RETRIEVAL-AUGMENTED GENERATION (RAG) PADA *CHATBOT* WHATSAPP UNTUK LAYANAN INFORMASI AKADEMIK INSTITUT TEKNOLOGI INDONESIA

Akmal Aufa Alim [1)], Melani Indriasari .[1)]
1) Program Studi Teknik Informatika Institut Teknologi Indonesia
E-mail: akmalalim28@gmail.com, melani.indriasari@yahoo.com

Abstrak

Perkembangan teknologi kecerdasan buatan (Artificial Intelligence/AI) telah menghadirkan berbagai inovasi dalam layanan pendidikan tinggi, salah satunya melalui penerapan chatbot sebagai media interaktif untuk penyediaan informasi akademik. Penelitian ini bertujuan untuk mengimplementasikan metode Retrieval–Augmented Generation (RAG) pada chatbot berbasis WhatsApp untuk mendukung layanan informasi akademik di Institut Teknologi Indonesia (ITI). Sistem dikembangkan menggunakan pendekatan CRISP–DM yang mencakup enam tahapan: business understanding, data understanding, data preparation, modeling, evaluation, dan deployment. Komponen utama RAG terdiri dari retriever, yang mencari konteks relevan melalui vector database FAISS, serta generator, yang memanfaatkan Large Language Model (LLM) pretrained untuk menghasilkan jawaban kontekstual. Backend sistem dibangun menggunakan FastAPI dan terhubung dengan MongoDB untuk penyimpanan log percakapan. Hasil implementasi menunjukkan bahwa chatbot mampu memberikan jawaban yang akurat, cepat, dan sesuai konteks terhadap pertanyaan pengguna terkait layanan akademik ITI. Dengan demikian, integrasi metode RAG pada chatbot berbasis WhatsApp dapat meningkatkan efisiensi layanan informasi akademik dan mengurangi beban administratif petugas kampus.

Kata kunci: *Chatbot, Retrieval–Augmented Generation (RAG), Artificial Intelligence, FAISS, WhatsApp, LLM, NLP.*

Pendahuluan

Perkembangan teknologi kecerdasan buatan (*Artificial Intelligence*/AI) semakin pesat dan telah memberikan dampak signifikan terhadap berbagai sektor kehidupan, termasuk bidang pendidikan tinggi. Salah satu implementasi AI yang banyak dikembangkan adalah *chatbot* atau sistem percakapan otomatis yang mampu berinteraksi dengan manusia melalui bahasa alami [1]. *Chatbot* saat ini telah banyak digunakan dalam layanan pelanggan, sistem informasi, hingga pendidikan karena kemampuannya memberikan respon cepat dan efisien tanpa keterbatasan waktu [2].

Dalam konteks pendidikan, khususnya perguruan tinggi, kebutuhan akan layanan informasi yang cepat, akurat, dan mudah diakses semakin meningkat. Calon mahasiswa maupun mahasiswa aktif sering kali membutuhkan informasi terkait pendaftaran, jadwal kuliah, administrasi akademik, serta layanan kampus lainnya. Namun, keterbatasan sumber daya manusia pada unit pelayanan sering kali menghambat efektivitas pemberian informasi secara langsung [3]. Oleh karena itu, diperlukan solusi berbasis teknologi yang mampu menjawab pertanyaan pengguna secara otomatis dan relevan sesuai konteks.

Salah satu pendekatan terbaru dalam pengembangan *chatbot* cerdas adalah *Retrieval–Augmented Generation* (RAG), yaitu metode yang menggabungkan pencarian informasi (*retrieval*) dan generasi bahasa alami (*generation*) untuk menghasilkan jawaban yang akurat dan kontekstual [4]. Dengan RAG, *chatbot* tidak hanya mengandalkan kemampuan model bahasa besar (*Large Language Model*/LLM), tetapi juga dapat menarik data dari sumber pengetahuan eksternal yang relevan, seperti dokumen akademik atau panduan kampus [5]. Pendekatan ini memungkinkan sistem memberikan informasi yang selalu terkini dan sesuai konteks tanpa perlu melatih ulang model.

