



Threat Detection Based on Log Analysis for Automating Security Information and Event Management (SIEM) Functionality

C.A. Hewagama
MS22909282
M.Sc. in CS

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.

Dr. Dharshana Kasthurirathna

Approved for MSc. Research Project:



MSc. 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is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources).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: Amal

C.A. Hewagama

Date: 03.11.2025

ABSTRACT

Threat detection based on log analysis for automating Security information and event management (SIEM) functionality

Chathura Amal Hewagama

MSc. in Cybersecurity

Supervisor: Dr. Dharshana Kasthurirathna

December 2025

The reliance of modern organizations on information systems continues to increase, making these infrastructures frequent targets for malicious activity. System logs represent a primary source of forensic evidence, yet the volume generated by large-scale environments renders manual inspection infeasible. Security Information and Event Management (SIEM) platforms automate log collection and correlation but remain limited in detecting evolving or previously unseen threats. As a result, there is growing interest in augmenting SIEM functionality through Natural Language Processing (NLP) and machine learning.

This study investigates lightweight Transformer models as candidates for log-based anomaly detection in SIEM contexts. Two compressed architectures, DistilBERT and TinyBERT, are evaluated under parameter-efficient adaptation strategies: frozen encoders with linear classification heads and Low-Rank Adaptation (LoRA). Log templates are extracted using the Drain algorithm to normalize unstructured log data, and experiments are conducted on two benchmark datasets, BGL and HDFS. A classical baseline using TF-IDF with Logistic Regression is also included for comparison. Evaluation covers both detection metrics (precision, recall, F1-score, PR-AUC, ROC-AUC) and efficiency metrics (latency, throughput, memory usage).

The scope of this research is limited to training and evaluation rather than live SIEM deployment. Its contribution lies in assessing the trade-offs between detection accuracy and computational efficiency across lightweight adaptation strategies, providing guidance on configurations most viable for integration into real-time SIEM pipelines.

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to Sri Lanka Institute of Information Technology for providing me with the opportunity and resources to pursue my studies. I am especially grateful to Dr. Anuradha Jayakody, Dr. Harinda Fernando, and Dr. Prasanna Haddela for their invaluable guidance and expertise throughout this journey.

A special thank you to my supervisor, Dr. Dharshana Kasthurirathna, for his continuous support, insightful feedback, and encouragement, which were crucial to the completion of this thesis.

Lastly, I want to extend my heartfelt thanks to my parents for their unwavering love, patience, and support, without which this accomplishment would not have been possible.

TABLE OF CONTENTS

DECLARATION	ii
ABSTRACT.....	iii
ACKNOWLEDGEMENT.....	iv
TABLE OF CONTENTS.....	v
List of Figures	viii
List of Tables	ix
Introduction	1
1.1 Research Problem	2
1.2 Research Questions	3
1.3 Research Objectives.....	4
Chapter 2 Literature review	5
2.1 Introduction to Logs and SIEM.....	5
2.2 Limitations of Traditional SIEM.....	6
2.3 AI/ML Approach for SIEM	7
2.4 Log Parsing Methods.....	9
2.5 Transformers for Log Anomaly Detection.....	11
2.6 Lightweight Transformers and Adaptation	13
2.7 Challenges and Gaps in Current Research	15
Chapter 3 Methodology	17
3.1 Research Design	17
3.2 Data Collection	18
3.3 Data Preprocessing	19
3.3.1 Exploratory Data Analysis	19
3.3.2 Tokenization.....	20
3.3.3 Template Extraction with Drain:	21
3.4 Model Selection: Lightweight Transformers.....	21
3.4.1 Adaptation Strategies	22
3.4.1 Traditional Baselines.....	23
3.5 Training and Validation	23
3.6 Evaluation Metrics	23
3.7 Experimental Design	24
3.8 Output and Reproducibility.....	24
3.9 <i>Scope and Limitations</i>	25
Chapter 4 Experimental Setup	26

4.1 Hardware and Software Environment	26
4.2 Dataset Preparation	27
4.3 Training Configuration and Reproducibility	28
4.4 Candidate Models and Adaptation Strategies	31
4.4.1 DistilBERT	31
4.4.2 TinyBERT	31
4.4.3 Adaptation Strategies	31
4.4.4 Traditional Baselines	32
4.5 Architectural Explanations	32
4.5.1 End-to-End Pipeline	32
4.5.2 Frozen Encoder with Linear Classification Head	33
4.5.3 Low-Rank Adaptation (LoRA)	33
4.6 Relevance to SIEM Contexts	33
Chapter 5 Results and Discussion	35
5.1 Overview	35
5.2 Experimental Results	36
5.2.1 Baseline Performance	36
5.2.2 DistilBERT Results	38
5.2.3 TinyBERT Results	41
5.2.4 Ablation Study (Template-Text Variant)	45
5.2.5 Comparative Summary	47
5.3 Result Analysis and Discussion	52
5.3.1 Accuracy and Detection Performance	52
5.3.2 Leakage Check Verification	54
5.3.3 Efficiency and Deployability	55
5.3.4 Generalization and and Template Dependence	57
5.3.5 Interpretation for SIEM Automation	59
5.3.6 Limitations	61
5.4 Chapter Summary	62
Chapter 6 Conclusion	64
6.1 Overview	64
6.2 Summary of Findings	64
6.3 Research Contributions	66
6.4 Limitations	67
6.5 Future Works	68
6.6 Concluding Remarks	70

References	72
Appendix	75

List of Figures

Figure 1 Theoretical framework of the research	17
Figure 2 Low-Rank Adaptation (LoRA) formulation	33

List of Tables

Table 1 Token length statistics for BGL and HDFS datasets.....	21
Table 2 System Specifications for Model Training and Evaluation	27
Table 3 Hyperparameters Used in Standard Runs.....	30
Table 4 Baseline results summary	38
Table 5 DistilBERT Performance on BGL (Frozen vs LoRA)	40
Table 6 Distilbert results (↑ indicates higher = better, ↓ indicates lower = better).....	40
Table 7 TinyBERT results on the BGL dataset (↑ = higher is better, ↓ = lower is better)	41
Table 8 TinyBERT results (↑ = higher is better, ↓ = lower is better).....	43
Table 9 Quantitative results for ablation test with Distilbert + LoRA(↑ = higher is better; ↓ = lower is better)	46
Table 10 Leakage-check status and interpretation.....	55