

Automated Identification of On-hold Self-admitted Technical Debt

Rungroj Maipradit*, Bin Lin[†], Csaba Nagy[†]
 Gabriele Bavota[†], Michele Lanza[†], Hideaki Hata*, Kenichi Matsumoto*

*Nara Institute of Science and Technology, Japan

[†]Software Institute, USI Università della Svizzera Italiana, Switzerland

Abstract—Modern software is developed under considerable time pressure, which implies that developers more often than not have to resort to compromises when it comes to code that is well written and code that just does the job. This has led over the past decades to the concept of “technical debt”, a short-term hack that potentially generates long-term maintenance problems. Self-admitted technical debt (SATD) is a particular form of technical debt: developers consciously perform the hack but also document it in the code by adding comments as a reminder (or as an admission of guilt). We focus on a specific type of SATD, namely “On-hold” SATD, in which developers document in their comments the need to halt an implementation task due to conditions outside of their scope of work (e.g., an open issue must be closed before a function can be implemented).

We present an approach, based on regular expressions and machine learning, which is able to detect issues referenced in code comments, and to automatically classify the detected instances as either “On-hold” (the issue is referenced to indicate the need to wait for its resolution before completing a task), or as “cross-reference”, (the issue is referenced to document the code, for example to explain the rationale behind an implementation choice). Our approach also mines the issue tracker of the projects to check if the On-hold SATD instances are “superfluous” and can be removed (i.e., the referenced issue has been closed, but the SATD is still in the code). Our evaluation confirms that our approach can indeed identify relevant instances of On-hold SATD. We illustrate its usefulness by identifying superfluous On-hold SATD instances in open source projects as confirmed by the original developers.

Index Terms—Self-admitted technical debt, empirical software engineering, issue

I. INTRODUCTION

Technical debt (TD) was first mentioned as a concept by Cunningham close to 30 years ago [1], when he wrote the following lines: “Shipping first time code is like going into debt. A little debt speeds development so long as it is paid back promptly [...] The danger occurs when the debt is not repaid. Every minute spent on not-quite-right code counts as interest on that debt. Entire engineering organizations can be brought to a stand-still under the debt.”

In simple words, TD is a short-term “hack” (often induced by industrial reality, which dictates that either time and/or money are short) with long-lasting consequences if not properly handled. Since developers naturally keep working on new parts and do not revisit something unless it is strictly necessary, very often TD results, in the long run, in low maintainability and poor performance [2].

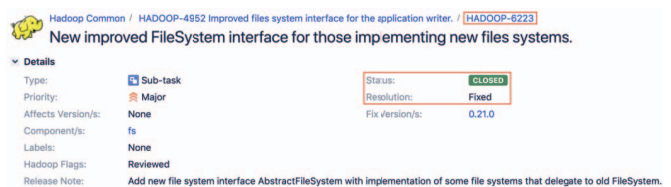
Potdar and Shihab extended the concept of TD to the notion of self-admitted technical debt (SATD) [3], performed intentionally by developers, but mentioned/admitted as comments in the source code. They found that SATD is present, depending on the system, in 2.4% to over 30% of the files and that only 26%-63% gets removed, i.e., a non-SATD often remains in the code. Zampetti *et al.* furthermore found that 20% - 50% of the removals were accidental and are even unintended [4].

Maldonado and Shihab categorized SATD into 5 types: design debt, defect debt, documentation debt, requirement debt, and test debt [5], with design debt and requirement debt being the most common ones. Xavier *et al.* also found that SATD not only manifests itself as comments in the source code, but is also present in issue reports [6].

We focus on a particular type of SATD, first introduced by Maipradit *et al.* [7]: “On-hold SATD”, defined as self-admitted technical debt due to a waiting condition for an external event to happen before the technical debt can be removed. In particular, this paper focuses on On-hold SATD with references to issues.

```
@InterfaceStability.Unstable /* return type will change to AFS once
HADOOP-6223 is completed */
public AbstractFileSystem getDefaultFileSystem() {
    return defaultFS;
}
```

(a) A SATD code comment referencing an issue



(b) A referenced issue report (<https://tinyurl.com/ybunu2dj>)

Fig. 1: Motivating Example

Motivating Example. Fig. 1-(a) shows code from Apache Hadoop. The comment in the code indicates that an action will be taken once a condition is fulfilled, i.e., the closing of issue 6223. As we see from Fig. 1-(b), the issue has in fact already been closed, but the On-hold SATD was not removed, thus creating confusion to anyone inspecting the code.

In essence, On-hold SATD are intentionally reminders left in the source code whose sole purpose is to be removed.

We present a large-scale empirical study to ascertain whether (i) On-hold SATD can be automatically detected, and (ii) it is possible to identify cases in which the On-hold SATD should be removed, since the “waiting condition” has been fulfilled, thus making the SATD a form of “wrong documentation” in the code. Besides quantitatively evaluating the approaches we built to identify and remove On-hold SATD instances, we also show its usefulness in practice by collecting feedback from developers of open source projects.

II. RELATED WORK

A. Empirical Studies on (self-admitted) Technical Debt

Storey *et al.* [8] studied how annotations in code comments (*e.g.*, TODO, FIXME) are used by developers to keep track of tasks. Several types of activities are supported by these annotations, *e.g.*, the usage of TODOs to ask questions to other developers during code comprehension. These annotations are a subset of the ones used nowadays to detect SATD.

Guo *et al.* [9] studied a specific technical debt instance to assessing its impact on the project costs. Their findings confirmed the harmfulness of technical debt, showing that the delayed task resulted in tripled implementation costs.

Klinger *et al.* [10] investigated how decisions to acquire technical debt are made within IBM by interviewing four technical architects. They found that technical debt is often due to imposed requirements to meet a specific deadline sacrificing quality. Also, the interviewed architects reported a lack of effective communication between technical and non-technical stakeholders involved in technical debt management.

Lim *et al.* [11] interviewed practitioners (35 in this case) to investigate their perspective on TD. They found that most of the participants were familiar with the notion of TD and they do consider it as a poor programming practice, but more as an *intentional decision to trade off competing concerns during development* [11]. Practitioners also highlighted the difficulty in measuring the cost of TD. Similarly, Kruchten *et al.* [12] reported their understanding of the technical debt in industry as the result of a four-year interaction with practitioners.

Spinola *et al.* [13] asked 37 practitioners to validate 14 statements about TD (*e.g.*, “*The root cause of most technical debt is pressure from the customer*” [14]). The statement achieving the highest agreement was “*If technical debt is not managed effectively, maintenance costs will increase at a rate that will eventually outrun the value it delivers to customers*”.

Kruchten *et al.* [15] provided theoretical foundations to the concept of TD by presenting the “technical debt landscape”, classifying TD as visible or invisible and highlighting the debt types causing evolvability and maintainability issues. Alves *et al.* [16] proposed an ontology of terms on technical debt.

Potdar and Shihab [3] introduced the notion of SATD by mining five software systems to investigate (i) the amount of SATD they contain, (ii) the factors promoting the introduction of the SATD, and (iii) how likely is the SATD to be removed. Bavota and Russo [17] performed a differentiated replication of that study involving a larger set of subject systems (159), confirming the findings of the original study.

