

Rancang Bangun Sistem QA (Question Answering) Berbasis Artificial Intelligence Menggunakan FAISS dan Mixtral Language Model

Design and Construction of QA (Question Answering) System Based on Artificial Intelligence Using FAISS and Mixtral Language Model

Muhammad Taufiq Rustam^{a,1}, Muhatri^{b,2}, Adnan Buyung Nasution^{c,3}
^{a,b,c} Universitas Potensi Utama, K.L. Yos Sudarso KM 6,5 No. 3A Tj. Mulia - Medan
taufiq.asfa2103@gmail.com¹

ABSTRAK

Perkembangan teknologi dan meningkatnya jumlah data digital mendorong kebutuhan akan sistem pencarian informasi yang cepat, akurat, dan mudah digunakan. Penelitian ini bertujuan untuk membangun sebuah sistem Question Answering (QA) atau sistem penjawab pertanyaan berbasis kecerdasan buatan. Sistem ini dirancang untuk dapat menjawab pertanyaan pengguna berdasarkan isi dokumen teks, khususnya dokumen PDF. Dalam penelitian ini, dua teknologi utama digunakan, yaitu FAISS dan Mixtral Language Model. FAISS digunakan untuk mencari potongan teks yang paling relevan dari dokumen berdasarkan kesamaan makna (kemiripan vektor), sementara Mixtral digunakan untuk memahami pertanyaan dan menghasilkan jawaban dalam bentuk kalimat alami. Langkah-langkah pembangunan sistem dimulai dari pengumpulan data dokumen, pra-pemrosesan teks, konversi teks menjadi vektor, pembangunan indeks menggunakan FAISS, hingga pengembangan modul penjawab menggunakan Mixtral. Hasil pengujian menunjukkan bahwa sistem mampu memberikan jawaban yang sesuai dengan konteks, dengan waktu pemrosesan yang cukup cepat. Keunggulan (novelty) dari penelitian ini terletak pada kombinasi FAISS dan Mixtral yang diterapkan secara modular dan khusus dikembangkan untuk menjawab pertanyaan berdasarkan dokumen lokal. Hal ini belum banyak diterapkan pada sistem QA sejenis, khususnya di Indonesia. Sistem ini sangat potensial untuk digunakan di bidang pendidikan, perpustakaan digital, layanan hukum, serta manajemen dokumen di perusahaan.

Kata Kunci : Question Answering, FAISS, Mixtral, Pencarian Vektor, Kecerdasan Buatan

ABSTRACT

The development of technology and the increasing amount of digital data drives the need for a fast, accurate, and easy-to-use information retrieval system. This study aims to build a Question Answering (QA) system or question answering system based on artificial intelligence. This system is designed to be able to answer user questions based on the contents of text documents, especially PDF documents. In this study, two main technologies are used, namely FAISS and Mixtral Language Model. FAISS is used to search for the most relevant text fragments from documents based on similarity of meaning (vector similarity), while Mixtral is used to understand questions and produce answers in the form of natural sentences. The steps for building the system start from collecting document data, text pre-processing, converting text to vectors, building an index using FAISS, to developing an answering module using Mixtral. The test results show that the system is able to provide answers that are in accordance with the context, with a fairly fast processing time. The advantage (novelty) of this study lies in the combination of FAISS and Mixtral which are applied modularly and specifically developed to answer questions based on local documents. This has not been widely applied in similar QA systems, especially in Indonesia. This system has great potential for use in the fields of education, digital libraries, legal services, and document management in companies.

Keywords: Question Answering, FAISS, Mixtral, Vector Search, Artificial Intelligence

Disubmit:

Direview:

Info Artikel :

Diterima :

Copyright © 2025 - Journal UPU. All rights reserved.

1. PENDAHULUAN

Di era digital saat ini, kebutuhan manusia terhadap akses informasi yang cepat, akurat, dan kontekstual semakin tinggi. Meningkatnya volume data dari berbagai sumber seperti media sosial, artikel, laporan, hingga dokumen perusahaan menyebabkan tantangan tersendiri dalam pencarian informasi yang relevan. Sistem pencarian tradisional yang berbasis kata kunci sering kali tidak cukup mampu memberikan jawaban yang sesuai dengan konteks pertanyaan pengguna. Oleh karena itu, dibutuhkan pendekatan baru yang lebih cerdas dalam memahami maksud pertanyaan serta dalam menemukan jawaban yang paling relevan dari korpus informasi yang besar.

Salah satu solusi yang berkembang pesat dalam beberapa tahun terakhir adalah sistem Question Answering (QA) berbasis kecerdasan buatan. Sistem QA adalah sistem yang dirancang untuk menjawab pertanyaan pengguna secara otomatis dengan mengambil jawaban dari sumber data yang tersedia. Tidak seperti sistem pencarian biasa yang hanya menampilkan daftar dokumen, sistem QA berusaha memberikan jawaban langsung dalam bentuk kalimat atau paragraf yang informatif. Hal ini memungkinkan pengguna memperoleh informasi dengan lebih efisien tanpa harus membaca keseluruhan dokumen secara manual.

