

BAB II

TINJAUAN PUSTAKA

2.1 Review Penelitian Terdahulu

“Model Penjawab Pertanyaan Otomatis Berdasarkan Peringkat Relevansi Kalimat Menggunakan Model BERT”. Penelitian ini bertujuan untuk mengembangkan sistem yang unggul dalam menjawab pertanyaan dengan relevansi kalimat yang lebih akurat dan sesuai dengan konteks. Kontribusi utama dari penelitian ini adalah kemajuan dalam pemrosesan bahasa alami (NLP) dengan mengintegrasikan model BERT, dataset SQuAD, dan evaluasi kinerja menggunakan metrik yang ketat. Evaluasi menunjukkan bahwa F1 Score dan Exact Match adalah metrik penting dalam evaluasi kinerja model. F1 Score, dengan nilai 0.6, mencerminkan keseimbangan antara presisi dan recall, menunjukkan bahwa model dapat memberikan jawaban yang relevan. Exact Match, dengan nilai 0.5, mengukur seberapa sering jawaban model cocok secara tepat dengan jawaban yang benar, menunjukkan kemampuan model memberikan jawaban yang benar untuk setengah dari kasus uji.

“Implementasi Question Answering System Tafsir Al-Azhar Menggunakan Langchain Dan Large Language Model Berbasis Chatbot Telegram”. Penelitian ini bertujuan untuk membuat sebuah sistem yang mampu menerima pertanyaan seputar Tafsir Al-Azhar dalam bentuk bahasa alami, serta menjawab pertanyaan tersebut dengan kata-kata yang mudah dipahami oleh pengguna. Adapun teknologi yang akan digunakan pada penelitian ini seperti Langchain dan Large Language Model, untuk pengimplementasiannya menggunakan chatbot telegram, karena telegram merupakan aplikasi yang sudah terkenal dan sudah banyak digunakan oleh banyak orang, dan juga telegram mempunyai interface yang ramah, sehingga tidak membuat pengguna kebingungan ketika menggunakannya. Sistem Question Answering ini diuji menggunakan UAT (Unit Acceptance Test) dan framework DeepEval, hasil dari pengujian UAT berupa skor akurasi sebesar 83.71%, sedangkan hasil dari pengujian menggunakan framework DeepEval, berupa skor hallucination sebesar 41 %, contextual precision sebesar 90 % dan contextual relevancy 79 %.

“Aplikasi Chatbot untuk Layanan Informasi Penerimaan Mahasiswa Baru”. Penelitian ini bertujuan membuat aplikasi chatbot untuk layanan informasi penerimaan calon mahasiswa baru. Aplikasi ini untuk mengatasi keterbatasan jam kerja dalam merespon pertanyaan- pertanyaan yang diajukan oleh calon mahasiswa baru sehingga informasi yang dibutuhkan dapat dengan mudah diakses. Sistem ini menggunakan metode waterfall dan Natural Language Processing (NLP) dan Information Retrieval (IR). Evaluasi dengan responden menunjukkan 95% fitur yang ada pada halaman admin dan 97% fitur yang ada pada halaman user berfungsi dengan baik.

“Rancangan Dan Implementasi Chatbot Layanan Informasi Pendaftaran Pascasarjana Di Perguruan Tinggi”. Penelitian ini bertujuan memanfaatkan Teknologi Chatbot untuk Pendaftaran mahasiswa Pascasarjana dengan melakukan Perancangan yang dapat diimplementasikan dengan menggunakan tools Dialogflow dengan metode Natural Language Processing (NLP) dan WhatsApp API. Berdasarkan tahapan pengujian yang telah dilakukan menunjukkan bahwa hasilnya aplikasi chatbot mampu menjawab pertanyaan-pertanyaan terkait pendaftaran Pascasarjana dengan akurasi sebesar 97,82% dengan menggunakan data sebanyak 17 kalimat pertanyaan tentang pendaftaran Pascasarjana.