Zazworka *et al.* [18] studied the overlap between the technical debt instances detected by automated tools and by manual inspection, finding very little overlap.

Maldonado and Shihab [19] used the TD classification by Alves *et al.* [16] to investigate the types of SATD more diffused in open source projects. They identified 33k comments in five software systems reporting SATD. These comments have been manually read by one of the authors who found as the vast majority of them (~60%) reported design debt.

Wehaibi *et al.* [20] studied the relationship between SATD and software quality, finding that files with SATD do not have more defects compared to files without SATD, but that changes in the context of SATD are more complex. Sierra *et al.* [21] conducted a survey about SATD research, categorizing it into: detection, comprehension, and repayment. They found a lack of research related to repayment and management of SATD.

B. Automatic detection/management of SATD

In the study by Potdar and Shihab [3], the authors identified SATD using 62 textual patterns. The patterns can be matched in code comments of a previously unseen project to identify SATD. Farias *et al.* [22] built on top of these 62 patterns and developed a model called CVM-TD (Contextualized Vocabulary Model for identifying TD) that exploits combinations of the patterns to identify different types of technical debt.

Maldonado *et al.* [23] presented an approach to automatically identify design and requirement SATD by applying Natural Language Processing (NLP) on code comments. A study performed on ten open source projects showed the superiority of their approach as compared to the state-of-the-art, represented at that time by the above-described pattern-based techniques. Wattanakriengkrai *et al.* [24] developed a classifier to identify design and requirements SATD using N-gram IDF and automated machine learning on Maldonado’s dataset. Comparing the result with the previous study [23], the classifier outperforms the NLP approach in both design and requirement. A similar idea has also been exploited by Huang *et al.* [25] that leveraged text-mining for SATD identification. Also in this case, the approach performed better than the pattern-based approach by Potdar and Shihab [3]. This approach is also available as an Eclipse plug-in [26].

Ren *et al.* [27] proposed an approach based on Convolution Neural Networks to classify code comments into SATD or non-SATD. An experiment performed on ten projects and 63k comments showed that their approach outperforms text mining techniques both for within-project and cross-project prediction.

Zampetti *et al.* [28] presented TEDIOUS (TEchnical Debt IdentificatiOn System), an approach to train a recommender to suggest developers writing new code when to self-admit design TD, or improve the code being written. TEDIOUS achieves an average precision of ~50%. Yan *et al.* [29] proposed a model to determine whether a change introduces SATD. They manually labeled changes that introduced SATD in the past and built a model exploiting 25 features to characterize SATD-introducing changes. An empirical study across ~100k changes reported an AUC for the model of 0.82.

TABLE I: Details of the projects in our dataset. SLOC is calculated on Java files using SLOCCounts [30].

Project	Version	ITS	SLOC	# Contributors	# Remaining comments that refer to issues	# Removed comments that refer to issues
Apache Ant	1.10.7	Bugzilla	144,966	47	27	22
Apache Camel	3.0.0	Jira	1,267,905	544	42	62
Apache Dubbo	2.7.4	Github	148,377	268	8	4
Apache Hadoop	2.10.0	Jira	1,885,604	239	272	269
Apache Jmeter	5.2.1	Bugzilla	142,030	19	116	136
Apache Kafka	2.4.0	Jira	319,990	606	24	21
Apache Log4j	1.2.17	Bugzilla	30,608	7	6	3
Apache Logging-log4j2	2.13.0	Jira	159,353	76	179	153
Apache Tomcat	10.0.0	Bugzilla	341,192	31	82	73
Mockito	3.3.10	Github	48,292	173	15	16
Total	-	-	4,488,317	2,010	771	759

TABLE II: Regular expressions to identify issue in comments

ITS	Regular expression
Bugzilla	<code>(?<![A-Za-z])(?:bug projectname bugzilla bz) [-] (?:#)?\d+(?:\.[0-9xX*]+)*</code> # issue IDs, e.g., Bug 34383 <code>https?:\\\/[\w._]*show_bug.cgi?id=\d+</code> # URLs, e.g., <code>https://bz.apache.org/bugzilla/show_bug.cgi?id=51687</code>
Github	<code>(?<![A-Za-z])(?:bug issues?) [-] (?:#)?\d+(?:\.[0-9xX*]+)*</code> # issue IDs, e.g., issue 55 <code>https?:\\\/github.com/[w._]*\issues\\\/\d+</code> # URLs, e.g., <code>https://github.com/apache/dubbo/issues/3251</code>
Jira	<code>(?<![A-Za-z])(?:bug projectname) [-] (?:#)?\d+(?:\.[0-9xX*]+)*</code> # issue IDs, e.g., HADOOP-7234

III. APPROACH

We aim to build a classifier which automatically detects On-hold SATD and indicates whether it is ready to be removed. To achieve this goal, we took the following four steps (Fig. 2): 1) issue reference detection, 2) dataset creation, 3) data pre-processing, and 4) On-hold SATD classification.

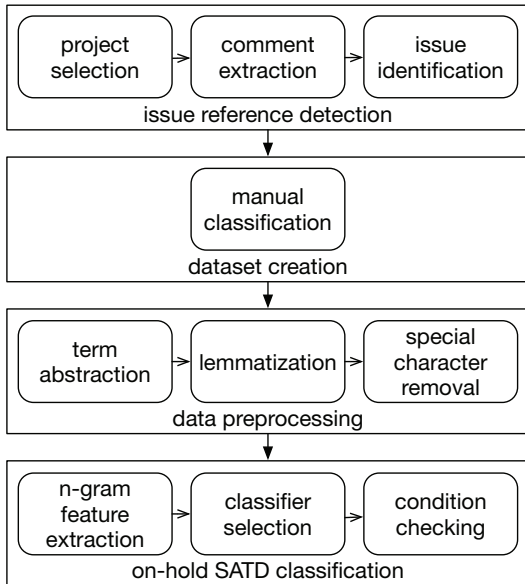


Fig. 2: Approach for On-hold SATD detection and removal

A. Issue Reference Detection

To detect On-hold SATD, our first step is to locate the code comments referring to issues.

1) *Project Selection*: We selected ten open source projects that consistently used for their entire change history a specific issue tracking system (ITS). This allowed us to run our study without the risk of missing important information due to the migration between different issue tracker systems (e.g., starting on JIRA and then moving to the GitHub issue tracker). Table I lists the projects used in this study.

2) *Comment Extraction*: We iterated over the commits in the repositories of the selected projects, and extracted all single line comments (e.g., `// ...`) and multi-line comments (e.g., `/* ... */`) from Java files. If multiple comments are next to each other (e.g., `/* ... */ // ...`) they are considered as a single block of comments. Comments from test files are ignored, as issue references there are most likely to serve as explanations of what developers are testing, instead of SATD.

3) *Issue Identification*: Issue references are identified using regular expressions to match issue IDs and issue URLs. Table II shows the regular expressions used for each issue tracking system. For each identified issue reference, we also recorded its life cycle: We iterated over the commit history and extracted the date when the issue reference was first introduced in the comment, and in case, when it was removed.

