Test Smells for Flaky Test Prediction

Pranay Reddy Juturu
pjuturu@stevens.edu
Stevens Institute of Technology
Hoboken, New Jersey, USA

Ashay Pable
apable@stevens.edu
Stevens Institute of Technology
Hoboken, New Jersey, USA

Diya Sanghvi
dsanghv1@stevens.edu
Stevens Institute of Technology
Hoboken, New Jersey, USA

ABSTRACT

xxx

CCS CONCEPTS

• **Software Engineering → AI for SE**;

KEYWORDS

software quality, mining software repositories

1 INTRODUCTION

Regression testing is a crucial step in software development, as it helps to ensure that software is delivered continuously with quality and minimal failures after changes to the production code. During this phase, developers rely on the test results to determine whether a program has a bug resulting from recent changes. However, the presence of flaky tests can make this evaluation unreliable. Flaky tests are a type of test with an intermittent behavior that alternates between passing and failing when executed in the same codebase, without any changes. This non-deterministic behavior frustrates developers, as it makes it challenging to identify and fix the root cause of the problem. Additionally, flaky tests are difficult to debug and can cause delays in the release cycles, halting the development process. Flaky tests can be a significant challenge in software development and identifying them is essential for ensuring the reliability and accuracy of test results. Dynamic and static approaches can be used to identify flaky tests, with each approach having its advantages and disadvantages. Dynamic approaches involve re-executing test cases a fixed number of times, which can be expensive and error-prone. It can also be difficult to determine how many executions are enough to identify flakiness accurately. Static approaches, on the other hand, do not require code re-execution and rely on machine learning methods to predict flakiness likelihood based on various features obtained from the code. Recently, an alternative approach for predicting flaky tests has been proposed based on identifying test smells. Test smells are associated with potential design problems in the test code, and their presence may impact software quality and lead to test flakiness. The alternative approach uses a set of predictors composed only of metrics collected statically, such as the size of the test case, the number of smells in the test code, and binary features related to the presence or absence of 19 test smells. The study found that this approach had better performance than the vocabulary-based model for cross-project prediction, achieving an F-measure of 0.83 with Random Forest.

GitHub is widely used for software development and version control. During development as the changes are made to the software, they are committed to the repository so that the team members can view it and can get to know the progress. In large software development, there could be several different kinds of changes that are committed, and sometimes viewing all the changes by all the members can be time-consuming.

Hence, we decided to make a feature that tells us what kind of changes are committed.

Suppose during the development of a mobile app one developer solves some bugs in the software and commits it. Then the commit should show a label with the bug written and the same goes for refactoring and adding features. In this way, without going in-depth, everyone will come to know about all the changes made.

The commits will be classified on the basis of how the code changes are reported using appropriate solutions.

- Static approaches do not need the code to be executed again. Models built using static features have many advantages and are less costly.
- Pinto et al. built a set of predictors considering that some patterns within the test code may be employed to identify flaky tests automatically.
- The authors came to the conclusion that the vocabulary-based strategy performs poorly when used across projects because it is context-sensitive and prone to overfitting.
- Considering this result, an alternative approach for flaky test prediction based on test smells is used. Test smells are associated with potential design problems in the test code.
- Test smells are a deviation from how tests should be created, arranged, and interacted with one another. That deviation can indicate issues with test design and negatively impact test performance.
- An open-source test smell detection tool, tsDetect is used. For each test case, this tool requires the identification of the corresponding production code to detect the test smells.

