PENGEMBANGAN CHATBOT SEBAGAI PENYEDIA INFORMASI MENGENAI UNIVERSITAS BINA NUSANTARA

A KNOWLEDGE-BASED CHATBOT FOR PROVIDING INFORMATION ABOUT BINA NUSANTARA UNIVERSITY

Topik: Text Mining

2501975684 Arvio Anandi 087764260779

2540122363 Matheus Prayogo 083871898657

2540133783 Steven Halim 085234566899



BINUS University

2024

Diperiksa oleh,

D5339 – Lili Ayu Wulandhari, S.Si., M.Sc., Ph.D

# DAFTAR ISI

[DAFTAR ISI 2](#_Toc185642356)

[BAB 1 PENDAHULUAN 4](#_Toc185642357)

[1.1 Latar Belakang 4](#_Toc185642358)

[1.2 Rumusan Masalah 5](#_Toc185642359)

[1.3 Tujuan & Manfaat 5](#_Toc185642360)

[1.4 Ruang Lingkup 6](#_Toc185642361)

[1.5 Metodologi 6](#_Toc185642362)

[1.5.1 Metode Pengumpulan Data 6](#_Toc185642363)

[1.5.2 Metode Perancangan/Pengembangan 6](#_Toc185642364)

[BAB 2 TINJAUAN REFERENSI 7](#_Toc185642365)

[2.1 Universitas Bina Nusantara 7](#_Toc185642366)

[2.2 Artificial Intelligence 8](#_Toc185642367)

[2.3 Chatbot 8](#_Toc185642368)

[2.4 Preprocessing 8](#_Toc185642369)

[2.5 Text Embedding 10](#_Toc185642370)

[2.6 ANN 11](#_Toc185642371)

[2.7 LSTM 12](#_Toc185642372)

[2.8 BERT 14](#_Toc185642373)

[2.9 Python 17](#_Toc185642374)

[2.10 Streamlit 17](#_Toc185642375)

[2.11 FastAPI 17](#_Toc185642376)

[2.12 RAG 18](#_Toc185642377)

[2.13 FAISS 18](#_Toc185642378)

[2.14 Performance Evaluation 19](#_Toc185642379)

[2.15 Database 20](#_Toc185642380)

[2.16 Indexing 20](#_Toc185642381)

[2.17 Typo Corrector 21](#_Toc185642382)

[2.18 Named Entity Recognition (NER) 23](#_Toc185642383)

[2.19 Related Works 23](#_Toc185642384)

[BAB 3 METODE PENELITIAN 42](#_Toc185642385)

[3.1 Kerangka Berpikir 42](#_Toc185642386)

[3.2 Metode Penelitian 44](#_Toc185642387)

[3.3 Perancangan UML 52](#_Toc185642388)

[3.3.1 Prompt Question 52](#_Toc185642389)

[3.3.2 Give Feedback 52](#_Toc185642390)

[3.3.3 Get Help 52](#_Toc185642391)

[3.4 Perancangan UI 53](#_Toc185642392)

[BAB 4 HASIL DAN PEMBAHASAN 55](#_Toc185642393)

[4.1 Deskripsi 55](#_Toc185642394)

[4.2 Penggunaan Chatbot 55](#_Toc185642395)

[4.3 Pengujian Model Chatbot 56](#_Toc185642396)

[4.3.1 Test Case 56](#_Toc185642397)

[4.3.2 Performa Model 56](#_Toc185642398)

[4.3.3 Contoh Output 58](#_Toc185642399)

[4.4 Diskusi 59](#_Toc185642400)

[BAB 5 KESIMPULAN DAN SARAN 60](#_Toc185642401)

[5.1 Kesimpulan 60](#_Toc185642402)

[5.2 Saran 61](#_Toc185642403)

[DAFTAR PUSTAKA 62](#_Toc185642404)

[LAMPIRAN 63](#_Toc185642405)

# BAB 1 PENDAHULUAN

## 1.1 Latar Belakang

Di tengah perkembangan pendidikan yang terus berubah, universitas semakin memanfaatkan teknologi untuk menyederhanakan dan meningkatkan berbagai aspek kehidupan mahasiswa. Salah satu area signifikan di mana teknologi dapat memainkan peran penting adalah dalam menyediakan informasi instan dan akurat kepada mahasiswa. Seiring dengan pertumbuhan ukuran dan kompleksitas universitas, kebutuhan akan sistem yang dapat diandalkan dan efisien untuk menanggapi pertanyaan mahasiswa menjadi sangat penting. Penelitian ini bertujuan untuk berkontribusi pada pengembangan layanan dukungan mahasiswa melalui pengembangan chatbot khusus yang difokuskan untuk menjawab pertanyaan terkait informasi universitas.

Universitas menghasilkan sejumlah besar informasi mulai dari program akademik dan rincian kursus hingga fasilitas kampus dan proses administratif. Mahasiswa, terutama mereka yang baru bergabung, sering merasa sulit untuk menjelajahi banyak informasi ini. Metode tradisional dalam mencari informasi, seperti mencari di situs web atau menghubungi kantor administrasi, bisa memakan waktu dan kadang tidak efisien. Chatbot yang dilengkapi dengan kemampuan pemrosesan bahasa alami menjadi solusi yang menjanjikan untuk mengatasi tantangan ini dengan memberikan jawaban yang cepat dan akurat terhadap pertanyaan mahasiswa.

Universitas saat ini menghadapi kesulitan dalam berkomunikasi secara efektif dengan berbagai kelompok audiens, termasuk calon/mahasiswa aktif, orang tua, fakultas, staf, alumni, dan lainnya, untuk menyampaikan informasi (Atmauswan, 2022). Jumlah pertanyaan yang diajukan oleh calon mahasiswa dapat memberikan dampak pada staff administrasi karena mereka sering ditanya pertanyaan umum mengenai masalah atau kendala dari calon mahasiswa. Hal ini dapat meningkatkan beban kerja staff administrasi karena mereka harus menjawab pertanyaan umum secara berulang-ulang (Nguyen, 2021). Oleh sebab itu peranan chatbot menjadi sangat penting pada universitas untuk menjawab pertanyaan yang bersifat umum dan berulang.

Dalam penelitian sebelumnya, banyak chatbot dalam dunia pendidikan telah difokuskan pada dua aspek utama: pertama, membantu pelajar dalam memahami materi pembelajaran dengan memberikan penjelasan, latihan, dan bimbingan secara interaktif (Kuhail, 2022); kedua, berperan sebagai teman atau pendamping bagi pelajar, menyediakan dukungan emosional, motivasi, atau bahkan hiburan (Nee, 2022).

Chatbot yang menyediakan informasi kampus memiliki kepentingan khusus karena dapat mengatasi kebutuhan mendesak pelajar terhadap akses cepat dan mudah ke informasi penting. Informasi yang tepat dan mudah diakses, seperti jadwal kuliah, lokasi kampus, atau prosedur administratif, menjadi krusial dalam pengalaman akademik mereka. Dengan chatbot yang dirancang yang memiliki knowledge spesifik mengenai hal-hal tersebut, pelajar dapat dengan mudah mengakses informasi kampus yang relevan dan tepat serta menghemat waktu dan upaya dalam pencarian informasi (Labadze, 2023).

**/TODO**

Berdasarkan penelitian terdahulu, terlihat bahwa penggunaan chatbot di dunia akademis bermanfaat bagi mahasiswa maupun calon mahasiswa suatu universitas, oleh sebab itu maka pada penelitian ini diajukan pengembangan chatbot khusus untuk Universitas Bina Nusantara yang akan bermanfaat dalam memberikan informasi mengenai kegiatan akademis di universitas tersebut.

## 1.2 Rumusan Masalah

1. Bagaimana mempersiapkan *data knowledge* yang berkaitan dengan Universitas Bina Nusantara
2. Bagaimana mengembangkan chatbot untuk mengintegrasikan pengetahuan tentang Universitas Bina Nusantara
3. Bagaimana mengimplementasikan model chatbot dalam platform online

## 1.3 Tujuan & Manfaat

Tujuan dari penelitian ini adalah sebagai berikut:

1. Mempersiapkan *data knowledge* yang berkaitan dengan Universitas Bina Nusantara
2. Mengembangkan chatbot untuk mengintegrasikan pengetahuan tentang Universitas Bina Nusantara menggunakan Large Language Model (LLM)
3. Mengimplementasi model chatbot dalam platform online menggunakan Streamlit

Manfaat dari penelitian ini adalah sebagai berikut:

1. Memudahkan pencarian informasi tentang Universitas Bina Nusantara: Pengembangan chatbot ini bertujuan untuk memberikan kemudahan akses informasi bagi mahasiswa, calon mahasiswa, dan masyarakat umum terkait berbagai aspek dari Universitas Bina Nusantara.
2. Meringankan beban kerja staf yang bertanggung jawab dalam menjawab pertanyaan khalayak umum tentang Universitas Bina Nusantara: Dengan adanya chatbot, beban kerja staf administrasi dalam menjawab pertanyaan berulang dapat dikurangi sehingga mereka dapat fokus pada tugas-tugas lain yang lebih strategis.
3. Pengembangan teknik dalam pembuatan model chatbot di Universitas Bina Nusantara: Penelitian ini berkontribusi pada pengembangan metode dan teknik baru dalam pembuatan model chatbot khusus untuk kebutuhan universitas.
4. Pengembangan arsitektur RAG agar lebih sesuai dengan kebutuhan universitas: Penelitian ini juga mengarah pada peningkatan dan penyesuaian arsitektur RAG sehingga lebih relevan dan efektif dalam konteks universitas.

Adapun kontribusi dari penelitian ini, adalah:

1. Kontribusi Praktis:
   1. Memudahkan pencarian informasi tentang Universitas Bina Nusantara: Pengembangan chatbot ini bertujuan untuk memberikan kemudahan akses informasi bagi mahasiswa, calon mahasiswa, dan masyarakat umum terkait berbagai aspek dari Universitas Bina Nusantara.
   2. Meringankan beban kerja staf yang bertanggung jawab dalam menjawab pertanyaan khalayak umum tentang Universitas Bina Nusantara: Dengan adanya chatbot, beban kerja staf administrasi dalam menjawab pertanyaan berulang dapat dikurangi sehingga mereka dapat fokus pada tugas-tugas lain yang lebih strategis.
2. Kontribusi Pengetahuan:
   1. Pengembangan teknik dalam pembuatan model chatbot di Universitas Bina Nusantara: Penelitian ini berkontribusi pada pengembangan metode dan teknik baru dalam pembuatan model chatbot khusus untuk kebutuhan universitas.
   2. Pengembangan arsitektur RAG agar lebih sesuai dengan kebutuhan universitas: Penelitian ini juga mengarah pada peningkatan dan penyesuaian arsitektur RAG sehingga lebih relevan dan efektif dalam konteks universitas.

## 1.4 Ruang Lingkup

1. Chatbot khusus bertujuan memenuhi kebutuhan informasi dari berbagai pemangku kepentingan di Universitas Bina Nusantara, termasuk calon dan mahasiswa aktif, orang tua, fakultas, staf, dan alumni.
2. Penyediaan informasi komprehensif terbatas pada topik-topik yang relevan dengan konteks universitas, seperti program akademik, kegiatan mahasiswa (organisasi/klub), akomodasi, prosedur penerimaan, rincian kursus, layanan dan fasilitas kampus, proses administratif, dan kebijakan umum universitas.
3. Chatbot terbatas untuk hanya memberikan informasi terkait Universitas Bina Nusantara, tanpa menyediakan informasi di luar cakupan topik terkait universitas.

## 1.5 Metodologi

### 1.5.1 Metode Pengumpulan Data

Metode pengumpulan data dalam pengembangan chatbot ini melibatkan dua tahap utama: pertama, pengumpulan data mengenai Universitas Bina Nusantara yang meliputi informasi terkait jurusan, mata kuliah, organisasi kemahasiswaan, biaya, dosen, serta alumni. Data ini akan diperoleh melalui berbagai sumber seperti situs web resmi universitas, dan dokumen resmi. Kedua, pengumpulan data untuk proses training model, seperti pertanyaan-pertanyaan yang mungkin diajukan oleh pengguna chatbot dan jawaban yang sesuai.