Selain itu, penggunaan *vector database* seperti FAISS (*Facebook AI Similarity Search*) berperan penting dalam mempercepat proses pencarian informasi yang relevan melalui representasi vektor teks [6]. Integrasi antara RAG, FAISS, dan LLM menjadikan *chatbot* mampu memahami konteks pertanyaan pengguna serta menampilkan jawaban yang bersumber dari data resmi institusi. *Platform WhatsApp* dipilih sebagai antarmuka utama karena tingkat penggunaannya yang tinggi di

Indonesia serta kemudahan aksesnya oleh masyarakat luas [7]. Dengan demikian, *chatbot* berbasis *WhatsApp* dapat menjadi media interaktif yang efektif dalam memberikan layanan informasi akademik.

Berdasarkan uraian tersebut, penelitian ini berfokus pada implementasi metode *Retrieval–Augmented Generation* (RAG) pada chatbot berbasis *WhatsApp* untuk mendukung layanan informasi akademik di Institut Teknologi Indonesia (ITI). Sistem ini diharapkan dapat meningkatkan efisiensi layanan informasi, mengurangi beban kerja petugas, serta memberikan pengalaman interaktif yang lebih baik bagi calon mahasiswa dan mahasiswa aktif ITI.

Studi Pustaka

Chatbot merupakan sistem berbasis kecerdasan buatan yang dirancang untuk berinteraksi dengan manusia melalui bahasa alami, baik dalam bentuk teks maupun suara [8]. Dalam konteks pendidikan, chatbot telah digunakan sebagai sarana layanan informasi akademik, membantu calon mahasiswa dalam memperoleh informasi seputar pendaftaran, program studi, hingga jadwal kegiatan kampus [9]. Penerapan chatbot di perguruan tinggi dinilai dapat meningkatkan efisiensi layanan dan mengurangi beban administratif petugas.

Kemampuan chatbot modern sangat dipengaruhi oleh kemajuan Natural Language Processing (NLP), yang memungkinkan sistem memahami, memproses, dan menghasilkan teks yang menyerupai bahasa manusia [10]. Salah satu pendekatan terbaru dalam NLP adalah *Retrieval–Augmented Generation* (RAG), yang menggabungkan dua pendekatan utama dalam sistem tanya jawab cerdas, yaitu *information retrieval* dan *text generation* [11].

Secara konseptual, RAG terdiri atas dua komponen utama, yaitu *retriever* dan *generator*. *Retriever* berfungsi mencari potongan teks atau dokumen paling relevan dari basis pengetahuan berdasarkan pertanyaan pengguna. Proses ini biasanya dilakukan menggunakan *vector similarity search* dengan bantuan *vector database* seperti FAISS (*Facebook AI Similarity Search*) [12]. Sementara itu, *generator* menggunakan hasil pencarian tersebut sebagai konteks untuk menghasilkan jawaban baru yang lebih akurat, kontekstual, dan alami.

Pada komponen *generator*, RAG memanfaatkan Large Language Model (LLM) yang telah dilatih sebelumnya (*pre-trained model*), seperti GPT, LLaMA, atau FLAN-T5 dan lain - lainnya, untuk merumuskan jawaban berdasarkan konteks hasil retrieval. Dengan demikian, model tidak perlu dilatih dari awal, tetapi mampu menghasilkan respon yang kaya dan relevan sesuai dengan domain data yang diberikan. Pendekatan ini memungkinkan sistem untuk menggabungkan kemampuan generatif LLM dengan keakuratan data terkini dari basis pengetahuan eksternal.

Kelebihan RAG dibandingkan model generatif murni (seperti GPT tanpa retrieval) adalah kemampuannya mengakses dan memanfaatkan sumber informasi eksternal secara dinamis, sehingga jawaban yang dihasilkan tidak hanya bergantung pada data pelatihan model, tetapi juga dapat disesuaikan dengan data institusi atau domain tertentu [13]. Pendekatan ini sangat bermanfaat untuk sistem chatbot di lingkungan akademik, di mana keakuratan dan kesesuaian informasi menjadi hal utama.