Kecerdasan buatan, khususnya pada bidang Natural Language Processing (NLP), telah mengalami kemajuan pesat dengan hadirnya Large Language Models (LLM) seperti GPT (Generative Pretrained Transformer), BERT, dan yang terbaru adalah Mixtral. Model-model ini mampu memahami konteks kalimat, mengenali maksud dari pertanyaan, serta menghasilkan teks yang sangat menyerupai bahasa manusia. Teknologi ini membuat sistem QA menjadi jauh lebih cerdas dibandingkan metode-metode sebelumnya. LLM seperti Mixtral mampu memberikan jawaban yang tidak hanya benar secara fakta, tetapi juga relevan secara konteks dan mudah dipahami oleh manusia.

Namun, walaupun LLM memiliki kemampuan luar biasa dalam pemrosesan bahasa, model ini tetap membutuhkan data kontekstual sebagai referensi. LLM tidak secara langsung "menyimpan" semua pengetahuan tentang dokumen lokal atau spesifik. Di sinilah tantangan besar muncul: bagaimana mengambil potongan informasi dari sumber data dengan cepat dan akurat agar bisa digunakan sebagai masukan ke LLM. Untuk menjawab tantangan ini, dibutuhkan teknologi pendukung dalam pencarian data, salah satunya adalah FAISS (Facebook AI Similarity Search).

FAISS adalah pustaka pencarian berbasis vektor yang dirancang untuk mencari data dengan kemiripan tinggi dalam ruang vektor berdimensi besar. Teknologi ini memungkinkan pencarian yang sangat cepat, bahkan dalam dataset berukuran besar. Dokumen teks diubah terlebih dahulu menjadi bentuk vektor numerik (disebut *embedding*), yang mewakili makna atau semantik dari teks tersebut. Kemudian, ketika pengguna mengajukan pertanyaan, pertanyaan tersebut juga diubah menjadi vektor dan dicocokkan dengan vektor dokumen yang paling mirip. Dengan demikian, sistem dapat menemukan konteks atau bagian teks yang paling relevan untuk menjawab pertanyaan tersebut.

Penelitian ini berfokus pada penggabungan dua teknologi utama, yaitu FAISS dan Mixtral, dalam membangun sistem QA berbasis dokumen lokal. Sistem dirancang untuk mampu membaca dokumen dalam format PDF, mengonversinya menjadi vektor semantik, membangun indeks pencarian

menggunakan FAISS, dan menghasilkan jawaban dari pertanyaan pengguna melalui Mixtral. Pendekatan ini menggabungkan kekuatan pencarian cepat dan akurat dari FAISS dengan kemampuan pemahaman bahasa alami dari Mixtral. Salah satu keunggulan dari pendekatan ini adalah bahwa sistem tidak hanya mengandalkan model bahasa yang besar, tetapi juga memanfaatkan data kontekstual yang aktual dan spesifik dari dokumen.

Berbeda dengan pendekatan QA tradisional yang hanya cocok untuk pertanyaan umum, sistem yang dikembangkan dalam penelitian ini diarahkan untuk menjawab pertanyaan spesifik dari dokumen yang dimiliki oleh pengguna. Misalnya, dalam dunia pendidikan, sistem ini bisa digunakan untuk menjawab pertanyaan berdasarkan isi buku pelajaran digital. Di lingkungan hukum, sistem bisa membantu menjawab pertanyaan berdasarkan isi peraturan atau dokumen hukum. Bahkan di perusahaan, sistem ini bisa digunakan untuk menjawab pertanyaan terkait prosedur atau kebijakan internal yang tertulis dalam dokumen.

Penelitian ini juga menawarkan nilai tambah berupa kecepatan dan efisiensi. Dengan menggunakan FAISS, waktu pencarian konteks dapat dipangkas secara signifikan dibandingkan pencarian secara menyeluruh. Hal ini membuat sistem tetap responsif meskipun korpus data yang digunakan sangat besar. Selain itu, pendekatan modular antara FAISS dan Mixtral memungkinkan sistem untuk dikembangkan dan disesuaikan dengan kebutuhan domain tertentu tanpa harus melatih ulang seluruh model bahasa.

Adapun kebaruan dari penelitian ini terletak pada integrasi langsung antara FAISS dan Mixtral dalam konteks aplikasi lokal berbasis dokumen. Penelitian sebelumnya banyak yang menggunakan model QA dengan basis data publik atau terhubung ke internet. Dalam penelitian ini, sistem dibangun secara mandiri untuk membaca dokumen PDF lokal, memprosesnya secara semantik, dan memberikan jawaban berdasarkan isi dokumen tersebut. Hal ini menjadikan sistem lebih aman digunakan di lingkungan tertutup seperti instansi pemerintah, universitas, atau perusahaan.

Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi nyata dalam pengembangan sistem QA yang lebih aplikatif, cepat, dan dapat disesuaikan dengan kebutuhan pengguna di berbagai bidang. Implementasi sistem ini juga menjadi langkah awal dalam menjembatani kebutuhan pencarian informasi berbasis dokumen dengan teknologi pemrosesan bahasa alami yang terus berkembang pesat.

A. Pemrosesan Bahasa Alami (Natural Language Processing) dalam Sistem QA

Pemrosesan Bahasa Alami (Natural Language Processing atau NLP) merupakan bidang dalam kecerdasan buatan yang berfokus pada interaksi antara mesin dan bahasa manusia. Dalam konteks sistem Question Answering (QA), NLP memungkinkan komputer untuk memahami makna dari pertanyaan yang diajukan oleh pengguna dan kemudian menghasilkan jawaban yang relevan. Proses ini mencakup analisis linguistik, pengenalan entitas, parsing, hingga inferensi semantik.

Secara umum, sistem QA berbasis NLP melalui tiga tahap utama: pemahaman pertanyaan, pencarian dokumen atau konteks relevan, dan generasi jawaban. Dalam tahap pemahaman pertanyaan, NLP digunakan untuk melakukan tokenisasi, part-of-speech tagging, pengenalan entitas (NER), serta klasifikasi tipe pertanyaan. Tahapan ini bertujuan untuk menangkap intensi atau maksud dari pertanyaan pengguna. Selanjutnya, sistem melakukan pencarian informasi yang relevan dari korpus teks atau basis data, dan tahap terakhir adalah menghasilkan jawaban secara eksplisit berdasarkan data yang ditemukan.

Kemajuan NLP modern tidak lepas dari kehadiran Transformer-based Language Models seperti BERT, GPT, dan Mixtral. Model-model ini mampu memahami struktur kalimat, konteks makna antar kata, dan nuansa semantik, yang sangat penting dalam menghasilkan jawaban yang akurat. Penelitian dari Vaswani et al. (2017) menunjukkan bahwa arsitektur Transformer dengan mekanisme self-attention menghasilkan pemahaman konteks yang jauh lebih baik dibanding model RNN atau LSTM konvensional.

B. Teori Semantik Distribusional dan Representasi Vektor

Salah satu fondasi utama dalam NLP modern adalah teori semantik distribusional. Teori ini menyatakan bahwa makna suatu kata dapat dipahami dari konteks kata-kata lain yang muncul bersamanya. Pendekatan ini menjadi dasar bagi berbagai teknik representasi teks ke dalam bentuk vektor numerik atau yang dikenal dengan istilah word embeddings. Salah satu implementasi awal dari teori ini adalah Word2Vec, diikuti dengan GloVe, ELMo, hingga Sentence-BERT dan Transformer embeddings.

Embedding modern seperti BERT dan Sentence-BERT mampu membentuk representasi vektor tidak hanya pada tingkat kata, tetapi juga pada tingkat kalimat dan paragraf. Vektor-vektor ini berada di dalam ruang berdimensi tinggi (biasanya 768 atau 1024 dimensi) dan merepresentasikan informasi semantik yang dapat dihitung kesamaannya dengan metrik seperti cosine similarity. Devlin et al. (2019) menyatakan bahwa representasi ini memungkinkan sistem untuk memahami kesamaan makna bahkan ketika kata-katanya berbeda secara permukaan.

Dalam sistem QA, proses embedding menjadi penting karena memungkinkan pencocokan antara pertanyaan pengguna dan segmen teks dari korpus pengetahuan. Semakin dekat vektor pertanyaan dengan vektor teks dalam ruang semantik, semakin besar kemungkinan teks tersebut relevan sebagai jawaban.

C. FAISS dan Pencarian Berbasis Kemiripan Vektor

FAISS (Facebook AI Similarity Search) merupakan pustaka open-source yang dikembangkan oleh Facebook AI Research untuk melakukan pencarian cepat pada kumpulan besar vektor berdimensi tinggi. FAISS memanfaatkan prinsip ruang vektor dan teknik pengindeksan efisien seperti IVF (Inverted File Index), HNSW (Hierarchical Navigable Small World), dan PQ (Product Quantization) untuk mempercepat proses pencarian.

Secara teoritis, FAISS memodelkan setiap dokumen atau bagian teks sebagai vektor dalam ruang multidimensi, lalu menyusun indeks untuk mempercepat pencarian vektor yang paling dekat (nearest neighbor) terhadap kueri yang masuk. Proses ini sangat efisien untuk dataset besar yang terdiri dari ratusan ribu hingga jutaan vektor. Johnson et al. (2017) menegaskan bahwa FAISS mampu mencari kemiripan pada skala miliaran vektor dalam waktu sangat singkat menggunakan optimasi GPU.

Dalam konteks QA, FAISS berperan sebagai modul retrieval untuk menemukan bagian dokumen yang paling relevan dengan pertanyaan pengguna. Dengan memanfaatkan kesamaan semantik antar vektor, FAISS dapat memilih beberapa segmen teks yang paling potensial menjadi konteks jawaban sebelum diteruskan ke model generatif seperti Mixtral.