From the 10 selected projects, we identified 1,530 comments containing issue references, among which 759 had already been removed, while the remaining 771 still remain in the latest commit by the date of data collection.

TABLE III: Regular expressions for term abstraction

String	ITS	Regular expression
abstractissueid	Bugzilla	(?![A-Za-z]) (? :bug projectname bugzilla bz) [-] (? :#) ?\d+ (? :\.[0-9xX*]+) * # issue IDs https?:\//\//[\w._/]*show_bug.cgi?id=\d+ # URLs
	Github	(?![A-Za-z]) (? :bug issues?) [-] (? :#) ?\d+ (? :\.[0-9xX*]+) * # issue IDs https?:\//\//github.com/[w._/]*\//issues\//\d+ # URLs
	Jira	(?![A-Za-z]) (? :bug projectname) [-] (? :#) ?\d+ (? :\.[0-9xX*]+) * # issue IDs https?:\//\//issues.apache.org\//jira\//browse\//(? :projectname)-\d+ # URLs
abstracturl	—	https?:\//\//(\www\.)?[-a-zA-Z0-9@:%_\+~#=#]{2,256}\.[a-z]{2,6}\b ([-a-zA-Z0-9@:%_\+~#=#&/=]*)

B. Dataset Creation

To build the On-hold SATD classifier, we created a dataset for training and testing, based on the issue-referring comments collected in our previous step. For each of the 1,530 comments, the first and the second author independently labeled whether it is an actual instance of On-hold SATD or, instead, it is used as cross-reference. We evaluated the inter-rater reliability with the Cohen’s kappa coefficient, and the score of 0.748 demonstrates a substantial agreement between the two labelers. The third author resolved labeling conflicts. As a result, we got 133 On-hold SATD and 1,397 cross-reference comments.

Table IV summarizes the annotation results. 133 (8.7%) of the issue-referring comments are instances of On-hold SATD.

TABLE IV: Statistics of annotated comments containing issue references

	On-hold SATD	Cross-reference	Total
Remaining comments	40	731	771
Removed comments	93	666	759
Total	133	1,397	1,530

C. Data Preprocessing

Before extracting features from the comments and feeding them into the classifier, we performed three preprocessing steps: 1) term abstraction, 2) lemmatization, and 3) special character removal.

1) *Term Abstraction*: For all the comments, we abstracted issue IDs and hyperlinks referring to issues to the string “abstractissueid”, while the hyperlinks unrelated to issues were abstracted to “abstracturl”. This is done to eliminate the impact of issue IDs and hyperlinks during classification, as we are not interested in their real content. Table III summarizes the regular expressions we used to extract relevant issue IDs and hyperlinks for different issue tracking systems.

2) *Lemmatization*: We applied lemmatization with the Spacy natural language processing tool [31], which normalizes words with the same root but different surfaces into the same

format [32]. For example, the words “sang”, “singing”, and “sings” will be converted into “sing”.

3) *Special character removal*: We removed all non-English and non-numeric characters using the regular expression $[^A-Za-z0-9]+$.

For our study we did not apply stop word removal, a commonly used text preprocessing step, as it might remove some keywords important for identifying On-hold SATD, such as “when” and “until”.

D. On-hold SATD Classification

After preprocessing, we extracted n-gram features from the comments and used them to train a classifier to identify On-hold SATD. We also checked issue status and issue resolution to determine whether an On-hold SATD comment is ready to be removed.

1) *N-gram Feature Extraction*: Similar to another SATD classification approach by Wattanakriengkrai *et al.* [24], we extracted n-gram features by applying n-gram IDF [33], [34]. N-gram IDF is a theoretical extension of IDF (Inverse Document Frequency). The traditional IDF approach assigns more weight to terms occurring in fewer documents, which does not work well for n-grams. For example, “Leonardo da is” might have higher weight than “Leonardo da Vinci”. N-gram IDF is designed to address this issue and can determine the dominant n-grams and extract key terms of any length [33], [34]. In this study, we extracted n-grams from SATD comments using the library n-gram weighting scheme [35] with default settings. We obtained the list of all valid n-gram terms containing up to 10-gram terms. In total, we receive 644 terms of n-grams.

2) *Classifier Selection*: After extracting the n-gram terms, we build a classifier to identify bug referencing comments into On-hold SATD or not. While there many different algorithms available for supervised classification, it is hard to decide which one to pick, as different datasets and hyper-parameter settings might both impact the performance of these algorithms. Automated machine learning addresses this problem by running multiple classifiers with different parameters to optimize performance. In this study, we used auto-sklearn [36], which includes 15 classification algorithms, 14 feature preprocessing and 4 data preprocessing techniques [36].

3) *Condition Checking*: After identifying the On-hold SATD using our classifier, our program automatically checks the referred issue status and resolution to decide whether the SATD is ready to be removed. In the issue tracking system, if the status of the referred issue is set to “resolved”, “closed”, or “verified”, and the field of resolution (if applicable) is set to “fixed”, we consider it ready for removal.

IV. STUDY DESIGN

The *goal* of this study is to evaluate the accuracy of our approach for On-hold SATD identification and removal. Moreover, we are interested in the evolution of On-hold SATD in open source projects. The *context* of the study consists of 1,530 code comments containing issue references, extracted from the previously presented 10 open source projects.

A. Research Questions

In this study, we answer the following three research questions (RQs):

- **RQ₁**: *What is the accuracy of our approach in identifying On-hold SATD?* This RQ investigates the performance of our classifier in identifying On-hold SATD. We also examined the impact of oversampling, different features and machine learning algorithms on the performance of our classifier:
 - **RQ_{1.1}**: *How do n-grams impact the performance of our classifier as compared to Bag-Of-Words features?*
 - **RQ_{1.2}**: *How does oversampling impact the performance of the classifier?*
 - **RQ_{1.3}**: *How do different machine learning algorithms impact the performance of the classifier?*
- **RQ₂**: *How does On-hold SATD evolve in open source projects?* To gain deeper insights on how On-hold SATD evolves in the projects, with this RQ we inspect the duration of existence of On-hold SATD in software projects, and the time it takes to address SATD after the relevant issue is resolved.
- **RQ₃**: *To what extent can our approach identify “ready-to-be-removed” On-hold SATD?* This RQ empirically evaluates the reliability of our approach in identifying On-hold SATD which should be removed, since it was already “paid back”.

B. Context Selection & Data Collection

In this study, we used the dataset presented in Section III-B, which contains 1,530 annotated comments containing issue references.

To answer RQ₁, we built a classifier using auto-sklearn with n-grams extracted by n-gram IDF [24] as features. N-grams were extracted from On-hold SATD comments only. N-grams from Cross-reference comments are not included because we want to extract important patterns to detect on-hold SATD, and we use these patterns to discriminate between On-hold SATD and Cross-reference. We performed a ten-fold cross validation: We divided the 1,530 issue-referring comments into

ten different sets, each one composed of 153 comments. Then, we iteratively used one set as the *test set*, while the remaining 1,377 comments were used for *training*.

