits and also connected to the reported SATD beautifully."

"The integration with Docker allowed the tool to be setup very easily on my system so I would have to say that 'addon' was fantastic. I found myself struggling to 1) Build the tool from the source code and 2) running the executable jar. Once I figured out how to download/install the image everything began to go smoothly for me. That integration is not necessarily a feature of the tool itself, so in that regard I would say I really liked the reports generated. The html + csv output were well designed and labels were intuitive. On top of that, I didn't have to do anything crazy to get the tool to run on my repo, just changed the repo url in the csv file which was quite simple. I appreciated that about the tool."

"I liked how fast are the SATD mining process and the correctness of the output"

"I like how the tool is analyze the comments with high accuracy of detect each comment that was add or removed. Also, the way of the data is stored, the user have 3 options to analyze the data excel, html report and Sql database. In MySql database there are multiple tables with different information the user can create customize report based on his requirements."

As for the suggested features of the tool, participants mainly mentioned 3 main features: (1) IDE integration, (2) issue troubleshooting, and (3) specifying the starting and ending commits. Participants did explicitly share their concerns during the survey as follows:

"I feel that a frontend GUI that would help feeding in the data and also a simple GUI that did the MYSQL query in the backend and for us it would just be a download button would make the tool much more friendly and intuitive to use." and *"A GUI would be good as everything right now is terminal based."*

"Add as a plugin and add more details about how to use the tool if a particular error is found."

"I think having a way to choose a starting and ending point would be great"

7. SATDBailiff limitations and upcoming features

In this section, we outline the known limitations to our tool, and the features we are planning on developing.

- As pointed out in Section 3, our current tool does not track comments introduced by commits with multiple parents. We have not yet performed any manual analysis to assess the extent to which this affects the tracking of current comments. In the future, we plan to perform a thorough examination to determine whether we should prioritize handling multiple parents.
- To reduce the false positiveness of our tool, as pointed out by the experiments, we are currently exploring various text processing techniques to preprocess the comments and reduce the effects of special characters that may interfere with our string matching.
- We are also investigating the use of pre-trained models that can detect whether comments are pointing out to the same technical debt. Such a model can help us avoid detecting two different comments as being the same one, in the case of removing one SATD and adding another one, for the same method, and the same commit.
- Since the management of technical debt has been heavily correlated with refactoring, we have already integrated the Refactoring Miner library, through its API, to run on the commits that we are also detecting the removal of SATD in. This will support existing and future studies that analyze the necessary changes needed to remove technical debt. Our tool currently reports all the refactorings that are associated with the classes and methods containing SATD comments. This is a recent feature that we are still testing, therefore we do not have any observations about it in this paper. We anticipate that this feature will help researchers in developing empirical evidence of the usefulness of refactoring in terms of managing technical debt.
- Upon performing the external validation, we found that the participants pointed out limitations that are mainly related to tool setup and usage. In the future, we plan to improve the setup and the usability of the tool by adding several features mentioned by the participants, including but not limited to, IDE integration, issue troubleshooting, and specifying the starting and ending commits. Further, the external validation is currently conducted with a group of software engineering students. As future work, we plan to perform another round of external validation with professional software engineers in industry to hear their perception.

8. Threats to validity

In this section, we identify potential threats to the validity of our approach and our experiments.

Threats to the validity of this tool include the limited manual evaluation and general lack of testing. An important potential threat relates to our manual classification. Since the manual verification of samples is a human intensive task and it is subject to personal bias, we mitigate this first by selecting SATD instances from an existing dataset. Then we performed the tracking of their existence by one author. Only 1,882 samples of SATD were recorded and addressed to determine the accuracy of SATDBailiff.

Another threat relates to the SATD instances that are extracted only from open source Java projects. Our results may not generalize to commercially developed projects, or to other projects using different programming languages. Another

threat concerns the generalization of SATD patterns used in this study. Since a method is considered holder of TD when a comment contains SATD, this may not generalize to other projects if they do not allow inline documentation (comments).