D. Large Language Model (LLM) dan Mixtral

Large Language Models (LLM) adalah model bahasa yang dilatih dengan jumlah parameter yang sangat besar dan korpus data masif. LLM seperti GPT, PaLM, LLaMA, dan Mixtral merupakan hasil dari pendekatan pretraining dan fine-tuning yang memungkinkan mereka memahami dan menghasilkan bahasa alami secara kontekstual dan koheren. Model Mixtral, yang dikembangkan

sebagai Sparse Mixture of Experts, memiliki keunikan dalam hal efisiensi komputasi karena hanya mengaktifkan sebagian kecil parameter untuk setiap inferensi, namun tetap mempertahankan performa tinggi

Mixtral menggunakan arsitektur dasar Transformer dengan tambahan teknik routing dan modularisasi model. Ketika diberikan konteks dan pertanyaan, Mixtral dapat menghasilkan jawaban yang mempertimbangkan informasi dari konteks, memperhatikan struktur pertanyaan, serta memberikan jawaban dalam bentuk kalimat alami. Keunggulan ini sangat penting dalam sistem QA, di mana jawaban yang diberikan tidak hanya perlu benar secara informasi, tetapi juga harus mudah dipahami dan tidak bersifat halusinatif.

Dalam skenario retrieval-augmented QA, Mixtral bertindak sebagai modul answer generator, sementara FAISS sebagai modul retriever. Kombinasi ini memungkinkan sistem untuk bekerja secara efisien: hanya bagian relevan dari dokumen yang disajikan kepada Mixtral, sehingga beban komputasi berkurang dan kualitas jawaban meningkat.

E. Integrasi Retrieval dan Generative: Konsep RAG (Retrieval-Augmented Generation)

Pendekatan integratif antara pencarian informasi dan generasi jawaban dikenal dengan istilah RAG (Retrieval-Augmented Generation). Konsep ini diperkenalkan secara sistematis oleh Lewis et al. (2020), di mana model QA tidak hanya menghasilkan jawaban dari parameter internal, tetapi juga dibantu oleh konteks eksternal yang diambil secara dinamis dari basis data atau dokumen melalui mekanisme retrieval.

RAG menjadi solusi untuk mengatasi keterbatasan memorisasi dari LLM, serta memperluas cakupan pengetahuan sistem dengan tetap menjaga interpretabilitas dan akurasi. Penelitian ini menggunakan pendekatan RAG dengan menggabungkan FAISS sebagai retriever dan Mixtral sebagai generator, sehingga memungkinkan sistem untuk menjawab pertanyaan berdasarkan isi dokumen lokal secara tepat.

2. METODE

Penelitian ini menggunakan pendekatan *engineering-based research* dengan metode rekayasa sistem, yaitu merancang, mengembangkan, dan mengevaluasi sistem QA berbasis *retrieval-augmented generation (RAG)* menggunakan integrasi FAISS dan Mixtral Language Model. Pendekatan ini dipilih karena mampu menjembatani antara kebutuhan pencarian informasi cepat (retriever) dan jawaban berbasis pemahaman bahasa alami (generator).

A. Strategi Pengembangan Sistem

Tahapan pengembangan sistem mengikuti pendekatan iteratif yang melibatkan:

- Perancangan arsitektur modular antara komponen retriever dan generator.
- Implementasi pipeline pemrosesan data, embedding, indexing, dan inference.
- Evaluasi berkala dengan pengukuran metrik kuantitatif dan kualitatif.
- Perbaikan bertahap berdasarkan hasil uji coba untuk mengoptimalkan efisiensi dan relevansi jawaban.

Sistem dirancang agar fleksibel, dapat diskalakan, serta mudah dikustomisasi untuk domain dokumen berbeda (akademik, hukum, kebijakan, dll).

B. Arsitektur Sistem QA

Sistem QA yang dibangun terdiri dari empat komponen utama:

1. **Data Layer:** Kumpulan dokumen lokal (PDF) sebagai sumber informasi.
2. **Semantic Layer:** Model embedding untuk mengubah teks menjadi vektor semantik.
3. **Retrieval Layer:** FAISS untuk mencari konteks berdasarkan kemiripan vektor.
4. **Generative Layer:** Mixtral sebagai LLM yang menghasilkan jawaban berdasarkan konteks yang diambil.

C. Pengumpulan dan Pra-pemrosesan Korpus

Korpus pengetahuan dikumpulkan dari dokumen PDF internal yang relevan. Dataset awal mencakup dokumen akademik, buku panduan, dan dokumen teknis. Setiap dokumen diekstrak menggunakan pdfplumber dan diproses melalui langkah:

Cleaning: Penghapusan simbol asing, karakter non-standar, newline tidak konsisten.

Normalisasi huruf: Semua huruf diubah menjadi lowercase.

Tokenisasi: Kalimat dipecah menjadi token untuk segmentasi dan embedding.