2 RELATED WORK

We have come across some papers that compare various methods to predict flaky tests. First, 'An Evaluation of Machine Learning Methods for Predicting Flaky Tests' utilizes Naive Bayes, Support Vector Machines (SVM), and Random Forests (RF) models and showed RF performed better when it comes to precision (> 90%) but provided very low recall (< 10%) as compared to NB (i.e., precision < 70% and recall >30%). They extracted test cases from multiple open-source projects and used the dataset. Second, 'FlakeFlagger: Predicting Flakiness Without Rerunning Tests' proposes a FlakeFlagger model that achieved a 60% average precision on 24 Java projects while the highest being 90% on spring-boot and going as low as 0% on others. Other methods compared in the paper include the

Study	Year	Method	Category	ML algo	Training Size	Result
An Evaluation of Machine Learning Methods for Predic	2020	ML	Identify flaky tests	Naive-Bayes, Support Vector Machines, and Random Forests	~2500	Precision: >90%, Recall: < 10%
FlakeFlagger: Predicting Flakiness Without Rerunning	2021	Mixed	Identify flaky tests	FlakeFlagger(own model), and a hybrid of vocabulary based approach and flakeflagger	21734	Precision 66%, Recall 74%
DeFlaker: Automatically Detecting Flaky Tests	2018	Traditional	Identify flaky tests	DeFlaker (own software) (used techniques like code coverage)	96 Java Projects	Recall: 95.5%
On the use of test smells for prediction of flaky tests	2021	ML	Identify flaky tests	Random Forest, Decision Tree, KNN, LR, LDA, Perceptron, SVM, Naive Bayes	2854	Precision: 83%, Recall: 83%

Figure 1: Related work review

DeFlake model, a vocabulary-based approach, and a combined approach between the flake-flagger and a vocabulary-based approach that yielded the highest average precision of 66%. One of the limitations in the paper appears to come from the fact that the authors had to supersample the data using SMOTE which some ML researchers are not very accepting of, Besides the 0% accuracy clearly shows that the model may not be useful as a general algorithm. Third, 'DeFlaker: Automatically Detecting Flaky Tests' introduces a software called deflaker, since this is an old paper, they do not use ML and try to use traditional software engineering techniques to solve the problem. However, they manage to achieve good accuracy on a given set of data, which might change if tested on a different dataset. They ran the software on 96 Java projects and claim to have a recall of 95%. This high accuracy gave DeFlaker a place in the task and is used as a base software for comparison to many other papers like the one above mentioned. They utilize a lot of techniques and use the result to analyze and create a list of all flaky tests like AST builder, code coverage recorder, test outcome monitor, etc. All these techniques lead to pretty impressive results on their tested projects. However, such traditional processes may fail on a project out of their set and may not be general solutions to the problem at hand. Fourth, is our base paper 'On the use of test smells for prediction of flaky tests', which utilizes test smells for the prediction of flaky tests with the help of ML models like Random Forest, Decision Tree, KNN, LR, LDA, Perceptron, SVM, Naive Bayes. They claim to have achieved 83% Precision and 83% Recall on a Random Forest algorithm which was the highest they could achieve. Their dataset is extracted data from existing open-source projects.

3 STUDY DESIGN

The objective of our study is to determine if test smells can be used to predict the existence of flaky tests.

To find the test smells in the code, the authors employed a program called tsDetect. Assertion roulette, conditional test logic, and mystery guest are just a few of the code patterns the program flags as being indicative of test smells. After that, the authors examined the test smells and took pertinent characteristics from the code. They used a feature selection technique known as mutual information to choose the most crucial attributes. The features with the highest mutual information score are chosen using this method, which assesses the dependencies between the features and the target variable (flaky tests). After feature selection, the authors used one-hot encoding to convert the data into a numerical format appropriate for machine learning techniques.

After studying the data we decided to drop empty columns as well as rows that contain some empty entries for categorical columns because these empty entries do not provide any information to the model. To increase the dataset's quality and guarantee that we are only using pertinent data to train our models, empty columns and rows should be removed. The columns 'Author Email', 'Author Name', 'Committer Email', 'Committer Name', 'Commit Message', 'Commit SHA', 'Filepath', and 'Line' are removed. These columns either have no bearing on our goal or include categorical items with blank. Once this was done we decided to drop columns with zero relevance to the target column Klass which is Binary making it a binary classification. The columns that have zero relevance to the target columns are 'App', 'Build time in minutes', 'TimeStamp', and 'Version'. We have used random forest, decision trees, Naive Bayes, KNN, LDA, and Logistic Regression models. Grid search with cross-validation is used to adjust the hyperparameters for each model and the model with the best validation set performance is chosen. The random forest provides the best results out of the tested models. We use k-fold cross-validation, where we divide the data into k subsets and use each subset as a validation set in turn, to guarantee that our results are not overfitting the training data. This procedure is repeated k times, and the average performance over all iterations is reported. The evaluation metrics we used are Accuracy, Precision, Recall, F1 score, Matthews correlation coefficient, and Area Under the Curve. The performance of each of the models is expressed and compared in the further sections of this paper. From studying the data we also concluded that the dataset only identifies if a test is flaky with a binary representation without the inclusion of any magnitude to provide the extent of flakiness, which might produce some limitations and could be scope for future experimentation. For embedding of text inputs we are using CountVectorizer and TfidfVectorizer from the sklearn.feature_extraction.text class.