To answer RQ_{1.1}, we ran a different classifier implementation on the dataset, using Bag-Of-Words (BOW) as features.

To answer RQ_{1.2}, we applied an oversampling technique (*i.e.*, SMOTE) to our training set, and then compared the results achieved by our classifier with/without oversampling.

To answer RQ_{1.3}, we built three variants of the classifier with different machine learning algorithms: Naive Bayes, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN).

To answer RQ₂, we inspected the removed issue-referring comments for both On-hold and cross-reference comments. We first checked the time interval between the introduction and the removal of these comments. Then, for the instances referring issues that have been solved, we compute the difference between the issue resolution time and the corresponding On-hold SATD removal event.

To answer RQ₃, we identified the On-hold SATD comments which are ready to be removed from the 40 still remaining On-hold issue-referring comments, based on the corresponding issue status and resolution, as described in Section III-D3. In total, we identified 10 “ready-to-be-removed” On-hold SATD comments. By the time we started working on RQ₃, 4 of 10 comments had already been removed by developers (three were removed thanks to code changes addressing the On-hold SATD, while one was removed due to the deletion of the file containing it). We reported the remaining six “ready-to-be-removed” On-hold SATD comments to the developers by creating issue reports in the respective issue tracker. In the issue report, we inform developers why the On-hold SATD comments should be removed and where they are located. An example of the issue reports can be seen in Fig. 3.

Description

In a research project, we analyzed the source code of Hadoop looking for comments with on-hold SATDs (self-admitted technical debt) that could be fixed already. An on-hold SATD is a TODO/FIXME comment blocked by an issue. If this blocking issue is already resolved, the related todo can be implemented (or sometimes it is already implemented, but the comment is left in the code causing confusions). As we found a few instances of these in Hadoop, we decided to collect them in a ticket, so they are documented and can be addressed sooner or later.

A list of code comments that mention already closed issues.

- A code comment suggests making the `setJobConf` method deprecated along with a mapped package `HADOOP-1230`. `HADOOP-1230` has been closed a long time ago, but the method is still not annotated as deprecated.

```
/**
 * This code is to support backward compatibility and break the compile
 * time dependency of core on mapred.
 * This should be made deprecated along with the mapped package HADOOP-1230.
 * Should be removed when mapped package is removed.
 */
```

Comment location: <https://github.com/apache/hadoop/blob/trunk/hadoop-common-project/hadoop-common/src/main/java/org/apache/hadoop/util/ReflectionUtils.java#L88>

Fig. 3: An example issue report.

C. Data Analysis

To answer RQ₁ we compare the precision, recall, F1-score, and area under the ROC curve (AUC) of each experimented approach in classifying issue-referring comments (as belonging or not to On-hold SATD) for the dataset of 1,530 comments.

TABLE V: Performance of classifiers in identifying On-hold SATD

	Original approach	BOW as feature	With Oversampling	Different ML algorithms		
	n-gram + auto-sklearn	BOW + auto-sklearn	n-gram + oversampling + auto-sklearn	n-gram + Naive Bayes	n-gram + SVM	n-gram + KNN
Precision	0.79	0.69	0.38	0.64	0.87	0.88
Recall	0.70	0.68	0.48	0.56	0.38	0.15
F1-score	0.73	0.67	0.41	0.59	0.51	0.25
AUC	0.97	0.94	0.87	0.81	0.95	0.76

TABLE VI: Statistical results of performance comparisons of classifiers

	P-value (Precision)	Effect size (Precision)	P-value (Recall)	Effect size (Recall)
<i>n-gram+auto-sklearn vs BOW+auto-sklearn</i>	< 0.01	0.48 (large)	0.32	-
<i>n-gram+auto-sklearn vs n-gram+oversampling+auto-sklearn</i>	< 0.01	0.92 (large)	0.01	0.67 (large)
<i>n-gram+auto-sklearn vs n-gram+Naive Bayes</i>	0.06	-	0.03	0.58(large)
<i>n-gram+auto-sklearn vs n-gram+SVM</i>	0.30	-	0.03	0.8 (large)
<i>n-gram+auto-sklearn vs n-gram+KNN</i>	0.30	-	0.03	1.0 (large)
<i>n-gram+Naive Bayes vs n-gram+SVM</i>	0.06	-	0.03	0.58 (large)
<i>n-gram+Naive Bayes vs n-gram+KNN</i>	0.30	-	0.03	1.0 (large)
<i>n-gram+SVM vs n-gram+KNN</i>	0.31	-	0.03	0.74 (large)

The comparisons are also performed via the Mann-Whitney test [37], with results intended as statistically significant at $\alpha = 0.05$. For RQ_{1.3}, to control the impact of multiple pairwise comparisons (e.g., the precision of auto-sklearn is compared with Naive Bayes, SVM, and KNN), we adjust p -values with Holm’s correction [38]. We estimate the magnitude of the differences by using the Cliff’s Delta (d), a non-parametric effect size measure [39]. We follow well-established guidelines to interpret the effect size: negligible for $|d| < 0.10$, small for $0.10 \leq |d| < 0.33$, medium for $0.33 \leq |d| < 0.474$, and large for $|d| \geq 0.474$ [39].

To answer RQ₂, we present via violin plots the life spans of both On-hold SATD and cross-reference comments, as well as the duration between the resolution of issues and the removal of corresponding SATD comments.

To answer RQ₃, we qualitatively analyze the developers’ feedback.

V. RESULTS

A. RQ₁: What is the accuracy of our approach in identifying On-hold SATD?

Table V reports the average precision, recall, F1-score, and AUC of each experimented classifier implementations during 10-fold evaluation.

Table VI reports the statistical results of comparisons between different classifier implementations.

Fig. 4 also shows the results of the 10-fold evaluation for each experimented classifier in terms of precision, recall, F1-Score, and AUC.

As can be seen from Table V, the precision, recall, and F1-score achieved by our approach (“n-gram + auto-sklearn”) are all between 0.7 to 0.8, while AUC is as high as 0.97. This result demonstrates the reliability of our approach in On-hold SATD detection.

To gain a better understanding of how our classifier works, we list the important n-gram features which frequently appear in On-hold SATD comments in Table VII.

TABLE VII: N-gram features which frequently appear in On-hold SATD comments

N-gram features	Frequency
‘after’, ‘abstractissueid’	20
‘once’, ‘abstractissueid’	18
‘for’, ‘now’	12
‘temporary’, ‘fix’	10
‘workaround’	10
‘this’, ‘be’, ‘a’, ‘temporary’	8
‘via’, ‘abstractissueid’	7
‘be’, ‘commit’	7
‘can’, ‘be’, ‘remove’	7
‘remove’, ‘after’, ‘abstractissueid’	5

These features help discriminate On-hold SATD from cross-reference. We can see that n-grams such as “*once abstractissueid*”, “*this be a temporary*”, and “*remove after abstractissueid*” are especially important for identifying On-hold SATD.

Additionally, we also illustrate some classification results in Table VIII. From the two true positive examples (correctly identified as On-hold SATD by our approach), we can clearly see the patterns including “*workaround*”, “*temporary fix*” and “*remove after abstractissueid*”, which can be related to Table VII. Therefore, it is not surprising that our classifier can correctly identify these On-hold SATD comments.