“Question Answering Al-Qur’an Menggunakan GPT-3.5 Berbasis Chatbot Telegram”. Penelitian ini mengembangkan chatbot berbasis Telegram untuk sistem tanya jawab (Question Answering System/QAS) terhadap Al-Qur’an menggunakan model GPT-3.5. Data Al-Qur’an dan terjemahan diolah menjadi embedding vektor menggunakan LangChain, yang kemudian disimpan dalam basis pengetahuan untuk pencarian relevansi (similarity search). LLM digunakan untuk menghasilkan jawaban terbaik berdasarkan konteks pertanyaan, dengan relevansi jawaban sebesar 98,3% dan tingkat halusinasi 22,5%. Chatbot ini memungkinkan pengguna mendapatkan jawaban yang akurat dan relevan, lengkap dengan referensi ayat. Penelitian ini menunjukkan efisiensi LLM dalam memahami pertanyaan bahasa alami manusia dan memberikan jawaban yang sesuai konteks keagamaan.

Tabel 2.1 Tabel Studi Literatur

Penelitian	Judul Perbandingan		
	Judul	Metode	Teknologi
(Sasongko Wibowo et al., 2024)	Model Penjawab Pertanyaan Otomatis Berdasarkan Peringkat Relevansi Kalimat Menggunakan Model BERT.	BERT	BERT dan dataset SQuAD
(Aji Bayu Permadi et al., 2024)	IMPLEMENTASI QUESTION ANSWERING SYSTEM TAFSIR AL-AZHAR MENGGUNAKAN LANGCHAIN DAN LARGE LANGUAGE MODEL BERBASIS CHATBOT TELEGRAM	LLM	FAISS, LangChain, Telegram Bot
(Novansyah Pratama et al., 2022)	Aplikasi Chatbot untuk Layanan Informasi Penerimaan Mahasiswa Baru	NLP & IR	SQL, Web-based Chatbot
(Ariestono et al., 2023)	RANCANGAN DAN IMPLEMENTASI CHATBOT LAYANAN INFORMASI PENDAFTARAN PASCASARJANA DI PERGURUAN TINGGI	NLP	Dialogflow, WhatsApp Messenger
(Elvino Dwi Saputra, 2024)	Question Answering Al-Qur'an Menggunakan GPT-3.5 Berbasis Chatbot Telegram	LLM	GPT-3.5, LangChain, Telegram Bot
(Anang Hardanto, 2025)	Implementasi Automated Question Answering Menggunakan AI dan N8N dalam Layanan Informasi PMB Politeknik Caltex Riau	NLP	GPT-3.5, Gemini, Deepseek, LangChain, Email

Berdasarkan tabel penelitian yang disajikan, terdapat lima penelitian terkait dengan sistem chatbot dan question answering yang menggunakan berbagai metode

dan teknologi. Penelitian pertama oleh Sasongko Wibowo et al. (2024) berfokus pada model pertanyaan berbasis relevansi kalimat menggunakan model BERT, dengan memanfaatkan teknologi BERT dan dataset SQuAD. Penelitian kedua dari Aji Bayu Permadi et al. (2024) mengimplementasikan sistem question answering tafsir Al-Azhar dengan menggunakan LangChain dan Large Language Model berbasis chatbot Telegram, memanfaatkan teknologi FAISS, LangChain, dan Telegram Bot.

Selanjutnya, Novransyah Pratama et al. (2022) mengembangkan aplikasi chatbot untuk layanan informasi penerimaan mahasiswa baru dengan pendekatan NLP dan IR, menggunakan teknologi SQL dan web-based chatbot. Penelitian keempat oleh Ariansono et al. (2023) merancang dan mengimplementasikan chatbot layanan informasi pendaftaran mahasiswa di perguruan tinggi menggunakan metode NLP dengan teknologi Dialogflow dan WhatsApp Messenger. Terakhir, Elvino Dwi Saputra (2024) mengembangkan question answering Al-Qur'an menggunakan GPT-3.5 dengan metode LLM, memanfaatkan teknologi GPT-3.5, LangChain, dan Telegram Bot. Kelima penelitian ini menunjukkan perkembangan dalam penggunaan teknologi kecerdasan buatan untuk sistem tanya-jawab dan layanan informasi otomatis dengan berbagai pendekatan dan platform implementasi.

Penelitian terakhir oleh Anang Hardanto (2025) mengimplementasikan automated question answering menggunakan AI dan N8N dalam layanan informasi PMB Politeknik Caltex Riau dengan pendekatan NLP, memanfaatkan teknologi GPT-3.5, Gemini, Deepseek, LangChain, dan Email. Keenam penelitian ini menunjukkan perkembangan dalam penggunaan teknologi kecerdasan buatan untuk sistem tanya-jawab dan layanan informasi otomatis dengan berbagai pendekatan, mulai dari model BERT, Large Language Model, hingga integrasi multi-platform seperti Telegram, WhatsApp, dan Email, serta penggunaan berbagai teknologi AI terkini seperti GPT-3.5, Gemini, dan Deepseek.

