Using Deep Learning to Identify Technical Debt

Levereging Self Admitted Technical Debt

Master's Thesis submitted to the
Faculty of Informatics of the *Università della Svizzera Italiana*
in partial fulfillment of the requirements for the degree of
Master of Science in Informatics
Major in Artificial Intelligence

presented by
Simone Giacomelli

under the supervision of
Prof. Dr. Gabriele Bavota
co-supervised by
Dr. Csaba Nagy

January 2021

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Simone Giacomelli
Lugano, 1 January 2021

Dedication

Abstract

Composed of four parts:
a: context
b: problem
c: proposed solution
d: results

max 400 words. I will write it once the intro has been approved

Acknowledgements

Ack

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Technical Debt (TD) is a metaphor between the financial concept and software development. A TD is contracted when a workaround or shortcut is taken during code implementation. Choosing an easy and quick solution over a slower and more correct one gives the benefit to save time and deliver the artifact faster.
The downsides of incurring in TD are many, one of the most intuitive is that further work on the affected parts will be more expensive and time consuming than on clean and healthy code. There are also indirect effects to contemplate: the software could misbehave in the domain it is operating, causing costs and damages as the effect of unexpected output or wrong behavior [49].
TD do not only appear in software projects but can also be found in many layers of a technology stack: for example, delaying an hardware upgrade or a maintenance can give an immediate benefit of less downtime or financial savings, but an increased cost for future unexpected downtime or failures [1].
Developers or managers can choose to incur in TD because of strict deadline, limited resources available or just plain laziness [20, 1]. Cunningham, who coined the technical debt metaphor, writes in "The WyCash portfolio management system" [8]: "A little debt speeds development so long as it is paid back promptly with a rewrite". Cunningham implies that one could benefit from a small amount of TD but its should be paid back as soon as possible.
TD can arise from intentional and unintentional decision, e.g. an inexperienced person could contract it without being aware of it [20]; In both cases it's often done for saving the limited available resources and shortening time-to-market [48]. For example, startups are highly pressured to quickly test products and ideas in order to save capitals and be faster than competitors. Besker at al. studied how startups incur in TD, which are the factors, challenges and benefits of intentional acquisition of TD; two of the regulating factors found by the authors are the experience of the developers and the software knowledge of the founders [4].
In contrast to the beneficial viewpoint of TD, Ron Jeffries argues that the metaphor

could be "perhaps too gentle", because it highlights the wise aspect of the choice of contracting a debt; the problem is that people also takes debt unwisely. Technical debt benefits a software project as long as it is handled before the bigger long term cost is realized [15].
******
It is useful to list some aspects of TD that form the basis for a tractable model of the phenomenon; many of these aspects comes from the already well known attributes of financial debt. The *principal* is the amount that needs to be payed to 'make things right'. The *interest amount* is the added cost in addition to the principal;

In order to better analyze and create a tractable model of TD, Alves et al. identifies three variables [34]:

Identified variables (3) Tracking and managing is important Few proposed tracking and managing examples

there is a cost paying the debt and there are interest costs when efforts are wasted coping with non optimal code.

Why detecting it is important

the satd is there
it's undocumented
the team is not aware of it

Tom et al. in their systematic literature review studied, among others, the benefits and drawbacks of incurring in TD [48]; the part we are now interested in is the drawbacks:
Increasing costs over time, such as the amount of effort required to deliver a certain amount of functionality
Work estimation becomes difficult
Developer productivity is negatively impacted
Becomes increasingly difficult to repay as decisions are affected by existing debt
Increased risk involved in modifications to the system
Change becomes prohibitively expensive to the point of bankruptcy, and a complete rewrite and new platform may become necessary
Decreased quality in the end product

Ci sono team che risolvo as-you-go, altri hanno strategie per rilevare, identificare e fixare. It has been studied that TD [Investigating the Impact of Design Debt on Software Quality] E' importante rilevare e monitorare i TD perche' questi aumentano il costo di un progetto software, costo dovuto dall'interesse che si deve pagare in relazione al TD.

1.1 Objectives and Results

Once the context is clear, I will explain the goal of the thesis and summarize the achieved results The goal of the thesis is to exploit SATD to train a model that can acquire the capacity to distinguish between TD-free code and code affected by TD. Using the comments in open source projects I will identify class methods noted as SATD. Through the vcs commit history I will identify when this comments disappear; the assumption here is that when the comment is removed from the code the SATD has been fixed. Many cases are excluded to minimize the probability of keeping a false positive, i.e. the SATD is not fixed but the comment is removed. For example, the simplest case excluded is when the code is exactly the same and the only change is the SATD comment removal. Another reason of exclusion is when the code is changed too much; in such case it would not be prudent to keep the sample in the dataset.

The achieved results tell us that it is difficult to predict with high accuracy on methods with big bodies. As the train dataset is limited to shorter and shorter method size, the accuracy grows.

1.2 Structure of the Thesis

A simple bullet list saying "Chapter 2 presents the state of the art, bla... Chapter 3 ..."

- Chapter 2 presents the state of the art
- Chapter 3 explains how to leverage SATD to train a Deep Learning model to detect TD
- Chapter 4 Empirical Study design
- Chapter 5 Results discussion
- Chapter 6 Threats to validity
- Chapter 7 Conclusion

Chapter 2

State of the Art

The following sections describe existing approaches to detect different types of technical debt.

2.1 Automatic Identification of Software Bugs

This section deals with tools that are capable of detecting bugs automatically. We divide bugs in three different categories:

- Memory related bugs, e.g. null pointer violation, memory leak and buffer overflow.
- Concurrency related bugs, e.g. critical race conditions, deadlock and unsafe concurrent data access.
- Semantic related bugs, when the fault arise from a contradiction to the intention of the programmer or the original design.

Different tools use different methods to detect bugs; the first big distinction to be made is between dynamic and static analysis.

Static analysis is performed with no execution of the program and relies only on the source code or on some object code.

Dynamic analysis executes the program and inspect its behaviour at run-time. It can be implemented in many flavours, for example: with instrumentation of the executable, with a virtual processor or taking the form of a scheduler. Two are the main factors that are often taken under observation, one is the performance loss in the execution and the second is the extension of the run-time monitor (both in quality and quantity).

We can define another distinction on how bugs are identified; many tools use rules to detect if some violation has occurred. The rules themselves are of two types: programming and statistical rules.

Programming rules are usually clear and squared, e.g. they derive from axioms, mathematical models or are manually defined.

Statistical rules are defined through a statistical analysis on multiple samples. Usually there is a training phase where observations are collected and correct rules (invariants) are refined.

Another means of detecting bugs is *model checking* where the tool verify correctness in a usually finite state system. This verification is performed using formal methods on typically formally specified system.

The last category for bugs finding techniques is the *annotation-based* tools. Those requires the annotation of programs to extract semantics and verify consistency and correctness.

What follows is the description for each selected tool.

PREfix [7] A static analyzer for finding dynamic programming errors

PREFix is a source code analyzer that detects a broad range of errors in C and C++ code. Its goal is to detect many runtime issues on real world programs, without dynamic analysis and instrumentation; only the source code text is used. PREfix can detect defects efficiently through a model that abstracts functions and their interactions; the analyzer traces execution paths handling multiple language features: pointers, arrays, structs, bit field operations, control flows statements and so on.

The method used by the tool is based on the simulation of individual functions. It employs a virtual machine that simulates the actions of each operator and function call. With the detailed tracking it can report defects information to the user so to easily characterize the detected error. The tool can be applied both to a complete program source or only a subset. This bottom up approach is particularly useful when the source code is not fully available (e.g. in the case of a thirty part library).

RacerX [9] RacerX: effective, static detection of race conditions and deadlocks

RacerX deals with complex multithread systems. It detects race conditions and deadlock using static analysis. It can infer from the source code the object lock that is assigned to a particular code block. It detects code that is used in a multi-threaded context; it also detects when the code endeavours in dangerous shared access.

RaceX uses annotations only to mark the code that deals with lock acquisition; this requirement keeps the burden on the user to the minimum to increase ease of adoption. The authors of RaceX report their experiences on the biggest problem about race detection: in large codebase there are massive amounts of unprotected variable access; the key point is to report only those that can actually cause problems. There is emphasis on two aspects:

- The first is to minimize the impact of reporting false positive in order to avoid

the users to discard the use of the tool; to achieve this, RaceX employs specific techniques to lower the impact of analysis mistakes.