If we take a look at the two false negative examples (On-hold SATD classified as cross-reference), we find that phrases like “*to be revisit*” and “*until abstractissueid*” are probably useful n-grams for identifying On-hold SATD. Due to absence or infrequent occurrence, these n-grams are not used as features for the classifier. Expanding the training set can be a potential way for addressing the n-gram feature limitations.

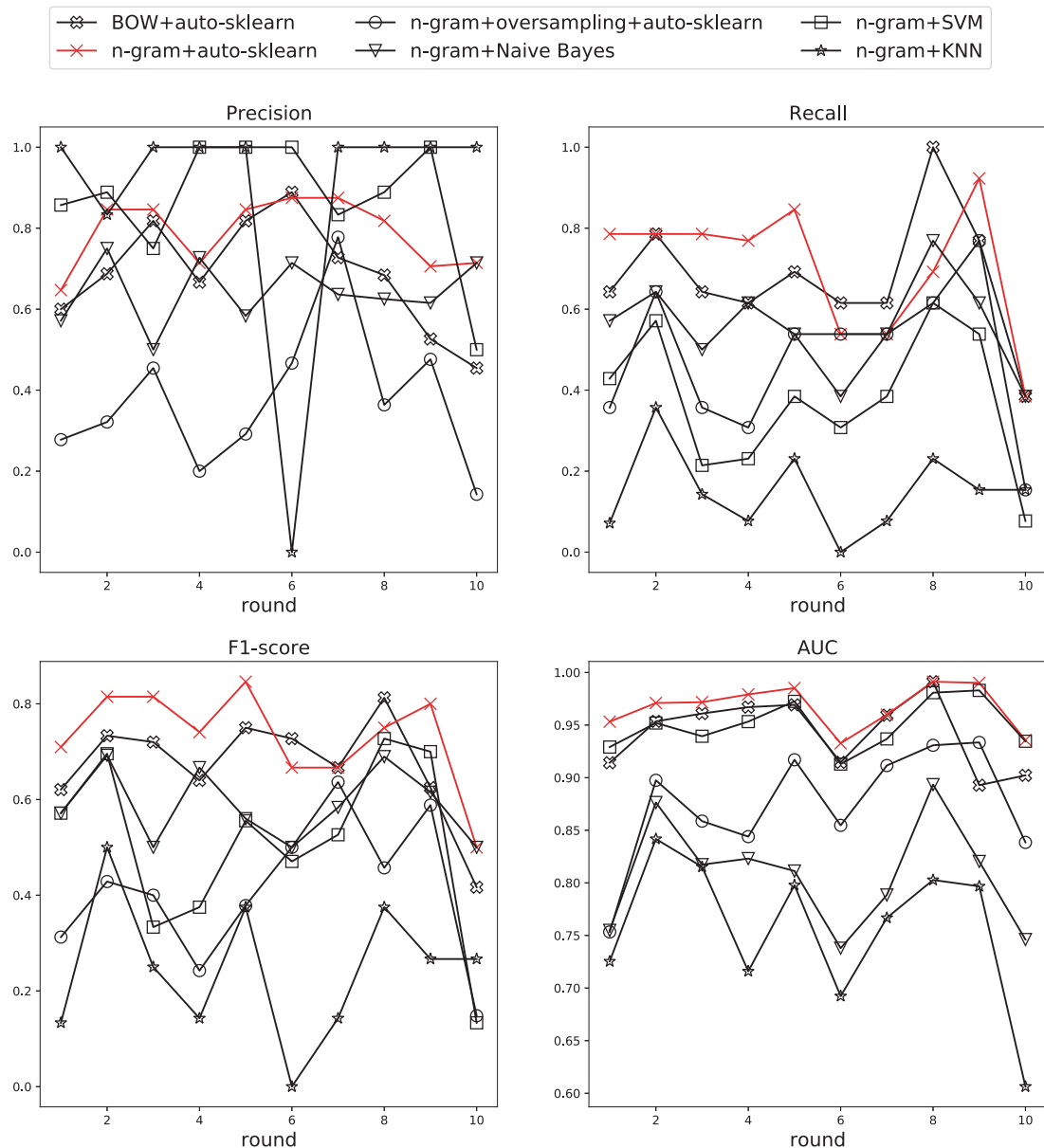


Fig. 4: Results of each round in 10-fold evaluation for different classifier implementations

In the two false positive examples (cross-reference classified as On-hold SATD), we can see that the n-gram terms like “*todo after abstractissueid*” and “*todo remove*” can be actually matched, and our classifier misclassified them into On-hold SATD. However, if we check the comments carefully, we can find that in the first sentence, it is clear that the issue has already been resolved, however, for some reason the developers decided to say “*keeping it for now*”, where “*it*” refers to `YarnException`. In the second sentence, what follows “*todo remove*” is actually not a reference to an issue, but a reference to a version. Some heuristic rules might help our classifier to better deal with these cases.

To understand how n-grams impact the performance of our

classifier as compared to Bag-Of-Words (BOW) features, we inspect the first two columns of Table V, and the first row of Table VI. Using n-grams as features leads to a higher precision with a statistically significant difference and a large effect size. As for the recall, while the average value is higher when using n-grams, the performed analysis does not indicate a statistically significant difference. We conclude that compared to BOW features, n-grams lead to a significantly higher precision.

To understand how oversampling impacts the performance of the classifier, we inspect the first and the third column of Table V, as well as the second row of Table VI.

TABLE VIII: Example of classification results of our approach

Type	Comment
True	TODO: workaround (filling fixed bytes), to remove after HADOOP-11938
Positive	... This is a temporary fix ... See the discussion on HDFS-1965.
False	TODO: Temporarily keeping ... This has <i>to be revisited</i> as part of HDFS-11029.
Negative	<i>placeholder for javadoc</i> to prevent broken links, <i>until HADOOP-6920</i>
False	TODO: after MAPREDUCE-2793 YarnException is probably not expected here anymore but keeping it for now ...
Positive	... (CAMEL-9657) [TODO] Remove in 3.

From the tables we can observe that the classifier obtains a statistically significant higher precision and recall when oversampling is not applied. Meanwhile, the effect sizes for both precision and recall comparisons are large. Indeed, after applying oversampling, the average precision, recall, F1-score, and AUC drop by around 40%, 20%, 30%, and 10%, respectively. We conclude that oversampling reduces the performance of our classifier in identifying On-hold SATD.

To understand how different machine learning algorithms impact the performance of the classifier, we inspect the first and the last three columns of Table V, as well as the last six rows of Table VI. From the tables we can see that all the implementations achieved comparable precisions (from 0.64 to 0.88). Indeed, there is no statistically significant difference in terms of precision among these implementations. However, the differences emerge when comparing recall. Using auto-sklearn achieves a significantly higher recall than classifiers using other machine learning algorithms (*i.e.*, Naive Bayes, SVM and KNN).