### 1.5.2 Metode Perancangan dan Pengembangan

Proses perancangan dan pengembangan chatbot ini akan melibatkan beberapa langkah:

1. Data yang telah dikumpulkan akan diproses dan dipersiapkan untuk digunakan dalam pelatihan model, termasuk proses pembersihan data dan pra-pemrosesan.
2. Text embedding menggunakan pre-trained model BERT untuk mendapatkan representasi vektor dari teks.
3. Data akan diindeks menggunakan teknik HNSW (Hierarchical Navigable Small World) untuk memungkinkan pencarian informasi yang cepat dan akurat.
4. Database akan dirancang dan diimplementasikan, terdiri dari Vector Database untuk menyimpan representasi vektor dari data dan Multimedia Database untuk menyimpan gambar atau data multimedia lainnya.
5. Typo corrector akan diterapkan untuk meningkatkan kualitas interaksi antara chatbot dan pengguna dengan mendeteksi dan memperbaiki kesalahan pengetikan dalam teks.
6. Model Question Answering (QA) yang telah dilatih akan diimplementasikan menggunakan arsitektur RAG (Retrieval-Augmented Generation) yang memungkinkan chatbot untuk menjawab pertanyaan pengguna berdasarkan informasi yang tersedia dalam database.

# BAB 2 TINJAUAN REFERENSI

## 2.1 Universitas Bina Nusantara

Universitas Bina Nusantara (BINUS University) didirikan pada tahun 1974 dan telah menjadi salah satu universitas swasta terkemuka di Indonesia. Dengan kampus utama yang terletak di Kemanggisan, Jakarta Barat, serta beberapa lokasi lain di Tangerang, Bekasi, Bandung, dan Malang, BINUS telah menjangkau berbagai kota besar di Indonesia.



Gambar 2.1 Kampus Anggrek (binus.ac.id)

BINUS University menawarkan berbagai program sarjana dan pascasarjana di berbagai bidang, termasuk bisnis, ilmu komputer, teknik, humaniora, ilmu sosial, seni, dan desain. Kurikulumnya selalu diperbarui sesuai dengan kebutuhan industri dan teknologi yang berkembang pesat, memastikan lulusan BINUS siap bersaing di pasar kerja yang kompetitif.

Selain fokus pada akademik, BINUS juga aktif dalam penelitian dan pengembangan, menjalin kemitraan dengan industri dan institusi lain untuk mendorong inovasi. Kegiatan mahasiswa juga menjadi aspek penting, dengan berbagai organisasi dan klub yang aktif dalam berbagai kegiatan akademik, seni, olahraga, dan sosial.

## 2.2 Artificial Intelligence

Artificial intelligence (AI) adalah kemampuan sebuah mesin seperti sistem komputer untuk melakukan aktivitas yang biasanya hanya bisa dilakukan oleh manusia karena tugas tersebut membutuhkan kemampuan berpikir dan pengambilan keputusan (Copeland, 2024). Tujuan utama dari AI adalah membuat mesin atau program komputer yang mampu memproses informasi, belajar dari pengalaman, dan mengadaptasi diri untuk menyelesaikan tugas-tugas tertentu dengan lebih efisien.

Semakin berkembangnya teknologi, AI juga semakin digunakan dan manfaat yang diberikan dari AI dapat dilihat dalam hal-hal seperti: proses automasi, repetitive learning, analisa data dengan akurasi yang tinggi. Di zaman sekarang, AI banyak digunakan dalam kegiatan sehari-hari, contohnya seperti Google search, voice assistant seperti Siri dan Alexa, algoritma advertising dan rekomendasi di sosial media, chatbots, dan lain-lain.

## 2.3 Chatbot

Chatbot adalah jenis aplikasi AI yang dirancang untuk berinteraksi dengan pengguna melalui percakapan atau chat. Tujuan utama chatbot adalah memberikan respon yang terarah dan relevan berdasarkan pertanyaan atau perintah yang diberikan oleh pengguna. Dengan menggunakan teknologi *Natural Language Processing* (NLP) dan *machine learning*, chatbot dapat memberikan pengalaman interaksi yang semakin mirip dengan percakapan manusia. (Labadze et al., 2023)

Penggunaan chatbot dalam bidang akademik semakin berkembang pesat, terutama dalam menjawab pertanyaan mahasiswa dan membantu proses pembelajaran. Chatbot dapat digunakan untuk menjawab pertanyaan umum terkait administrasi, jadwal kuliah, dan materi pelajaran. Dalam lingkungan pendidikan, chatbot membantu mengurangi beban kerja staf dengan menangani pertanyaan yang sering diajukan oleh mahasiswa, sehingga staf dapat fokus pada tugas-tugas yang lebih kompleks dan membutuhkan interaksi manusia.

## 2.4 Preprocessing Teks

Sebelum mengolah data teks, perlu dilakukan pemrosesan kata (preprocessing). Tujuan dari pemrosesan ini adalah mengubah teks menjadi sesuatu yang dapat dimengerti oleh komputer yaitu data berupa angka. langkah dari preprocessing ini adalah:

1. Tokenisasi: Data teks dipecah menjadi unit-unit yang lebih kecil, seperti kata-kata atau sub-kata, yang disebut token. Tokenisasi memungkinkan model untuk memahami struktur dan makna dari teks yang kompleks. Contoh tokenisator yang umum digunakan adalah WordPiece dan Byte-Pair Encoding (BPE).

Contoh Tokenisasi:

text = 'We are looking forward to hiking on Everest Mountain next year'

words = word\_tokenize(text)

print(words)

# ['We', 'are', looking', 'forward', to', 'hiking', 'on', 'Everest', Mountain', next', 'year']

1. Pembersihan Teks: Tahap ini melibatkan penghapusan karakter khusus, tanda baca, dan token yang tidak relevan atau tidak diinginkan, seperti URL, emotikon, atau tanda baca.
2. Normalisasi: Normalisasi teks dilakukan untuk memastikan konsistensi dalam representasi kata-kata, seperti mengubah semua huruf menjadi huruf kecil atau menggantikan variasi kata-kata yang mirip dengan satu bentuk standar.
3. Tokenisasi Khusus: Beberapa model memerlukan tokenisasi khusus untuk tanda-tanda awal dan akhir kalimat, atau tokenisasi untuk token khusus yang menandakan awal dan akhir teks.
4. Padding dan Truncation: Data teks sering memiliki panjang yang berbeda-beda. Untuk memproses batch data secara efisien, seringkali diperlukan padding (penambahan token khusus) untuk membuat semua sampel dalam batch memiliki panjang yang sama. Sebaliknya, jika data teks terlalu panjang, dilakukan truncation (pemotongan) untuk memastikan semua sampel memiliki panjang yang seragam.
5. Encoding: Setiap token direpresentasikan dalam bentuk vektor numerik. Proses ini disebut encoding, dan representasi vektor yang dihasilkan seringkali disesuaikan dengan model yang digunakan, misalnya menggunakan WordPiece atau BPE encoding.
6. Membangun Input Sequences: Data teks kemudian dibangun menjadi urutan token yang sesuai dengan format yang diperlukan oleh model. Ini bisa termasuk menambahkan token khusus untuk tanda-tanda awal dan akhir kalimat, serta tokenisasi untuk token khusus yang mungkin diperlukan oleh model tertentu. (Siino et al., 2023)

Pada penelitian ini, salah satu alat yang digunakan dalam preprocessing adalah Typo Corrector, yang berfungsi untuk memperbaiki kesalahan ketik pada tahap awal tokenisasi. Proses ini membantu meningkatkan kualitas input, sehingga model BERT dapat menerima teks yang lebih bersih dan bebas dari kesalahan ketik. Setelah proses tokenisasi, teks juga melalui tahap padding dan truncation. Model BERT memiliki batasan panjang input, sehingga jika teks melebihi batas maksimum, teks akan dipotong, dan jika terlalu pendek, padding akan ditambahkan. Selain itu, token khusus seperti [CLS] untuk menandai awal dan [SEP] untuk pemisah segmen digunakan agar teks dapat diproses dengan struktur yang sesuai. Typo Corrector memastikan bahwa semua teks yang sudah dikoreksi bebas dari kesalahan ketik, sehingga meminimalkan gangguan dalam proses berikutnya.

Akhirnya, teks yang telah melewati typo correction dan tokenisasi dikonversi menjadi ID token dan siap diproses oleh BERT. Setiap token dikonversi menjadi vektor numerik yang sesuai dengan embedding model. Jika teks terdiri dari dua segmen, token [SEP] memisahkan kedua segmen tersebut, dan teknik masking bisa digunakan dalam pelatihan model untuk memprediksi token tertentu yang disembunyikan. Dalam penelitian ini, Typo Corrector juga diterapkan untuk meningkatkan interaksi chatbot dengan pengguna, membantu chatbot memahami input pengguna dengan lebih baik sehingga memberikan respons yang lebih relevan. Dengan adanya Typo Corrector, kesalahan ketik pada input pengguna dapat dikoreksi secara otomatis sebelum teks diproses oleh chatbot, meningkatkan efektivitas komunikasi dan mengurangi risiko kesalahpahaman akibat kesalahan ketik.

## 2.16 Typo Corrector

*Typo corrector* dalam konteks *machine learning* adalah model yang dirancang untuk mendeteksi dan memperbaiki kesalahan pengetikan dalam teks. model ini bekerja dengan menganalisis teks masukan dan menyarankan ejaan yang paling mungkin benar untuk kata-kata yang salah eja. Model ini dapat dilatih dengan kumpulan data kata-kata yang salah eja beserta koreksinya, sehingga membantu model untuk mengkhususkan diri dalam tugas ini dan meningkatkan akurasinya (Jurafsky, 2024). Model yang digunakan dalam penelitan ini adalah *spell correction english base,* sebuah model yang dibangun oleh sebuah akun huggingface bernama oliverguhr. Model *spell correction english base* ini didasarkan pada arsitektur transformer, yang merupakan jenis model deep learning yang dikenal karena efektivitasnya dalam tugas pemrosesan bahasa alami. Model transformer terdiri dari arsitektur encoder-decoder dengan mekanisme perhatian, yang memungkinkannya untuk menangkap hubungan antara kata-kata dalam sebuah urutan.

Transformer merupakan sebuah arsitektur *neural network* yang memperkenalkan mekanisme perhatian yang memungkinkan model untuk menangkap hubungan yang lebih kuat antara kata-kata dalam teks. Dengan demikian, model transformer dapat mengatasi batasan arsitektur sekuensial tradisional yang cenderung sulit dalam menangkap hubungan jarak jauh dalam teks.

Transformer terdiri dari dua bagian utama, yaitu encoder dan decoder. Encoder bertugas untuk memproses input teks dan menghasilkan representasi kontekstual untuk setiap kata dalam teks. Hal ini dicapai melalui beberapa lapisan mekanisme perhatian yang memungkinkan encoder "melihat" seluruh teks secara bersamaan. Selanjutnya, decoder menggunakan representasi kontekstual yang dihasilkan oleh encoder untuk memprediksi keluaran, seperti dalam terjemahan mesin atau generasi teks.

Salah satu keunggulan utama dari transformer adalah kemampuannya untuk mengatasi masalah vanishing gradient yang sering terjadi pada jaringan saraf rekuren (Recurrent Neural Networks, RNNs). Dengan mekanisme perhatian dan penggunaan blok residual, transformer dapat menjaga informasi yang relevan dari input asli selama proses pelatihan, sehingga meningkatkan kinerja model dalam tugas-tugas NLP.

Dalam kasus koreksi ejaan, model *spell correction english base* dilatih pada dataset teks bahasa Inggris yang besar untuk mempelajari hubungan antara kata-kata yang dieja dengan benar dan kesalahan ejaan umum. Selama pelatihan, model belajar untuk memprediksi ejaan yang benar dari sebuah kata yang dieja salah, dengan memperhitungkan konteks kata-kata di sekitarnya.

Ketika sebuah kata atau frasa yang dieja salah dimasukkan ke dalam model koreksi ejaan, model pertama-tama akan melakukan tokenisasi pada input menjadi urutan token, masing-masing mewakili kata atau subkata. Model kemudian memproses token-token ini melalui beberapa lapisan mekanisme *self attention* dan jaringan saraf *feedforward* untuk menangkap konteks dan semantik dari input. Akhirnya, model menghasilkan versi yang sudah diperbaiki dari input dengan memprediksi ejaan yang benar dari setiap token berdasarkan representasi yang dipelajarinya.

Sebagai contoh, jika frasa "lets do a comparsion" dimasukkan, model mungkin melakukan tokenisasi menjadi ["lets", "do", "a", "comparison"] dan menghasilkan versi yang sudah diperbaiki seperti "let's do a comparison" berdasarkan pengetahuannya tentang ejaan dan tata bahasa Inggris.