Our dataset consists of data coming from a variety of famous GitHub projects and extracting information, like commit messages, analyzing tests, LOC, etc.

Our workflow will consist of information extraction from Github repositories, specifically the commit messages, and analyzing tests to identify test smells. This information is then used to feed data into our classifier once the text from the commit messages is tokenized and embedded. This is then fed into the trained model to classify the data point into flaky or not flaky.

4 EXPERIMENTAL RESULTS

RQ1 – How accurately can we predict test flakiness based on test smells in the test cases?

By first training and then evaluating the classifiers, the prediction model was developed. Every classifier had a reasonable performance archive except the Naive Bayes classifier with an accuracy of 65%. Random Forest classifier had higher accuracy and precision of 83%. The collected results demonstrate that test smell-based models, with precision values ranging from 75% to 83%, perform reasonably well in predicting test flakiness. The trained classifiers were tested using the flaky tests included in the idFlakies dataset to confirm the model's performance in the cross-project environment.

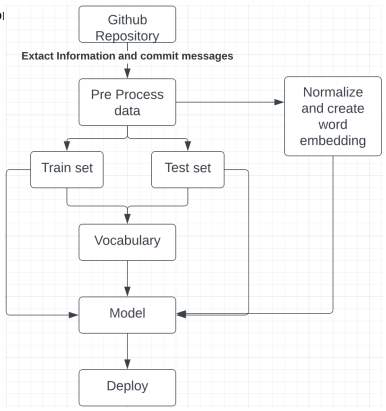

Figure 2: Flowchart

classifier	acc	precision	recall	f1	mcc	auc	TP	FN
randomForest	0.836331	0.836912	0.836331	0.836402	0.672862	0.905892	NaN	NaN
decisionTree	0.836331	0.836361	0.836331	0.836343	0.672106	0.863780	NaN	NaN
naiveBayes	0.652878	0.738687	0.652878	0.610184	0.368766	0.783951	NaN	NaN
svc	0.751799	0.752165	0.751799	0.751188	0.502338	0.829971	NaN	NaN
knn	0.812950	0.812908	0.812950	0.812918	0.625101	0.812445	NaN	NaN
logisticRegression	0.793165	0.793934	0.793165	0.792616	0.585713	0.873632	NaN	NaN
perceptron	0.776978	0.777729	0.776978	0.776340	0.553177	0.864558	NaN	NaN
lda	0.782374	0.783548	0.782374	0.781608	0.564326	0.861758	NaN	NaN

Figure 3: Training result parameters

In the intra-project scenario, the performance of all the classifiers is dropped. Logistic regression attained the highest score. It accurately identified 26 out of 9 flaky tests. In an inter-project scenario, the classifier's performance declined more sharply. With recall values ranging from 48% to 55%, the classifiers do not differ significantly from one another, with Naive-Bayes reaching a value of 14% by accurately identifying 17 out of 103 flaky tests. The results collected demonstrate that the smells can be utilized to predict flakiness. But, in the inter-project scenario, performance suffers significantly. The findings demonstrate that the performance of the smell-based models is equivalent to, and occasionally even superior to vocabulary model. The classifier's performance ranges from 11% to 55%. This led to the conclusion that smells are reliable indicators of flakiness.

RQ2 – Which attributes are the most strongly associated with test flakiness prediction?