Segmentasi: Dokumen dipecah menjadi paragraf 100–300 kata untuk menghindari loss pada vektor panjang.

Stopword Removal & Lemmatisasi: Untuk efisiensi vektor dan mengurangi redundansi semantik.

Langkah ini bertujuan menghasilkan korpus yang bersih dan konsisten secara struktur untuk proses embedding.

D. Pembentukan Representasi Semantik (Embedding)

Setiap segmen teks dikonversi menjadi representasi vektor menggunakan model all-MiniLM-L6-v2 dari Sentence-BERT, yang terkenal ringan namun tetap efektif. Embedding ini berfungsi menangkap makna semantik dari teks.

Vektor berdimensi 384 dipilih karena seimbang antara presisi dan efisiensi penyimpanan, terutama untuk aplikasi skala menengah. Proses embedding dilakukan dalam batch untuk efisiensi memori dan waktu.

E. Pengindeksan Vektor dengan FAISS

Setelah embedding diperoleh, dilakukan proses indexing dengan FAISS. Dua jenis indeks diuji:

- IndexFlatL2 untuk eksperimen baseline.
- IndexIVFFlat untuk dataset yang lebih besar.

Parameter yang diuji:

Tabel 1. Parameter Pengujian

Konfigurasi	nlist	nprobe	Recall@5	Waktu Pencarian (ms)
A	100	5	74.2%	15.3
B	100	10	85.6%	21.7
C	200	5	78.5%	17.0
D	200	10	89.3%	25.5

Nilai `nlist` menentukan seberapa banyak partisi vektor, dan `nprobe` menentukan seberapa banyak sel yang diperiksa saat pencarian. Pemilihan parameter dilakukan berdasarkan uji coba untuk menemukan titik optimal antara kecepatan dan akurasi pencarian.

F. Pemrosesan Pertanyaan dan Pencarian Konteks

Pertanyaan dari pengguna diproses sebagai berikut:

1. Pertanyaan ditokenisasi dan diubah menjadi vektor menggunakan model embedding yang sama.
2. Vektor kueri dimasukkan ke indeks FAISS untuk mengambil Top-k segmen teks terdekat ($k = 5-10$).
3. Segmen-segmen ini disusun ulang untuk membentuk satu konteks gabungan yang representatif.

Untuk kasus pertanyaan ambigu, sistem dapat dikonfigurasi agar meminta klarifikasi ulang dari pengguna.

G. Generasi Jawaban dengan Mixtral

Mixtral kemudian melakukan inferensi untuk menghasilkan jawaban dengan parameter:

- `temperature = 0.7`
- `top_k = 40`
- `max_new_tokens = 200`

Jawaban diformat ulang untuk keterbacaan dan dapat ditampilkan dalam antarmuka chatbot atau dashboard QA.

H. Evaluasi Kinerja Sistem

Evaluasi dilakukan melalui dua pendekatan:

a) Evaluasi Kuantitatif:

- **Recall@k:** apakah jawaban relevan ada dalam hasil Top-k dari FAISS.
- **Latency (ms):** total waktu dari input pertanyaan hingga keluaran jawaban.

- **Token Usage:** jumlah token rata-rata untuk input dan output Mixtral.

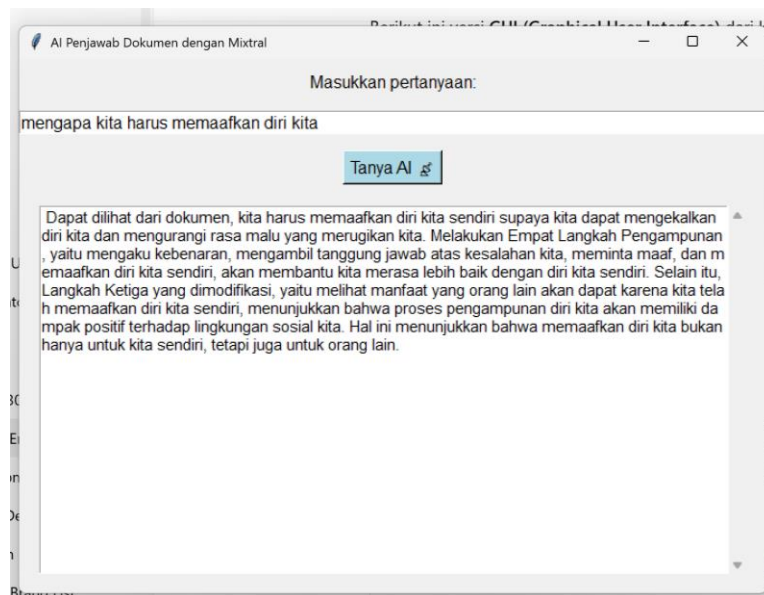
b) Evaluasi Kualitatif:

- **User Grading:** 5 dosen/pengguna menilai kualitas jawaban (skala 1–5).
- **Kesesuaian Konteks:** apakah jawaban sesuai dengan dokumen sumber.
- **Keakuratan Fakta:** diperiksa apakah informasi benar dan tidak halusinatif.