- The second is the speed of the tool: the authors keep the time of execution for the analysis under deep scrutiny; they claim that for a codebase of 1.8 LOC the time required is between 2-14 minutes.

The tool has been found capable of finding severe problems in huge projects like Linux, FreeBSD and also in a large closed source commercial software.

Purify [18] Fast Detection of Memory Leaks and Access Errors. Winter 1992 USENIX Conference

Purify is a dynamic analysis tool for software testing and quality assurance . It instruments the object files generated by the compiler (the software dates back to 1992 and the supported platform is Sun Microsystem's SPARC); the process that acts on the object files include also third-party libraries. Purify detects multiple errors: memory leaks, access error, reading uninitialized memory. The injected instructions check every read and write memory operations; the slow down of the target is under three times in respect to the non-instrumented execution time. Enabling the use of Purify is as simple as adding a word in the makefile; the generated overhead is inside the limit of tolerance of developers and it allows to detect bugs early in the development cycle.

Valgrind [38] Valgrind: a framework for heavyweight dynamic binary instrumentation

Valgrind is a framework available for many Linux, Android and Darwin architecture. It is the foundation where many tools are built upon. All these tools, several are already included with the standard distribution, help to build more correct and faster program and are categorized as dynamic analysis tools.
The eight supplied tools can be divided in the following groups: memory error detector, cache profiler, thread error detector, heap profiler. There are also two additional tools that are provided to illustrate how to use the framework works and how to use the core low level infrastructure to implement instrumentation.
Basically, the core implements a synthetic CPU that asks the selected tool how to instrument the code and then continues and coordinates the execution. All the instructions are simulated and the memory access is sandboxed; this includes also the third party library linked into the executable. There are ways to manage and suppress every output generated to avoid clutter and unwanted error reports.
It is advised to enable the debug info into the executable; without them Valgrind is unable to determine which function is the owner of a specific instruction, as such it will produce almost useless error and profiling messages. It is also advised to use minimal compiler optimization to avoid incurring into false positive error reports.
The slowdown of the execution depends on the specific instrumentation of the selected tool, and it is roughly between four to fifty times of the original speed.

CP-Miner [28] CP-Miner: Finding Copy-Paste and Related Bugs in Large-Scale Software Code

CP-Miner stands for Copy and Pasted code Miner. Zhenmin et al. found that a significant portion of source code in many widely used open source projects are duplication made with copy and paste. This practice introduces bugs, the main reason being that programmers leave identifier untouched instead of renaming them consistently to match the new code context. When the label do not exists in the new place it will be detected with a compilation error; on the other hand, if the identifier exists in the new context it will not be detected by the compiler and it will introduce a hidden bug very hard to detect.
CP-Miner tolerates modification to the pasted code. In order to be detected, the segments do not need to be identical; they can also contain insertions or modifications. The tool is capable of detecting duplicated code but not all detection are true positive; nonetheless it is able to report many significant duplicates with hidden issues. The authors detected 28 copy-paste related bugs in the Linux kernel code base and 23 in FreeBSD; these bugs were reported and most of them where previously unknown to the project team. The analysis is done infra-project and targeted the following large projects: Linux, FreeBSD, PostgreSQL and Apache HTTP server.
The main contributions of the paper are: scalability, bugs detection and statistical study of the copy-pasted code.

- *Scalability* is a strong point of CP-Miner because the technique used in it allow to quickly and efficiently scan large projects including operating system code. For example, it took 20 minutes to find 150,000-190,000 copy-pasted segments respectively in the FreeBSD and Linux kernel; such fragments account for roughly one fifth of the code base. At the time, both projects had more the three million lines of code.

- *Bugs detection* was found to be very effective because of the positive response from the open source project maintainers; most of the reported bugs were not found by static or dynamic analysis detection tools.

- The authors conducted a *statistical study* of the copy-pasted code to give an overview of the phenomenon; the majority of the copied code is between 5 and 16 statements. Around 50 percent of the code has only two copies but around 7 percent has more than eight duplicates. Roughly 12 percent of the copy-pasted code segments are whole functions. Kernel modules are affected with different concentration, depending on the module under analysis: drivers, arch and crypt modules have high copy and paste segments than other parts of the project.

D. Engler's [10] Bugs as Deviant Behaviour: A General Approach to Inferring Errors in Systems Code

The approach taken by Engler et al is to extract from source code beliefs and properties that must or could hold. Through static analysis it is automatically detected two different types of beliefs: MUST beliefs and MAY beliefs. The first one is something that must certainly hold, for example the dereference of a pointer holds the credence that the reference is not null and must be valid. The second one is statistical by nature; the code is observed and searched for patterns that suggests beliefs, for example a call to function "x" followed by a call to function "y". The probabilistic nature of the MAY belief comes when validating it: at first it is treated as a MUST belief then a search yields all uses in the source code of such belief (i.e. the use of "x" and "y"). If it turns out that such pattern is respected most of the times then the belief is probably valid, otherwise it is treated as a coincidence and, as such, discarded. MUST beliefs bear no doubt about their validity and represent internal consistency. All contradictions from both MUST beliefs and valid MAY beliefs are reported as errors (i.e. bugs). The authors leverage their prior work [11] where they used static analysis to fix manual defined rules for specific system; for example a call to spin_lock(l) must be paired with a following spin_unlock(l). Such patterns were previously specified by hand; with the current work they enable a system to infer the same (and more) rules automatically. They reported that the automatic system is able to detect all the manual rules plus a considerable additional amount that goes from ten to one hundred patterns more. The general idea underling this paper, is that the source code contains intrinsic information about what is correct; finding errors in real system means exploiting what is intended as "correct". Sadly, most of the times these rules are not documented, not formalized or if they are available, they are present in informal and unusable format. With this work the authors use static analysis to extract the beliefs that the programmers infused in the source code, without the need of a priori knowledge. Manually performing the task of extracting correct behaviours and rules from source code is usually a hard, difficult and daunting experience, particularly in view of multiple releases and big code bases. Engler et al show of being able to apply this techniques to complex systems as Linux and OpenBSD operating systems. The results are hundreds of detected contradictions (i.e. errors or bugs) reported; many of them have been assessed and resulted in kernel patches.

DIDUCE [17] Tracking Down Software Bugs Using Automatic Anomaly Detection

The paper introduces a tool written by the authors called DIDUCE: Dynamic Invariant Detection ∪ Checking Engine. It is based on instrumentation of Java programs so to observe their behaviour at run-time. During the lifetime of the target Java process, DIDUCE gather information and collects hypothesis of invariants; those are the rules that the program should obey. As violation to the invariants are encountered the tool relaxes the hypothesis allowing for different behaviour. Every invariant has a confidence level, the process previously described updates it; then the user can go through all the anomalies reported ordered by their rank. The tool is intended to be used in the

discovery of the root cause of bugs and as an aid in better understanding the program under analysis. The query of this ranking can be done, for example, just before a crash occurs: inspecting what DIDUCE detects as anomalies often lead to the discovery of the root cause of the problem.

The paper describes also the findings during the application of DIDUCE to four Java real-life programs, one of which is the JSSE Library (Java Secure Sockets Extension); the bug under analysis was found during the development of a proxy server to be applied to the JSSE Library. The problem was that using the proxy server triggered unexpected behaviour in the JSSE internals. Thanks to DIDUCE the programmer was able to find the issue in the core of the library: it was reported with high confidence that method read() returned a different value than usual. This method returns the number of bytes read from the stream and Java specs clearly define that the method can also return with a buffer not fully filled. A common Java programmers pitfall is to ignore this result and skip the loop on read() to obtain all expected data. DIDUCE reported a violated invariant with high confidence because the result from the method was always 74 and in one instance (just before the Exception) the number was smaller; this information made apparent that using the proxy triggered a bug present in JSSE Library. The authors' experience suggests that discovering bugs using their proposed methodology is simple across many different kind of programs, shown in the paper with four compelling cases.

AccMon [57] AccMon: Automatically Detecting Memory-related Bugs via Program Counter-based Invariants

The contributions of this paper are two innovative ideas: the first is a novel statistical approach to detect bugs in memory related issues, the second is a novel architectural extension to decrease the overhead of the monitoring process.