Pada penelitian ini, *Typo Corrector* digunakan untuk meningkatkan interaksi chatbot dengan pengguna. Alat ini membantu chatbot memahami input pengguna dengan lebih baik, sehingga dapat memberikan respon yang lebih akurat dan relevan. Dengan adanya *Typo Corrector*, kesalahan ketik yang mungkin terjadi saat pengguna memasukkan teks akan dikoreksi secara otomatis sebelum diproses oleh chatbot. Hal ini akan membantu meningkatkan keefektifan komunikasi antara pengguna dan chatbot, serta mencegah terjadinya kesalahpahaman akibat kesalahan ketik.

## 2.17 Named Entity Recognition (NER)

Named Entity Recognition (NER) adalah teknik dalam pemrosesan bahasa alami yang digunakan untuk mengidentifikasi dan mengklasifikasikan entitas penting dalam teks, seperti nama orang, lokasi, organisasi, tanggal, dan lain-lain. NER berfungsi untuk mengekstraksi informasi spesifik dengan cara mengenali kata atau frasa yang memiliki makna khusus dalam konteks tertentu. Teknik ini umumnya terdiri dari dua langkah utama: identifikasi entitas (menemukan kata atau frasa yang relevan dalam teks) dan klasifikasi entitas (menentukan jenis atau kategori dari kata atau frasa tersebut). Dengan NER, proses analisis data teks menjadi lebih efisien, terutama dalam aplikasi seperti penambangan data, analisis sentimen, dan pencarian informasi, karena memungkinkan sistem untuk menyoroti informasi inti dari kumpulan teks yang besar dengan cepat.

## 2.5 Text Embedding

Text embedding adalah teknik dalam pemrosesan bahasa alami (Natural Language Processing/NLP) yang digunakan untuk merepresentasikan teks sebagai vektor numerik. Teknik ini memanfaatkan model pembelajaran mesin, seperti Word2Vec, GloVe, atau BERT, yang telah dilatih pada korpus teks besar untuk memahami hubungan semantik antara kata-kata dalam teks. Saat teks dimasukkan ke dalam model embedding, kata-kata dalam teks diubah menjadi vektor numerik berdasarkan representasi yang telah dipelajari oleh model. Proses ini memungkinkan kata-kata dengan makna atau konteks yang serupa untuk direpresentasikan sebagai vektor yang mendekati satu sama lain dalam ruang vektor, mencerminkan hubungan semantik antar-kata.

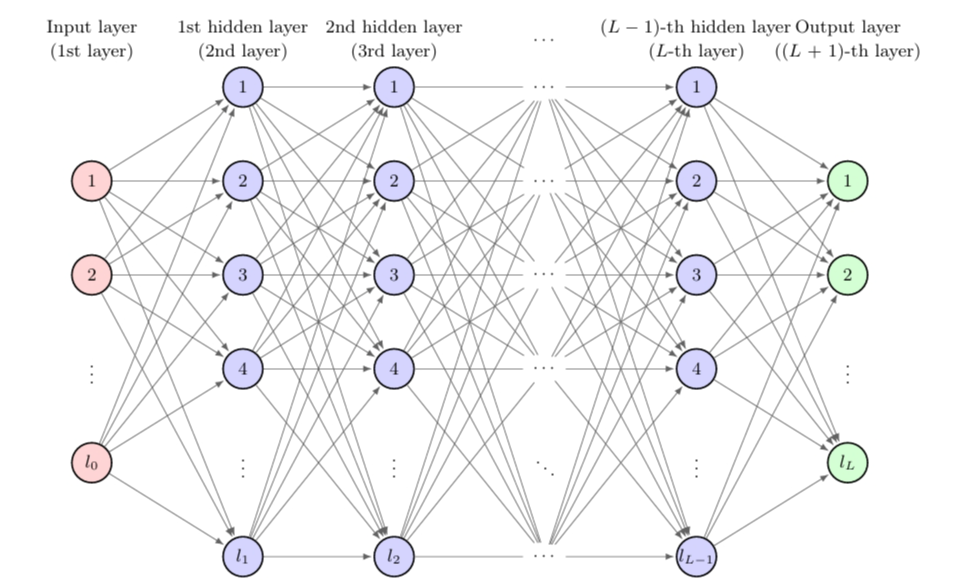
Teknik embedding ini memanfaatkan berbagai pendekatan, seperti neural networks atau metode statistik, untuk mengekstrak fitur penting dari kata-kata, termasuk makna, konteks, dan relasi antar-kata. Representasi vektor ini menyimpan informasi semantik yang kaya, memungkinkan model pembelajaran mesin untuk memproses teks dengan cara yang lebih efektif. Dengan menggunakan ruang vektor semantik, kata-kata dengan makna atau konteks serupa ditempatkan berdekatan, sementara kata-kata dengan perbedaan semantik yang besar ditempatkan lebih jauh dalam ruang multidimensi ini. Hasilnya, text embedding sangat berguna dalam berbagai aplikasi NLP, seperti klasifikasi teks, penerjemahan mesin, dan pencarian informasi.

Pada penelitian ini, teknik embedding dilakukan dengan menggunakan model BERT yang telah dilatih sebelumnya (pre-trained model). Model seperti BERT mampu menghasilkan representasi teks yang kaya akan konteks dan makna berkat kemampuannya dalam memahami hubungan antar-kata secara bidireksional. Pemanfaatan model pre-trained memungkinkan efisiensi dalam menghasilkan representasi teks tanpa perlu melatih model dari awal, sehingga mempercepat proses dan meningkatkan akurasi analisis.

Pada penelitian ini, teknik embedding menggunakan pre-trained model, khususnya model BERT, untuk menghasilkan representasi dari data. Penggunaan pre-trained model seperti BERT memungkinkan representasi teks yang kaya akan konteks dan makna, karena model tersebut memahami hubungan antar kata secara bidireksional dalam teks.(Morriz et al., 2023)

## 2.6 Artificial Neural Network (ANN)

ANN adalah model komputasi yang terdiri dari unit-unit pemrosesan yang disebut neuron. Struktur ANN melibatkan lapisan-lapisan neuron yang saling terhubung. Proses pembelajaran dalam ANN dilakukan melalui penyesuaian *weight* antar neuron dengan menggunakan algoritma seperti backpropagation. ANN memiliki berbagai aplikasi, mulai dari *voice recognition* dan gambar hingga *natural language processing* dan *robotics*. Kemampuan ANN untuk belajar dan beradaptasi membuatnya sangat berguna dalam bidang-bidang yang memerlukan analisis data yang kompleks dan dinamis (Jentzen et al., 2023).



Gambar 2.2 Neural Network ((Jentzen et al., 2023)

Activation function adalah fungsi matematis yang diterapkan pada setiap neuron dalam neural network untuk menghasilkan output. Dua fungsi aktivasi yang umum digunakan adalah:

1. Sigmoid

Fungsi sigmoid menghasilkan output dalam rentang nilai antara 0 dan 1.

|  |  |
| --- | --- |
|  | (2.1) |

1. tanh (Hyperbolic Tangent)

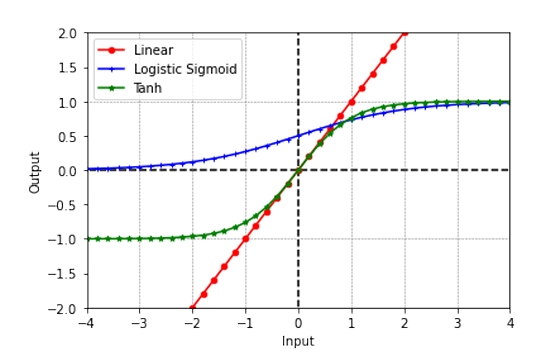
Fungsi tanh menghasilkan output dalam rentang nilai antara -1 dan 1.

|  |  |
| --- | --- |
|  | (2.2) |

1. ReLU (Rectified Linear Unit)

Fungsi ReLU menghasilkan output yang sama dengan inputnya jika inputnya positif, dan 0 jika inputnya negatif.

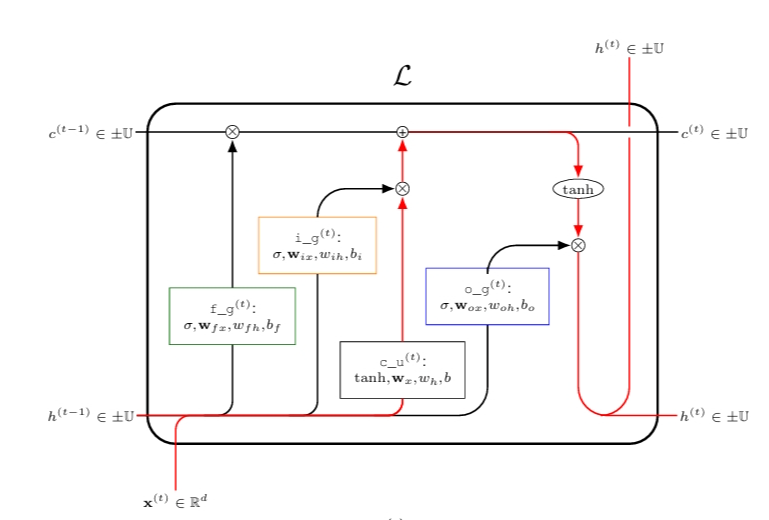
|  |  |
| --- | --- |
|  | (2.3) |



Gambar 2.3 Activation Function (Dubey et al., 2022)

## 2.7 LSTM

LSTM (*Long Short-Term Memory*) adalah jenis neural network yang dirancang untuk mengatasi masalah vanishing gradient, dimana terdapat gate kontrol untuk mengatur aliran informasi di dalam sel memori sehingga dapat mempertahankan informasi dalam jangka waktu yang lebih lama. Vanishing gradient adalah fenomena yang sering terjadi dalam neural network, terutama pada network yang sangat dalam. Ini terjadi ketika gradien dari loss function menurun secara signifikan saat mundur ke layer-layer awal dari network. Dalam kasus ini, ketika gradien menjadi sangat kecil, atau "menghilang," itu mengakibatkan parameter-parameter pada layer-layer awal menjadi terlalu sedikit diperbarui selama proses training, sehingga network gagal belajar dengan baik dari data pelatihan. Hal ini dapat mengakibatkan kinerja network menjadi stagnan atau bahkan memburuk seiring dengan bertambahnya kedalaman network. Dengan menggunakan gate kontrol yang terdiri dari forget gate, input gate, dan output gate, LSTM mampu mengatur aliran informasi di dalam sel memori dengan lebih efektif.



Gambar 2.4 LSTM (Elsworth Steven, 2020)

LSTM didesain untuk menghadapi *Vanishing Gradient Problem* pada RNN (*Recurrent Neural Network*). *Vanishing Gradient Problem* adalah suatu fenomena yang terjadi saat gradient yang digunakan untuk memperbarui parameter terlalu kecil pada proses *backpropagation* (Hochreiter et al., 1997). LSTM memperkenalkan beberapa *gate* pada LSTM ialah:

* *Forget gate*, adalah gerbang yang menentukan informasi apa yang perlu di lupakan di memori sel, formula *forget gate* ialah:

|  |  |
| --- | --- |
|  | (2.4) |

dengan:

* + σ : fungsi aktivasi sigmoid
  + : matrix *weight* pada *forget gate*
  + : matrix *bias* pada *forget gate*
  + : *state* sebelumnya
  + : input pada *state* sekarang
* *Input gate,* adalah gerbang yang menentukan informasi apa yang akan disimpan, formula *input gate* ialah:

|  |  |
| --- | --- |
|  | (2.5) |

dengan:

* + : matrix *weight* pada *input gate*
  + : matrix *bias* pada *input gate*
* *Output gate,* adalah gerbang yang menentukan *output* dari *state* sekarang, formula *output gate* ialah:

|  |  |
| --- | --- |
|  | (2.6) |

dengan:

* + : matrix *weight* pada *output gate*
  + : matrix *bias* pada *output gate*

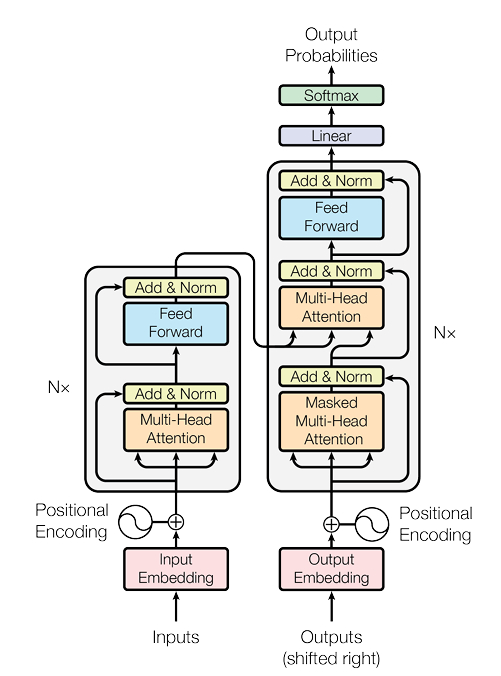
Kemampuan LSTM untuk mempertahankan/mengingat informasi selama periode waktu yang lama membuatnya cocok untuk digunakan dalam tugas *time-series forecasting*, *machine translation*, *natural language processing*, dan lain-lain.