2.2 Landasan Teori

2.2.1 Profil Politeknik Caltex Riau (PCR)

Politeknik Caltex Riau (PCR) adalah institusi pendidikan tinggi vokasi yang berlokasi di Pekanbaru, Riau. Didirikan pada tahun 2001 oleh kerja sama antara PT Chevron Pacific Indonesia dan Pemerintah Provinsi Riau, PCR bertujuan untuk menyediakan pendidikan vokasional berkualitas tinggi dalam bidang teknologi dan bisnis. PCR memiliki reputasi sebagai salah satu politeknik terbaik di Indonesia, dengan fokus pada pembelajaran berbasis praktik dan kesiapan kerja lulusan.



Gambar 2.1 Logo Politeknik Caltex Riau

PCR memiliki fasilitas lengkap seperti laboratorium modern, ruang kelas berbasis teknologi, pusat inovasi, serta sistem informasi akademik digital. Lembaga ini juga aktif menjalin kerja sama dengan industri nasional dan internasional untuk memastikan bahwa kurikulumnya tetap relevan dan lulusannya siap bersaing di dunia kerja. Dengan tenaga pengajar yang kompeten dan berpengalaman di bidangnya, PCR terus berkomitmen untuk menjadi motor penggerak dalam mencetak tenaga profesional di bidang teknologi dan bisnis terapan.

2.2.2 Visi dan Misi Politeknik Caltex Riau (PCR)

Adapun visi dan misi dari Politeknik Caltex Riau adalah:

1) Visi:

Diakui sebagai politeknik unggul yang mampu bersaing dalam bidang teknologi dan bisnis terapan pada tingkat nasional maupun ASEAN tahun 2031.

2) Misi:

- a. Menyelenggarakan sistem pendidikan vokasi bidang teknologi dan bisnis yang berkualitas serta relevan dengan tantangan nasional maupun ASEAN.
- b. Menciptakan budaya akademik dan budaya organisasi yang berkarakter dan bermartabat.

- c. Melaksanakan penelitian dan menyebarluaskan hasilnya untuk pengembangan bidang teknologi dan bisnis terapan.
- d. Melaksanakan pengabdian kepada masyarakat dengan menyebarluaskan ilmu pengetahuan, teknologi, dan budaya organisasi.

2.2.3 Penerimaan Mahasiswa Baru (PMB)

Penerimaan Mahasiswa Baru (PMB) merupakan salah satu proses penting dalam sistem pendidikan tinggi yang berfungsi sebagai gerbang awal bagi calon mahasiswa untuk memasuki jenjang pendidikan di perguruan tinggi. PMB mencakup berbagai tahapan mulai dari sosialisasi, pendaftaran, seleksi, hingga registrasi ulang mahasiswa yang diterima. Proses ini dirancang untuk menjaring individu yang memiliki potensi dan minat belajar sesuai dengan program studi yang ditawarkan oleh institusi. Melalui proses ini, PCR dapat menyaring calon mahasiswa yang selaras dengan visi dan misi institusi, yakni menghasilkan lulusan yang unggul dalam bidang teknologi dan bisnis terapan, serta mampu bersaing di tingkat nasional dan ASEAN.

Proses PMB di PCR dilakukan secara sistematis dan terstruktur, dengan memanfaatkan teknologi informasi untuk meningkatkan efisiensi dan transparansi. Sistem informasi PMB yang berbasis online memungkinkan calon mahasiswa dari berbagai daerah untuk mengikuti proses pendaftaran secara mudah dan cepat. Hal ini sejalan dengan komitmen PCR untuk memberikan pelayanan pendidikan yang modern dan responsif terhadap kebutuhan zaman.

2.2.4 *Natural Language Processing* (NLP)

Natural Language Processing (NLP) adalah cabang dari kecerdasan buatan (*Artificial Intelligence*) yang berfokus pada interaksi antara komputer dan bahasa manusia alami. Tujuan utama NLP adalah untuk memungkinkan komputer memahami, menafsirkan, menghasilkan, dan merespons teks atau ucapan dalam bahasa yang digunakan manusia sehari-hari (Faturohman & Arifin, 2024). Proses ini digunakan dalam beragam aplikasi, mulai dari sistem terjemahan otomatis,

chatbot, deteksi sentimen, pencarian cerdas, hingga sistem pertanyaan dan jawaban otomatis (*question answering systems*).