Untuk media interaksi, integrasi dengan *WhatsApp API* dinilai efektif karena tingkat penggunaannya yang tinggi di Indonesia serta kemudahan akses bagi calon mahasiswa [14].

Sementara itu, *FastAPI* digunakan sebagai *backend framework* untuk menghubungkan model RAG, *vector database*, dan WhatsApp API, sehingga sistem dapat berjalan secara efisien, responsif, dan mudah diintegrasikan dengan komponen lain seperti dashboard admin berbasis Streamlit [15].

Metodologi Penelitian

Metodologi yang digunakan dalam penelitian ini adalah CRISP–DM (Cross Industry Standard Process for Data Mining), yang terdiri dari enam tahapan utama, yaitu Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment. Metodologi ini dipilih karena mampu memberikan kerangka kerja sistematis dalam pengembangan

sistem berbasis kecerdasan buatan, khususnya chatbot Retrieval–Augmented Generation (RAG) untuk layanan informasi akademik di Institut Teknologi Indonesia.

Gambar 1. Tahapan CRISP-DM

Business Understanding

Tahap ini bertujuan untuk memahami permasalahan utama yang dihadapi oleh pihak Institut Teknologi Indonesia (ITI), khus usnya pada layanan informasi akademik dan penerimaan mahasiswa baru (PMB). Permasalahan yang diidentifikasi meliputi keterbatasan waktu dan sumber daya petugas dalam memberikan respon terhadap pertanyaan calon mahasiswa. Tujuan utama penelitian ini adalah membangun chatbot berbasis RAG yang mampu menjawab pertanyaan calon mahasiswa secara akurat, cepat, dan kontekstual melalui platform WhatsApp.

Data *Understanding*

Pada tahap ini dilakukan pengumpulan dan pemahaman terhadap data yang akan digunakan sebagai sumber informasi chatbot. Data diperoleh dari dokumen resmi ITI seperti panduan PMB, brosur program studi, jadwal akademik, panduan KRS, serta informasi administrasi kampus. Data tersebut kemudian dianalisis untuk memahami struktur, format, dan jenis informasi yang akan digunakan sebagai knowledge base sistem.

Data *Preparation*

Tahap ini mencakup proses data *cleaning* dan *preprocessing*, seperti penghapusan karakter khusus, normalisasi teks, dan segmentasi dokumen menjadi potongan teks kecil (*chunks*). Setiap potongan teks kemudian diubah menjadi representasi vektor melalui proses text embedding menggunakan model embedding modern (misalnya sentence-transformers). Hasil embedding disimpan dalam vector database FAISS (Facebook AI Similarity Search) untuk mendukung pencarian berbasis kesamaan semantik.

Modeling

Tahapan ini merupakan inti dari pengembangan sistem chatbot. Model yang digunakan adalah Retrieval–Augmented Generation (RAG), yang terdiri dari dua komponen utama:

- Retriever: berfungsi mencari konteks paling relevan dari knowledge base berdasarkan pertanyaan pengguna dengan menggunakan FAISS.
- Generator: menggunakan model Large Language Model (LLM) yang sudah ada (pre-trained), seperti GPT atau LLaMA, untuk menghasilkan jawaban yang kontekstual dan alami berdasarkan hasil pencarian retriever.

Integrasi sistem dilakukan menggunakan FastAPI sebagai backend, yang menghubungkan modul RAG, vector database, dan WhatsApp API.

Evaluation

Tahap ini dilakukan untuk mengevaluasi kinerja chatbot berdasarkan dua aspek:

- Kuantitatif, meliputi akurasi dan relevansi jawaban berdasarkan perbandingan dengan data uji.
- Kualitatif, melalui kuesioner pengguna untuk menilai tingkat kepuasan, kemudahan penggunaan, serta kecepatan respon chatbot dalam memberikan informasi akademik.