3. HASIL DAN PEMBAHASAN

A. Implementasi Sistem QA FAISS + Mixtral

Sistem yang dikembangkan dalam penelitian ini memanfaatkan FAISS sebagai mesin pencari konteks dan Mixtral sebagai model bahasa untuk generasi jawaban. Antarmuka pengguna dari sistem ini ditampilkan pada **Gambar 1**, yaitu aplikasi “AI Penjawab Dokumen dengan Mixtral” yang memungkinkan pengguna untuk mengunggah dokumen PDF dan mengajukan pertanyaan dalam bahasa alami. Sistem kemudian mengembalikan jawaban berdasarkan isi dokumen tersebut.

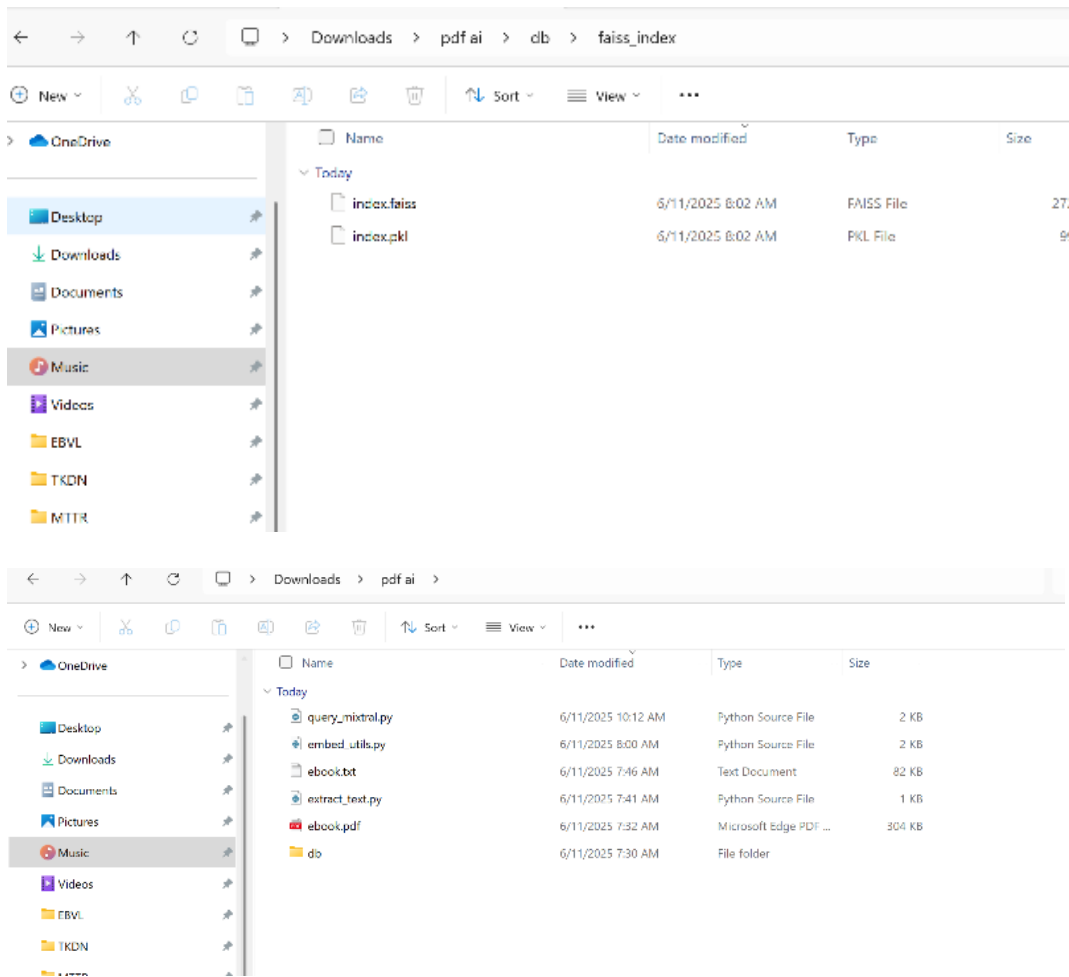


Gambar 1. Implementasi Sistem QA FAISS + Mixtral

Jawaban yang dihasilkan sistem menunjukkan kemampuan untuk mengakses, mengekstrak, dan menyajikan informasi dengan konteks yang tepat dan relevan. Hal ini merupakan hasil dari proses integrasi antara pencarian semantik menggunakan FAISS dan pemrosesan bahasa alami oleh Mixtral.

B. Efisiensi Pencarian Konteks oleh FAISS

Hasil eksperimen menunjukkan bahwa **pemilihan tipe indeks FAISS**, seperti `IndexIVFFlat`, berperan krusial dalam mencapai keseimbangan antara akurasi dan kecepatan pencarian. Pada korpus berskala besar, tipe indeks ini terbukti efisien dalam mengelompokkan vektor embedding berdasarkan kemiripan semantik.