The first is called PC-based invariant detection (PC stands for program counter); it leverages the observation that most programs access memory location mostly from the same instructions. Being probabilistic in nature, it can detect memory access anomalies that deviate from the baseline; they usually are the causes for bugs, stack overflows and many other memory-related issues. We can see that there are tho phases: one where the statistical data are gathered and the baseline rules (e.g. invariants) are formed; the other phase put to use the collected rules and check for violations that will be reported. These two phases are intended to be used in multiple runs but also in the same single long-running execution.

The second contribution is called Check Look-aside Buffer (CLB); it aims to lower the burden of the dynamic process monitor activities to decrease the overhead needed. The authors report their experiments on performance: in the worst case analyzed the loss of speed of the process is less than 3 times. In other tools, this slowdown can be of one order of magnitude greater.

The effectiveness of the authors' contribution are tested through AccMon, a tool

they developed in order to implement the idea of a PC-based invariant detection; such experiments shows that the proposed novel ideas are sometime capable of finding more bugs in respect to other tools such as Purify or CCured. Then, in conjunction with previous work from the same authors (iWatcher) it is demonstrated the effectiveness of CLB in lowering the burden of the overhead.

The authors report many other advantages in using AccMon in respect to other tools:

- the analysis is not done on the values of the variables, thus It can detect also bugs that do not violate value-based invariants.
- in the current form, it uses source code to achieve compiler-based optimization but it can directly work with binaries without the need of compilation.
- it does not need type information; given it's statistical nature it can detect anomalies just using abstract memory pointers.
- it's possible to switch off the monitoring activities dynamically at runtime, with almost no overhead. The authors states that AccMon can be used in production runs.

Liblit's [29] Bug isolation via remote program sampling

The underling observation that pave the road for this paper is that often the user community of a program has more raw throughput of running and executing it than the developers. In other words, the number of executions that the team responsible for the program can apply for testing is dwarfed by the number of executions that the community can or will bring up to bear. The authors propose an infrastructure for gathering data from the user's execution to a central information store; then they propose a process, called automatic bug isolation, to analyze gathered data in order to provide information to the developers to help find and fix bugs. This infrastructure shows multiple benefits:

- Use a vast amount of data that is generated by the execution of the program by the user community; it is usually discarded and do not contribute to better the quality and the experience for the user itself.
- Enabling the collection of information helps to draw a clearer picture on the effective use of the program and drive better decisions about development roadmap
- Map and define feature usage statistics
- Avoid issues related to manual feedback generated from an user intervention: usually the user is unsophisticated and non technical; it's ability to have a positive impact on the bug reporting is limited. The benefits of automatically gather this information are many and varied.

Designing an infrastructure that was able to scale was a non-trivial process; there are two main issues to address.

One is to make the lowest impact on the performance on the program execution. It is very important to be respectful of the resources used in gathering debug information. To achieve this goal, the authors employ sampling; they also address a technique to conduct fair sampling.

The other one is a craftiness in gathering and periodically sending data to the central system: even collecting a small amount of information have an huge impact on scalability.

The authors then focus on the data analysis phase. They propose three different application with increasing level of sophistication:

- They show how to share the burden of assertion across the use base so to inflict upon each user only a small fraction of the checks.
- They show how to start from a large set of predicates (predicate guessing) and shrinking it down over time to reveal the smallest set that can deterministically predict a bug.
- They show how to use linear regression to isolate non-determinstc bugs; in other words they shrink the set of predicates that has the highest correlation with the failure.

ESC/Java [12] Extended Static Checking for Java

This tool perform static analysis on Java source code looking for errors and warnings to report. It provides specific Java annotation to formally express design decisions. ESC handles and warns about multiple common programming errors, e.g. null dereference, index out of bounds, types errors; it also warns about concurrency errors, like race conditions and deadlock. Aided by the custom annotations, ESC employs an engine to decode the semantics of the program and apply techniques to automatically prove theorems; doing so it is able to report potential bugs that are not detected by the type system and that are detected at runtime.

One core requirement imposed on ESC is the modularity of checking; in other words, it can work on pieces of code (i.e. methods) in isolation to the rest of the code. This restriction was chosen for scalability reasons even if the downside is the need of custom annotations. The authors argue that the cost of using the annotations are not an hard overhead: when developers are engaged in manual code review they need information that usually come from unstructured sources (e.g. natural language comments) and they already sustains the burden of gathering additional knowledge not present in the raw code.

It is evident throughout all the paper the importance the authors put on the tradeoff between the cost of the annnotation process and the benefit of true errors feedback; this sentence taken from the paper's introduction sums the core of this tradeoff: "if the

checker finds enough errors to repay the cost of running it and studying its output, then the checker will be cost-effective, and a success". Infact, they enumerates two important features needed by an ideal static checker :

- soundness - if the program has errors, it will find some
- completeness - every error reported is a true positive

ESC do not seek to honor these two features for the very belief quoted above. The authors observe that the alternative processes used to achieve software quality (testing and code reviews) do not possess any of the features of the ideal static checker.

At paper time of writing, ESC was used for two years on multiple kind of programs and was proven effective in finding meaningful bugs. The performance was adequate for interactive use on most methods. Even if it was proven of real usefulness, the users' feedback suggest that the cost of annotating was high and the number of warnings was excessive. It must be said that the annotation were added after the development of the project and not during the evolution of it; this is commonly know by developers to be a dreaded task.
At this point it is unclear if the tool delivers a positive tradeoff between cost and benefit; the experience of the authors during internal use of the tool was encouraging. They believe that ESC is already a valid tool to be used in classrooms: it enforces good design, modularity and verification.

VeriSoft [14] Model checking without a model: An analysis of the heart-beat monitor of a telephone switch using VeriSoft

The goal of VeriSoft is to detect problems in a concurrent reactive system (CRS) through the exploration of its state space. The *reactive* word stands to describe the continuous interaction of the system with the environment. The issues detected are: deadlocks, assert violations, livelocks. A CRS is composed of two parts: a finite number of processes written in arbitrary code (e.g. C, C++, Java, tcl, and so on) and a finite number of communication objects (TCP connections, semaphores, shared memory, and so on). The need of such tool arise from the difficulty of writing a robust and reliable CRS; it is well accepted that concurrent systems are prone to unexpected issues, difficult to track, test and to reproduce.

What VeriSoft does is a systematic state space exploration; it defines the state space as a directed graph: the nodes are the global states and the edges are the transitions between states. It follows that each global state should be uniquely identified and this is one of the core issue that VeriSoft solves with an original combination of algorithms; the author calls it "an efficient state-less search".

The paper reports the analysis conducted with VeriSoft on a software owned by Lucent Technologies: "Heart-Beat Monitor" (HBM). Such software monitors the status of a telephone switch elements and determines their state based on the propagation delays of messages sent through those elements. HBM is an important piece of software

because it has a big impact on the switch performance due to its influence on the switch routing.

The experience of the authors is reported as successful: they were able to find errors in the documentation and in the software itself. Subsequently they modified the code to strengthen some properties and tested them again with VeriSoft; after another run, as desired, VeriSoft reported satisfactory results. The development team of HBM decided to integrate the code changes from the authors for the next commercial release.

VeriSoft acts as a scheduler and has complete control over non-determinism so it can reproduce any interesting scenario (i.e. those that during the automatic tests led to errors and issues). One other benefit is that there is no need to describe the model with specific languages: it relies on the exercise of the actual code. One downside is that it cannot detect cycles in the graph of the global spaces and, as such, it can only detect violation of safety properties.

JPFinder [19] Applying Model Checking in Java Verification

Java PathFinder (Jpf) is a prototype translator from the Java language to Promela (Process Meta Language). Promela runs on SPIN (Simple Promela Interpreter). SPIN is a general tool to find concurrency problems and verify the correctness of a system.

It is reported that Jpf is not the first attempt of Java-to-Promela translator; in addition to this other work, Jpf can handle a significant number of features of the Java language. At the same time, many other are missing. The paper describe the issues due the impedance between Java and Promela; they come in two flavors: performance issues and missing feature issues.

Jps provides the programmer with Java static methods to annotate the source code with assertions; those assertions will be checked with the SPIN model checker.

The authors used a Chess game server written in Java as the test subject for finding synchronization bugs. They did not use the original source code but wrote a simplified abstraction in Java composed with 16 classes and roughly 1400 lines of code; it is reported that this Java program was non trivial and the development was done without thinking about formal verification. Then the authors fed the Java semplification to Jpf and were able to find a bug that was later confirmed.

