Automated Analysis of Commenting Styles and Documentation Practices: A Data-Driven Approach to Software Quality and Maintainability

K.A.H.P.Sathyangani
(Reg. No.:MS24063296)

A THESIS
SUBMITTED TO
SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

December 2025

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

5th January 2026

Dr. Junius Anjana

Approved for MSc. Research Project:

MSc in IT Programme Co-ordinator, SLIIT

Approved for MSc:

Head of Graduate Studies, FoC, SLIIT

DECLARATION

This is to certify that the work is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources). The work has not previously been submitted in any form to the Sri Lanka Institute of Information Technology or to any other institution for assessment for any other purpose.

Sign: ……Hansini.....

K.A.H.P.Sathyangani

Date: ...2025/12/14...

ABSTRACT

K.A.H.P.Sathyangani
MSc. in Information Technology
Supervisor: Dr. Junius Anjana
December 2025

Software maintainability is strongly influenced by the quality of code comments, which guide developers in understanding system functionality and behaviour. Poorly written, missing, or ambiguous comments reduce productivity and increase the cost of maintenance. The current study introduces an automated, data-driven approach to evaluating comment quality in Java projects. The proposed solution, implemented as a Java-based tool named Comment Quality Analyser, automatically scans source files, extracts comments, and evaluates them using four quality dimensions: grammatical correctness, readability, understandability, and meaningfulness. The tool integrates LanguageTool for grammatical analysis, the Flesch Reading Ease metric for readability, heuristic rules for understandability, and a Jaccard-similarity-based algorithm for measuring semantic alignment between comments and code identifiers. The results are presented through JSON reports and an interactive HTML dashboard that visualises the quality distribution across files. Real-world validation was conducted using the Apache Commons IO open-source repository, containing over 100 comments. Experimental results indicate that the system provides consistent scoring with an average accuracy of 86 % when compared with manual reviews. The proposed framework contributes to improving software documentation practices and offers a foundation for further research integrating Natural Language Processing (NLP) and Machine Learning (ML) to enhance software maintainability analysis.

ACKNOWLEDGEMENT

First and foremost, I express my heartfelt gratitude to my supervisor, Dr Junius Anjana, for his continuous guidance, patience, and constructive feedback throughout this research. His expertise and motivation were invaluable in shaping the direction of this study.

I also extend my appreciation to the lecturers and staff of the Sri Lanka Institute of Information Technology (SLIIT) for their academic support and resources provided during my MSc programme. Special thanks are due to my colleagues at Sysco Labs for their encouragement, technical discussions, and support during implementation and testing. Finally, I dedicate this work to my family and friends for their unwavering love, understanding, and belief in my potential throughout this journey.

TABLE OF CONTENTS

List of Figures

List of Tables