## 2.8 BERT

BERT (Bidirectional Encoder Representations from Transformers) adalah sebuah model bahasa yang menggunakan Arsitektur Transformer dan memiliki kemampuan untuk memahami konteks dan hubungan antar kata dalam teks secara mendalam. Arsitektur Transformer adalah sebuah arsitektur jaringan neural yang memanfaatkan self-attention mechanism untuk memproses data dalam jangkauan konteks yang lebih luas. Dalam Arsitektur Transformer, terdapat dua komponen utama: encoder dan decoder (Devlin et al., 2019). Dalam konteks BERT, hanya encoder yang digunakan, encoder bertanggung jawab untuk menghasilkan representasi vektor dari teks input.

Sebelum teks dimasukkan ke dalam model BERT, teks akan menjalani langkah-langkah preprocessing terlebih dahulu:

1. Tokenisasi: Teks dipecah menjadi token-token kecil, seperti kata-kata atau sub-kata, yang disebut dengan tokenisasi. Setiap token memiliki ID numerik yang unik, dan tokenisasi memungkinkan BERT untuk memahami struktur teks dengan lebih baik.
2. Pemotongan (Truncation) dan Padding: BERT memiliki batasan panjang input yang dapat diterima. Jika teks melebihi panjang maksimum yang ditentukan, maka teks tersebut perlu dipotong (truncated) atau ditambahkan padding sehingga memiliki panjang yang sesuai dengan kebutuhan model.
3. Penambahan Token Khusus: BERT menggunakan token khusus [CLS] untuk menandai awal teks dan [SEP] untuk menandai akhir teks atau pemisahan antara dua teks yang berbeda.
4. Konversi ke ID Token: Setiap token dalam teks diubah menjadi ID token yang sesuai untuk representasi numerik dalam model.
5. Segmentasi: Jika teks terdiri dari dua bagian atau lebih (seperti pasangan kalimat dalam tugas pemahaman bahasa), teks tersebut perlu dipisahkan menjadi segmen-segmen yang terpisah. BERT menggunakan token [SEP] untuk memisahkan segmen-segmen tersebut.
6. Masking: BERT menggunakan teknik masking, di mana beberapa token acak dalam teks diubah menjadi token [MASK] untuk melatih model dalam memprediksi token yang di-mask.
7. Embedding: Setiap token dalam teks diubah menjadi representasi vektor numerik sesuai dengan embedding yang telah dipelajari oleh model.



Gambar 2.5 Arsitektur Transformers (Vaswani et al., 2017)

Arsitektur Transformer memungkinkan BERT untuk mempertimbangkan konteks dari kedua arah; kiri dan kanan, untuk lebih memahami konteks suatu teks. Hal ini dicapai melalui penggunaan self-attention mechanism, di mana setiap kata dalam teks diberi bobot berdasarkan hubungannya dengan kata-kata lain dalam kalimat. Bobot ini menggambarkan seberapa pentingnya setiap kata dalam konteks kalimat secara keseluruhan.

Transformer memungkinkan BERT untuk melakukan pelatihan secara unsupervised, di mana model mempelajari representasi kata dari teks yang tidak berlabel secara langsung. Selain itu, model BERT yang telah dilatih sebelumnya dapat dimodifikasi dengan menambahkan lapisan output tambahan, memungkinkan BERT untuk digunakan dalam berbagai tugas, seperti menjawab pertanyaan dan melakukan inferensi bahasa, tanpa perlu mengubah arsitektur secara signifikan.

## 2.9 Python

Python adalah bahasa pemrograman high-level dan object-oriented (Kuhlman, 2012) yang dirancang oleh Guido van Rossum dan diperkenalkan pada tahun 1991. Bahasa pemrograman ini sangat diminati di industri (Tank, 2022) terutama dalam tugas terkait Pemrosesan Bahasa natural (*Natural Language Processing*/NLP). Python digunakan secara luas dalam berbagai keperluan seperti pengembangan web, perangkat lunak, scripting, sistem database, sistem pembelajaran mesin, analisis data, otomatisasi, dan lain-lain.

Salah satu alasan utama popularitas Python adalah kemampuannya yang luas dan beragam, termasuk kehandalannya dalam tugas-tugas terkait NLP. Kelebihan lainnya adalah kemudahan pembacaan kode (*readability*) yang tinggi, di mana Python mengikuti struktur sintaksis Bahasa Inggris, serta keberadaan library yang luas.

## 2.10 Streamlit

Streamlit adalah sebuah framework open-source dalam bahasa pemrograman Python yang dirancang untuk memfasilitasi pengembangan aplikasi web interaktif dengan cepat dan mudah. Framework ini sangat populer karena kemampuannya dalam membuat aplikasi web yang dapat menampilkan dan berinteraksi dengan data secara efisien. Salah satu keunggulan utama dari Streamlit adalah antarmuka pengguna yang responsif dan kemudahan penggunaannya yang memungkinkan pengembang untuk fokus pada logika aplikasi dan model. Streamlit dapat melakukan pembaruan otomatis saat kode berubah dan dukungan yang kuat untuk visualisasi data.

Streamlit memiliki beberapa fitur yang dapat membantu infrastruktur chatbot. Misalnya, Streamlit menyediakan caching otomatis untuk mempercepat pemrosesan data dan mencegah pengulangan proses yang lambat setiap kali aplikasi diperbarui. Ini sangat berguna untuk aplikasi yang memerlukan pemuatan data berat atau pemrosesan simulasi yang kompleks.

## 2.11 FastAPI

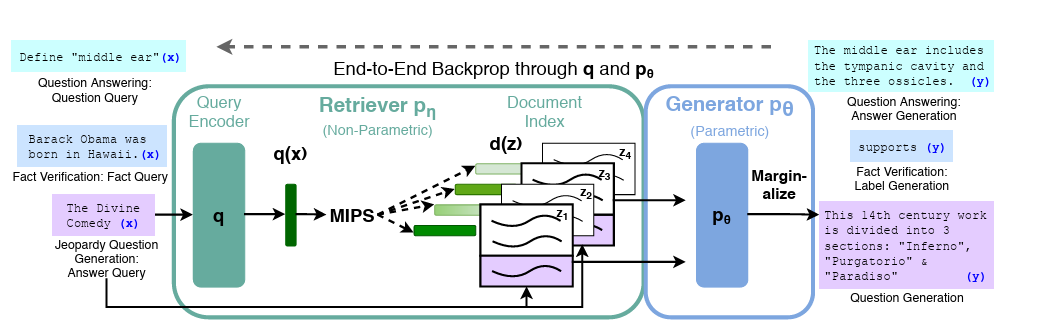
FastAPI adalah framework web modern yang ditulis dengan Python untuk membangun API (Application Programming Interface) dengan performa tinggi dan efisiensi yang optimal. Diperkenalkan oleh Sebastián Ramírez pada tahun 2018, FastAPI dirancang untuk memanfaatkan fitur-fitur asinkron Python yang memungkinkannya menangani request secara paralel, sehingga sangat cepat dan cocok digunakan pada aplikasi berskala besar (Ramírez, 2018). Salah satu keunggulan utama FastAPI adalah penggunaan tipe anotasi Python (type hints) yang membantu developer mendefinisikan struktur data dengan lebih jelas dan memvalidasi input secara otomatis.

FastAPI mengadopsi standar OpenAPI dan JSON Schema untuk dokumentasi otomatis, yang memungkinkan developer mendapatkan dokumentasi API interaktif secara *real-time* tanpa usaha tambahan. Selain itu, FastAPI mendukung *asynchronous programming* yang mempercepat penanganan request, membuatnya sangat efektif untuk aplikasi yang membutuhkan skalabilitas tinggi seperti sistem *microservices* dan aplikasi berbasis *machine learning.*

## 2.12 RAG

*Retrieval-Augmented Generation* (RAG) merujuk pada pendekatan dalam NLP, di mana model AI menggabungkan 2 pendekatan dalam menjawab pertanyaan: *retrieval* atau mengambil pengetahuan melalui sumber eksternal seperti basis pengetahuan atau korpus teks besar dan *generation atau* menghasilkan jawaban dari pengetahuan yang diambil. RAG memungkinkan model untuk memanfaatkan keunggulan dari kedua pendekatan tersebut sehingga memberikan fleksibilitas yang besar dalam memproses bahasa alami, memungkinkan model untuk menghasilkan jawaban yang lebih tepat dan informatif dalam berbagai konteks dan tugas (Lewis et al., 2020).

Dalam bidang akademik, RAG sangat bermanfaat untuk membantu mahasiswa dan peneliti mengakses informasi yang relevan dengan cepat. Misalnya, chatbot yang menggunakan RAG dapat menjawab pertanyaan mahasiswa tentang topik kuliah ataupun tentang topik terkait universitas. Dengan mengakses database akademik dan artikel penelitian terbaru, RAG dapat menghasilkan jawaban yang tidak hanya akurat tetapi juga didukung oleh sumber yang kredibel



Gambar 2.6 Arsitektur RAG (Lewis et al., 2020)

## 2.13 FAISS

Facebook AI Similarity Search (FAISS) adalah *library* yang dibuat untuk melakukan pencarian kesamaan yang efisien dalam dokumen. FAISS memanfaatkan representasi vektor dari titik data untuk menemukan tetangga terdekat, yaitu item yang serupa dengan *query*. Konsep utama FAISS adalah menggunakan representasi vektor dari titik data dan melakukan pencarian tetangga terdekat mendekati untuk menemukan item yang serupa (Johnson, 2017). Algoritma pencarian ini memungkinkan FAISS bekerja dengan dataset sangat besar, sehingga ideal untuk aplikasi skala besar seperti sistem rekomendasi, pengenalan gambar, dan deteksi plagiarisme. Selain itu, FAISS mendukung berbagai tipe indeks dan metode pencarian, termasuk algoritma k-means clustering dan hierarchical clustering, yang memberikan fleksibilitas tinggi dalam penanganan data.

FAISS dirancang untuk kinerja tinggi, memanfaatkan kemampuan komputasi dari CPU dan GPU. Ini membuatnya sangat efisien dalam menangani pencarian dalam dataset besar dengan waktu yang singkat. FAISS juga menyediakan berbagai alat dan API yang mudah digunakan untuk integrasi dengan proyek-proyek machine learning dan data science. Dengan dukungan untuk berbagai bahasa pemrograman seperti Python dan C++, FAISS menjadi pilihan populer di kalangan peneliti dan praktisi industri. Penggunaan FAISS dalam aplikasi nyata telah membuktikan bahwa pencarian kesamaan berbasis vektor dapat dilakukan dengan cepat dan akurat, menjadikannya alat berharga untuk berbagai kebutuhan analisis data dan pengolahan informasi.

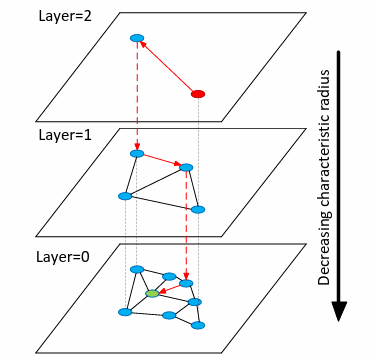
## 2.14 Database

Database adalah kumpulan data yang terorganisir secara terstruktur agar dapat diakses, dimanipulasi, dan dikelola dengan efisien. Ada banyak bentuk Database, diantaranya ialah Multimedia Database dan Vector Database, Multimedia Database adalah database yang ditujukan untuk menyimpan jenis data multimedia, seperti gambar (Hobbs et al., 1999). Database ini memungkinkan chatbot untuk mengakses dan menyediakan informasi multimedia kepada pengguna jika diperlukan, seperti dalam menjawab pertanyaan atau memberikan ilustrasi visual tentang informasi yang diminta. Vector Database dirancang untuk menyimpan data vektor dengan efisien. Dalam konteks komputasi, vektor adalah entitas yang didefinisikan oleh panjang dan arah. Database ini terutama berguna dalam konteks penyimpanan data dengan banyak dimensi yang dapat diwakili dalam bentuk vektor, Vector Database memungkinkan penyimpanan dan pengelolaan data vektor dengan cara yang efisien, serta menyediakan kemampuan untuk melakukan operasi seperti pencarian, pemrosesan, dan analisis spasial (Taipalus Toni, 2024).