CMC [37] CMC: A pragmatic approach to model checking real code

Model checking is very hard in practice; it usually involves the use of a specific domain language to describe the model and then a model checker. This common approach to model checking is very hard to endure in practice: it exposes the age-old dualism of having two parallel systems. On one hand we have the actual implementation, on the other hand we have the abstraction that represents the model of the implementation. Having two distinct bodies opens the following issues:

- The model could exhibit issues that are not present in the actual implementation.

- The implementation could show bugs that are not present or detected in the model.

- The need to maintain both systems coping with the impedance of two different ways of expressing meaning, behaviour and intent.

Complex systems often hides rare but nasty bugs that arise only after many weeks of continuous run; this is a major issue with such systems. Explicit checkers can help in this scenario; they search a huge state space without wasting the resources for repeated parts of usual testing.

The first contribution reported by the authors is CMC: C and C++ model checker. It works directly on the implementation without the need to create a separate model to be checked. CMC needs some adaptations on the code, some are just good programming practices (e.g. asserts, specifying the environment), other are changes required specifically by CMC: one is for handling the non-determinism and another one is to handle the initialization functions and event handlers. The tool works by directly executing and scheduling the system under analysis. It needs to store and load the whole state space and as such it handles techniques to cope with the *state explosion problem*: simple heuristics help to prune a huge amount of states.

The second contribution is the application of CMC to three implementations of AODV routing protocol. The actual goal of CMC is to check network code implementations, but the ultimate goal is to check a broader range of programs. During the experimentation the authors, through CMC, were able to find 34 errors many of whom were meaningful errors. Actually, their work exposed also a bug with the AODV specification itself (that was later acknowledged in the RFC 3561 citing the first author of this paper).

Network code is of core importance for the stability of a system; it is prone to many issues that undermine correctness, e.g. packet loss, hardware errors, security attacks.

CMC proved to work well, given the results on three different implementation of a routing protocol. For a wider acquisition of CMC it essential for the authors, to lessen the burden of the code adaptation and automate it as much as possible.

2.2 Code Smells and Anti-patterns

This section explains briefly what code smells and anti-patterns are. The following sections contain the most interesting contribution from the literature on the subject.

An Anti-pattern is a common poor solution to a design problem. It presents itself in

object-oriented based systems where a developer applies a solution that is usually ineffective or worse, harmful. The term stems from its correct and desirable counter part: design pattern. Design pattern is defined as a reusable solution to a commonly occurring problem within a given context in software design.

Anti-patterns are well known to have negative effects on software projects: they hinder code comprehension, and increase maintenance costs.

A code smell is any aspect of the source code that hint to a deeper problem; in other words they are symptoms of a potential bigger issue. It does not always indicate that a real problem exists but it suggests to look closer and inspect if something more profound is present.

It must be said that neither anti-patterns nor code smells are strictly bugs: in fact, they do not imply incorrect results and they do not stop the program from functioning. Nevertheless, for the risks specified above, automatic detection of code smells and anti-patterns received a lot of attention.

In the following sections, we will report several contributions from the literature that start to be available in earnest from roughly 1995.

2.2.1 Four inspirational books description

There are four books that inspired many automatic detection techniques. The first editions of these books span from year 1995 to 1999. The following paragraphs describe them briefly.

Webster [54] Pitfalls of object-oriented development

This book analyzes the cycle of object oriented programming shedding light on its weaknesses and shortcomings. It compares and explains OOP and previous programming techniques. It provides insight and counsel on how to avoid OOP risks.

Riel [43] Object-oriented Design Heuristics

The author provide the reader with metrics to understand the quality of the object-oriented software. The book explains guidelines to help make better decisions on the design of the OO system. The sixty recommendations in this book are language-independent and help the reader to evaluate the quality of a software design.

Fowler [13] Refactoring: improving the design of existing code

The objective of this book is to give practical refactoring strategies to apply on projects so to improve the design of existing systems. It describes many code smells and for each it explains the appropriate actions to take to fix the problem. The first edition dates back to 1999.

Brown et al. [6] AntiPatterns: Refactoring Software, Architectures, and Projects in Crisis

The reader finds the detailed description of 40 anti-patterns divided in three categories: managerial, architectural and developmental. Each anti-pattern entry is a card with many key aspects, e.g. name, description, root causes, symptoms and an optional anecdotal evidence.

2.2.2 Last decade proposed approaches

Many proposal for design flaws identification have been done in the last decade; many of them have some of their roots in previous mentioned books.

Travassos et al. [50] Detecting defects in object-oriented designs: using reading techniques to increase software quality

Travassos et al. created a set of techniques to identify manually defects in order to improve software quality. These practices help individuals to read object oriented code and assess, through a predefined taxonomy. The authors conducted an empirical study on these techniques and reported their feasibility.

jCOSMO [52] Java quality assurance by detecting code smells

This paper presents an approach to automate the detection of code smells. The authors assess that code smells are not precise and formal and need human intuition to appreciate them; the outcome of this observation is that a tool to automatically detect code smells needs to be user configurable. They developed a tool based on these ideas, jCOSMO, that comes already configured to detect two specific code smells (instanceof and typecast). The customization deals with three aspects on code smells: inclusion of new ones, exclusion and fine tuning for more precise definitions. The users of jCOSMO can benefit of automatic detection and graphical visualization of the results.

Simon et al. [45] Metrics based refactoring

The authors agree that the developer is the last authority with the power to decide where to apply refactoring techniques. One of their contributions is providing them with metrics to support subjective perception regarding code smells. They believe that a key issue is helping developers with tools that support human intuition. The authors demonstrate that metrics are effective in finding and pointing to places where code anomalies are detected. These are the presented refactorings: move method, move attribute, extract class and inline class.

Marinescu [33] Detection strategies: Metrics-based rules for detecting design flaws

Marinescu observed that there are multiple problems using quality metrics to improve software quality. Often the definition of the metrics are imprecise, confusing or incomplete. Another issue is the interpretation of the metrics; they seldom provide a

model that help to correctly understand and apply them to a concrete situation. Using the metrics in isolation leads to excessive detail and it becomes difficult to use this information to investigate design flaws. In other words, an isolate measure can be helpful to identify the presence of an anomaly but it does not point to the cause; this leaves the developer without a meaningful insight on how to handle the refactoring. The author calls this 'bottom-up' approach. Marinescu proposes a novel method called *detection strategy* to overcome this problem; it's a 'top-down' approach that starts from an abstract high level goal and drives the investigation of design trait that conforms to the strategy. This technique was applied to ten detection strategies and used on industrial case studies; the result of the experiment proved that the method is applicable and usable in practice.

Lanza and Marinescu [27] Object-oriented metrics in practice: using software metrics to characterize, evaluate, and improve the design of object-oriented systems

The ideal reader of this book, as defined by the book itself, is a fluent programmer who concretely deals with the maintenance and evolution of a complex large application. This text uses metrics in a practical way to improve software. It uses them in three phase

- Characterizing the design. The goal is to picture a panorama of the design of the software system. It shows how to use tho metrics-based visualization techniques, *Overview Pyramid* and *Polymetric Views*: Overview Pyramid gives in one shot an overview of the complexity, coupling and inheritance; Polymetric Views shows entities and their relationships supplied with metrics.

- Evaluating the design. This phase serves to better understand and assess the design of the application. *Detection strategies* [33] supply a tool to detect flawed design. *Class Blueprint* is a powerful tool to visualize information about classes, i.e. control-flow and access structure; the goal of this tool is to closely inspect the design flaws detected.

- Design disharmonies. There are three categories of disharmonies: Identity, Collaboration and Classification.

The authors, after the identification of disharmonies, propose also insight on how to improve the design through refactoring.

Munro [36] Product metrics for automatic identification of "bad smell" design problems in java source-code

Munro's focus is on automation of bad smell detection in Java source code. The underling motivation of the paper is to enhanced the process for finding places where to apply refactoring. He begins with a precise definition of code smell, building up on the informal definition from Fowler and Beck. The core idea is to analyze the description

of the code smell and translate it in measurable attributes that are quantifiable. This process is divided in three phases: informal definition analysis, extract possible quantifiable measures and establishment of rules that are used on the metrics to identify the smells. The author applies this process to two specific smells: Lazy Class and Temporary Field. For example, he extracts quality measures as NOM (number of methods), WMC (weighted methods per class), LOC (line of code) and CBO (coupling between objects).

