# Layanan Web Machine Learning dan Manajemen Beban untuk Automatic Indonesian News Generation System = Machine Learning Web Service and Load Management for Automatic Indonesian News Generation System

Sulthan Afif Althaf, author



---

Abstrak

Large Language Model (LLM) generatif merupakan jenis model machine learning yang dapat diaplikasikan dalam industri jurnalisme, khususnya dalam proses pembuatan dan validasi berita. Namun, LLM memerlukan sumber daya yang besar untuk operasionalnya serta membutuhkan waktu proses inferensi yang relatif lama. Penelitian ini bertujuan untuk mengembangkan layanan web machine learning yang memanfaatkan LLM generatif untuk proses pembuatan dan validasi berita. Tujuan lainnya adalah menciptakan sistem dengan mekanisme manajemen beban yang efisien untuk meminimalkan waktu inferensi. Pengembangan melibatkan beberapa tahap, yakni analisis kebutuhan stakeholder, perancangan desain dan arsitektur, implementasi, serta evaluasi. Dalam implementasi layanan web machine learning, pengembangan ini berfokus pada manajemen GPU untuk meningkatkan kecepatan proses inferensi LLM. Selain itu, dilakukan implementasi design pattern untuk meningkatkan skalabilitas dalam penambahan model machine learning. Untuk manajemen beban, dikembangkan dua mekanisme, yaitu load balancer dan scheduler. Implementasi load balancer memanfaatkan NGINX dengan metode round-robin. Sedangkan untuk scheduler, digunakan RabbitMQ sebagai antrean, dengan publisher menerima permintaan dan subscriber mendistribusikan permintaan ke layanan yang tersedia. Berdasarkan API Test, layanan ini berhasil melewati uji fungsionalitas dengan waktu respons API sekitar 1-2 menit per permintaan. Evaluasi performa pada kedua mekanisme manajemen beban menunjukkan tingkat keberhasilan 100%, dengan waktu respon rata-rata meningkat seiring dengan peningkatan jumlah request per detik. Pengelolaan beban dengan load balancer menghasilkan waktu respon yang lebih cepat, sementara pengelolaan beban dengan scheduler menghasilkan mekanisme yang lebih efektif pada proses koneksi asinkron.
.....Generative Large Language Model (LLM) is a type of machine learning model that can be applied in the journalism industry, especially in the process of news generation and validation. However, LLM requires large resources for its operation and requires a relatively long inference process time. This research aims to develop a machine learning web service that utilizes generative LLM for news generation and validation. Another goal is to create a system with an efficient load management mechanism to minimize inference time. The development involves several stages, namely stakeholder needs analysis, design and architecture, implementation, and evaluation. In the implementation of machine learning web services, this development focuses on GPU management to increase the speed of the LLM inference process. In addition, the implementation of design patterns is done to improve scalability in adding machine learning models. For load management, two mechanisms are developed: load balancer and scheduler. The load balancer implementation utilizes NGINX with the round-robin method. As for the scheduler, RabbitMQ is used as a queue, with the publisher receiving requests and the subscriber distributing requests to available services. Based on the API Test, the service successfully passed the functionality test with an API response time of about 1-2 minutes per request. Performance evaluation on both load management mechanisms showed a

100% success rate, with the average response time increasing as the number of requests per second increased. The use of a load balancer results in faster response times, while load management with a scheduler results in a more effective mechanism for asynchronous connection processes.