

INTISARI

Saat ini teknologi *machine learning* telah banyak digunakan dalam berbagai bidang mulai dari mesin pencarian (*search engine*) seperti Google Search, analisis data untuk merekomendasikan suatu produk, konten iklan yang akan ditampilkan, asisten virtual (*virtual assistant*), pembuatan *autonomous vehicle*, hingga analisis suatu media sosial. Telah banyak metode *machine learning* yang dikembangkan dan dapat dikelompokkan menjadi tiga jenis: metode *supervised learning*, *unsupervised learning*, serta *semi-supervised learning*. Algoritme yang ada atau disediakan untuk *machine learning* pun beragam. Salah satunya adalah yang penulis gunakan, yaitu algoritme *Extreme Gradient Boosting* (XGBoost).

Pada penelitian ini, dilakukan sebuah perancangan sistem terdistribusi agar dapat melakukan prediksi akun bot pada media sosial Twitter dengan kata kunci yang digunakan, dalam penelitian ini yaitu "#pilpres". Pendekatan yang dilakukan berbasis *machine learning* dengan menggunakan algoritme XGBoost.

Secara garis besar, sistem terdistribusi terdiri atas beberapa bagian, yaitu *crawler* data dari Twitter yang menggunakan *library* untuk bahasa pemrograman Python bernama Twint untuk diimplementasikan ke sistem, *message broker* RabbitMQ untuk melakukan *load-balancing* dengan cara mengirim dan mendistribusikan pesan yang berupa data mentah menuju ketiga *classifier* untuk diolah dan diklasifikasi jenis akun media sosial Twitter tersebut, dan masing-masing *classifier* sekaligus menyimpan data yang sudah diolah dan diprediksi ke dalam bentuk CSV.

Sistem terdistribusi ini bertugas untuk melakukan deteksi jenis akun pada media sosial Twitter. Kemudian performa dari *classifier* diukur dan kemudian dilakukan perbandingan antara performa sistem yang menggunakan satu *classifier* dalam menganalisis 3000 twit dan performa sistem yang menggunakan tiga *classifier* secara bersamaan dengan jumlah twit yang sama. Hasil dari pengujian menunjukkan bahwa sistem dengan satu *classifier* memerlukan waktu 55,44 detik dan mengkonsumsi RAM sebanyak 6,26 MB. Sedangkan bila menggunakan tiga *classifier*, sistem menghabiskan waktu rata-rata 18,53 detik dan mengkonsumsi RAM sekitar 5.6 MB.

Kata kunci: *Machine Learning, XGBoost, Distributed System, Load Balancing, RabbitMQ, Text Processing, Supervised Learning, Twitter, Stemming, Vectorization*

ABSTRACT

Nowadays, machine learning technology has been widely used in various fields ranging from search engines such as Google Search, data analysis to recommend a product, ad content to be displayed, virtual assistants, making autonomous vehicles, and a social media analysis. Many machine learning methods have been developed. They can be grouped into three categories: supervised learning, unsupervised learning, and semi-supervised learning. There are many algorithms available for the learning process, one of which I use is the XGBoost Supervised Machine Learning algorithm.

This research explains about the design of a distributed system in order to predict bot accounts on Twitter social media with the keyword "#pilpres". The approach taken is based on machine learning by using the Extreme Gradient Boosting (XGBoost) algorithm.

Broadly speaking, the distributed system consists of several main parts: designing data crawlers from Twitter using a Python library called Twint to be implemented in the system, designing a RabbitMQ broker to carry out load-balancing by sending and distributing messages in the form of raw data to the third classifier to be processed and classified the type of Twitter account, and each classifier simultaneously stores data that has been processed and predicted in the form of CSV.

The performance of the classifiers and the comparison between the performance of one classifier in analyzing 3000 tweets and the performance of three classifiers with the same number of tweets were measured. The results of the test showed that the system with one classifier took 55.44 seconds and consumed 6.26 MB of RAM. While the system using three classifier spent an average of 18.53 seconds and consumed about 5.6 MB of RAM.

Keywords: Machine Learning, XGBoost, Distributed System, Load Balancing, RabbitMQ, Text Processing, Supervised Learning, Twitter, Stemming, Vectorization