

INTISARI

IMPLEMENTASI *RETRIEVAL AUGMENTED GENERATION* PADA KLASIFIKASI ALAMAT IPv4 BERBASIS *CLASSLESS INTER-DOMAIN* *ROUTING* DENGAN *DOCKER DEPLOYMENT OPTIMIZATION*

Saadah Mardatillah
21/473000/SV/18792

Retrieval Augmented Generation (RAG) merupakan pendekatan yang menggabungkan kemampuan *retrieval* dan *generation* berbasis pencarian data berupa dokumen atau *database* dengan model generatif LLM dalam menghasilkan respons yang relevan terhadap konteks. Proyek akhir ini mengimplementasikan arsitektur *chatbot* berbasis RAG untuk melakukan klasifikasi alamat IPv4 berbasis *Classless Inter-Domain Routing* (CIDR) dan melakukan perbandingan dengan sistem *Non-RAG*. Sistem dikembangkan dan dijalankan melalui pendekatan *containerization* menggunakan Docker untuk mendukung proses *deployment* berjalan konsisten dan mendukung pengelolaan dependensi setiap komponen secara terisolasi. Pengujian dilakukan terhadap sistem RAG dan *Non-RAG* menggunakan dataset sebanyak 10, 20, dan 40 pertanyaan untuk mengukur performa klasifikasi dan konsumsi sumber daya. Hasil pengujian menunjukkan bahwa sistem RAG unggul dengan akurasi 84%, presisi 93%, *recall* 89%, dan relevansi rata-rata 98%, dibandingkan sistem *Non-RAG* yang menunjukkan hasil akurasi 69%, presisi 92%, *recall* 73%, dan relevansi 91%. Waktu respons sistem RAG lebih lambat 4,45 detik dibanding *Non-RAG* 3,14 detik. Hal ini dikarenakan adanya proses *embedding* dan *retrieval*. Sistem RAG menunjukkan stabilitas yang lebih baik pada penggunaan CPU dengan beban 1,00%, sedangkan *Non-RAG* mencapai 1,294%. Proses *containerization* pada sistem RAG juga menunjukkan kompleksitas lebih tinggi dengan ukuran *image* 6,67 GB dan waktu *build* 259,1 detik, dibandingkan *Non-RAG* dengan ukuran *image* 75,7 MB dan waktu *build* 16,5 detik.

Kata kunci: *Retrieval Augmented Generation*, IPv4, CIDR, *Chatbot*, Docker

ABSTRACT

IMPLEMENTATION OF RETRIEVAL AUGMENTED GENERATION IN IPv4 ADDRESS CLASSIFICATION BASED ON CLASSLESS INTER- DOMAIN ROUTING WITH DOCKER DEPLOYMENT OPTIMIZATION

Saadah Mardatillah
21/473000/SV/18792

Retrieval Augmented Generation (RAG) is an approach that combines the capabilities of retrieval and generation by utilizing external data sources, such as documents or databases, along with a generative large language model (LLM) to produce contextually relevant responses. This final project implements a RAG-based chatbot architecture to classify IPv4 addresses using Classless Inter-Domain Routing (CIDR) and compares its performance with a Non-RAG system. The system is developed and deployed using a containerization approach with Docker to ensure consistent deployment and isolated dependency management for each component. The evaluation was conducted on both RAG and Non-RAG systems using datasets consisting of 10, 20, and 40 questions to measure classification performance and resource consumption. The results show that the RAG system outperforms the Non-RAG system, achieving an accuracy of 84%, precision of 93%, recall of 89%, and an average relevance score of 98%, while the Non-RAG system achieved an accuracy of 69%, precision of 92%, recall of 73%, and relevance of 91%. The average response time of the RAG system is 4,45 seconds, which is slower compared to the Non-RAG system at 3,14 seconds, due to additional processes such as query embedding, vector database search, and contextualization. In terms of CPU usage, the RAG system demonstrates better stability with a lower CPU load of 1,00%, compared to 1,294% in the Non-RAG system. Furthermore, the containerization process of the RAG system is more complex, requiring a build time of 259.1 seconds and resulting in an image size of 6,67 GB, whereas the Non-RAG system only requires 16,5 seconds with an image size of 75,7 MB.

Keywords: Retrieval Augmented Generation, IPv4, CIDR, Chatbot, Containerization.