Moha et al. [35] DECOR: A method for the specification and detection of code and design smells
The authors split their contributions in three parts.

The first is DECOR (DEtection and CORrection), it's a method that systematically and formally describes the process to detect code and design smells. It is based on previous work in the field and it leverages the experience and fill the gap of missing features, for example: explicitness on how to specify the detection algorithms, opacity of the technique used, completeness of the analysis on smells description and others.

The process is defined through five well defined steps: description analysis, specification, processing, detection and validation. The correction part of DECOR, as stated by the authors, is for future work and it is not present in the paper.

The second contribution is DETEX (DETection EXpert); the authors revisit their previous detection technique through the lenses of DECOR and name it DETEX. It uses a DSL to specify smells with high level abstraction and it automatically generates the algorithms for the actual search process.

The third contribution is an empirical evaluation of DETEX; consistently with the fifth step of DECOR, Moha et al. provides evidence of the application of DETEX on four anti-patterns relating the results in the form of precision and recall. The target of these experiments are eleven open source projects.

Tsantalis and Chatzigeorgiou [51] Identification of move method refactoring opportunities

This paper has a narrow but highly focused contribution: define a process to identify Feature Envy code smells to enable Move Method refactoring.

The authors observe that high coupling and low cohesion are well known indicators of low quality software design; those features are often linked to unwanted outcomes: low maintainability, low productivity and high bug density.

We can split the proposed process in two parts: identification and ranking.

The identification phase serves to locate candidate methods that could benefit from a move method refactoring. For each method it is evaluated which class could be the destination; this part is driven with distance measures between entities (class, methods and attributes).

The ranking phase serves to avoid overloading the developer with suggestions. This

part uses cohesion and coupling as measures to drive the sorting; the metric employed (called Entity Placement) works evaluating the effect of the refactoring without actually modifying the source code. It is well specified that this is not a fully automated approach because the designer is ultimate responsible on which refactoring should be applied. As one could expect, there are cases where a move method is not a good choice (e.g. moving unit test methods into the target class).

The researchers evaluate their approach with quantitative analysis through the application on two open source-projects. They also track the evolution of the metrics between multiple iteration of refactoring. They then ask a third party to assess the conceptual validity of the proposed refactoring. Lastly they also report on the computational cost of their approach.

Ligu et al. [30] Identification of refused bequest code smells

The authors propose a method to identify the Refused Bequest anti-pattern. In synthesis such design smell happens when polymorphism is badly used, for example when a subclass overrides all the superclass methods. The detection of the smell is achieved with both static and dynamic analyses.

- Static analysis is used to identify those class hierarchies that are candidates to potential anomalies.

- Dynamic analysis is employed through the exercise of unit testing. The methods of the descendant classes are injected with instructions that intentionally raise an exception; the execution of tests will verify which methods are called or not called, and the result will determine a score towards or away good design. The merge of tests result, from the dynamic analysis, with other structural data generates an output that represents the smell strength.

The authors developed an Eclipse plug-in to incorporate the process described above.

2.2.3 Code smell detection formulated as an optimization problem

Kessentini et al. [21] Deviance from perfection is a better criterion than closeness to evil when identifying risky code

This paper proposes a novel method in detecting bad smells. The idea that fueled this work is inspired by artificial immune systems; such systems behave in the following way: the more something is detected as different the more it is considered extraneous. Based on this assumptions, the authors generate a set of detectors that are able to measure various manifestations of anomalies (smells). Thanks to this measurements, they can evaluate how far the system under inspection is from normalcy.

To be able to create the detectors they need to elect some model to be representative of what normality is; this elected model was taken from the project JHotDraw (by Erich Gamma) that represents, ideally, good design and good programming practices.

The authors report that the results outperform the state of the art and their developed tool is able to detect a good mix of bad smells.

Kessentini et al. [22] A cooperative parallel search-based software engineering approach for code-smells detection

The authors' proposed approach use P-EA (Parallel Evolutionary Algorithms) where multiple algorithms cooperate to find a consensus on the common goal of finding the bad smells. Those algorithms use different adaptations: fitness functions, change operator and solution representations. The cooperation between algorithms happens during the parallel execution in multiple iterations and it is not just a result of one final consensus.

To test the effectiveness of the implementation of the approach, the authors make an empiric comparative evaluation with: random search and two other methods not based on meta heuristics. The experiment is based on a benchmark of nine large open-source systems; the reported results shows that the approach is better of the state of the art.

Boussaa et al. [5] Competitive coevolutionary code-smells detection

The authors propose a novel approach to finding bad smells: they use two populations with their CCEA (Competitive Co-evolutionary Algorithm) search and they employ the use of a code-base sample that contains bad smells.

The first population goal is to maximize the detection of bad smells thanks to the generation of rules based on quality metrics.

The second population goal is to maximize the number of synthetic bad smells that the first populations is missing.

The two populations behave similarly to a machine learning GAN framework (Generative Adversarial Network).

The evaluation of the ideas is conducted on four systems through an existing benchmark. The authors reports the statistical analysis: CCEA shows great promise in its performance compared to random and single population approaches.

Sahin et al. [44] Code-smell detection as a bilevel problem

The proposed idea and implementation is based on a bilevel optimization. In such formulation there is an outer problem (upper-level) and an inner problem (lower-level).

The upper-level optimization goal is to maximize the detection of bad smells in a sample dataset through the generation of rules based on quality metrics.

The lower-level maximizes the generation of new artificial bad smell samples that

are not detected by the counter part.

The evaluation of the system was performed with 31 runs on nine open source-projects; seven bad smells were detected with an average of more than 86% in terms of precision and recall.

2.2.4 Non binary classification

The previous section identified code flaws in a binary class classification (smell/clean), while the following paragraphs focuses on those analyses that take care of borderline classes.

Khomh et al. [23] A bayesian approach for the detection of code and design smells

The nature of bad smells contains a measure of uncertainty due to its natural language definition. The authors propose an approach to handle this uncertainty with BBNs (Bayesian Belief Networks).

Through a systematic process of their own making, the authors convert classic detection rules to a BBNs probabilistic model, using the Blob anti-pattern as their test bench.

They use two open-source projects as targets for the evaluation of the model: GanttProject and Xerces. The authors make a comparison between their model and DECOR and show that it returns the same defective classes plus ordering them by importance.

The last contribution is about exploiting bad smells historical information (in the sense of alternative pre-made dataset) in order to train a machine learning model using Weka; the authors show that this calibration increases the quality of the detection.

Oliveto et al. [39] Numerical signatures of antipatterns: An approach based on B-Splines

This paper proposes to overcome two limitations of previous detection technique: the first limitation is the binary classification, there is no in-between or continuous classification of bad smells (e.g. DECOR by Moha et al. [35]). The second limitation is the need of expert knowledge to fuel the detection model (e.g. BBNs by Khom el al. [23]). To overcome these limitations, the authors propose ABS (Antipattern identification using B-Splines). It creates signature of anti-patterns using quality metrics; then it uses the B-spline to create an abstraction of such metrics. This process is applied both to known codes containing bad smells and to unseen unclassified source code; the distance with the B-splines of known anti-patterns measure the similarity to known anti-patterns. In reference to the second limitation, the authors observe that their technique needs only a dataset but no human intervention and tuning.

2.2.5 Usage of historical data for code smells

These two last papers of the section exploit the use of source version control as the basis for their contribution.

Ratiu et al. [42] Using history information to improve design flaws detection

The authors propose an idea to make use of historical data on source code to increment performance on the detection of bad smells. The underling concept is that the evolution of a system can give useful feedback to better determine and analyze the last state of the system. This paper uses the foundation of *design strategies* (Marinescu [33]) adding the concept of a system that evolves through time. The *history* is defined as a sequence of states of the same entity (e.g. system, class and method). The history is used to calculate and evaluate the entity measures *persistence* and *stability*.

The contributions of this paper are: (1) definition of a measure to show how persistent a smell is and how much maintenance effort it absorbed (2) show the improvement in accuracy detecting two class smells (3) describe the valuable information extracted from the history of the anomalies.