We also inspected which machine learning algorithm was adopted by auto-sklearn after automatic classifier selection. The records show that in 9 of the ten rounds of 10-fold evaluation Extra Trees was adopted, while the remaining one adopted Random Forest. That is, these two machine learning algorithms would potentially be a good choice for identifying On-hold SATD when automatic selection of the classifier is not possible.

B. RQ₂: How does On-hold SATD evolve in open source projects?

To answer RQ₂, we first looked into the life span of removed issue-referring comments for On-hold SATD and cross-reference comments separately. The life span distributions can be found in Fig. 5.

The median life span of On-hold SATD comments is 42 days, while it is 119.5 days for cross-reference comments. That is, overall, the median life span of cross-reference comments is almost three times of that of On-hold SATD.

Indeed, while both types of comments contain issue references, only On-hold SATD requires maintenance actions from developers. Cross-reference comments stay much longer as they are usually used for documentation purposes.

We then investigated how long it takes to address On-hold SATD comments after the corresponding issues are resolved, and plotted the duration distribution in Fig. 6.

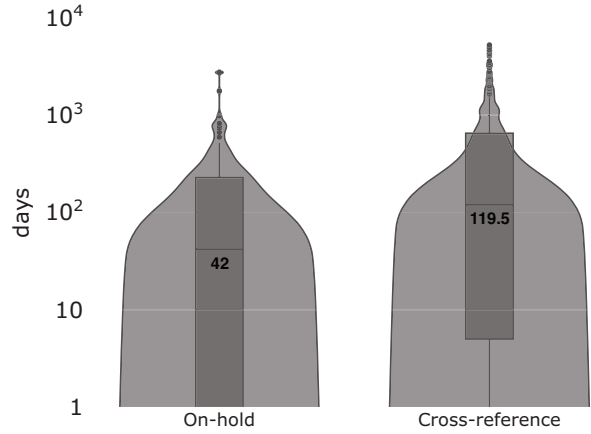


Fig. 5: Distribution of life spans of removed issue-referring comments

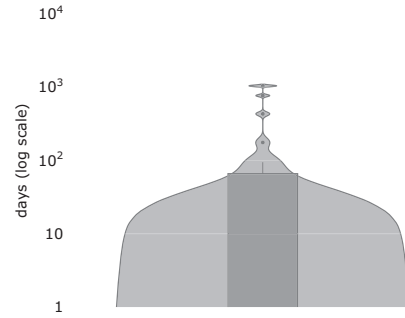


Fig. 6: Distribution of days needed to address SATD comments after issues were resolved

Around 53% of On-hold SATD were removed within the same day when the issue was resolved. However, it takes longer than one year to remove 13% of On-hold SATD.

Additionally, we observed that some developers did not wait until the issue was resolved to address On-hold SATD comments. In fact, from a total of 93 removed On-hold SATD comments, we found that only 30 of them were removed after the issues were resolved. The corresponding issues of 9 On-hold SATD comments are still open or have the resolution set to “wontfix”. 54 On-hold SATD comments were removed before the issues were resolved, although these issues have been resolved in the meantime.

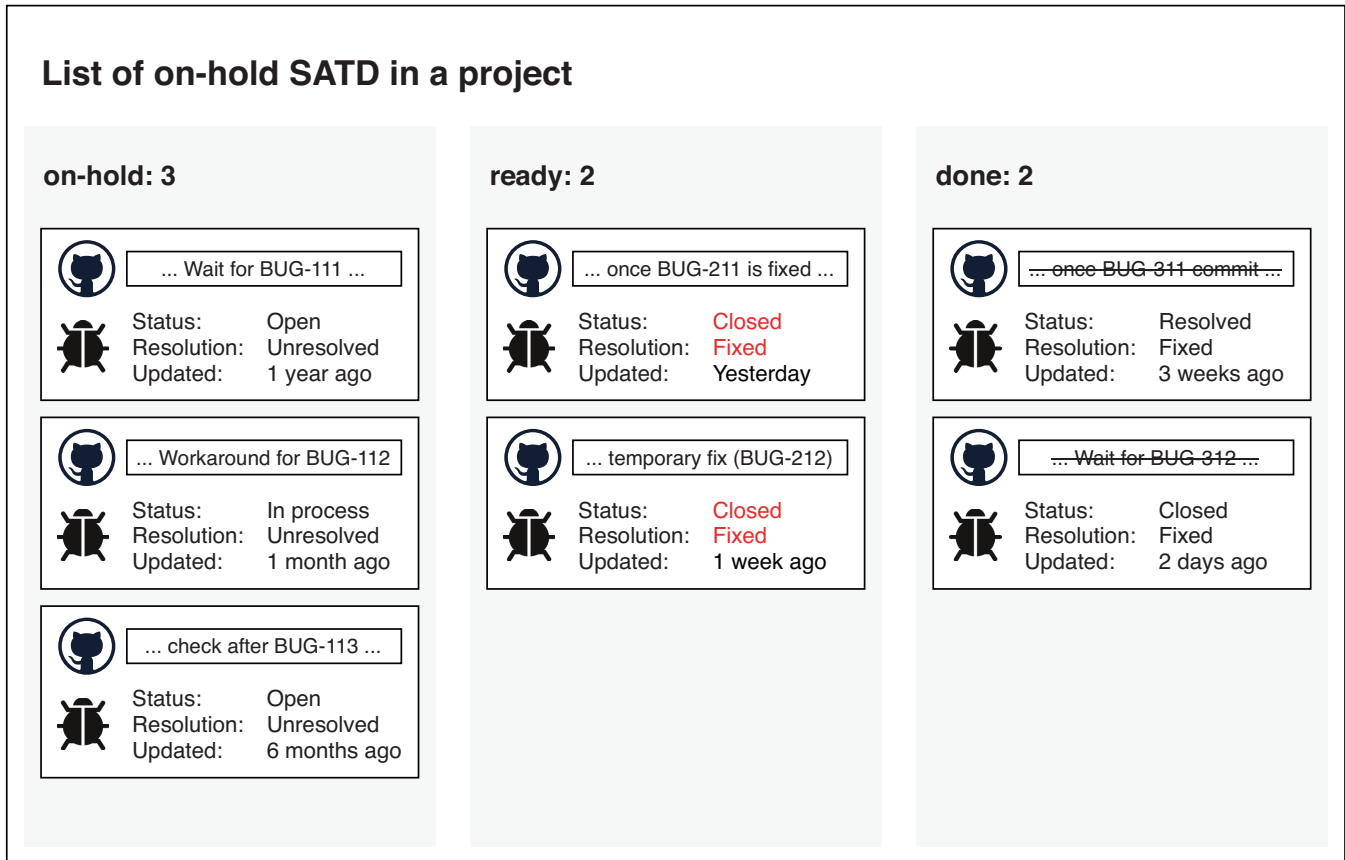


Fig. 7: A mockup of On-hold SATD identification tool

C. RQ₃: To what extent can our approach identify “ready-to-be-removed” On-hold SATD?

To understand how well our approach performs in identifying “ready-to-be-removed” On-hold SATD comments, we reported six identified cases to developers in three issue reports, as these six cases correspond to three subsystems of the Apache Hadoop project (two for Hadoop Common, one for Hadoop HDFS, and three for Hadoop YARN). By the time of writing, we have received the feedback from the developers about the two “ready-to-be-removed” On-hold SATD comments in the Hadoop Common subsystem.