The reliance on the existing tool (i.e., SATD Detector) is another threat to validity due to the possibility of introducing false positive results. Thus, the binary classification of SATD comments resulting from the SATD Detector might have an impact on our findings.

9. Conclusion and future work

This paper presents a preview of our tool, SATDBailiff, and discusses its benefits. The tool aims at offering an unmatched ability to extract SATD instances and the development operations upon them from Git repositories. This paper discussed the benefits that a high quality empirical history would have for the further study of SATD.

A high level of accuracy was achieved with SATDBailiff, however there is still opportunity for improvement. In their nature, source code differencing tools may not always record modifications that reflect the true nature of a developer's intentions [26]. It is because of these inaccuracies that solving each and every edge case is infeasible for the purpose of data generation. Some attempts to fix these edge-cases were made, however their large numbers and unpredictability made it a relatively futile task. To compensate, a list of known edge cases is included on the tool's website to detail known areas where inaccuracies may appear. As Edit Scripts are more reliably instantiated from these differences, more accurate detection of SATD-impacting operations will be possible.

Edit scripts were generated for this tool using the Histogram and Myers algorithms made available through JGit. Using a differencing tool like GumTree [26] or a hybrid approach [27] may produce more accurate results. However, GumTree currently does not offer the tracking of comment changes and does not plan on implementing that functionality.¹⁰ Modification of the edit script generation methodology may also have positive impacts on the project's runtime which may currently be excessive for larger project.

Some other issues encountered include the handling of non-English language source comments. None of the projects used to validate SATDBailiff contained any known instances of these comments, however the larger dataset of 30 projects released alongside the tool does contain these instances. Ideally, a tool would be able to detect a non-English comment before classifying it as SATD, but no attempt was made to solve this issue as it only represents a small subset of the addressed projects.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] P. Kruchten, R.L. Nord, I. Ozkaya, Technical debt: from metaphor to theory and practice, *IEEE Softw.* 29 (6) (2012) 18–21.
- [2] W. Cunningham, The wycash portfolio management system, *ACM SIGPLAN OOPS Messenger* 4 (2) (1992) 29–30.
- [3] A. Potdar, E. Shihab, An exploratory study on self-admitted technical debt, in: 2014 IEEE International Conference on Software Maintenance and Evolution, IEEE, 2014, pp. 91–100.
- [4] G. Bavota, B. Russo, A large-scale empirical study on self-admitted technical debt, in: *Proceedings of the 13th International Conference on Mining Software Repositories*, 2016, pp. 315–326.
- [5] S. Wehaibi, E. Shihab, L. Guerrouj, Examining the impact of self-admitted technical debt on software quality, in: 2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER), vol. 1, IEEE, 2016, pp. 179–188.
- [6] Z. Liu, Q. Huang, X. Xia, E. Shihab, D. Lo, S. Li, Satd detector: a text-mining-based self-admitted technical debt detection tool, in: *Proceedings of the 40th International Conference on Software Engineering: Companion Proceedings*, 2018, pp. 9–12.
- [7] E.d.S. Maldonado, E. Shihab, N. Tsantalis, Using natural language processing to automatically detect self-admitted technical debt, *IEEE Trans. Softw. Eng.* 43 (11) (2017) 1044–1062.
- [8] E.d.S. Maldonado, E. Shihab, Detecting and quantifying different types of self-admitted technical debt, in: 2015 IEEE 7th International Workshop on Managing Technical Debt (MTD), 2015, pp. 9–15.
- [9] E.D.S. Maldonado, R. Abdalkareem, E. Shihab, A. Serebrenik, An empirical study on the removal of self-admitted technical debt, in: 2017 IEEE International Conference on Software Maintenance and Evolution (ICSME), 2017, pp. 238–248.
- [10] F. Zampetti, A. Serebrenik, M. Di Penta, Was self-admitted technical debt removal a real removal? An in-depth perspective, in: 2018 IEEE/ACM 15th International Conference on Mining Software Repositories (MSR), 2018, pp. 526–536.
- [11] M. Iammarino, F. Zampetti, L. Aversano, M. Di Penta, Self-admitted technical debt removal and refactoring actions: co-occurrence or more?, in: 2019 IEEE International Conference on Software Maintenance and Evolution (ICSME), IEEE, 2019, pp. 186–190.
- [12] Q. Huang, E. Shihab, X. Xia, D. Lo, S. Li, Identifying self-admitted technical debt in open source projects using text mining, *Empir. Softw. Eng.* 23 (05 2017), <https://doi.org/10.1007/s10664-017-9522-4>.
- [13] M.A. de Freitas Farias, M.G. de Mendonça Neto, M. Kalinowski, R.O. Spínola, Identifying self-admitted technical debt through code comment analysis with a contextualized vocabulary, *Inf. Softw. Technol.* 121 (2020) 106270.
- [14] F. Zampetti, A. Serebrenik, M. Di Penta, Automatically learning patterns for self-admitted technical debt removal, in: 2020 IEEE 27th International Conference on Software Analysis, Evolution and Reengineering (SANER), IEEE, 2020, pp. 355–366.