Palomba et al. [40] Mining version histories for detecting code smells

This paper shows how to exploit changes on source code to achieve bad smell identification. The history of changes are extracted through the versioning system.

A novel approach is proposed, called HIST (Historical Information for Smell deTection); it detects five classes of code smells: Divergent Change, Shotgun Surgery, Parallel Inheritance, Feature Envy and Blob.

Using the source history of a project is the only way to detect some bad smells; for example, the Parallel Inheritance smell definition cannot be decoupled from tracking the changes in the code. In other words, the very nature of such smell needs to be able to analyze the changes through time of the system.

There are other kinds of smells that do not strictly need historical data but can benefit from using it; for example, Divergent Change smell and Shotgun Surgery have literature that shows detection approach using last-snapshot information only.

The authors compared the accuracy and recall of HIST with alternative approach and their approach tend to perform better. It is reported that HIST is able to detect smells missed by others (the recall is in range 58% and 100%). The previous evaluation was achieved with an empirical study on twenty Java projects; it comprised accuracy and recall calculated against a manually-produced oracle.

A second empirical study represents the closing edge in the loop of the detection process: feedback from the developers of the projects. The goal was to assess at what

extent the programmers agreed on the smell detected by HIST: 75% of the anomalies detected were reported as true problems by the people contacted by the authors (20 developers of 4 projects). This paper makes available a comprehensive replication package.

2.3 Self-Admitted Technical Debt

Throghout all this section Technical debt will (could) be shortened to TD and foo bar satd

Potdar and Shihab [41] An exploratory study on self-admitted technical debt

The authors conducted an empirical study on four open source projects, focusing on three main research questions; they aimed to:

- RQ1 finding the concentration of SATD in the projects
- RQ2 discovering the reasons for introducing the SATD
- RQ3 calculating the per of SATD removal after its introduction

They found that 2.4% to 31% of the files contained SATD. An interesting finding is that the experienced practitioners are the most likely to introduce SATD. On the other hand, a counter-intuitive discovery is that the amount of SATD has correlation with neither complexity nor time pressure. The removal ratio was found to be roughly between 0.26 and 0.63.

Storey et al. [47] TODO or to bug: Exploring How Task Annotations Play a Role in the Work Practices of Software Developers

This empirical study has the goal to shed light on how the developers behave on personal and team tasks, in respect to source code annotations (i.e. comments). The authors analyze the relations that annotations have with common used tools like, e.g. wikis, issue and bug trackers. In order to do so, they gathered and combined data coming from a mix of methods, divided in phases:

- Phase 1. It was conducted a survey targeting users of Eclipse IDE. The topic was about annotations: if they did write them, which types and the use of them.
- Phase 2. The authors did contextual interviews with developers on three open source projects. Then augmented the answer from the interview with direct analysis on many versions of the source code related to the annotation in question.

In the conclusion is reported how these finding could be useful to improve tooling and software process.

Guo et al. [16] Tracking technical debt - An exploratory case study

This paper aims to highlight and make evident the effects of technical debt on the cost and management of a software project. Through the tracking of a single delayed task in a real project, the authors analyze the effect of such technical debt. They created a framework for the explicit management of TD and then applied it, with a simulation, to the real scenario under scrutiny. The objective of this study is:

- determine technical debt effects on the project and evaluate their impact
- after the application of the simulation, determine if the framework provided real gain and uncover benefits.

The results of this simulation made clear that careful planning and analysis of TD is of high importance: in retrospect the cost of the delayed task almost tripled the cost for the project.

Klinger et al. [24] An enterprise perspective on technical debt

This short but meaningful study see the design of an interview to four IBM technical architect. One of the idea was to broad the view about TD from the perspective of the single developer to the perspective of an enterprise. Starting from the premise that TD can be leveraged as a financial asset (i.e. incur in TD today to gain competitive advantage and repay tomorrow) the study and the interviews are structured to assess how an enterprise handle TD from these standpoints:

- how decisions to acquire TD are conducted
- what is the leverage earned contracting TD

The following are some of the findings:

- two different sources of unintentional contraction of TD: from non-technical stakeholders (e.g. fixing a stringent release date at the expense of software quality) and from external forces (e.g. changes in the market and acquisitions).
- the process of acquiring TD was informal. The decision had no written records or written analysis on the impact, effects and expectations of such choice
- scarcity of knowledge and awareness on the consequences of taking on TD, insufficient channels of communication and lack of a common vocabulary to express contracted costs

Kruchten et al. [25] Technical debt: From metaphor to theory and practice

This article expands the original metaphor of technical debt by Cunningham in search of a better definition that enables reasoning on a variety of technical debt. The authors want to lay a theoretical foundation to help the challenge of dealing with TD. These are the main points covered by this work:

- TD Landscape. It's a possible organization of the many aspects of software improvement. It divides between visible elements (e.g. new features and defects) and mostly invisible (e.g. architecture and code). The idea is that TD is limited to the invisible part

- Tackling of TD. The authors reason about the root causes of TD (e.g. carelessness, lack of education and poor processes) and describe which steps can effectively handle TD (e.g. awareness, explicit management, understand what tools can and cannot do, nurture architecture, documentation)

- Unified theory. It is observed that the challenge is making the right sequence of changes to improve the software; in respect to this, perhaps the financial and economic models could be the underling layer to the TD landscape (i.e. express all the changes tied to their cost and value over time).

Lim et al. [31] Technical debt: towards a crisper definition report on the 4th international workshop on managing technical debt

Lim et al. conducted an interview study on 35 practitioners aimed to define the perceived characteristics of technical debt and in what context TD was encountered. What emerged is most of the teams know well TD and it is an unavoidable necessity in the business reality. Because its certainty one key factor is active management of it: recognition, tracking, analysis, cogent decisions and prevention of worst consequences.

The participants were queried with both specific and open questions. Aside of general demographic questions, they were asked to describe an example of TD with its properties, causes, effects and benefits. The answer pointed to a different root cause than sloppy programming, poor discipline. Most of the testimony acknowledged that TD was acquired through an intentional decision; some of those decisions were the results of short-term thinking, yielding to the pressure of the moment. The negative effects of TD were perceived as long term (e.g. the fear to change code expecting to break other parts of the system). In some cases it was clear that the benefits were far repaid, in others it was not clear if the balance was positive. The respondents provided many examples of situations that provided the crucible for TD (e.g. contracts with string deadline, exploiting market opportunity windows).

The interviewees reported some of their strategies to handle TD:

- do nothing. In those parts where low maintenance is required, it's safer to leave things as they are

- establish a policy to use development resource to fix TD (5 to 10 percent on total resources)

- communication and open dialog about TD between all parties involved (technical and non-technical stakeholders and customers)

- make TD explicit and visible to all the developers (e.g. through audits) and keep track of the discoveries.

Zazworka et al. [56] A case study on effectively identifying technical debt This paper conducted a study to compare manual and automatic technical debt detection. The manual detection was implemented through a questionnaire undertaken by five developers in the same team. The automatic detection was performed using three stable and established tools. All participants reported different debt (except in on case) so there is almost no consensus in the human component, on the other hand, the results show a good overlap between manual and automatic detection regarding defect debt. Human intervention is still needed for the other types of debt: documentation, design, testing and usability debt; they were, for the most part, unrecognized by automatic tool.

Spinola et al. [46] Investigating technical debt folklore: Shedding some light on technical debt opinion

The goal of this paper is to provide some guidance on new research questions about TD. Exploiting the folklore extracted from grey literature, the authors gather 14 statements; then they proceeded to survey to 37 practitioner asking their level of agreement/disagreement. The most agreed upon statement 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".

The underling observation of this paper is that common belief, traditional stories and customs (i.e. folklore) can help the discovery of interesting topics; then, the agreement of knowledgeable people on those concepts could give a measure of value and worthiness and guide possible future research.

Alves et al. [3] Towards an ontology of terms on technical debt

Alves et al. proposed an ontology of terms on technical debt. They developed a *lightweight domain ontology*, designed the quality criteria, conducted a systematic literature mapping and finally submitted the result to a specialist for an evaluation.

The main contribution of this work is to gather information that was spread out and organize a common vocabulary for the field of TD. This common ground wants to help researchers and practitioners evolving the Technical Debt Landscape [**izurieta2012organizing**]. The first contribution is the organization of 13 types of TD: architecture, build, code, defect, design, documentation, infrastructure, people, process, requirement, service, test automation and test debt. The second contribution consists in the organization of indicators themselves; these indicators were used to support identification of the TD.