Perkembangan NLP mengalami akselerasi signifikan sejak munculnya teknik berbasis deep learning, terutama dengan diperkenalkannya model transformer seperti *BERT* dan *GPT*, yang memungkinkan analisis konteks kata secara lebih akurat dan efisien. Model-model ini dilatih pada data skala besar dan mampu mempelajari representasi semantik kata dan kalimat dalam berbagai konteks. Dengan kemajuan NLP, mesin kini dapat lebih baik dalam memahami bahasa manusia, yang sangat krusial dalam pengembangan teknologi seperti *chatbot*, *voice assistant*, dan sistem informasi berbasis teks.

Sebelum teks dianalisis oleh model NLP, tahap penting yang harus dilakukan adalah pre-processing atau pra-pemrosesan teks. Pre-processing bertujuan untuk membersihkan data teks dari elemen-elemen yang tidak relevan agar hasil analisis menjadi lebih akurat dan efisien. Langkah-langkah umum dalam tahap ini meliputi case folding (mengubah semua huruf menjadi huruf kecil), tokenisasi (memecah kalimat menjadi kata-kata), stopword removal (menghapus kata-kata umum yang tidak bermakna penting), stemming (mengembalikan kata ke bentuk dasarnya), dan normalisasi (menyeragamkan variasi penulisan kata). Proses ini tidak hanya meningkatkan kualitas data, tetapi juga mengurangi dimensi data dan mempercepat kinerja algoritma NLP. Menurut Alyami dan Alzahrani (2020), pemilihan dan penerapan teknik pra-pemrosesan yang tepat sangat berpengaruh terhadap performa model, khususnya dalam tugas seperti analisis sentimen atau klasifikasi teks.

2.2.5 *Automated Question Answering (AQA)*

Automated Question Answering (AQA) merupakan cabang dari kecerdasan buatan yang bertujuan untuk mengembangkan sistem yang mampu memahami dan merespons pertanyaan secara otomatis dengan jawaban yang relevan dan tepat. Sistem AQA menggabungkan pendekatan dari pemrosesan bahasa alami (*Natural Language Processing/NLP*), pengambilan informasi (*Information Retrieval*), dan pembelajaran mesin (*Machine Learning*) untuk menganalisis pertanyaan pengguna dan menghasilkan jawaban dari sumber data yang tersedia, baik dalam bentuk

terstruktur maupun tidak terstruktur (Fernando, Kristanto, & Anditya, 2023). Dalam pendekatan berbasis NLP modern, sistem AQA memanfaatkan model pre-trained seperti BERT atau GPT untuk memahami konteks semantik dari pertanyaan dan dokumen sumber, sehingga meningkatkan akurasi dan relevansi jawaban. Dalam praktiknya, AQA digunakan dalam berbagai aplikasi seperti chatbot, sistem pencarian cerdas, asisten digital, dan layanan pelanggan otomatis yang semakin populer karena efisiensi dan skalabilitasnya. Dengan demikian, AQA berperan penting dalam meningkatkan akses informasi secara cepat dan efisien di era digital saat ini.

2.2.6 Artificial Intelligence (AI)

Menurut (Nur & Sahyuni, 2025) Kecerdasan buatan atau lebih dikenal dengan *Artificial Intelligence* (AI) adalah program komputer yang dirancang dan dibangun untuk dapat meniru kecerdasan manusia, termasuk kemampuan pengambilan keputusan, logika, dan karakteristik kecerdasan lainnya. AI digunakan untuk mengotomatisasi pengambilan keputusan, pengenalan pola, hingga interaksi dengan pengguna. Dalam konteks proyek akhir ini, AI digunakan untuk memahami pertanyaan dan memberikan jawaban secara otomatis melalui sistem QA (*Question Answering*).

