MASARYK UNIVERSITY

FACULTY OF INFORMATICS

Source Code Quality Impact on Pull Requests Acceptance

Master's Thesis

ONDŘEJ KUHEJDA

Advisor: Assistant professor Bruno Rossi

Department of Computer Systems and Communications

Brno, Spring 2022

MUNI
FI

Declaration

Hereby I declare that this paper is my original authorial work, which I have worked out on my own. All sources, references, and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Ondřej Kuhejda

Advisor: Assistant professor Bruno Rossi

Acknowledgements

I would like to thank my supervisor, Bruno Rossi, for his guidance throughout the whole process.

Computational resources were supplied by the project "e-Infrastruktura CZ" supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Abstract

TODO

Keywords

code quality, pull request, static code analysis

Contents

1 Introduction EXPORT

Does code quality influence the acceptance of pull requests? Although many project maintainers view the code quality as the most important factor regarding the pull request acceptance [1], a recent study shows that presence of quality flaws in the code does not influence the acceptance or rejection of pull requests [2].

To the best of my knowledge, that is the only study that investigated if quality issues affects the acceptance of pull requests. Lenarduzzi et al. analyzed 28 well-known Java projects and applied several statistical techniques to find the relation between code quality and pull request acceptance. The quality of the pull requests was evaluated using the open-source tool called PMD. This tool is able to perform static analysis of the code and detect common programming flaws, such as code-smells, anti-patters, or code-style violations. Traditional statistical techniques did not find any connection between the code quality and pull request acceptance. Because of that, they trained several machine learning models to predict the acceptance based on the quality issues found in the code. However, machine learning models yielded similar results as traditional techniques.

Unfortunately, the study performed by Lenarduzzi et al. analyzed only projects written in Java. Moreover, 22 out of 28 analyzed projects were from the Apache Software Foundation. This is the major threat to generalizability of their findings. To address these shortcomings, I analyzed one hundred projects from five different programming languages — Python, Java, Kotlin, Haskell, and C/C++ (twenty projects per language). For each programming language, different tool for static analysis was used to evaluate the code quality. I applied similar techniques as Lenarduzzi et al. to analyze the relation between the code quality and pull request acceptance. Furthermore, the regression models were created to predict the time required to close a pull request using the code quality. Subsequently, the retrieved results were compared between individual languages. Specially, the research questions I am trying to answer are following:

RQ_1 Which code issues are typically introduced by the pull requests?

RQ$_2$ Are there some particular issues/code smells that affect the pull request acceptance?

RQ$_3$ Is there a relationship between the source code quality and the pull request acceptance?

RQ$_4$ Does code quality influence the time it takes to close a pull request?

RQ$_5$ Is code quality impact higher in projects that are using some particular programming language?

This thesis first discuss the code quality and its link to pull-based development model. After that, the various factors that influence the pull request acceptance are mentioned. Next part of the thesis is dedicated to the research design. At first, I state how the data about the project was retrieved and which methods were used to evaluate the code quality of individual pull requests. Then, the used statistical methods are being introduced. Furthermore, the retrieved results are evaluated separately for each programming language. Towards the end, results are compared between languages and potential threats to validity are discussed. Finally, I summarize my findings and compare them to the findings of Lenarduzzi et al. Moreover, the potential extensions of my work are proposed.

Bibliography

1. GOUSIOS, G.; ZAIDMAN, A.; STOREY, M.; DEURSEN, A. Work Practices and Challenges in Pull-Based Development: The Integrator's Perspective. In: *International Conference on Software Engineering*. IEEE, 2015, vol. 1, pp. 358–368. Available from DOI: 10.1109/ICSE.2015.55.
2. LENARDUZZI, V.; NIKKOLA, V.; SAARIMÄKI, N.; TAIBI, D. Does code quality affect pull request acceptance? An empirical study. *Journal of Systems and Software*. 2021, vol. 171, pp. 110806. Available from DOI: 10.1016/j.jss.2020.110806.