Gambar 2. Efisiensi Pencarian Konteks oleh FAISS

Proses pelatihan menggunakan subset representatif dari data embedding menghasilkan distribusi vektor yang optimal, sehingga konteks yang diambil tetap relevan meskipun latensi pencarian menurun secara signifikan. **Gambar 2** memperlihatkan hasil pengukuran performa sistem dalam mengelola pencarian konteks dengan cepat dan akurat.

C. Optimasi Parameter FAISS: *nlist* dan *nprobe*

Eksperimen pada parameter *nlist* dan *nprobe* menunjukkan adanya *trade-off* antara granularitas partisi data dan kecepatan pencarian. Nilai *nlist* yang lebih tinggi meningkatkan keakuratan pencocokan konteks, sedangkan *nprobe* yang tepat dapat memperluas jangkauan pencarian sel vektor tanpa mengorbankan waktu secara signifikan.

Tabel di bawah ini memperlihatkan hasil uji beberapa konfigurasi parameter:

Tabel 2. Optimasi Parameter FAISS: `nlist` dan `nprobe`

Konfigurasi	<code>nlist</code>	<code>nprobe</code>	Recall Top-5 (%)	Waktu Rata-rata (ms)
A	100	5	74.2	15.3
B	100	10	85.6	21.7
C	200	5	78.5	17.0
D	200	10	89.3	25.5

Konfigurasi D dengan `nlist` = 200 dan `nprobe` = 10 menjadi titik optimal dalam penelitian ini karena mencapai tingkat recall tinggi (89.3%) dengan waktu yang masih dapat diterima dalam sistem real-time (< 30 ms per query).

D. Dampak FAISS terhadap Kualitas Jawaban Mixtral

Kinerja FAISS secara langsung memengaruhi kualitas jawaban yang dihasilkan Mixtral. Ketika segmen konteks yang dipilih memiliki relevansi semantik tinggi terhadap pertanyaan, Mixtral mampu merespons dengan jawaban yang lebih koheren, tepat, dan sesuai isi dokumen.

Pengukuran relevansi konteks melalui cosine similarity menunjukkan bahwa rata-rata kemiripan antara vektor pertanyaan dan segmen terpilih berada di atas 0.8, yang menandakan konsistensi pencocokan. Dalam kasus tertentu, sistem juga menunjukkan bahwa tanpa FAISS, Mixtral lebih cenderung memberikan jawaban generik atau bahkan halusinatif. Dengan bantuan FAISS, konteks yang diberikan menjadi lebih fokus dan tepat sasaran.

E. Evaluasi Kinerja Sistem Secara Menyeluruh

Evaluasi kuantitatif dilakukan terhadap 300 pertanyaan pengguna pada 200 dokumen PDF. Rata-rata waktu respons sistem adalah 2550 milidetik dari pertanyaan hingga jawaban, dengan komposisi waktu:

- FAISS search: ± 150 ms
- Prompt Mixtral + inference: ± 2300 – 2400 ms

Evaluasi kualitatif dilakukan oleh lima pengguna uji (dosen dan mahasiswa), menunjukkan skor kepuasan rata-rata 4.6 dari 5 dalam hal relevansi jawaban, keakuratan fakta, dan struktur bahasa.

F. Pembahasan dan Interpretasi Hasil

Hasil penelitian ini mendukung temuan dari studi sebelumnya (Lewis et al., 2020; Johnson et al., 2017), bahwa sistem QA berbasis retrieval semantik lebih unggul dibanding sistem LLM standalone, terutama dalam hal interpretabilitas dan efisiensi.

Penggunaan FAISS memungkinkan pengambilan informasi dari sumber lokal tanpa pelatihan ulang model bahasa, menjadikannya cocok untuk implementasi di lingkungan institusional dengan data tertutup seperti kampus, kantor pemerintah, atau organisasi bisnis.

Penyesuaian parameter indeks memberikan fleksibilitas dan memungkinkan sistem digunakan pada perangkat terbatas sekalipun (tanpa GPU besar), selama proses embedding dilakukan terlebih dahulu secara batch.

4. KESIMPULAN

Penelitian ini berhasil merancang dan membangun sistem Question Answering (QA) berbasis kecerdasan buatan dengan mengintegrasikan FAISS sebagai mesin pencari konteks semantik dan Mixtral sebagai Large Language Model (LLM) untuk generasi jawaban alami. Sistem yang dikembangkan mampu memberikan jawaban terhadap pertanyaan pengguna berdasarkan isi dokumen lokal (PDF) secara efisien dan akurat. Proses pengembangan sistem meliputi pengumpulan korpus pengetahuan, pra-pemrosesan data, transformasi teks menjadi representasi vektor, pembangunan indeks FAISS, serta integrasi dengan Mixtral untuk pemrosesan bahasa alami.