To identify associations between attributes and flakiness, we used sklearn.feature_selection.mutual_info_classify method of scikit-learn that allows us to select features and in the future experiment by eliminating the least relevant features to optimize the model. The function calculates Mutual Information, which is the measure of the mutual dependence between two random variables. We use MI over Correlation between each attribute as it is more versatile and can capture non-linear relationships, while correlation is limited to linear relationships. **The equation for MI between a feature X and a target variable y is:**

classifier	acc	precision	recall	f1	mcc	auc	TP	FN
randomForest	0.685714	NaN	0.685714	NaN	NaN	NaN	24.0	11.0
decisionTree	0.657143	NaN	0.657143	NaN	NaN	NaN	23.0	12.0
naiveBayes	0.571429	NaN	0.571429	NaN	NaN	NaN	20.0	15.0
svc	0.657143	NaN	0.657143	NaN	NaN	NaN	23.0	12.0
knn	0.514286	NaN	0.514286	NaN	NaN	NaN	18.0	17.0
logisticRegression	0.742857	NaN	0.742857	NaN	NaN	NaN	26.0	9.0
perceptron	0.714286	NaN	0.714286	NaN	NaN	NaN	25.0	10.0
lda	0.657143	NaN	0.657143	NaN	NaN	NaN	23.0	12.0

```
tatistics.process == 'traditional') & (classifierStatisti
```

classifier	acc	precision	recall	f1	mcc	auc	TP	FN
randomForest	0.541667	NaN	0.541667	NaN	NaN	NaN	65.0	55.0
decisionTree	0.550000	NaN	0.550000	NaN	NaN	NaN	66.0	54.0
naiveBayes	0.141667	NaN	0.141667	NaN	NaN	NaN	17.0	103.0
svc	0.550000	NaN	0.550000	NaN	NaN	NaN	66.0	54.0
knn	0.508333	NaN	0.508333	NaN	NaN	NaN	61.0	59.0
logisticRegression	0.475000	NaN	0.475000	NaN	NaN	NaN	57.0	63.0
perceptron	0.475000	NaN	0.475000	NaN	NaN	NaN	57.0	63.0
lda	0.475000	NaN	0.475000	NaN	NaN	NaN	57.0	63.0

Figure 4: Vocabulary-based approach

	position	token	information_gain	total_ocurences	total_flaky_occurences	total_nonflaky_occurences
0	0	loc	2.544574e-01	2777	1377	1400
1	1	assertionRoulette	8.323976e-02	1389	968	421
2	2	smellsCount	2.705301e-02	2655	1356	1299
3	3	sleepyTest	1.948161e-02	112	105	7
4	4	generalFixture	1.600704e-02	267	61	206
5	5	duplicateAssert	1.552284e-02	376	269	107
6	6	constructorInitialization	1.094276e-02	68	63	5
7	7	printStatement	1.056366e-02	58	55	3
8	8	sensitiveEquality	5.852377e-03	129	95	34
9	9	lazyTest	5.490260e-03	1788	817	971
10	10	resourceOptimism	4.252596e-03	75	17	58
11	11	conditionalTestLogic	4.217346e-03	356	219	137
12	12	unknownTest	2.110196e-03	544	234	310
13	13	verboseTest	1.771423e-03	7	7	0
14	14	magicNumberTest	1.108724e-03	411	227	184
15	15	mysteryGuest	5.519103e-04	124	71	53
16	16	eagerTest	2.573414e-04	970	496	474
17	17	redundantAssertion	9.909595e-08	8	4	4
18	18	defaultTest	0.000000e+00	0	0	0
19	19	emptyTest	0.000000e+00	0	0	0
20	20	ignoredTest	0.000000e+00	0	0	0