¹⁰ <https://github.com/GumTreeDiff/gumtree/issues/39>.

- [15] L. Xavier, F. Ferreira, R. Brito, M.T. Valente, Beyond the code: mining self-admitted technical debt in issue tracker systems, arXiv preprint, arXiv: 2003.09418, 2020.
- [16] Y.S. Nugroho, H. Hata, K. Matsumoto, How different are different diff algorithms in git?, *Empir. Softw. Eng.* 25 (1) (2020) 790–823.
- [17] E.W. Myers, An o(nd) difference algorithm and its variations, *Algorithmica* 1 (1986) 251–266.
- [18] B. Cohen, <https://alfeddenzo.livejournal.com/170301.html>.
- [19] E.A. AlOmar, M.W. Mkaouer, A. Ouni, Can refactoring be self-affirmed? An exploratory study on how developers document their refactoring activities in commit messages, in: *Proceedings of the 3rd International Workshop on Refactoring*, IEEE, 2019.
- [20] E.A. AlOmar, M.W. Mkaouer, A. Ouni, Toward the automatic classification of self-affirmed refactoring, *J. Syst. Softw.* (2020) 110821.
- [21] E.A. AlOmar, A. Peruma, M.W. Mkaouer, C. Newman, A. Ouni, M. Kessentini, How we refactor and how we document it? On the use of supervised machine learning algorithms to classify refactoring documentation, *Expert Syst. Appl.* 167 (2021) 114176.
- [22] A. Martini, T. Besker, J. Bosch, Technical debt tracking: current state of practice: a survey and multiple case study in 15 large organizations, *Sci. Comput. Program.* 163 (2018) 42–61.
- [23] V. Frick, T. Grassauer, F. Beck, M. Pinzger, Generating accurate and compact edit scripts using tree differencing, in: *2018 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 2018, pp. 264–274.
- [24] F. Zampetti, C. Noisieux, G. Antoniol, F. Khomh, M. Di Penta, Recommending when design technical debt should be self-admitted, in: *2017 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, IEEE, 2017, pp. 216–226.
- [25] T. Amanatidis, N. Mittas, A. Moschou, A. Chatzigeorgiou, A. Ampatzoglou, L. Angelis, Evaluating the agreement among technical debt measurement tools: building an empirical benchmark of technical debt liabilities, *Empir. Softw. Eng.* 25 (5) (2020) 4161–4204.
- [26] J. Falleri, F. Morandat, X. Blanc, M. Martinez, M. Monperrus, Fine-grained and accurate source code differencing, in: *ACM/IEEE International Conference on Automated Software Engineering, ASE '14*, Vasteras, Sweden, September 15–19, 2014, 2014, pp. 313–324.
- [27] J. Matsumoto, Y. Higo, S. Kusumoto, Beyond gumtree: a hybrid approach to generate edit scripts, in: *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, 2019, pp. 550–554.