AI-POWERED LOG ANALYSIS AND THREAT DETECTION USING NLP

أحمد السيد أحمد البهجي
أحمد محمد توفيق رجب
عبدالله أحمد محمد السعيد الششتاوي
محمود فرج حسين مصطفى المكبر
مهند محمد السيد عبدالكريم

AI-Powered Anomaly Detection in Log Files

Introduction

In the rapidly evolving cybersecurity landscape, detecting anomalies in system logs is crucial for identifying potential threats and mitigating security risks. This case study explores the development of an AI-driven anomaly detection model leveraging Natural Language Processing (NLP) and machine learning techniques. The model is designed to employ Isolation Forest and TF-IDF vectorization to classify log entries as normal or anomalous.

Problem Statement

Traditional methods of log analysis rely on predefined rules and manual inspection, which can be inefficient and prone to human error. The primary challenge is to develop an automated, scalable, and efficient solution for detecting anomalies in log files without prior knowledge of attack patterns.

Methodology

The project follows this approach:

Data Collection: Log entries are obtained from various system log sources.

Preprocessing: The logs undergo cleaning by removing timestamps, converting text to lowercase, and eliminating special characters.

Feature Extraction:

- **Text Representation**: TF-IDF (Term Frequency-Inverse Document Frequency) vectorization is applied to transform textual logs into numerical representations.
- **Time-Based Features**: Differences between log timestamps are computed to capture temporal patterns.

Anomaly Detection: An Isolation Forest model is planned to be trained on the feature set to identify anomalous log entries.

Result Classification: Logs will be classified as either 'Normal' or 'Anomaly' based on the model's predictions.

Expected Results and Analysis

The model aims to efficiently process system logs and flag potential anomalies. Anticipated outcomes include:

- Effective capture of variations in log text through TF-IDF vectorization, improving anomaly detection.

- Identification of abnormal activity patterns such as log bursts or sudden gaps in event sequences using time-based features.

- Efficient detection of outliers by the Isolation Forest model with minimal reliance on predefined attack signatures.

Conclusion and Future Work

This project explores the potential of AI-powered log analysis for proactive cybersecurity threat detection. Future enhancements may include:

- Incorporating deep learning techniques for improved log pattern recognition.
- Expanding the dataset with diverse log sources for better generalization.
- Implementing real-time anomaly detection with streaming log processing.

By leveraging machine learning and NLP, this approach aims to enhance the security posture of organizations, reduce manual effort, and improve threat detection capabilities.