Table IX lists these two instances of On-hold SATD reported to the developers in JIRA issue tracking system (*link hidden for double blind review*).

TABLE IX: Two “ready-to-be-removed” On-hold SATD comments which received developers’ feedback

#	Ready On-hold SATD
1	“/* return type will change to AFS once HADOOP-6223 is completed */”
2	“... This should be made deprecated along with the mpared package HADOOP-1230. ...”

For the first case, the return type had already changed to AFS, and the resolution of the referred issue “HADOOP-6223”

had been set to “resolved”. In the issue report, we suggested that this On-hold SATD comment should be removed. Developers agree that it can be removed:

“I think this is correct finding. Would you like to put a patch for this”

Later on, the patch we submitted got integrated into the repository.

Regarding the second case, we found that the referred issue “HADOOP-1230” had also been resolved. Thus, we suggested that developers could apply corresponding changes (*i.e.*, making the `setJobConf` method deprecated). The developers agreed that the action should be taken but it is a rather complicated fix, thus recommending a new JIRA issue thread:

“...we need to update the document in a separate jira.”

“... Given that is a bigger subject than this fix, we should discuss on that separately ...”

Overall, the two cases for which we have already received feedback on indicate the practical value of our approach for On-hold SATD identification and removal.

D. Replication

To facilitate replication, we released our dataset in our online appendix, which can be accessed at <https://tinyurl.com/onholdissue>. The spreadsheet file of our dataset contains three sheets: removed comments, remaining comments, and identified “ready-to-be-removed” On-hold SATD. For all the comments in our dataset, we include the comment context, code file path, line number, referred issue, and our annotation (On-hold SATD or cross-reference). For removed comments, we also include when the code comment was introduced and removed. For the “ready-to-be-removed” On-hold SATD, there are also the status and the resolution of the corresponding issues.

VI. TOWARDS A ON-HOLD SATD RECOMMENDER

Our findings can serve as guideline for developers writing reference issues in code comments:

- Developers should check SATD comments referring to issues which had already been resolved, as we reported that 13% of comments were removed with a delay of more than one year.
- When the code comments refer to issues, developers should clearly mention the intention in the comments, *i.e.*, whether the issue is used for documentation or to denote the condition on which one is waiting on.

While we plan on expanding our work to analyze more projects and to include also other issue tracking systems, we believe that our work can be synthesized into a recommender system for On-Hold SATD. In Fig. 7 we depicted a mock-up of such a recommender.

The tool would report the list of On-hold SATD comments, ready to be addressed On-hold SATD comments, and removed On-hold SATD. Each item would include comments, links to the original comments and to the pertaining issue (including its status and duration).

VII. THREATS TO VALIDITY

Threats to *construct validity* concern the relation between the theory and the observation, and in this work they are mainly due to the measurements we performed:

- *Imprecisions in the identification of issue references in comments.* We used the regular expressions in Table II to mine issue references in code comments. The regular expressions have been defined and tested by the first author, and are customized for each of the issue trackers used by the subject systems.
- *Subjectivity/errors in the manual classification.* To mitigate this threat, the first two authors independently classified the 1,530 issue-referencing comments as On-hold SATD or as cross-reference. Then, the third author resolved the conflicts.

Threats to *external validity* concern the generalizability of results. Rather than going large-scale, we preferred to work on a set of ten well-known Java open source projects and to manually validate all issue-referencing comments we found in

them in such a way to increase the reliability of the presented data. Other systems should be included in the analysis to allow for a broader generalizability of our conclusions. Also, the results of RQ₃ are based on only two feedback we received from developers, thus do not allowing any sort of generalizability but only serving as pointers for qualitative analysis.

VIII. CONCLUSION

Since the definition of the term “technical debt” by Cunningham three decades ago [1], researchers have investigated the phenomenon, leading to the understanding that it is its creeping, barely visible nature that leads to maintenance and evolvability problems down the road. Developers cannot be faulted for the introduction of technical debt, as software industry functions under great time and budget pressure, and compromises have to be made to meet said time and budget constraints. Indeed, developers often admit that they are creating technical debt, which led to the term “self-admitted technical debt” (SATD) coined by Potdar and Shihab [3].

A particular type of SATD is the one we named “On-hold” SATD, where a developer has to make a compromise or halt development because of an external condition. Human nature dictates that often On-hold SATD is simply forgotten about.

We performed an empirical study to understand whether On-hold SATD can be automatically detected: We analyzed ten open source projects, and found that 8% of the comments referring to issues are On-hold SATD. To identify On-hold SATD, we developed a classifier using n-gram and auto-sklearn, resulting in an average precision of 0.79, an average recall of 0.70, an average F_1 -score of 0.73, and an average AUC of 0.97. In short, On-hold SATD can indeed be detected automatically in a fairly reliable way.

To understand how On-hold SATD evolves, we looked into life-span of removed issue-referring comments. We found that the median life-span of On-hold comments is 42 days. This is certainly beyond the horizon of human short-term memory, and indeed we found that after the issues were resolved, 13% of On-hold SATD takes longer than one year to remove. To evaluate the reliability in identifying On-hold SATD which should be removed, we collected feedback from developers from open source projects. Developers agreed with our findings that the reported On-hold SATD should be fixed or removed.

The next logical step is thus the design and implementation of the recommender system we described in Section VI and aimed at facilitating the identification, understanding, and resolution of On-hold SATD instances.

ACKNOWLEDGEMENT

We gratefully acknowledge the financial support of Japan Society for the Promotion of Science for the JSPS KAKENHI Grant No. 16H05857 and 20H05706, and the Swiss National Science Foundation for the project SENSOR (SNF-JSPS Project No. 183587).