Figure 5: Features

$$MI(X, y) = \sum p(x, y) * log(p(x, y)/(p(x) * p(y)))$$

While PCA Is a great Dimensionality tool, its usage as a feature selection is controversial due to the information loss it causes, as well as the reduced flexibility to manually select attributes to eliminate. We identified that the most relevant feature in the dataset was loc (lines of Code) followed by assertion roulette (test smell), while the least relevant feature turned out to be redundant assertion (which is a test smell). Whereas, IgnoredTest, emptyTest, and defaultTest have no relation with an absolute zero score, thus indicating the need to drop the three columns. The dataset only identifies if a test is flaky with a binary representation without the inclusion of any magnitude to provide the extent of flakiness, which might produce some limitations and could be scope for future

classifier	acc	precision	recall	f1	mcc	auc	VP	FN
randomForest	0.971223	0.971579	0.971223	0.971199	0.942660	0.989448	NaN	NaN
decisionTree	0.928058	0.928324	0.928058	0.928083	0.856181	0.928241	NaN	NaN
naiveBayes	0.951439	0.951548	0.951439	0.951416	0.902767	0.950959	NaN	NaN
smo	0.967626	0.967975	0.967626	0.967599	0.935440	0.992377	NaN	NaN
knn	0.929856	0.930938	0.929856	0.929889	0.860675	0.930736	NaN	NaN
logisticRegression	0.967626	0.967975	0.967626	0.967599	0.935440	0.994179	NaN	NaN
perceptron	0.965827	0.965952	0.965827	0.965811	0.931625	0.991794	NaN	NaN
lda	0.866906	0.871221	0.866906	0.866858	0.738208	0.876601	NaN	NaN

Figure 6: Vocabulary-based approach

experimentation.

RQ3 – How does the test smell-based approach compare with the existing vocabulary-based approach?
In the vocabulary-based approach, the values of VP (True Positive) and FN (False Negative) for the classification metrics are NaN (Not a Number). This is due to the dataset used for evaluation in the vocabulary-based approach only containing flaky tests and excluding non-flaky tests. As a result, this dataset lacks True Negatives (TN) and False Negatives (FN), making it impossible to calculate VP and FN.

Precision, recall, and F1-score for the vocabulary-based technique cannot be determined in the absence of TN and FN data. The accuracy and AUC numbers for this technique are the only ones that the authors have reported. We trained the classifiers with the training and testing dataset using the vocabulary-based approach. The vocabulary-based strategy performs better than the smell-based approach: the best F1 score for vocabulary-based models is 97% (Random Forest), while the score for the smell-based approach is 83%. (Random Forest). The disparity is greater when MCC is analyzed. The best outcome for the smell-based technique was 0.66, and the best result for the vocabulary-based approach was 0.94. This score takes into consideration true and false positives, as well as negatives. The cross-project validation results, however, demonstrate that the test smell-based strategy yields superior outcomes. The test smell-based strategy yielded 74% of recall (LR) in the intra-project context, while the vocabulary-based approach only managed to reach 57%. (KNN).

Using the training and testing datasets, the performance of the vocabulary-based models is superior to that of the test smell-based models. Yet, the smell-based approach achieves noticeably higher outcomes in the intra-project and inter-project contexts in the cross-project validation scenario.

5 THREATS TO VALIDITY

Construct Validity: The degree to which a concept is operationalized (i.e., how it is measured) accurately reflects the intended construct is known as its construct validity. The relation between test smells and the flaky test is the construct that interests this study's researchers. To operationalize this concept, a variety of test smells, including the assertion roulette and conditional test logic were considered. There is a potential threat to the identification of the flaky tests.
The most widely used metrics in the machine learning (ML) community were adopted to evaluate the classifiers in order to reduce this threat, which can aid in improving the study's generalizability and reliability. However, the authors utilized a program called tsDetect to find test smells in the code during the pre-processing stage of the test code. The production class, a vital component of the codebase that the tests are testing, has occasionally been missed by tsDetect. As a result, there were instances where test smells could not be extracted from the code, which could jeopardize the study's findings.

Internal Validity: The degree to which a research study accurately ascertains the link between the independent variable and the dependent variable is referred to as internal validity. The existence of confounding variables, which are variables that can impact the relationship between the independent and dependent variables, is one factor that could endanger internal validity. Confounding variables in this study may include things like the size and complexity of the codebase, the level of experience of the developers, and the particular programming language employed. The authors employed statistical techniques like logistic regression and decision trees to account for the effects of these confounding variables and identify the association between test smells and flaky tests in order to address this possible danger.

