INTISARI

Analisis dan Implementasi Deteksi Log Hadoop *Distributed File System* Berbasis Transformer FP16

Ridwan Akmal
21/480411/SV/19632

Dalam era *big data* seperti saat ini, sistem manajemen data skala besar seperti Hadoop *Distributed File System* (HDFS) menjadi pemegang peran penting dalam pengolahan dan penyimpanan data dalam jumlah besar. Namun, tantangan utama dari sistem ini adalah mendeteksi anomali secara efektif untuk mencegah potensi gangguan operasional dan ancaman keamanan. Penelitian ini mengusulkan metode berbasis transformer untuk mendeteksi anomali pada log HDFS dengan membandingkan performa tiga model *pretrained*, yaitu ELECTRA, MiniLM, dan DistilBERT, yang dioptimisasikan dengan pendekatan *mixed precision training* FP16. Eksperimen dilakukan pada dataset Loghub dengan melakukan beberapa tahapan seperti pemetaan *Event Id* ke *log message*, *sampling*, tokenisasi, pelatihan, evaluasi model, serta integrasi model dan web. Hasil eksperimen menunjukkan bahwa ELECTRA menjadi model terbaik di berbagai metrik seperti *loss*, *accuracy*, *precision, recall* serta F1-*score* tertinggi yaitu 0.99860. Model ini juga memiliki *load time* tercepat (0.2286 detik) dan *inference time* tercepat (0.0241 detik). Berdasarkan penelitian ini model berbasis transformer khususnya ELECTRA dapat diimplementasikan secara efektif untuk deteksi anomali pada log HDFS.

Kata kunci: Deteksi anomali, *Hadoop Distributed File System*, ELECTRA, MiniLM, DistilBERT

ABSTRACT

Analysis and Implementation of Hadoop Distributed File System Log Detection Based on Transformer FP16

Ridwan Akmal
21/480411/SV/19632

In the era of big data, large-scale data management systems such as the Hadoop Distributed File System (HDFS) play a crucial role in processing and storing vast amounts of data. However, a major challenge in these systems is effectively detecting anomalies to prevent operational disruptions and security threats. This study proposes a transformer-based approach for anomaly detection in HDFS logs by comparing the performance of three pretrained models: ELECTRA, MiniLM, DistilBERT and optimized using the mixed precision training FP16 approach. Experiments were conducted on the Loghub dataset, involving several stages such as mapping Event IDs to log messages, sampling, tokenization, model training, model evaluation, and integration with a web-based system. The experimental results show that ELECTRA outperforms the other models across various metrics, achieving the highest scores in loss, accuracy, precision, recall, and an F1-score of 0.99860. Additionally, ELECTRA demonstrated the fastest load time (0.2286 seconds) and inference time (0.0241 seconds). Based on this study, transformer-based models, particularly ELECTRA, can be effectively implemented for anomaly detection in HDFS logs.

Keyword: Anomaly Detection, Hadoop Distributed File System, ELECTRA, MiniLM, DistilBERT