2.2.7 N8N



Gambar 2.2 Logo N8N

N8N (dibaca: "n-eight-n") adalah *platform* otomatisasi alur kerja (*workflow automation*) berbasis *source-available* yang memungkinkan pengguna untuk mengintegrasikan berbagai aplikasi, layanan *web*, dan proses bisnis tanpa perlu menulis kode dalam jumlah besar. Dengan antarmuka visual berbasis node, N8N

mendukung pendekatan *low-code*, di mana pengguna dapat membangun sistem yang kompleks melalui penyambungan blok fungsional seperti *API calls*, transformasi data, dan pemrosesan logika bisnis (n8n.io Docs, 2023). N8N dirancang *modular* dan *extensible*, memungkinkan integrasi lebih dari 300 aplikasi termasuk *database*, *platform cloud*, CRM, dan layanan notifikasi, serta mendukung eksekusi skrip kustom melalui *node Function* berbasis JavaScript (n8n Community, 2023). Dengan arsitektur alur kerja yang dapat dijadwalkan atau dipicu oleh peristiwa tertentu (*webhook*), n8n telah menjadi alat yang populer dalam membangun sistem integrasi otomatis di berbagai sektor, termasuk bisnis, IT, dan riset data.

2.2.8 Langchain

LangChain adalah *framework open-source* yang dikembangkan oleh Harrison Chase pada tahun 2022 untuk membangun aplikasi berbasis *Large Language Models (LLMs)* dengan pendekatan modular dan terintegrasi. *Framework* ini mengatasi keterbatasan fundamental dari model bahasa standar yang hanya mengandalkan dataset pelatihan tetap dengan menyediakan kemampuan untuk mengintegrasikan sumber data eksternal, basis pengetahuan *real-time*, dan alur kerja otomatis (Rizki et al., 2024). *LangChain* menggunakan arsitektur modular yang terdiri dari komponen-komponen seperti *chains* untuk sekuens operasi, *agents* untuk *decision-making*, *memory management* untuk konteks percakapan, dan dukungan *Retrieval-Augmented Generation (RAG)* yang memungkinkan model mengakses informasi dari dokumen eksternal untuk menghasilkan respons yang lebih akurat dan kontekstual. Dalam implementasi sistem *Question Answering*, *LangChain* telah terbukti efektif dalam memfasilitasi pembangunan chatbot dengan tingkat akurasi yang tinggi, seperti yang ditunjukkan dalam penelitian (Lubis et al., 2024) yang mencapai akurasi 86% dalam sistem *Question Answering* berbasis Telegram menggunakan *LangChain* dan *LLM*. *Framework* ini menyediakan tools seperti *document loaders*, *text splitters*, *vector stores*, *retrieval chains*, dan *question answering chains* yang memungkinkan pembangunan sistem QA yang dapat memahami pertanyaan kompleks dan

menghasilkan jawaban berdasarkan informasi yang diambil dari basis pengetahuan eksternal.

2.2.9 Gemini

Gemini adalah nama dari generasi terbaru *Large Language Model* (LLM) yang dikembangkan oleh Google DeepMind sebagai bagian dari evolusi sistem kecerdasan buatan multimodal. Gemini dirancang untuk menggabungkan kemampuan pemrosesan teks, gambar, audio, dan kode ke dalam satu arsitektur terpadu, dengan kemampuan penalaran dan pemahaman konteks yang lebih dalam dibandingkan model sebelumnya seperti Bard atau PaLM 2. Gemini juga merupakan upaya Google dalam merespons perkembangan pesat model bahasa besar seperti GPT dari OpenAI, dengan menekankan pada kemampuan integratif dan efisiensi performa lintas media (Shen et al., 2023).

Secara teknis, Gemini dibangun dengan pendekatan *transformer-based architecture* dan mendukung *multimodal learning*, yang memungkinkan sistem memahami konteks kompleks dari berbagai jenis input data. Dalam implementasinya, Gemini digunakan dalam berbagai aplikasi seperti asisten cerdas, sistem *question answering*, pengembangan kode otomatis, hingga sistem pemrosesan visual berbasis AI. Salah satu fitur unggulan Gemini adalah kemampuannya untuk mengintegrasikan berbagai domain data ke dalam satu pipeline inferensi, yang secara signifikan meningkatkan relevansi, konsistensi, dan efisiensi dalam menjawab permintaan pengguna yang kompleks. Dengan pendekatan ini, Gemini mewakili langkah besar dalam pengembangan AI yang lebih terintegrasi dan adaptif terhadap kebutuhan komunikasi manusia yang multimodal.