REFERENCES

- [1] W. Cunningham, "The wycash portfolio management system," in *Addendum to the Proceedings on Object-Oriented Programming Systems, Languages, and Applications (Addendum)*, ser. OOPSLA '92. New York, NY, USA: Association for Computing Machinery, 1992, p. 29–30. [Online]. Available: <https://doi.org/10.1145/157709.157715>
- [2] E. Lim, N. Taksande, and C. Seaman, "A balancing act: What software practitioners have to say about technical debt," *IEEE Software*, vol. 29, no. 6, pp. 22–27, 2012.
- [3] A. Potdar and E. Shihab, "An exploratory study on self-admitted technical debt," in *Proceedings of the 2014 IEEE International Conference on Software Maintenance and Evolution*, ser. ICSME '14. USA: IEEE Computer Society, 2014, p. 91–100. [Online]. Available: <https://doi.org/10.1109/ICSME.2014.31>
- [4] F. Zampetti, A. Serebrenik, and 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.
- [5] E. d. S. Maldonado and 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.
- [6] L. Xavier, F. Ferreira, R. Brito, and M. T. Valente, "Beyond the code: Mining self-admitted technical debt in issue tracker systems," *arXiv preprint arXiv:2003.09418*, 2020.
- [7] R. Maipradit, C. Treude, H. Hata, and K. Matsumoto, "Wait for it: identifying "on-hold" self-admitted technical debt," *Empirical Software Engineering*, 2020. [Online]. Available: <https://doi.org/10.1007/s10664-020-09854-3>
- [8] M.-A. Storey, J. Ryall, R. I. Bull, D. Myers, and J. Singer, "Todo or to bug: Exploring how task annotations play a role in the work practices of software developers," in *Proceedings of the 30th International Conference on Software Engineering*, ser. ICSE '08, 2008, pp. 251–260.
- [9] Y. Guo, C. Seaman, R. Gomes, A. Cavalcanti, G. Tonin, F. da Silva, A. Santos, and C. Siebra, "Tracking technical debt - an exploratory case study," in *Software Maintenance (ICSM), 2011 27th IEEE International Conference on*, 2011, pp. 528–531.
- [10] T. Klinger, P. Tarr, P. Wagstrom, and C. Williams, "An enterprise perspective on technical debt," in *Proceedings of the 2Nd Workshop on Managing Technical Debt*, ser. MTD '11, 2011, pp. 35–38.
- [11] E. Lim, N. Taksande, and C. Seaman, "A balancing act: What software practitioners have to say about technical debt," *Software, IEEE*, vol. 29, no. 6, pp. 22–27, 2012.
- [12] P. Kruchten, R. L. Nord, I. Ozkaya, and D. Falessi, "Technical debt: Towards a crisper definition report on the 4th international workshop on managing technical debt," *SIGSOFT Softw. Eng. Notes*, vol. 38, no. 5, pp. 51–54, 2013.
- [13] R. Spinola, N. Zazworka, A. Vetró, C. Seaman, and F. Shull, "Investigating technical debt folklore: Shedding some light on technical debt opinion," in *Managing Technical Debt (MTD), 2013 4th International Workshop on*, 2013.
- [14] K. S. Rubin, *Essential Scrum: A Practical Guide to the Most Popular Agile Process*, 1st ed. Addison-Wesley Professional, 2012.
- [15] P. Kruchten, R. L. Nord, and I. Ozkaya, "Technical debt: From metaphor to theory and practice," *IEEE Software*, vol. 29, no. 6, pp. 18–21, 2012.
- [16] N. Alves, L. Ribeiro, V. Caires, T. Mendes, and R. Spinola, "Towards an ontology of terms on technical debt," in *Managing Technical Debt (MTD), 2014 Sixth International Workshop on*, 2014, pp. 1–7.
- [17] G. Bavota and B. Russo, "A large-scale empirical study on self-admitted technical debt," in *Proceedings of the 13th International Conference on Mining Software Repositories, MSR 2016, Austin, TX, USA, May 14–22, 2016*, 2016, pp. 315–326.
- [18] N. Zazworka, R. O. Spínola, A. Vetro', F. Shull, and C. Seaman, "A case study on effectively identifying technical debt," in *Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering*, ser. EASE '13, 2013, pp. 42–47.
- [19] E. da S. Maldonado and E. Shihab, "Detecting and quantifying different types of self-admitted technical debt," in *7th IEEE International*
- [20] S. Wehaibi, E. Shihab, and 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, 2016, pp. 179–188.
- [21] G. Sierra, E. Shihab, and Y. Kamei, "A survey of self-admitted technical debt," *Journal of Systems and Software*, vol. 152, pp. 70–82, 2019. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0164121219300457>
- [22] M. A. d. F. Farias, M. G. d. M. Neto, A. B. d. Silva, and R. O. Spínola, "A contextualized vocabulary model for identifying technical debt on code comments," in *2015 IEEE 7th International Workshop on Managing Technical Debt (MTD)*, 2015, pp. 25–32.
- [23] E. d. S. Maldonado, E. Shihab, and N. Tsantalis, "Using natural language processing to automatically detect self-admitted technical debt," *IEEE Transactions on Software Engineering*, vol. 43, no. 11, pp. 1044–1062, 2017.
- [24] S. Wattanakriengkrai, R. Maipradit, H. Hata, M. Choetkiertikul, T. Sunetnanta, and K. Matsumoto, "Identifying design and requirement self-admitted technical debt using n-gram idf," in *2018 9th International Workshop on Empirical Software Engineering in Practice (IWESEP)*, 2018, pp. 7–12.
- [25] Q. Huang, E. Shihab, X. Xia, D. Lo, and S. Li, "Identifying self-admitted technical debt in open source projects using text mining," *Empirical Software Engineering*, vol. 23, no. 1, pp. 418–451, 2018. [Online]. Available: <https://doi.org/10.1007/s10664-017-9522-4>
- [26] Z. Liu, Q. Huang, X. Xia, E. Shihab, D. Lo, and S. Li, "Satd detector: A text-mining-based self-admitted technical debt detection tool," in *2018 IEEE/ACM 40th International Conference on Software Engineering: Companion (ICSE-Companion)*, 2018, pp. 9–12.
- [27] X. Ren, Z. Xing, X. Xia, D. Lo, X. Wang, and J. Grundy, "Neural network-based detection of self-admitted technical debt: From performance to explainability," *ACM Trans. Softw. Eng. Methodol.*, vol. 28, no. 3, Jul. 2019. [Online]. Available: <https://doi.org/10.1145/3324916>
- [28] F. Zampetti, C. Noiseux, G. Antoniol, F. Khomh, and M. Di Penta, "Recommending when design technical debt should be self-admitted," in *2017 IEEE International Conference on Software Maintenance and Evolution (ICSM)*, 2017, pp. 216–226.
- [29] M. Yan, X. Xia, E. Shihab, D. Lo, J. Yin, and X. Yang, "Automating change-level self-admitted technical debt determination," *IEEE Transactions on Software Engineering*, vol. 45, no. 12, pp. 1211–1229, 2019.
- [30] D. A. Wheeler, "Sloccount user's guide," 2004.
- [31] M. Honnibal and I. Montani, "spacy - industrial-strength natural language processing in python," <https://spacy.io/>, 2017, (Accessed on 13/04/2019).
- [32] D. Jurafsky and J. H. Martin, "Speech and language processing," 2009.
- [33] M. Shirakawa, T. Hara, and S. Nishio, "Idf for word n-grams," *ACM Trans. Inf. Syst.*, vol. 36, no. 1, Jun. 2017. [Online]. Available: <https://doi.org/10.1145/3052775>
- [34] —, "N-gram idf: A global term weighting scheme based on information distance," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2015, p. 960–970. [Online]. Available: <https://doi.org/10.1145/2736277.2741628>
- [35] M. Shirakawa, "N-gram weighting scheme," Jul 2017. [Online]. Available: <https://github.com/iwnsew/ngweight>
- [36] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2962–2970. [Online]. Available: <http://papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning.pdf>
- [37] W. J. Conover, *Practical nonparametric statistics*, 3rd ed. Wiley New York, 1998.
- [38] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, pp. 65–70, 1979.
- [39] R. J. Grissom and J. J. Kim, "Effect sizes for research: A broad practical approach," *Mahwah, NJ: Earlbaum*, 2005.