Deployment

Tahap terakhir adalah implementasi sistem ke lingkungan nyata. Chatbot diintegrasikan dengan WhatsApp API sebagai frontend agar dapat diakses langsung oleh calon mahasiswa. Sistem ini juga dilengkapi dengan MongoDB sebagai penyimpan log percakapan dan data pengguna. Hasil implementasi kemudian digunakan untuk mendukung layanan informasi akademik ITI secara berkelanjutan dan dapat dikembangkan lebih lanjut di masa depan.

Hasil dan Pembahasan

Penelitian ini menghasilkan rancangan sistem *chatbot* WhatsApp berbasis *Retrieval Augmented Generation* (RAG) untuk mendukung layanan penerimaan mahasiswa baru (PMB) di Institut Teknologi Indonesia. Sistem ini dirancang untuk memberikan respons otomatis yang relevan terhadap pertanyaan calon mahasiswa melalui integrasi antara model bahasa besar (LLM), basis pengetahuan kampus, dan antarmuka WhatsApp sebagai media interaksi utama.

Arsitektur sistem terdiri dari empat komponen utama, yaitu WhatsApp *Interface*, *Backend FastAPI, Vector Database* (FAISS), dan Model LLM *Generator*.

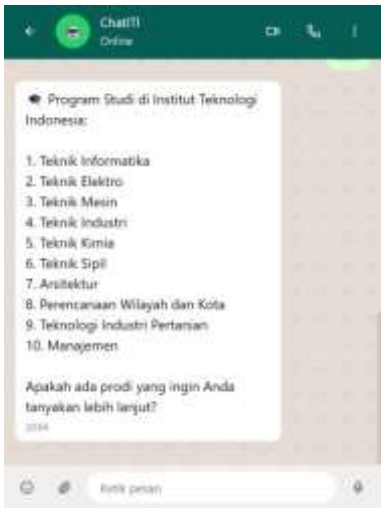
1. WhatsApp *Interface* berfungsi sebagai media komunikasi antara pengguna dan sistem melalui integrasi dengan WhatsApp *Business* API atau platform serupa. Komponen ini menjadi gerbang utama untuk menerima dan mengirim pesan.
2. *Backend* FastAPI bertugas mengatur logika sistem, termasuk pemrosesan pesan, komunikasi antar-komponen, dan pengelolaan permintaan ke model RAG. FastAPI dipilih karena ringan, cepat, dan mudah diintegrasikan dengan layanan eksternal.
3. *Vector Database* (FAISS) menyimpan informasi penting terkait PMB seperti program studi, jadwal pendaftaran, dan biaya kuliah. Dokumen ini dikonversi menjadi representasi vektor menggunakan teknik embedding agar dapat dilakukan pencarian semantik secara efisien.
4. Model LLM *Generator* bertugas menghasilkan respons bahasa alami berdasarkan hasil pencarian dokumen relevan. Model yang digunakan merupakan model bahasa besar yang sudah ada, seperti GPT atau Llama 3, tanpa pelatihan ulang, sehingga dapat langsung dimanfaatkan untuk inferensi (*inference-based generation*).

Alur kerja sistem dimulai ketika pengguna mengirim pesan melalui WhatsApp. Pesan tersebut diterima oleh *backend* untuk diproses melalui *pipeline* RAG.

1. Sistem melakukan tahap *retrieval* untuk menemukan dokumen atau informasi paling relevan dari vector *database* menggunakan FAISS.
2. Dokumen hasil pencarian dikombinasikan dengan konteks pertanyaan pengguna.
3. Data gabungan tersebut kemudian dikirim ke model LLM untuk menghasilkan respons yang sesuai konteks.
4. Hasil respons dikirim kembali kepada pengguna melalui WhatsApp API dengan format pesan yang natural dan mudah dipahami.

Proses ini memungkinkan chatbot memberikan jawaban yang kontekstual dan tepat sasaran, sekaligus menjaga efisiensi waktu dalam pelayanan informasi.