## 2.15 Indexing

Mencari informasi tertentu dalam kumpulan data yang besar akan memakan waktu dan sumber daya yang besar. untuk meminimalisir sumber daya dan waktu pencarian informasi, maka digunakan *indexing, indexing* mengatur data sedemikian rupa untuk memudahkan pencarian item yang relevan dengan cepat. Indexing bekerja dengan membuat struktur data tambahan yang dirancang khusus untuk pencarian cepat. Struktur ini berfungsi seperti direktori, yang menunjuk ke lokasi data sebenarnya. Dengan memanfaatkan index, program tidak perlu mencari setiap titik data dalam *database*. Sebaliknya, index akan mengarahkan pencarian ke area yang paling menjanjikan untuk menemukan informasi yang dibutuhkan.

Salah satu teknik Indexing adalah Indeks *Hierarchical Navigable Small World* (HNSW), HNSW merupakan teknik pengindeksan yang populer untuk pencarian similaritas secara efisien dalam ruang berdimensi tinggi. Metode ini membangun struktur grafik berlapis dimana titik data yang serupa saling terhubung sebagai tetangga. Saat pencarian dilakukan, algoritma bergerak menyusuri grafik ini, dengan prioritas menelusuri sub-ruang yang kemungkinan besar berisi item yang paling mirip dengan query. Penelusuran tertarget ini secara signifikan mengurangi jumlah perbandingan yang diperlukan dibandingkan pencarian menyeluruh, menjadikan HNSW pilihan menarik untuk aplikasi yang membutuhkan pengambilan item serupa secara cepat dan akurat dari kumpulan data yang besar (Malkov Yu, 2016).



Gambar 2.7 HNSW (Malkov Yu, 2016)

Teknik lain yang sering digunakan adalah cosine similarity, sebuah metode untuk mengukur kesamaan antara dua vektor dalam ruang berdimensi tinggi. Cosine similarity menghitung derajat kemiripan berdasarkan sudut kosinus antara dua vektor, di mana nilai kemiripan berkisar antara -1 hingga 1. Nilai 1 menunjukkan bahwa kedua vektor memiliki arah yang sama (identik), sementara nilai 0 menunjukkan tidak ada kesamaan arah, dan -1 berarti vektor tersebut berlawanan. Teknik ini sangat berguna dalam tugas pengambilan informasi dan pencarian berbasis teks karena tidak memperhitungkan panjang vektor tetapi hanya sudut antara mereka, sehingga cocok untuk data yang terdistribusi dalam ruang vektor yang besar (Steck et al, 2024). Dalam kombinasi dengan metode seperti HNSW, cosine similarity memungkinkan pencarian similaritas yang efisien, terutama dalam konteks data besar yang membutuhkan kecepatan dan akurasi tinggi.

## 2.18 Performance Evaluation

Evaluasi kinerja chatbot akan dilakukan menggunakan metrik *precision* untuk mengukur seberapa baik chatbot dalam memberikan respon yang akurat terhadap pertanyaan pengguna. Metrik *precision* mengukur proporsi jumlah jawaban yang benar dari semua jawaban yang diberikan oleh chatbot. Evaluasi ini akan dilakukan dengan mempertimbangkan empat kategori pertanyaan yang berbeda, yaitu bahasa gaul, bahasa informal, typo, dan bahasa formal. Dalam setiap kategori, beberapa pertanyaan uji akan di-*prompt* ke dalam model chatbot, dan jawaban yang dihasilkan akan dievaluasi berdasarkan kebenaran dan relevansinya terhadap pertanyaan. Hasil evaluasi akan memberikan gambaran tentang kemampuan chatbot dalam menjawab pertanyaan dari berbagai kategori pertanyaan.

Dua belas pertanyaan uji akan di *prompt* ke dalam model dan performa model akan diuji menggunakan metrik presisi. Metrik presisi merupakan ukuran evaluasi yang digunakan dalam pemrosesan informasi dan pengambilan keputusan, dalam hal ini termasuk ke dalam konteks klasifikasi dan informasi relevansi. Presisi mengukur proporsi jumlah hasil positif yang diklasifikasikan secara benar dari semua jumlah hasil yang diklasifikasikan sebagai positif oleh model. Berikut adalah definisi formal serta rumus dari metrik tersebut:

* Presisi adalah rasio antara jumlah true positive (TP) dan jumlah semua *instance* yang diklasifikasikan sebagai positif (TP + false positive (FP))

|  |  |
| --- | --- |
|  | (2.7) |

Dalam rumus tersebut, penjelasan istilah-istilah di dalamnya antara lain:

* True Positive (TP): Jumlah *instance* positif yang diklasifikasikan dengan benar.
* False Positive (FP): Jumlah *instance* negatif yang salah diklasifikasikan sebagai positif.