Maldonado and Shihab [32] Detecting and quantifying different types of self-admitted technical debt

The contribution of this paper is the classification of SATD types in four open source

projects. The authors manually classified 33093 comments; these are the findings with the range of presence across projects: design debt (42% - 84%), requirement debt (5% - 45%), defect debt (4% - 9%), test debt (0% - 7%) and finally documentation debt (0% - 5%).

The projects were chosen in the Java realm in different domains with well commented sources: Apache Ant, Apache JMeter, ArgoUml, Columba and JFreeChart. Using JDeodorant as comments extractor, the authors gathered more than 166K comments. This number decreased to roughly 33K thanks to processing and filtering of those comment with low likelihood of being SATD. Such operation was conducted through four simple heuristics that targeted the following cases: license comments (removal), commented source code (removal), javadoc (removal), multi line comments instead of block comments (joining). The classification process made evident that one SATD can belong to multiple categories (e.g. a design debt can also be a defect debt at the same time). For the sake of clarity this paper will associate only one class to the SATD; in case of ambiguity between multiple types, the more meaningful one was chosen. The set of possible SATD classes is taken from Alves et al. [3] paper. It is observed that not all 13 original TD classes are found in the selected open source projects; Maldonado and Shihab argues that some of the technical debt are not likely to be reported in written comments (e.g. people and infrastructure debt). The authors reports that the personal bias and subjectivity can be threats to internal validity (the manual classification was executed by the first author). Other factors on internal validity: quantity and quality of comments could be affected by biased filtering. About external validity, the authors consider the domain of the projects: it is diverse but all of them are open source Java projects; thus, the results may not generalize to other languages or market segments.

Wehaibi et al. [55] Examining the impact of self-admitted technical debt on software quality

This empirical study on fie open source projects wants to explore the relation between self-admitted technical debt and defects in source code. The results reported is that there is an increase of defects after the introduction of SATD. It's also clear that introducing a SATD makes much harder the changes to the related code.

2.4 TD and machine learning

–this section will reference the paper from CSABA (techdebt conference 2020 related paper)–
(c&p) Several detection techniques and tools have been proposed in the literature. Recently, the adoption of machine learning techniques to detect bad smells became a trend

[20]

Khom et al. [27] and [28] Khomh et al. proposed a Bayesian approach which initially converts existing detection rules to a probabilistic model to perform the predictions [27]. Khom et al. [28] extend [27] by the introduction of Bayesian Belief Networks, improving the accuracy of the detection.

Maiga et al. Maiga et al. proposed an SVM-based approach that uses the feedback information provided by practitioners [35, 36]

Amorim et al. [3] Amorim et al. [3] presented an experience report on the effectiveness of Decision Trees for detecting bad smells. They choose these classifiers due to their interpretability [3]. Thus, most of the proposed works focus on only one classifier. They were also trained in a dataset composed of few systems and, consequently, the results may be positive towards their aproach due to overfitting.

Fontana et al. [21] Fontana et al. evaluated different machine learning algorithms on a set of different systems [21]. Their work was later extended and refined, providing a larger comparison of classifiers [20]. The notorious impact of this work was the incredible performance re- ported. Even naive algorithms were able of achieving great results using a small training dataset. This draws attention to possible drawbacks and limitations of their work, which was later reported by Di Nucci et al. [15]

Di Nucci et al. [15] Di Nucci et al. [15]. They replicated the study and verified that the reported performance was highly biased by the dataset and the procedures adopted, such as unrealistic balanced dataset, in which one third of the instances were smelly.

Daniel Cruz at al. –the paper itself- Detecting Bad Smells with Machine Learning Algorithms: an Empirical Study. Bad smells are symptoms of bad design choices implemented on the source code. They are one of the key indicators of technical debts, specifically, design debt. To manage this kind of debt, it is important to be aware of bad smells and refactor them whenever possible. Therefore, several bad smell detection tools and techniques have been proposed over the years. These tools and techniques present different strategies to perform detections. More recently, machine learning algorithms have also been proposed to support bad smell detection. However, we lack empirical evidence on the accuracy and efficiency of these machine learning based techniques. In this paper, we present an evaluation of seven different machine learning algorithms on the task of detecting four types of bad smells. We also provide an analysis of the impact of software metrics for bad smell detection using a unified approach for interpreting the models' decisions. We found that with the right optimization, machine

learning algorithms can achieve good performance (F1 score) for two bad smells: God Class (0.86) and Refused Parent Bequest (0.67). We also uncovered which metrics play fundamental roles for detecting each bad smell.

2.5 Summing Up (WAS 2.4)

Here I explain why what I did is different

Chapter 3

Using Deep Learning to Detect Technical Debt

Here I describe the approach.

I'm going to start from a short overview, then I go into details

3.1 Mining SATD Instances and their Fixes

3.2 The Deep Learning Model

3.3 Hyperparameter Tuning

Chapter 4

Empirical Study Design

I'm going to start with a short description about the goal/research questions

4.1 Context Selection

I'll explain the dataset used for the evaluation

4.2 Data Collection and Analysis

How I compute the results

4.3 Replication Package

A link to a repo with all data and code

Chapter 5

Results Discussion

. . . results discussion . . .

5.1 Quantitative Results

Report precision/recall for different confidence thresholds

5.2 Qualitative Results

Discuss interesting cases in which your approach succeeds/fails

Chapter 6

Threats to Validity

. . . Discussion on the limitations of my study . . .

Chapter 7

Conclusion

. . . conclusion . . .

Bibliography

[1] Eric Allman. “Managing technical debt”. In: *Communications of the ACM* 55.5 (2012), pp. 50–55.

[2] Nicolli SR Alves et al. “Identification and management of technical debt: A systematic mapping study”. In: *Information and Software Technology* 70 (2016), pp. 100–121.

[3] Nicolli SR Alves et al. “Towards an ontology of terms on technical debt”. In: *2014 Sixth International Workshop on Managing Technical Debt*. IEEE. 2014, pp. 1–7.

[4] Terese Besker et al. “Embracing technical debt, from a startup company perspective”. In: *2018 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE. 2018, pp. 415–425.

[5] Mohamed Boussaa et al. “Competitive coevolutionary code-smells detection”. In: *International Symposium on Search Based Software Engineering*. Springer. 2013, pp. 50–65.

[6] WJ Brown et al. “AntiPatterns: Refactoring Software, Architectures, and Projects in Crisis.(1998)”. In: *Google Scholar Google Scholar Digital Library Digital Library* ().

[7] William R Bush, Jonathan D Pincus, and David J Sielaff. “A static analyzer for finding dynamic programming errors”. In: *Software: Practice and Experience* 30.7 (2000), pp. 775–802.

[8] Ward Cunningham. “The WyCash portfolio management system”. In: *ACM SIGPLAN OOPS Messenger* 4.2 (1992), pp. 29–30.

[9] Dawson Engler and Ken Ashcraft. “RacerX: effective, static detection of race conditions and deadlocks”. In: *ACM SIGOPS operating systems review* 37.5 (2003), pp. 237–252.

[10] Dawson Engler et al. “Bugs as deviant behavior: A general approach to inferring errors in systems code”. In: *ACM SIGOPS Operating Systems Review* 35.5 (2001), pp. 57–72.

[11] Dawson Engler et al. *Checking system rules using system-specific, programmer-written compiler extensions*. Tech. rep. STANFORD UNIV CA COMPUTER SYSTEMS LAB, 2000.

[12] Cormac Flanagan et al. “Extended static checking for Java”. In: *Proceedings of the ACM SIGPLAN 2002 Conference on Programming language design and implementation*. 2002, pp. 234–245.

[13] Martin Fowler. *Refactoring: improving the design of existing code*. Addison-Wesley Professional, 2018.

[14] Patrice Godefroid, Robert S Hanmer, and Lalita Jategaonkar Jagadeesan. “Model checking without a model: An analysis of the heart-beat monitor of a telephone switch using verisoft”. In: *Proceedings of the 1998 ACM SIGSOFT international symposium on Software testing and analysis*. 1998, pp. 124–133.

[15] Yuepu Guo, Rodrigo Oliveira Spínola, and Carolyn Seaman. “Exploring the costs of technical debt management–a case study”. In: *Empirical Software Engineering* 21.1 (2016), pp. 159–182.

