ABSTRACT

EVALUATION ON PREDICTION ACCURACY AND COMPUTATIONAL TIME FOR DETECTING ATTACKS WITH PYSPARK

By

Arif Fadillah Ramadhan

18/429286/PA/18677

With the growing amount of *data, attacks* and network traffic across the internet, *intrusion detection systems* have become a popular and useful strategy to *detect anomalies* and *attacks*. The background behind this research was to see the ability of *PySpark* in improving computational time when classifying *attacks* in a labelled *dataset. Data* helps to improve products and services, increase customer satisfaction, maximise profitability, and operate more effectively. It also enables computing and *machine learning* to extract value in an optimal manner from *data*. With this research, these parties will be able to suggest better services to their users.

Explained in this research is the investigation of the feasibility of applying one of the *big data technologies*, *Apache Spark*, to classify different *attacks* and *detect anomalies* as well as improving the *computational time.* This research employs *machine learning algorithms* on a *labelled* dataset of the KDD-CUP-99. This process is all put together in a created python program using the implementation of *PySpark* and performing the cross-validation technique. The research also performs *dataset preprocessing* as well as creating a similar application without *PySpark* for comparison.

The results demonstrate that employing *big data technologies* adds several benefits to *data anomaly detection* than *traditional machine learning environments* in terms of improving *computational time.*

Keywords: *prediction accuracy, computational time, attack, detection, data, machine learning, anomalies, PySpark.*

INTISARI

EVALUASI TERHADAP AKURASI PREDIKSI DAN WAKTU KOMPUTASI UNTUK MENDETEKSI SERANGAN MENGGUNAKAN PYSPARK

Oleh

Arif Fadillah Ramadhan

18/429286/PA/18677

Dengan meningkatnya jumlah *data*, *serangan*, dan jaringan di internet, sistem *deteksi* intrusi telah menjadi strategi yang populer dan berguna untuk mendeteksi data *anomali* dan *serangan*. Latar belakang penelitian ini adalah untuk melihat kemampuan *PySpark* dalam meningkatkan *waktu komputasi* ketika mengklasifikasikan serangan pada dataset yang berlabel. Data membantu meningkatkan produk dan layanan, meningkatkan kepuasan pelanggan, memaksimalkan profitabilitas, dan beroperasi lebih efektif. Ini juga memungkinkan *komputasi* dan *pembelajaran mesin* untuk mengekstrak nilai secara optimal dari *data*. Dengan penelitian ini, pihak-pihak tersebut akan dapat menyarankan layanan *data* yang lebih baik kepada penggunanya.

Dijelaskan dalam penelitian ini adalah investigasi kelayakan penerapan salah satu teknologi big *data*, *Apache Spark*, untuk mengklasifikasikan berbagai *serangan* dan mendeteksi *anomali* serta meningkatkan *waktu komputasi*. Penelitian ini menggunakan algoritma *pembelajaran mesin* pada dataset berlabel bernama KDD-CUP-99. Proses ini disatukan dalam program python yang dibuat menggunakan implementasi *PySpark* dan metode cross-validasi. Penelitian ini juga melakukan proses pembersihan dan membuat aplikasi python yang serupa dengan tidak menggunakan *PySpark* untuk perbandingan.

Hasilnya menunjukkan bahwa penggunaan teknologi big *data* menambahkan beberapa manfaat pada kecepatan deteksi *anomali* data daripada pendekatan *pembelajaran mesin* tradisional dalam hal meningkatkan *waktu komputasi.*

Kata kunci: *akurasi prediksi, waktu komputasi, serangan, deteksi, data, pembelajaran mesin, anomali, PySpark.*