2.2.10 Deepseek

DeepSeek adalah proyek open-source yang mengembangkan sistem *Large Language Model* (LLM) berbasis arsitektur transformer dengan kemampuan tinggi dalam pemrosesan teks dan kode secara bersamaan. Model ini dirancang dengan prinsip keterbukaan, efisiensi, dan kemampuan adaptasi terhadap kebutuhan industri maupun akademik. DeepSeek pertama kali diperkenalkan sebagai bagian

dari tren pengembangan model bahasa besar yang mendukung pembelajaran multitugas (*multi-task learning*), yang memungkinkan model tidak hanya memahami bahasa alami tetapi juga mampu melakukan pemrograman otomatis dan pemrosesan instruksi kompleks (DeepSeek AI, 2024).

DeepSeek dibangun di atas *decoding-optimized architecture* dan dilatih menggunakan data dalam jumlah besar yang mencakup bahasa alami dan kode pemrograman dari berbagai bahasa seperti Python, JavaScript, dan C++. Salah satu fitur menonjol dari DeepSeek adalah kemampuannya dalam *multi-domain inference*, yakni menafsirkan dan menanggapi perintah pengguna dalam bentuk narasi, instruksi teknis, maupun sintaksis kode. Proyek ini juga berkontribusi pada ekosistem AI terbuka karena model dan data pelatihannya dipublikasikan melalui GitHub dan Hugging Face, sehingga dapat diakses dan diadaptasi secara bebas oleh komunitas riset dan pengembang industri (DeepSeek AI GitHub, 2024). DeepSeek juga menawarkan varian seperti **DeepSeek Coder** untuk tugas-tugas *code generation*, *bug fixing*, dan *code explanation* berbasis LLM.

2.2.11 Gmail



Gambar 2.3 Logo Gmail

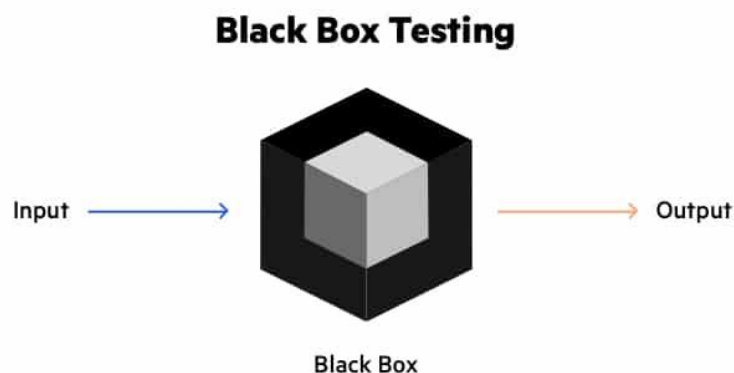
Gmail adalah layanan surat elektronik (*email*) berbasis web yang dikembangkan oleh Google dan secara resmi diluncurkan ke publik pada tahun 2004. Gmail merupakan salah satu platform komunikasi digital paling dominan di dunia, dengan lebih dari 1,8 miliar pengguna aktif global (Google, 2023). Layanan ini menawarkan penyimpanan cloud gratis, pencarian email berbasis algoritma, pengelompokan percakapan (*threaded conversation*), serta integrasi yang erat

dengan berbagai produk Google seperti Google Calendar, Google Drive, dan Google Workspace.

Secara teknis, Gmail beroperasi di atas protokol standar komunikasi email seperti IMAP, POP3, dan SMTP, yang memungkinkannya digunakan baik melalui antarmuka web maupun klien email lainnya. Gmail juga mendukung penerapan API melalui Google Workspace API, yang memungkinkan pengembang membangun sistem otomatisasi seperti pengiriman email notifikasi, sinkronisasi data pengguna, atau integrasi ke dalam sistem chatbot dan aplikasi web. Dengan menggunakan Gmail API, sistem eksternal dapat mengakses kotak masuk, mengirim, dan mengelola email pengguna secara terprogram, selama pengguna memberikan izin melalui protokol otorisasi OAuth 2.0 (Google Developers, 2024).

Dalam konteks pengembangan sistem informasi, Gmail sering digunakan sebagai *endpoint* dalam pengiriman informasi otomatis karena keandalannya, skalabilitasnya, dan kompatibilitasnya dengan sistem berbasis cloud dan IoT. Selain itu, Gmail juga memiliki fitur keamanan tinggi, termasuk deteksi spam berbasis AI dan sistem otentikasi dua langkah (*two-step verification*), yang menjadikannya pilihan utama untuk komunikasi digital yang aman dan efisien.

