UNIVERSITAS
GADJAH MADA
**Deteksi Anomali Log Sistem Menggunakan DistilBERT-Based Variational Autoencoder Dengan Pendekatan
Semi-Supervised Learning**
Mohammad Fajar Maulid, Dr.techn. Guntur Budi Herwanto, S.Kom., M.Cs.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

DETEKSI ANOMALI LOG SISTEM MENGGUNAKAN DISTILBERT-*BASED* VARIATIONAL AUTOENCODER DENGAN PENDEKATAN *SEMI-SUPERVISED LEARNING*

Oleh

Mohammad Fajar Maulid
22/498024/PA/21477

Log sistem dan aplikasi berperan penting dalam pemantauan operasional dan deteksi serangan siber. Namun, besarnya volume data, keragaman format, serta kelangkaan data berlabel menjadikan pendekatan deteksi anomali *supervised* kurang praktis, terutama dalam mendeteksi anomali yang baru atau jarang terjadi. Penelitian ini mengusulkan kerangka kerja *semi-supervised learning* menggunakan *DistilBERT-based Variational Autoencoder* (DistilBERT-VAE). Dalam arsitektur ini, DistilBERT digunakan untuk mengekstraksi representasi semantik dan dependensi kontekstual sekuens log secara ringan dan efisien, sedangkan VAE bertugas memodelkan distribusi ruang laten dari perilaku log normal. Penentuan batas anomali (*threshold*) dilakukan secara dinamis dengan mengevaluasi nilai *reconstruction error* menggunakan metode *Extreme Value Theory* (EVT). Berdasarkan pengujian pada dataset sistem BGL dan HDFS, model usulan terbukti memberikan performa pendeteksian dan efisiensi yang sangat baik. Pada dataset BGL, pemadatan dimensi dari BERT ke DistilBERT berhasil mempertahankan keseimbangan deteksi dengan nilai F1-Score mencapai 0,9820, meskipun uji hipotesis menunjukkan tidak ada perbedaan yang signifikan. Pada dataset HDFS, DistilBERT-VAE terbukti secara nyata mengungguli arsitektur dasar BERT-VAE dengan F1-Score sebesar 0,9791. Dari sisi efisiensi operasional, model usulan mampu memangkas *floating-point operations* (FLOPs) hingga 50% serta mempecepat waktu komputasi ekstraksi fitur hingga 2,28 kali lipat pada data BGL dan 1,90 kali lipat pada HDFS. Hasil penelitian ini mengonfirmasi bahwa integrasi DistilBERT dan VAE menawarkan solusi sistem deteksi anomali log yang andal, akurat, tangguh terhadap keterbatasan label data, dan sangat efisien untuk diterapkan pada skala operasional besar.

Kata-kata kunci : Deteksi anomali log, Log sistem, *DistilBERT*, *Variational Autoencoder*, *Transformer*, *Semi-Supervised Learning*.

UNIVERSITAS
GADJAH MADA
**Deteksi Anomali Log Sistem Menggunakan DistilBERT-Based Variational Autoencoder Dengan Pendekatan
Semi-Supervised Learning**
Mohammad Fajar Maulid, Dr.techn. Guntur Budi Herwanto, S.Kom., M.Cs.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

ANOMALY DETECTION IN SYSTEM LOGS USING A DISTILBERT-BASED VARIATIONAL AUTOENCODER WITH A SEMI-SUPERVISED LEARNING APPROACH

By

Mohammad Fajar Maulid
22/498024/PA/21477

System and application logs play a crucial role in operational monitoring and cyberattack detection. However, the massive volume of data, diverse formats, and scarcity of labeled data make supervised anomaly detection approaches less practical, especially in detecting novel or rare anomalies. This study proposes a semi-supervised learning framework using a DistilBERT-based Variational Autoencoder (DistilBERT-VAE). In this architecture, DistilBERT is utilized to extract semantic representations and contextual dependencies of log sequences in a lightweight and efficient manner, while the VAE is responsible for modeling the latent space distribution of normal log behavior. The determination of the anomaly threshold is performed dynamically by evaluating the reconstruction error using the Extreme Value Theory (EVT) method. Based on evaluations using the BGL and HDFS system datasets, the proposed model is proven to deliver excellent detection performance and efficiency. On the BGL dataset, the dimensional reduction from BERT to DistilBERT successfully maintains detection balance with an F1-Score reaching 0.9820, although hypothesis testing indicates no significant difference. On the HDFS dataset, DistilBERT-VAE is proven to significantly outperform the baseline BERT-VAE architecture with an F1-Score of 0.9791. In terms of operational efficiency, the proposed model is able to cut floating-point operations (FLOPs) by up to 50% and accelerate feature extraction computational time by 2.28 times on the BGL data and 1.90 times on the HDFS data. The results of this study confirm that the integration of DistilBERT and VAE offers a log anomaly detection system solution that is reliable, accurate, resilient to data label scarcity, and highly efficient for large-scale operational deployment.

Keywords : Log anomaly detection, System log, DistilBERT, Variational Autoencoder, Transformer, Semi-Supervised Learning.