## 2.19 Related Works

|  |  |  |  |
| --- | --- | --- | --- |
| **Referensi** | **Tujuan** | **Hasil Penelitian** | **Perbedaan** |
| (Kurni et al., 2023) | Mengembangkan sistem otomatis canggih untuk sektor pendidikan, yang mampu merespons pertanyaan pengguna atas nama manusia. Berbeda dengan chatbot yang sudah ada yang terutama mengandalkan basis data lokal, sistem yang diusulkan bertujuan untuk memanfaatkan basis data lokal dan web. Tujuan utama melibatkan skalabilitas, kemudahan penggunaan, dan interaktivitas tinggi. Makalah ini menggunakan berbagai teknik seperti pembelajaran mesin, pemrosesan bahasa alami (NLP), pencocokan pola, dan algoritma pemrosesan data untuk meningkatkan kinerja sistem dalam memberikan jawaban komprehensif terhadap berbagai pertanyaan pengguna. | Chatbot yang dijelaskan dalam makalah tersebut berhasil dibangun untuk menjawab pertanyaan dari perguruan-perguruan tinggi dan mahasiswanya. Implementasinya melibatkan penggunaan kecerdasan buatan dan pembelajaran mesin untuk memungkinkan sistem memahami dan merespons pertanyaan pengguna | Penelitian ini menggunakan data gabungan dari database lokal dan juga data dari web atau internet.    Sedangkan skripsi penulis akan menggunakan database lokal dari Universitas Bina Nusantara mengingat tujuan dari chatbot ini dibangun dalam ruang lingkup Universitas Bina Nusantara. |
| (Vani Bhat, 2024) | Paper ini bertujuan meningkatkan kecerdasan chatbot restoran dengan mengintegrasikan Retrieval-Augmented Generation (RAG) dan Large Language Model (LLM). Fokus utama adalah pada pengembangan chatbot yang dapat berinteraksi dalam bahasa alami sambil mengoptimalkan konteks restoran untuk memenuhi permintaan pelanggan yang meningkat pasca-COVID. | Hasil penelitian menunjukkan bahwa pendekatan ini meningkatkan kemampuan chatbot dalam memberikan respon yang nuansial dan relevan, diukur dengan skor Bilingual Evaluation Understudy (BLEU) sebesar 0.60. Evaluasi menunjukkan bahwa model mampu menghasilkan jawaban yang lebih berpengetahuan, beragam, dan relevan dibandingkan dengan model-state-of-the-art lainnya, dengan skor informasi rata-rata dalam rentang 0.6-0.8. | Paper ini berfokus pada pengembangan chatbot untuk industri restoran, menggunakan RAG untuk meningkatkan interaksi dan respon dalam konteks pelayanan pelanggan.  Sebaliknya, paper peneliti mengembangkan chatbot untuk Universitas Bina Nusantara, juga dengan metode RAG, tetapi ditujukan untuk memberikan informasi akademis terkait program studi, pendaftaran, dan kegiatan di lingkungan pendidikan. Dengan demikian, konteks aplikasi dan jenis informasi yang diberikan oleh kedua chatbot ini sangat berbeda; satu berorientasi pada layanan restoran, sedangkan yang lainnya berfokus pada kebutuhan edukasi. |
| (Dharwadkar & Deshpande, 2018) | Tujuan paper ini adalah mengatasi kurangnya informasi pengguna tentang pengobatan dan gejala penyakit, sehingga mereka tidak perlu pergi langsung ke rumah sakit untuk masalah kecil. Solusinya adalah menggunakan medical ChatBot berbasis Natural Language Processing. Pengguna dapat mengajukan pertanyaan melalui ChatBot tanpa harus datang langsung ke rumah sakit. Sistem ini menggunakan Google API untuk konversi suara-teks dan teks-suara, memungkinkan pengguna menerima jawaban melalui aplikasi Android dengan tujuan menganalisis sentimen pelanggan terkait layanan kesehatan. | Paper ini menyimpulkan bahwa sistem yang dikembangkan mampu memberikan hasil yang akurat, terutama dalam memprediksi penyakit berdasarkan gejala dengan menggunakan algoritma SVM dan Pemrosesan Bahasa Alamiah (NLP). Sistem juga efektif dalam mengambil informasi obat dari API dan menampilkan nama-nama obat. Kemajuan teknis ini membuatnya bermanfaat untuk institusi medis dan rumah sakit. Meskipun demikian, untuk memastikan kesuksesan dan penerapan yang lebih luas, perlu dilakukan analisis lebih lanjut terkait dampak dan efektivitas sistem ini dalam praktik medis sehari-hari. | Penelitian ini fokus pada pengembangan sistem chatbot yang lebih berfokus untuk memprediksi penyakit berdasarkan gejala dengan menggunakan algoritma SVM dan pemrosesan bahasa alamiah (NLP).  Sedangkan penelitian penulis bertujuan untuk mengembangkan sistem chatbot sebagai wadah pencarian informasi seputar Universitas Bina Nusantara menggunakan pre-trained model BERT. |
| (S. Hofstätter, J. Chen, K. Raman, and H. Zamani, 2023) | Memperkenalkan FiD-Light, sebuah model yang dirancang untuk meningkatkan efisiensi dari model generasi yang diperkaya dengan pengambilan. Model generasi yang diperkaya dengan pengambilan menyediakan jawaban teks terhadap kueri bersama dengan item provenans yang diambil dari basis pengetahuan. Namun, model-model ini kompleks dan mengalami kesulitan dalam menangani input yang panjang. FiD-Light bertujuan untuk mengatasi tantangan tersebut dengan meningkatkan efisiensi sambil tetap mempertahankan efektivitas. | Hasil dari kesimpulan ini adalah bahwa model FiD-Light, dilengkapi dengan kemampuan reranking listwise otoregresif, berhasil mengatasi keterbatasan efisiensi dan fleksibilitas yang diamati dalam model generasi yang diperkaya dengan pengambilan state-of-the-art sebelumnya, yaitu FiD. Dengan menyesuaikan arsitektur model FiD untuk mengompresi informasi yang disampaikan ke decoder, pengurangan signifikan dalam latensi inferensi berhasil dicapai, hanya dengan pengurangan efektivitas yang sederhana. Pengurangan ini dapat dikurangi dengan menggunakan backbone T5 yang lebih besar, akhirnya menghasilkan hasil optimal Pareto pada enam tugas KILT. Sistem multi-tugas berbasis FiD-Light juga mencapai hasil baru yang signifikan secara state-of-the-art untuk metrik pengambilan dan generasi yang digabungkan pada tugas-tugas tersebut dibandingkan dengan metode sebelumnya di papan peringkat publik. Temuan ini menunjukkan bahwa mencapai efektivitas tertinggi tidak selalu memerlukan peningkatan skala, sehingga memungkinkan lebih banyak peneliti untuk berkontribusi pada masalah ini di masa depan. | Makalah ini mengutamakan peningkatan efisiensi dan efektivitas model generasi yang diperkaya dengan pengambilan, terutama dengan memperkenalkan FiD-Light. Model tersebut dirancang untuk mengatasi kendala-kendala dalam model sebelumnya dengan mengadopsi kemampuan reranking listwise otoregresif. Fokus utamanya adalah pada pengurangan latensi inferensi tanpa mengorbankan efektivitas, yang terbukti melalui hasil eksperimen yang menunjukkan pencapaian state-of-the-art dalam tugas-tugas pengetahuan yang beragam.  Penelitian oleh penulis, di sisi lain, berbeda dalam pendekatannya. Dalam makalah penulis, penekanan diberikan pada pengembangan chatbot yang spesifik untuk memberikan informasi tentang Universitas BINUS. Makalah penulis menggunakan implementasi RAG, Koreksi Ejaan, dan FAISS untuk memastikan pengambilan informasi yang akurat dan efisien. Dengan menggunakan teknik-teknik ini dan dengan implementasi melalui platform Streamlit, penulis menyajikan solusi yang lebih terfokus dan langsung terhadap kebutuhan informasi spesifik pengguna terkait universitas Bina Nusantara. |
| (L. Patrick, P. Ethan, P. Aleksandra, et al. 2020) | Mengeksplorasi resep fine-tuning berbasis umum untuk generasi yang diperkaya dengan pengambilan (RAG). RAG adalah model yang menggabungkan memori parametrik dan non-parametrik untuk generasi bahasa. Dalam makalah ini, diperkenalkan model RAG di mana memori parametriknya adalah model seq2seq yang sudah dilatih sebelumnya, sedangkan memori non-parametriknya adalah indeks vektor padat dari Wikipedia yang diakses dengan penarik neural yang sudah dilatih sebelumnya. Tujuan utama adalah untuk meningkatkan kemampuan model dalam mengakses dan memanipulasi pengetahuan serta menyediakan provenans untuk keputusan yang diambil, terutama dalam tugas-tugas yang sangat bergantung pada pengetahuan. Makalah ini juga bertujuan untuk membandingkan dua formulasi RAG yang berbeda dan mengevaluasi kinerja mereka pada berbagai tugas NLP yang intensif pengetahuannya. Dengan demikian, tujuan utama makalah ini adalah untuk meningkatkan kinerja model RAG dalam menghasilkan bahasa yang lebih spesifik, beragam, dan faktual dibandingkan dengan model seq2seq parametrik yang ada. | Temuan utama dari makalah ini adalah bahwa model Retrieval-Augmented Generation (RAG) menetapkan hasil baru yang sangat baik pada berbagai tugas Open-Domain QA. RAG mampu memadukan fleksibilitas generasi dari pendekatan "closed-book" (hanya parametrik) dan kinerja dari pendekatan berbasis pengambilan "open-book". Berbeda dengan model REALM dan T5+SSM, RAG mencapai hasil yang kuat tanpa pre-training "salient span masking" yang mahal dan khusus. RAG juga berhasil bersaing dengan sistem DPR QA tanpa perlu reranker atau pembaca ekstraktif untuk mencapai kinerja state-of-the-art. Selain itu, RAG menunjukkan keunggulan dalam menghasilkan jawaban bahkan ketika mungkin untuk mengekstraknya, dengan mampu menghasilkan jawaban yang benar bahkan ketika jawaban yang benar tidak terdapat dalam dokumen yang diambil.  Dalam tabel 1, dapat dilihat bahwa RAG mencapai state-of-the-art baru pada berbagai tugas Open-Domain QA, bahkan melebihi kinerja model-model sebelumnya seperti REALM dan T5+SSM. RAG juga menunjukkan kemampuan untuk menghasilkan jawaban yang benar bahkan ketika jawaban tersebut tidak terdapat dalam dokumen yang diambil, yang merupakan keunggulan dibandingkan dengan pendekatan ekstraktif. Ini menunjukkan potensi besar dari pendekatan RAG dalam meningkatkan kinerja pada tugas-tugas yang sangat bergantung pada pengetahuan. | Makalah ini membahas pengembangan model Retrieval-Augmented Generation (RAG) untuk meningkatkan kemampuan akses dan manipulasi pengetahuan dalam tugas-tugas NLP yang intensif pengetahuan. Penelitian ini bertujuan untuk menggabungkan memori parametrik dan non-parametrik untuk generasi bahasa dengan menggunakan model RAG.  Di sisi lain, penelitian penulis berfokus pada pengembangan chatbot yang spesifik untuk memberikan informasi tentang Universitas BINUS kepada pengguna. Implementasi penelitian penulis melibatkan penggunaan teknologi RAG, Koreksi Ejaan, dan FAISS, serta dideploy menggunakan Streamlit untuk interaksi pengguna. |
| (H. Wenyu, L. Mirella, V. Pavlos, et al. 2023) | Meningkatkan kualitas generasi jawaban dalam tugas-tugas yang membutuhkan pengetahuan yang intensif, seperti generative question answering. Model generasi yang digunakan perlu mampu mengambil informasi dari sumber pengetahuan eksternal untuk mendukung pembangkitan jawaban. Namun, model-model terkini cenderung terlalu fokus pada pengambilan informasi yang relevan dengan label, daripada informasi yang relevan dengan pertanyaan dan jawaban. Makalah ini bertujuan untuk mengatasi masalah ini dengan menggabungkan pengkodean jawaban yang kaya melalui Dense Knowledge Similarity (DKS) dan Retriever as Answer Classifier (RAC). Tujuan utamanya adalah untuk meningkatkan kualitas generasi jawaban dalam tugas-tugas open domain question answering dan conversation dengan memperhatikan kedua aspek generasi dan pengambilan informasi. | Hasil dari penelitian ini menunjukkan pengenalan sebuah kerangka kerja pelatihan baru untuk model generasi yang diperkaya dengan pengambilan informasi, yang bertujuan untuk melatih retriever yang lebih kuat dan robust. Kerangka kerja pelatihan baru ini menggabungkan dua tujuan terkait pengetahuan, yaitu RAC dan DKS, untuk membantu mengarahkan retriever dalam mengambil informasi yang relevan dengan pengetahuan, bukan hanya informasi yang relevan dengan label. Evaluasi dilakukan pada dua dataset KIG terkenal, yaitu MSMARCO dan KILT Wizard of Wikipedia, dan hasilnya menunjukkan peningkatan yang konsisten dari metode yang diusulkan baik pada metrik pengambilan informasi maupun metrik generasi dibandingkan dengan model dasar. Hasilnya menunjukkan bahwa penggunaan metode yang diusulkan menghasilkan peningkatan yang signifikan pada kualitas pengambilan informasi, dan juga berhasil mengurangi risiko overfitting yang disebabkan oleh fokus terlalu besar pada informasi yang relevan dengan label. Penelitian ini juga mengemukakan bahwa metode yang diusulkan dapat memberikan kontribusi yang signifikan pada komunitas penelitian dalam era Large Language Models (LLMs), terutama dalam mengurangi halusinasi yang dihasilkan oleh LLMs. Sebagai saran untuk penelitian mendatang, peneliti mengusulkan untuk menggabungkan penggunaan informasi berbasis teks dengan pengetahuan terstruktur, seperti database dan knowledge graphs, serta menyelesaikan beberapa tantangan penting yang terkait dengan pengembangan metode ini. | Penelitian pertama fokus pada meningkatkan kualitas pengambilan informasi pada tugas generative question answering dengan menggabungkan Dense Knowledge Similarity (DKS) dan Retriever as Answer Classifier (RAC). Hasilnya menunjukkan peningkatan yang signifikan pada beberapa dataset.  Sementara itu, penelitian penulis mengembangkan chatbot untuk memberikan informasi tentang Universitas BINUS menggunakan teknologi Retrieval-Augmented Generation (RAG),Spell Correction, dan FAISS, serta dideploy menggunakan Streamlit.  Meskipun keduanya menggunakan RAG, fokus dan tujuan keduanya berbeda, dengan penelitian pertama lebih terfokus pada peningkatan pengambilan informasi pada tugas generative question answering, sementara penelitian kedua lebih menitikberatkan pada pengembangan aplikasi praktis untuk memberikan informasi kepada pengguna. |