[16] Yuepu Guo et al. “Tracking technical debt—An exploratory case study”. In: *2011 27th IEEE international conference on software maintenance (ICSM)*. IEEE. 2011, pp. 528–531.

[17] Sudheendra Hangal and Monica S Lam. “Tracking down software bugs using automatic anomaly detection”. In: *Proceedings of the 24th International Conference on Software Engineering. ICSE 2002*. IEEE. 2002, pp. 291–301.

[18] Reed Hastings and Bob Joyce. *Fast Detection of Memory Leaks and Access Errors. Winter 1992 USENIX Conference*. 1992.

[19] Klaus Havelund and Jens Ulrik Skakkebæk. “Applying model checking in Java verification”. In: *International SPIN Workshop on Model Checking of Software*. Springer. 1999, pp. 216–231.

[20] Konrad Hinsen. “Technical debt in computational science”. In: *Computing in Science & Engineering* 17.6 (2015), pp. 103–107.

[21] Marouane Kessentini, Stéphane Vaucher, and Houari Sahraoui. “Deviance from perfection is a better criterion than closeness to evil when identifying risky code”. In: *Proceedings of the IEEE/ACM international conference on Automated software engineering*. 2010, pp. 113–122.

[22] Wael Kessentini et al. “A cooperative parallel search-based software engineering approach for code-smells detection”. In: *IEEE Transactions on Software Engineering* 40.9 (2014), pp. 841–861.

[23] Foutse Khomh et al. “A bayesian approach for the detection of code and design smells”. In: *2009 Ninth International Conference on Quality Software*. IEEE. 2009, pp. 305–314.

[24] Tim Klinger et al. “An enterprise perspective on technical debt”. In: *Proceedings of the 2nd Workshop on managing technical debt*. 2011, pp. 35–38.

[25] Philippe Kruchten, Robert L Nord, and Ipek Ozkaya. “Technical debt: From metaphor to theory and practice”. In: *Ieee software* 29.6 (2012), pp. 18–21.

[26] Philippe Kruchten et al. “Technical debt: towards a crisper definition report on the 4th international workshop on managing technical debt”. In: *ACM SIGSOFT Software Engineering Notes* 38.5 (2013), pp. 51–54.

[27] Michele Lanza and Radu Marinescu. *Object-oriented metrics in practice: using software metrics to characterize, evaluate, and improve the design of object-oriented systems*. Springer Science & Business Media, 2007.

[28] Zhenmin Li et al. “CP-Miner: Finding copy-paste and related bugs in large-scale software code”. In: *IEEE Transactions on software Engineering* 32.3 (2006), pp. 176–192.

[29] Ben Liblit et al. “Bug isolation via remote program sampling”. In: *ACM Sigplan Notices* 38.5 (2003), pp. 141–154.

[30] Elvis Ligu et al. “Identification of refused bequest code smells”. In: *2013 IEEE International Conference on Software Maintenance*. IEEE. 2013, pp. 392–395.

[31] Erin Lim, Nitin Taksande, and Carolyn Seaman. “A balancing act: What software practitioners have to say about technical debt”. In: *IEEE software* 29.6 (2012), pp. 22–27.

[32] Everton da S Maldonado and Emad Shihab. “Detecting and quantifying different types of self-admitted technical debt”. In: *2015 IEEE 7th International Workshop on Managing Technical Debt (MTD)*. IEEE. 2015, pp. 9–15.

[33] Radu Marinescu. “Detection strategies: Metrics-based rules for detecting design flaws”. In: *20th IEEE International Conference on Software Maintenance, 2004. Proceedings.* IEEE. 2004, pp. 350–359.

[34] Antonio Martini, Terese Besker, and Jan Bosch. “Technical Debt tracking: Current state of practice: A survey and multiple case study in 15 large organizations”. In: *Science of Computer Programming* 163 (2018), pp. 42–61.

[35] Naouel Moha et al. “Decor: A method for the specification and detection of code and design smells”. In: *IEEE Transactions on Software Engineering* 36.1 (2009), pp. 20–36.

[36] Matthew James Munro. “Product metrics for automatic identification of" bad smell" design problems in java source-code”. In: *11th IEEE International Software Metrics Symposium (METRICS’05)*. IEEE. 2005, pp. 15–15.

[37] Madanlal Musuvathi et al. “CMC: A pragmatic approach to model checking real code”. In: *ACM SIGOPS Operating Systems Review* 36.SI (2002), pp. 75–88.

[38] Nicholas Nethercote and Julian Seward. “Valgrind: a framework for heavyweight dynamic binary instrumentation”. In: *ACM Sigplan notices* 42.6 (2007), pp. 89–100.

[39] Rocco Oliveto et al. “Numerical signatures of antipatterns: An approach based on b-splines”. In: *2010 14th European Conference on Software Maintenance and Reengineering*. IEEE. 2010, pp. 248–251.

[40] Fabio Palomba et al. “Mining version histories for detecting code smells”. In: *IEEE Transactions on Software Engineering* 41.5 (2014), pp. 462–489.

[41] Aniket Potdar and Emad Shihab. “An exploratory study on self-admitted technical debt”. In: *2014 IEEE International Conference on Software Maintenance and Evolution*. IEEE. 2014, pp. 91–100.

[42] D Rapu et al. “Using history information to improve design flaws detection”. In: *Eighth European Conference on Software Maintenance and Reengineering, 2004. CSMR 2004. Proceedings.* IEEE. 2004, pp. 223–232.

[43] A.J. Riel. *Object-oriented Design Heuristics*. Object oriented technology. Addison-Wesley Publishing Company, 1996. ISBN: 9780201633856. URL: https://books.google.it/books?id=oHkhAQAAIAAJ.

[44] Dilan Sahin et al. “Code-smell detection as a bilevel problem”. In: *ACM Transactions on Software Engineering and Methodology (TOSEM)* 24.1 (2014), pp. 1–44.

[45] Frank Simon, Frank Steinbruckner, and Claus Lewerentz. “Metrics based refactoring”. In: *Proceedings fifth european conference on software maintenance and reengineering*. IEEE. 2001, pp. 30–38.

[46] Rodrigo O Spı**********nola et al. “Investigating technical debt folklore: Shedding some light on technical debt opinion”. In: *2013 4th International Workshop on Managing Technical Debt (MTD)*. IEEE. 2013, pp. 1–7.

[47] Margaret-Anne Storey et al. “TODO or to bug: Exploring How Task Annotations Play a Role in the Work Practices of Software Developers”. In: *2008 ACM/IEEE 30th International Conference on Software Engineering*. IEEE. 2008, pp. 251–260.

[48] Edith Tom, Aybuke Aurum, and Richard Vidgen. “A consolidated understanding of technical debt”. In: (2012).

[49] Edith Tom, AybüKe Aurum, and Richard Vidgen. “An exploration of technical debt”. In: *Journal of Systems and Software* 86.6 (2013), pp. 1498–1516.

[50] Guilherme Travassos et al. “Detecting defects in object-oriented designs: using reading techniques to increase software quality”. In: *ACM Sigplan Notices* 34.10 (1999), pp. 47–56.

[51] Nikolaos Tsantalis and Alexander Chatzigeorgiou. “Identification of move method refactoring opportunities”. In: *IEEE Transactions on Software Engineering* 35.3 (2009), pp. 347–367.

[52] Eva Van Emden and Leon Moonen. “Java quality assurance by detecting code smells”. In: *Ninth Working Conference on Reverse Engineering, 2002. Proceedings.* IEEE. 2002, pp. 97–106.

[53] Alberto Villar, Santiago Matalonga, and Montevideo Uruguay. “Definiciones y Tendencia de Deuda Técnica: Un Mapeo Sistemático de la Literatura.” In: *CIbSE*. 2013, pp. 29–42.

[54] Bruce F Webster. *Pitfalls of object-oriented development*. M & T Books, 1995.

[55] Sultan Wehaibi, Emad Shihab, and Latifa 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.

[56] Nico Zazworka et al. “A case study on effectively identifying technical debt”. In: *Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering*. 2013, pp. 42–47.

[57] Pin Zhou et al. “AccMon: Automatically detecting memory-related bugs via program counter-based invariants”. In: *37th International Symposium on Microarchitecture (MICRO-37’04)*. IEEE. 2004, pp. 269–280.