External Validity: The generalizability of research findings may be constrained by factors that pose a threat to external validity. Four open-source projects were utilized by the authors to gather data for their study, although other software projects in different domains or with various characteristics might not be comparable to these projects. The results may not apply to projects created in other programming languages, for instance, since all of the projects utilized in the study were written in Java. The study's projects were small to medium-sized, and the authors also pointed out that they might not accurately represent the features of larger software projects because of their size. The authors emphasized the limits of their study and suggested that future research should investigate whether or not their findings apply to other software projects and domains in order to address this possible threat.

Conclusion Validity: The degree to which the inferences made from the data accurately reflect the underlying relationships between the variables under investigation is referred to as conclusion validity. The authors of the research thoroughly analyzed their data and assessed how well their classifier models performed using the relevant statistical methods. They also talked about some of the study's possible drawbacks, namely the use of a single dataset and the scant number of projects examined. They did not, however, point out any particular problems or worries regarding the inferences made from the data.

6 CONCLUSION

According to the study, a number of test smells, such as assertion roulette and conditional test logic, were strongly linked to flaky tests. Based on these smells, the classifiers created using machine learning approaches were highly accurate in

predicting flaky tests. The accuracy of the test smell detection tool and the small size and scope of the dataset utilized in the study are two major risks to the validity of the study, the authors point out. The authors contend that additional study is required to both explore additional variables that can affect test flakiness in software testing and to confirm the efficacy of test smells as predictors of flaky tests. Overall, the study offers insightful information about the use of test smells as a viable method for enhancing the accuracy and effectiveness of software testing.

7 FUTURE WORK

A promising solution to the problem of finding and anticipating flaky tests in software testing is presented by the authors. More study in this area is still needed. The current study focused on a particular set of test smells, but there might be additional clues that might be used to spot and foretell problematic testing. To increase the accuracy of shaky test prediction models, future studies could include other test smells or other elements, such as code complexity or ambient elements. Despite the study's encouraging predictions for flaky tests, it is crucial to comprehend how employing test smells impacts the entire software testing process. Future studies could look into how using test smells affects testing effectiveness, efficiency, and overall software product quality. The tsDetect tool was used in the investigation to find test smells in the codebase. Nonetheless, this tool's accuracy and dependability could yet be enhanced. Future studies could concentrate on creating more advanced tools or enhancing currently available technologies more correctly and effectively detect test scents. The study does not address the fundamental reasons for flakiness, even if it offers a means to recognize and anticipate flaky tests. Future studies could look at the underlying factors that contribute to flaky testing and devise methods to stop them from happening in the first place.

REFERENCES

[1] Azeem Ahmad. 2020. An evaluation of machine learning methods for predicting flaky tests. In *27th Asia-Pacific Software Engineering Conference (APSEC 2020) Singapore (virtual), December 1, 2020.*

[2] Abdulrahman Alshammari, Christopher Morris, Michael Hilton, and Jonathan Bell. 2021. FlakeFlagger: Predicting Flakiness Without Rerunning Tests. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. 1572–1584. https://doi.org/10.1109/ICSE43902.2021.00140 ISSN: 1558-1225.

[3] Jonathan Bell, Owolabi Legunsen, Michael Hilton, Lamyaa Eloussi, Tifany Yung, and Darko Marinov. 2018. D e F laker: automatically detecting flaky tests. In *Proceedings of the 40th International Conference on Software Engineering*. ACM, Gothenburg Sweden, 433–444. https://doi.org/10.1145/3180155.3180164

[4] B. H. P. Camara, M. A. G. Silva, A. T. Endo, and S. R. Vergilio. 2021. On the use of test smells for prediction of flaky tests. In *Brazilian Symposium on Systematic and Automated Software Testing*. 46–54. https://doi.org/10.1145/3482909.3482916 arXiv:2108.11781 [cs].

[4] [2] [3] [1]