| (Windiatmoko, Y., Rahmadi, R., & Hidayatullah, A. F. (2021) | Mengembangkan sistem cerdas berbasis kecerdasan buatan untuk perguruan tinggi, khususnya chatbot, guna membantu dalam berbagai tugas yang dibutuhkan oleh mahasiswa atau calon mahasiswa. Chatbot ini diharapkan dapat memberikan informasi yang diperlukan seperti informasi kurikulum, pendaftaran mahasiswa baru, jadwal perkuliahan, informasi nilai mahasiswa, serta tambahan fitur untuk jadwal ibadah Muslim dan ramalan cuaca. Pengembangan chatbot ini dilakukan dengan menggunakan model pembelajaran mendalam (deep learning) yang diadopsi oleh model kecerdasan buatan untuk mereplikasi kecerdasan manusia dengan skema pelatihan yang spesifik. Teknologi pembelajaran mendalam ini berdasarkan pada RNN (Recurrent Neural Network) dengan skema penyimpanan memori yang spesifik untuk model pembelajaran mendalam, khususnya dengan penggunaan LSTM (Long Short Term Memory) yang telah terintegrasi dalam kerangka kerja RASA. LSTM dikenal efisien dalam penyimpanan memori yang diperlukan namun juga menghapus memori yang tidak diperlukan. Penggunaan platform Facebook dipilih karena pengguna Facebook telah mencapai 60,8% dari total populasi di Indonesia. | Hasil dari penelitian ini menunjukkan pengembangan sistem chatbot untuk membantu menjawab semua pertanyaan yang diajukan oleh masyarakat kepada universitas. Meskipun chatbot belum digunakan secara operasional di kampus, kinerja dan kemudahan penggunaan belum dapat diukur. Namun, performa chatbot diukur melalui akurasi dalam menghasilkan kalimat-kalimat yang dapat dibaca dengan benar untuk menjawab pengguna. Selain itu, waktu tanggapan dari model inferensi chatbot cukup cepat, dalam waktu kurang dari satu detik. Hasil evaluasi model chatbot menunjukkan hasil yang cukup baik mendekati skor sempurna dari presisi, recall, dan F1. Meskipun demikian, terdapat beberapa penurunan dalam skor presisi dan recall pada beberapa intent tertentu. Penelitian ini masih berada dalam tahap pengembangan awal dengan data simulasi yang cukup. Implementasi model pembelajaran mendalam dilakukan dengan mengimplementasikan klasifikasi intent dan pengenalan entitas pada Rasa NLU serta kebijakan dialog pada Rasa Core, kemudian diintegrasikan dengan Facebook Messenger. Selanjutnya, chatbot ini perlu dikembangkan dengan memperluas data percakapan dan fitur lebih lanjut untuk membantu otomatisasi tugas di universitas, serta menuju pada chatbot domain terbuka untuk implementasi yang lebih luas dalam pembelajaran mendalam. Terdapat juga kebutuhan untuk mengimplementasikan chatbot dalam platformnya sendiri, tanpa batasan dan lebih dapat disesuaikan dengan kebutuhan. | Paper ini bertujuan untuk mengembangkan sistem cerdas berbasis kecerdasan buatan, terutama chatbot, untuk digunakan di lingkungan perguruan tinggi. Tujuan dari chatbot ini adalah untuk memberikan berbagai informasi yang dibutuhkan oleh mahasiswa atau calon mahasiswa, seperti informasi kurikulum, pendaftaran mahasiswa baru, jadwal perkuliahan, informasi nilai mahasiswa, serta fitur tambahan seperti jadwal ibadah Muslim dan ramalan cuaca. Pengembangan chatbot dilakukan menggunakan model pembelajaran mendalam (deep learning) dengan menggunakan LSTM (Long Short Term Memory) yang terintegrasi dalam kerangka kerja RASA. Selain itu, penggunaan platform Facebook dipilih karena pengguna Facebook telah mencapai 60,8% dari total populasi di Indonesia.  Sementara itu, dalam paper penulis, penulis menggunakan metode Retrieval-Augmented Generation (RAG) untuk membuat chatbot yang bertujuan untuk memberikan informasi mengenai Universitas Bina Nusantara. Chatbot ini dilengkapi dengan koreksi ejaan dan menggunakan FAISS (Facebook AI Similarity Search) untuk pencarian informasi. Metode RAG memungkinkan chatbot untuk menghasilkan jawaban yang lebih relevan dengan menggunakan model bahasa seperti GPT (Generative Pre-trained Transformer) untuk generasi dan model pencarian seperti BERT untuk meningkatkan relevansi informasi yang dihasilkan. Dengan demikian, paper penulis menunjukkan pendekatan yang lebih terintegrasi dan canggih dalam mengembangkan chatbot yang dapat memberikan informasi dengan lebih akurat dan relevan kepada pengguna, dibandingkan dengan pendekatan yang digunakan dalam paper pertama. |
| (A. A. Nahdatul, H. C. H. Mohamad, Z. Azaliza, B. Zirawani, 2019) | Mengembangkan sistem Chatbot yang dikenal sebagai UNISEL Bot untuk membantu departemen pemasaran dalam memberikan informasi secara interaktif untuk tujuan pemasaran. Masalah yang ada saat ini adalah informasi tidak disajikan secara interaktif, menyajikan informasi secara manual menggunakan portal dan kertas terasa rumit, dan tidak ada dukungan pelanggan real-time untuk membantu dalam pertanyaan dan jawaban. Oleh karena itu, proyek ini bertujuan untuk mengembangkan sistem asisten pemasaran Chatbot untuk institusi akademik swasta yang dikenal sebagai UNISEL Bot. Pengembangan sistem Chatbot ini diharapkan dapat membantu departemen pemasaran menggunakan cara pemasaran yang lebih cerdas dan interaktif, misalnya; untuk menerima pertanyaan umum dari mahasiswa dan memberikan umpan balik secara real-time serta mendorong orang untuk berinteraksi dengan teknologi terbaru. Metodologi Agile digunakan dalam pengembangan sistem Chatbot ini. Pengumpulan data kualitatif menggunakan metode wawancara dengan mahasiswa dan staf Universitas diimplementasikan. Banyak diagram disajikan dalam teks ini untuk menggambarkan alur proses sistem UNISEL Bot. Sistem UNISEL Bot terdiri dari tujuh modul utama termasuk Ask Question, Feedback, Registration, Event, Appointment, Survey, dan Map. Di masa depan, sistem Chatbot ini diharapkan dapat bekerja secara efektif untuk menggantikan metode manual layanan pelanggan tradisional dan juga membantu dalam mengumpulkan data pengguna untuk membangun data analitik. | Hasil paper tersebut adalah pengembangan UNISEL Bot, sebuah sistem Chatbot yang bertujuan untuk membantu departemen pemasaran dalam memberikan informasi secara interaktif untuk tujuan pemasaran di Universitas UNISEL. Sistem ini terdiri dari tujuh modul utama: Ask Question, Appointment, Registration, Survey, Feedback, Map, dan Event. Modul Ask Question menyediakan layanan percakapan langsung yang memungkinkan pengguna untuk mendapatkan jawaban cepat terkait pertanyaan tentang universitas. Modul Registration memfasilitasi proses pendaftaran mahasiswa baru secara online. Modul Feedback memungkinkan pengguna untuk memberikan komentar dan umpan balik tentang universitas. Modul Event memberikan informasi tentang acara-acara universitas yang akan datang. Modul Appointment memungkinkan pengguna membuat janji dengan dosen atau pejabat universitas tertentu. Modul Survey membantu admin dalam membuat survei sederhana. Terakhir, modul Map memberikan informasi tentang peta universitas dan memungkinkan pengguna untuk menemukan lokasi fasilitas seperti fakultas, masjid, asrama, atau kafe. Penelitian ini menunjukkan bahwa penggunaan Chatbot seperti UNISEL Bot dapat meningkatkan interaksi pengguna dengan informasi universitas secara efisien dan interaktif, serta menggantikan metode tradisional dalam menyajikan layanan pelanggan. | Paper ini mengembangkan UNISEL Bot untuk pemasaran interaktif di institusi akademik, menangani masalah penyajian informasi yang kurang interaktif. Metode Agile digunakan dengan pengumpulan data kualitatif. UNISEL Bot terdiri dari modul seperti pertanyaan, umpan balik, registrasi, acara, janji, survei, dan peta, dengan harapan menggantikan layanan pelanggan manual.  Sementara itu, paper penulis menggunakan RAG untuk membuat Chatbot Bina Nusantara yang canggih dengan koreksi ejaan dan pencarian informasi FAISS. Ini menunjukkan pendekatan yang lebih terintegrasi dalam menyajikan informasi yang akurat dan relevan kepada pengguna. |
| (Le Hoanh Su, T. D. H., Thi-Yen-Linh, T., Thi-Duyen-Ngoc, N., Bao-Tuyen, L., & Ha-Phuong-Truc, N., 2020) | Mengembangkan dan menerapkan chatbot berbasis kecerdasan buatan yang dapat mendukung proses penerimaan dan bimbingan karir di perguruan tinggi. Penelitian ini bertujuan untuk mengeksplorasi, menganalisis, dan memahami kesulitan, kekurangan, dan masalah yang muncul dalam konseling karir dan dukungan penerimaan mahasiswa. Metode yang digunakan meliputi pembuatan dataset yang mendukung bimbingan karir dan kegiatan penerimaan mahasiswa, serta desain dan integrasi chatbot ke dalam sistem sekolah untuk mendukung proses konseling penerimaan. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi dalam mengatasi dan meningkatkan kinerja konsultasi penerimaan universitas. | Tesis tersebut mengidentifikasi masalah dalam konseling karir dan pendaftaran mahasiswa di universitas. Dengan membangun dataset dan menerapkan teknologi AI seperti pemrosesan bahasa alami dan pembelajaran mesin, tesis tersebut berupaya meningkatkan kinerja konsultasi penerimaan universitas. Dengan chatbot yang dibangun, Le Hoanh Su, et al menawarkan solusi untuk mengatasi tantangan tersebut dan memperbaiki layanan akademik. | Tesis "Membangun Chatbot AI untuk Mendukung Penerimaan dan Bimbingan Karir di Perguruan Tinggi" mengeksplorasi, menganalisis, dan memahami kesulitan, kekurangan, dan masalah yang muncul dalam konseling karir dan dukungan penerimaan mahasiswa. Tesis ini menyimpulkan bahwa membangun chatbot adalah solusi yang sempurna untuk mendukung proses ini, karena chatbot dapat tersedia 24/7 selama 365 hari dalam setahun. Chatbot juga dapat menerima dan memproses permintaan dari mahasiswa/orangtua secara otomatis dan cepat, serta memberikan balasan secara seragam, terutama dioptimalkan dalam skenario-skenario yang repetitif. Selanjutnya, tesis ini merancang dan mengintegrasikan chatbot ke dalam sistem sekolah untuk mendukung proses konseling penerimaan.  Sementara itu, paper penulis menggunakan metode Retrieval-Augmented Generation (RAG) untuk membuat chatbot yang bertujuan untuk memberikan informasi tentang Universitas Bina Nusantara, dilengkapi dengan koreksi ejaan dan menggunakan FAISS (Facebook AI Similarity Search) untuk pencarian informasi. Dengan pendekatan ini, paper penulis mencoba untuk mengatasi masalah komunikasi yang efektif antara universitas dan calon mahasiswa dengan memanfaatkan teknologi canggih dalam pemrosesan bahasa alami dan pencarian informasi. Dalam kedua kasus, baik tesis maupun paper, chatbot digunakan sebagai alat untuk mendukung proses penerimaan mahasiswa di perguruan tinggi, meskipun dengan pendekatan dan fokus yang berbeda. |
| (Muhyidin, A., & Setiawan, M. A. F., 2021) | Mengembangkan media komunikasi hubungan masyarakat berbasis chatbot yang mampu mengatasi keterbatasan layanan dalam berinteraksi dan memberikan informasi tentang UNY kepada masyarakat, (2) memastikan kualitas perangkat lunak yang dikembangkan sesuai dengan model ISO/IEC 25010. Metode penelitian yang digunakan adalah Research and Development (R&D) dengan model pengembangan perangkat lunak Agile: Scrum. Hasil dari penelitian ini adalah (1) media komunikasi hubungan masyarakat berbasis chatbot yang dinamakan UNYSA sebagai layanan informasi tentang UNY, (2) UNYSA memenuhi standar kualitas ISO/IEC 25010 pada karakteristik kesesuaian fungsional dengan skor ahli pengembang perangkat lunak sebesar 0,86 (sebagian besar fitur berhasil diimplementasikan), ahli materi sebesar 98,82% (Sangat Layak), UAT sebesar 83,95% (Sangat Baik); efisiensi kinerja memiliki respons laten 15 ms (lebih cepat dari kecepatan standar: 2 detik); portabilitas 100% (mendukung semua aplikasi frontend yang direncanakan); kegunaan sebesar 83,81% (Sangat Layak); kehandalan sebesar 92% (Sangat Layak), dan maintainabilitas dengan indeks maintainability sebesar 83,21 (sangat mudah untuk dipelihara). | Penelitian ini menghasilkan chatbot UNYSA (UNY Smart Assistant) sebagai media komunikasi hubungan masyarakat di sekitar kampus Universitas Negeri Yogyakarta. Chatbot ini dapat diakses melalui berbagai aplikasi, seperti WhatsApp, Telegram, Line, Messenger, dan Google Assistant. UNYSA mampu menjawab berbagai pertanyaan seputar registrasi, biaya kuliah, akademik, beasiswa, profil, dan program studi dengan cepat dan akurat. Hasil pengujian menunjukkan bahwa chatbot ini memenuhi standar kualitas ISO/IEC 20510 dengan baik. | Penelitian ini bertujuan untuk mengembangkan media komunikasi berbasis chatbot untuk UNY dan memastikan kualitas perangkat lunak sesuai dengan standar ISO/IEC 25010. Fokusnya adalah memperbaiki layanan interaksi dan penyediaan informasi tentang UNY kepada masyarakat umum.  Di sisi lain, penelitian penulis menggunakan metode RAG untuk menciptakan chatbot yang memberikan informasi tentang Universitas Bina Nusantara dengan fitur tambahan spell correction dan FAISS. Dua penelitian ini mewakili upaya yang berbeda dalam pengembangan chatbot untuk keperluan informasi universitas dengan pendekatan yang berbeda pula. |
| (Shchegoleva, L., & Burdin, G., 2021) | Mengembangkan dan memperkenalkan prototipe sistem tanya jawab yang dapat otomatis mengatasi pertanyaan dari calon mahasiswa selama proses penerimaan di Universitas Negri Petrozavodsk. Ini bertujuan untuk meningkatkan efisiensi dan responsivitas dalam memberikan informasi kepada calon mahasiswa melalui penggunaan teknologi pemrosesan bahasa alami dan sistem pencarian yang canggih. | Paper ini membahas prototipe sistem tanya jawab untuk mengotomatisasi interaksi dengan calon mahasiswa selama kampanye penerimaan Universitas Negeri Petrozavodsk. Pendekatan utama dalam otomatisasi meliputi klasifikasi pertanyaan, aturan pembentukan template jawaban, dan model vektor mesin pencari untuk kumpulan jawaban yang mungkin. Inti dari sistem tanya jawab adalah metode pemrosesan bahasa alami, termasuk tokenisasi, lemmatisasi, dan analisis morfologis serta sintaksis. Prototipe sistem tanya jawab diimplementasikan dalam bentuk chatbot untuk jaringan sosial. Pengujian menunjukkan tingkat kepresisian sebesar 60% dalam menjawab pertanyaan yang diterima melalui email, dengan penelitian lanjutan sedang dilakukan untuk meningkatkan metrik tersebut. | Paper ini membahas prototipe sistem tanya jawab untuk mengotomatisasi interaksi dengan calon mahasiswa selama kampanye penerimaan di Universitas Negri Petrozavodsk. Pendekatan utama meliputi klasifikasi pertanyaan, pembentukan aturan respons, dan model vektor mesin pencari untuk jawaban yang mungkin. Intinya adalah metode pemrosesan bahasa alami. Prototipe ini diimplementasikan sebagai chatbot untuk jaringan sosial.  Sementara itu, dalam paper penulis, RAG digunakan untuk membuat chatbot yang membantu memberikan informasi tentang Universitas Bina Nusantara. Dilengkapi dengan fitur spell correction dan FAISS, chatbot ini membantu memberikan informasi yang akurat dan responsif kepada pengguna. Meskipun tujuannya sama untuk mengotomatisasi layanan informasi, pendekatan yang digunakan dan implementasinya berbeda, dengan fokus pada penggunaan RAG dan fitur tambahan dalam penyediaan layanan informasi yang lebih efektif. |