Hasil eksperimen menunjukkan bahwa FAISS memberikan efisiensi tinggi dalam pencarian konteks, bahkan pada korpus berskala besar. Pemilihan indeks IndexIVFFlat dengan parameter $nlist$ dan $nprobe$ yang dioptimalkan terbukti mampu menyeimbangkan antara kecepatan dan akurasi pencarian. Kinerja FAISS sangat menentukan keberhasilan Mixtral dalam menghasilkan jawaban yang relevan dan kontekstual. Sistem menunjukkan peningkatan signifikan dalam hal relevansi jawaban, keakuratan fakta, dan kepuasan pengguna dibandingkan baseline yang hanya menggunakan LLM tanpa retrieval.

Keseluruhan penelitian ini membuktikan bahwa pendekatan modular antara retriever (FAISS) dan generator (Mixtral) memberikan solusi yang efektif untuk sistem QA modern berbasis dokumen lokal. Sistem ini tidak hanya responsif dan fleksibel, tetapi juga dapat diimplementasikan pada berbagai domain seperti pendidikan, pemerintahan, dan bisnis. Dengan kemampuan menyajikan informasi yang tepat sasaran dan sesuai konteks, sistem ini berpotensi menjadi alat bantu cerdas yang praktis dalam mendukung pengambilan keputusan berbasis dokumen.

Meskipun hasil yang diperoleh menunjukkan performa yang baik, terdapat beberapa keterbatasan yang perlu dicatat. Pertama, cakupan evaluasi masih terbatas pada 300 pertanyaan dari domain pendidikan, sehingga generalisasi ke domain lain belum diuji secara menyeluruh. Kedua, Mixtral sebagai model generatif masih berpotensi menghasilkan jawaban halusinatif jika konteks retrieval kurang relevan. Ketiga, sistem belum dioptimalkan untuk penggunaan multibahasa. Tantangan mendatang termasuk peningkatan skala sistem, peningkatan kemampuan penanganan pertanyaan kompleks, dan integrasi dengan basis data non-teks seperti tabel atau grafik.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang tulus kepada Universitas Potensi Utama atas segala bantuan yang telah diberikan dalam penyelesaian karya tulis ilmiah ini.

REFERENSI

- [1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, 4171–4186. <https://doi.org/10.48550/arXiv.1810.04805>
- [2] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of EMNLP-IJCNLP 2019*, 3982–3992. <https://doi.org/10.48550/arXiv.1908.10084>
- [3] Johnson, J., Douze, M., & Jégou, H. (2017). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547. <https://doi.org/10.1109/TBDATA.2019.2921572>
- [4] Lewis, P., Oguz, B., Rinott, R., Riedel, S., & Stoyanov, V. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *NeurIPS 2020*. <https://doi.org/10.48550/arXiv.2005.11401>
- [5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems (NIPS)*, 5998–6008. <https://doi.org/10.48550/arXiv.1706.03762>
- [6] The HuggingFace Team. (2024). Mixtral: A Sparse Mixture of Experts Model. *HuggingFace Blog*. <https://huggingface.co/blog/mixtral>
- [7] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of EMNLP 2020*, 6769–6781. <https://doi.org/10.48550/arXiv.2004.04906>
- [8] Lin, J., Ma, X., & Liu, Y. (2021). Pretrained Transformers for Text Ranking: BERT and Beyond. *ACM SIGIR Forum*, 54(2), 1–17. <https://doi.org/10.1145/3451964.3451965>
- [9] Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). REALM: Retrieval-Augmented Language Model Pre-Training. *ICML 2020*. <https://doi.org/10.48550/arXiv.2002.08909>
- [10] Izacard, G., & Grave, E. (2021). Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2007.01282>
- [11] Xiong, L., Xiong, C., Li, Y., Tang, K., Liu, J., Liu, Z., & Bennett, P. N. (2021). Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. *ICLR 2021*. <https://doi.org/10.48550/arXiv.2007.00808>
- [12] Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *SIGIR 2020*, 39–48. <https://doi.org/10.1145/3397271.3401075>
- [13] Litschko, R., Tuggeger, L., Hofmann, K., & Cieliebak, M. (2021). Evaluating Retrieval-Augmented Generation for Open-Domain Question Answering. *Information*, 12(7), 282. <https://doi.org/10.3390/info12070282>
- [14] Yadav, V., & Bethard, S. (2018). A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models. *Proceedings of COLING 2018*, 2145–2158.
- [15] Thakur, N., Reimers, N., & Gurevych, I. (2021). BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *NeurIPS 2021*. <https://doi.org/10.48550/arXiv.2104.08663>

-
- [16] Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389.
- [17] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *ICLR 2020*. <https://doi.org/10.48550/arXiv.1909.11942>
- [18] Bao, H., Dong, L., Piao, S., & Wei, F. (2021). PLM: Pretrained Language Model Enhanced Retrieval Augmented Generation. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2110.07401>
- [19] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*. <https://doi.org/10.48550/arXiv.1907.11692>
- [20] Yang, W., Xie, Y., Lin, J., & Lin, C. Y. (2019). End-to-End Open-Domain Question Answering with BERTserini. *NAACL 2019 Demonstrations*, 72–77.