Gambar 2. Desain *UI* awal *Chatbot* pada WhatsApp



Gambar 3. Desain *UI* awal *Chatbot* pada WhatsApp

Kesimpulan

Penelitian ini merancang sistem *chatbot* berbasis *Retrieval–Augmented Generation* (RAG) yang diintegrasikan dengan platform WhatsApp sebagai sarana layanan informasi akademik di Institut Teknologi Indonesia. Melalui penerapan metode CRISP–DM, proses pengembangan dilakukan secara sistematis, mulai dari pemahaman kebutuhan bisnis hingga perancangan sistem yang mencakup pengolahan data, penyimpanan vektor menggunakan FAISS, serta integrasi model *Large Language Model* (LLM) untuk menghasilkan jawaban yang kontekstual dan relevan.

Hasil rancangan menunjukkan bahwa kombinasi antara RAG, vector database, dan FastAPI dapat membentuk arsitektur sistem yang efisien serta mudah diimplementasikan. Antarmuka WhatsApp dipilih karena kemudahan akses dan tingkat penggunaan yang tinggi di kalangan masyarakat, sehingga mendukung terciptanya layanan informasi akademik yang lebih interaktif, cepat, dan efisien.

Ke depan, sistem ini berpotensi dikembangkan lebih lanjut dengan menambahkan fitur analisis percakapan, integrasi ke *dashboard* admin berbasis Streamlit, serta pengujian kinerja sistem secara empiris untuk memastikan efektivitas chatbot dalam mendukung pelayanan akademik di perguruan tinggi.

Daftar Pustaka

[1] A. Nurrahman and F. S. Putra, "Perancangan Chatbot Menggunakan Artificial Intelligence Markup Language (AIML) untuk Layanan Informasi Akademik," Jurnal Teknologi dan Sistem Informasi, vol. 8, no. 2, pp. 120–127, 2022.

[2] R. S. Fadhilah, A. F. Rahman, and N. F. Azzahra, "Analisis dan Implementasi Chatbot sebagai Media Layanan Informasi Akademik," Jurnal Informatika dan Teknologi Komputer (JITK), vol. 6, no. 1, pp. 45–52, 2023.

[3] H. Prasetyo and R. Lestari, "Penggunaan Chatbot untuk Layanan Akademik di Perguruan Tinggi Menggunakan Teknologi NLP," Jurnal Sistem Informasi dan Komputerisasi Administrasi Perkantoran (JSIKAP), vol. 5, no. 3, pp. 155–162, 2021.

[4] Y. Huang and J. X. Huang, "A Survey on Retrieval-Augmented Text Generation for Large Language Models," arXiv preprint arXiv:2404.10981, 2024.

[5] W. Yu, "Retrieval-Augmented Generation across Heterogeneous Knowledge," Proceedings of NAACL Student Research Workshop (SRW), 2022.

[6] M. Douze et al., "The FAISS Library," arXiv preprint arXiv:2401.08281, 2024.

[7] Meta, "WhatsApp Business Platform – Meta for Developers," 2025. [Online]. Available: https://developers.facebook.com/docs/whatsapp

[8] Shawar, B. A., & Atwell, E. (2017). Chatbots: Are they really useful? International Journal of Computers and Applications, 39(9), 975–984. https://doi.org/10.1080/1206212X.2017.1307637

[9] Nurrahmi, H., & Hidayat, R. (2022). Pemanfaatan chatbot berbasis artificial intelligence dalam pelayanan informasi akademik di perguruan tinggi. Jurnal Teknologi Informasi dan Komunikasi, 10(2), 145–153.

[10] Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine, 13(3), 55–75. https://doi.org/10.1109/MCI.2018.2840738

[11] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., … Riedel, S. (2020). Retrieval-Augmented Generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (NeurIPS 2020). https://arxiv.org/abs/2005.11401

[12] Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

[13] Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. In Proceedings of the 16th Conference of the European Chapter of the ACL (EACL 2021) (pp. 874–880). https://arxiv.org/abs/2007.01282

[14] Kurniawan, D., & Maulana, A. (2021). Integrasi chatbot berbasis WhatsApp API sebagai media komunikasi akademik. Jurnal Teknologi dan Sistem Informasi, 9(2), 220–228.

[15] Tiwana, A. (2023). Building efficient APIs with FastAPI. O'Reilly Media.