2.2.12 Black Box Testing



Gambar 2.4 Black Box Testing

Black Box Testing adalah metode pengujian yang bertujuan untuk memastikan bahwa semua fungsi dalam perangkat lunak bekerja sebagaimana mestinya sesuai dengan kebutuhan fungsional yang telah ditetapkan (Puspitasari & Hidayat, 2021). Penguji hanya mengevaluasi input dan output dari sistem

berdasarkan spesifikasi yang telah ditentukan. Tujuan utama dari Black Box Testing adalah untuk memastikan dan mengoptimalkan pembangunan sebuah perangkat lunak atau aplikasi sebelum diserahkan kepada user untuk digunakan (Wijaya et al., 2023). Pengujian ini digunakan untuk memastikan bahwa setiap workflow dalam sistem bekerja sesuai.

2.2.13 BERTScore

BERTScore adalah metode evaluasi berbasis model transformer yang digunakan untuk menilai kesamaan semantik antara dua teks. Metode ini diperkenalkan sebagai alternatif yang lebih akurat dibandingkan metrik tradisional seperti BLEU dan ROUGE, yang hanya membandingkan kesamaan kata secara langsung. BERTScore bekerja dengan mengubah teks menjadi representasi vektor menggunakan model bahasa seperti BERT, RoBERTa, atau ALBERT, kemudian mengukur kesamaan antara vektor kata dari teks prediksi dan teks referensi menggunakan cosine similarity.

Pendekatan ini memungkinkan BERTScore untuk memahami hubungan semantik antar kata, termasuk sinonim dan variasi frasa, sehingga lebih efektif dalam menilai kualitas teks yang dihasilkan oleh model bahasa. Hasil evaluasi BERTScore biasanya dilaporkan dalam tiga metrik utama, yaitu *Precision*, *Recall*, dan *F1 Score*, yang masing-masing mengukur kesesuaian, kelengkapan, dan keseimbangan antara jawaban chatbot dan referensi. Dengan kemampuannya dalam menangkap makna teks secara lebih mendalam, BERTScore menjadi salah satu standar evaluasi yang banyak digunakan dalam pengujian model berbasis kecerdasan buatan, termasuk AQA dan sistem pemrosesan bahasa alami lainnya. Berikut ini merupakan perhitungan dari setiap nilai yang digunakan dalam BERTScore:

Rumus pertama menghitung nilai Recall dalam BERTScore. Recall ini mengukur sejauh mana kata-kata dalam teks referensi xxx dicocokkan dengan kata-kata dalam teks prediksi. Perhitungan ini mempertimbangkan bobot invers frekuensi dokumen (IDF) untuk memberikan prioritas lebih tinggi pada kata-kata yang lebih jarang muncul dalam dataset, serta memilih nilai kesamaan tertinggi

antara representasi vektor kata dari teks prediksi dan referensi menggunakan dot product.

$$R_{BERT} = \frac{\sum_{x_i \in x} \text{idf}(x_i) \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j}{\sum_{x_i \in x} \text{idf}(x_i)}$$

Gambar 2. 5 Perhitungan Nilai Recall BERTScore

Rumus kedua menghitung nilai Precision dalam BERTScore. Precision ini mengukur sejauh mana kata-kata dalam teks prediksi sesuai dengan kata-kata dalam teks referensi. Seperti halnya Recall, Precision juga mempertimbangkan bobot IDF dan menggunakan cosine similarity dari representasi vektor kata.

$$P_{BERT} = \frac{\sum_{\hat{x}_j \in \hat{x}} \text{idf}(\hat{x}_j) \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j}{\sum_{\hat{x}_j \in \hat{x}} \text{idf}(\hat{x}_j)}$$

Gambar 2. 6 Perhitungan Nilai Precision BERTScore

Rumus ketiga adalah perhitungan F1 Score yang merupakan rata-rata harmonis dari Precision dan Recall. F1 Score digunakan untuk memberikan keseimbangan antara kedua metrik tersebut, sehingga memberikan gambaran menyeluruh tentang kualitas kesamaan semantik antara teks prediksi dan referensi.

$$F_{BERT} = 2 \cdot \frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}$$

Gambar 2. 7 Perhitungan Nilai F1 Score BERTScore