Sebagai rangkuman, makalah-makalah tersebut memiliki beberapa tujuan yang berbeda terkait pengembangan sistem chatbot atau aplikasi berbasis kecerdasan buatan untuk berbagai keperluan di lingkungan pendidikan dan juga kesehatan. Salah satu tujuannya adalah untuk meningkatkan interaktivitas antara pengguna dengan sistem, baik dalam hal memberikan informasi mengenai program studi, kurikulum, maupun bantuan dalam hal pendaftaran dan konseling karir. Beberapa makalah juga bertujuan untuk meningkatkan kualitas pelayanan kesehatan di lingkungan kampus dengan menyediakan sistem ChatBot berbasis pemrosesan bahasa alami yang dapat memberikan informasi medis dan menganalisis sentimen pengguna terkait pelayanan kesehatan. Selain itu, beberapa makalah berfokus pada pengembangan model chatbot yang mampu memberikan jawaban spesifik dan komprehensif terhadap berbagai pertanyaan pengguna melalui penggabungan teknik generasi teks dan pengambilan informasi. Secara umum, tujuan dari semua makalah tersebut adalah untuk mengoptimalkan pemanfaatan teknologi dalam meningkatkan layanan dan interaksi di lingkungan pendidikan dan pelayanan kesehatan.

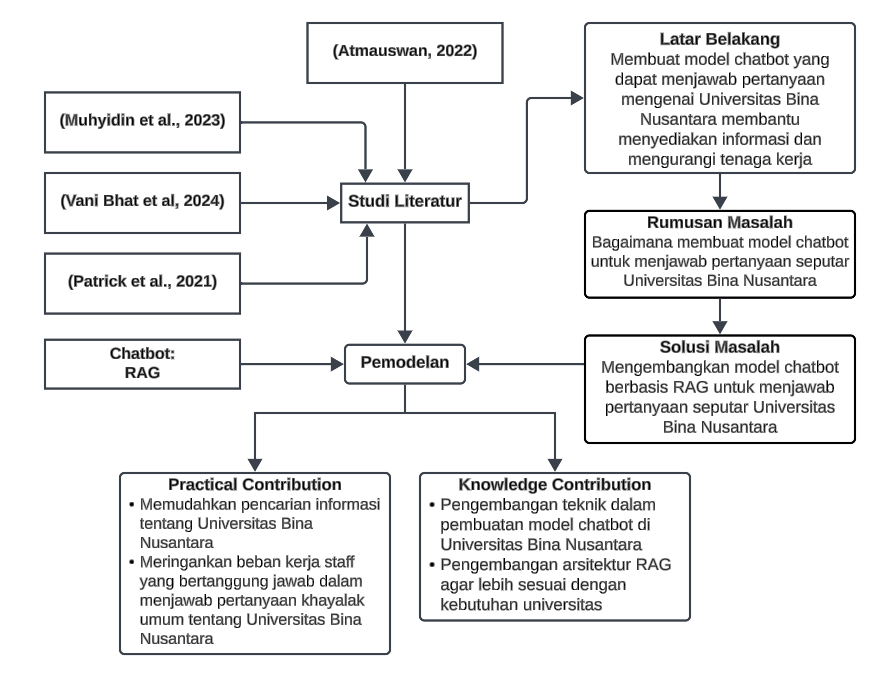
Hasil penelitian dari beberapa makalah diatas juga menunjukkan bahwa pengembangan dan penerapan chatbot di lingkungan perguruan tinggi dan juga pada sektor medis telah memberikan kontribusi positif dalam menyediakan informasi interaktif, meningkatkan layanan akademik dan medis seperti konseling karir, pendaftaran mahasiswa, dan juga penyediaan informasi medis. Chatbot-chatbot dari berbagai hasil penelitian diatas menggunakan kecerdasan buatan dan pembelajaran mesin untuk merespons pertanyaan pengguna, memprediksi penyakit berdasarkan gejala, dan memberikan informasi obat dengan akurat.

Penelitian yang telah dijabarkan membahas mengenai pengembangan chatbot dalam berbagai konteks, khususnya di lingkungan perguruan tinggi. Penelitian-penelitian tersebut juga menunjukkan variasi dalam pendekatan dan tujuan. Beberapa penelitian menggunakan data gabungan dari database lokal dan internet, sementara yang lain menggunakan database lokal dari institusi mereka masing-masing. Fokusnya juga bervariasi, dengan beberapa penelitian lebih menekankan aspek prediksi penyakit, pengembangan model generasi, dan pengembangan sistem cerdas umum, sementara penelitian ini fokus pada memberikan informasi yang komprehensif tentang institusi pendidikan Universitas Bina Nusantara. Meskipun beragam dalam pendekatan dan tujuan, semua penelitian ini bertujuan untuk meningkatkan layanan informasi dan interaksi melalui pengembangan chatbot.

# BAB 3 METODE PENELITIAN

## 3.1 Kerangka Berpikir

Kerangka berpikir adalah suatu alat bantu yang digunakan untuk menyusun dan mengorganisir konsep, ide, dan argumen secara sistematis, sehingga dapat memberikan gambaran yang jelas dan terstruktur mengenai topik yang dibahas. Kerangka berpikir akan digambarkan pada Gambar 3.1 dengan penjelasan bagian-bagiannya di bawahnya.



Gambar 3.1 Kerangka Berpikir

Dalam penelitian ini, peneliti menggunakan beberapa referensi yang dijadikan landasan dalam studi literatur, di antaranya yaitu:

Atmauswan (2022) membahas pembuatan chatbot universitas yang menyediakan jawaban akurat dan efisien untuk calon mahasiswa mengenai prosedur penerimaan, program studi, dan informasi beasiswa. Muhyidin et al. (2023) mengembangkan media komunikasi berbasis chatbot untuk UNY dengan memastikan kualitas perangkat lunaknya sesuai dengan standar ISO/IEC 25010, yang bertujuan untuk meningkatkan interaksi dan penyediaan informasi mengenai UNY kepada masyarakat. Vani Bhat et al. (2024) menyoroti peran komputer dan pentingnya chatbot berbasis teks dengan mengimplementasikan pengenalan pola untuk respons terdefinisi pada input pengguna, berfokus pada teknik desain dan implementasi chatbot untuk memahami potensinya di berbagai bidang. Patrick et al. (2021) mengembangkan model Retrieval-Augmented Generation (RAG) yang bertujuan meningkatkan akses dan manipulasi pengetahuan dalam tugas NLP intensif, dengan menggabungkan memori parametrik dan non-parametrik untuk menghasilkan bahasa melalui model RAG.

Referensi-referensi ini memberikan dasar teoritis dan empiris yang kuat bagi penelitian ini, terutama dalam konteks pengembangan model chatbot berbasis RAG (Retrieval-Augmented Generation) dalam ruang lingkup Universitas.

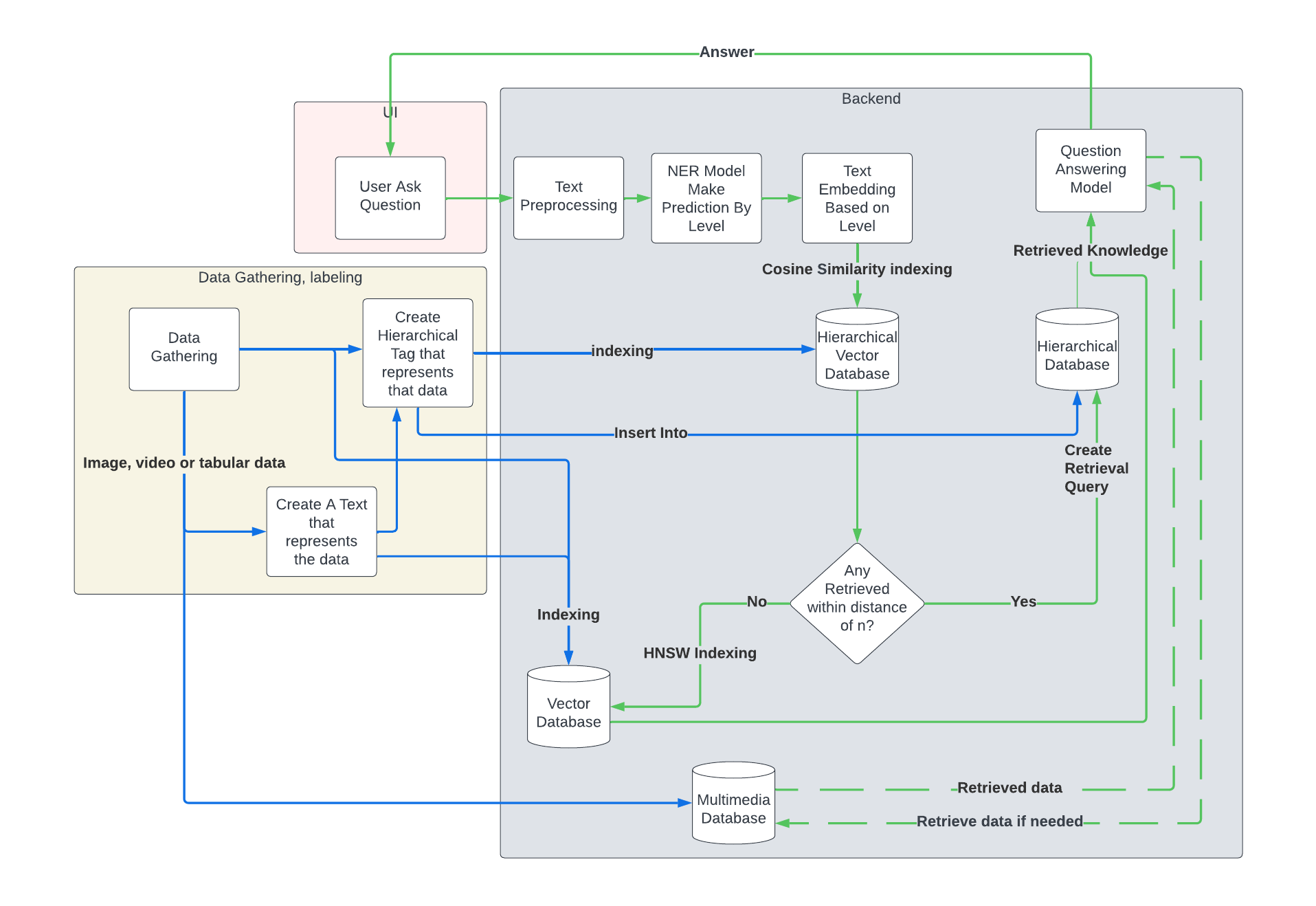
Sebagaimana yang dijelaskan pada BAB I, Latar belakang penelitian ini berfokus pada Universitas Bina Nusantara, yang bertujuan untuk membangun sebuah chatbot menggunakan data mengenai Universitas Bina Nusantara untuk memberikan informasi umum mengenai Universitas Bina Nusantara. Data yang digunakan diambil dari kumpulan website Universitas Bina Nusantara yang disatukan menjadi sebuah *text file.* Terdapat juga beberapa kendala ketika membangun chatbot ini, seperti kurangnya data dalam ukuran yang besar untuk melatih chatbot sehingga langkah selanjutnya yang mungkin dapat diambil adalah dengan menambahkan data dan wawasan baru ke dalam model chatbot.

Rumusan masalah dalam penelitian ini adalah untuk mengimplementasikan pembuatan model chatbot yang mampu menjawab pertanyaan seputar Universitas Bina Nusantara dengan memberikan informasi yang relevan, akurat, dan lengkap sehingga memudahkan pengguna dalam mengakses informasi penting terkait universitas ataupun menyederhanakan proses pencarian informasi, menghemat waktu mahasiswa, dan mengurangi beban kerja staf universitas.

Pada tahap implementasi, data yang telah diproses akan digunakan untuk melatih model dengan embedding teks menggunakan BERT, diindeks menggunakan HNSW, disimpan dalam Vector dan Multimedia Database, dilengkapi dengan typo corrector, dan diimplementasikan dalam model QA berbasis arsitektur RAG untuk chatbot yang dapat menjawab pertanyaan pengguna secara akurat. Dari hasil implementasi tersebut, penelitian ini memberikan kontribusi dalam dua aspek utama, yaitu kontribusi praktis dan kontribusi pengetahuan.

## 3.2 Metode Penelitian

Metode penelitian yang digunakan pada penelitian ini, dipaparkan pada gambar 3.2, Diagram ini menunjukkan alur proses secara keseluruhan, mulai dari interaksi pengguna hingga menghasilkan jawaban. Setiap komponen dalam diagram akan dijelaskan lebih lanjut pada sub bab 3.2.1. Tanda panah hijau mewakili proses saat pengguna memberikan pertanyaan, sedangkan tanda panah biru menggambarkan langkah-langkah persiapan data yang dilakukan sebelum model di *deploy*.



Gambar 3.2 Metode Penelitian

Proses penyiapan data bertujuan untuk memastikan kelancaran seluruh proses utama pada aplikasi, yaitu saat user memberi pertanyaan, keseluruhan kedua proses tersebut meliputi:

1. *Data Gathering* dan *Labeling*

Proses Data Gathering merupakan tahapan penting dalam pengembangan chatbot, di mana data yang dikumpulkan akan menjadi pengetahuan bagi chatbot tersebut. Pengetahuan ini akan berisi informasi umum terkait Universitas Bina Nusantara.

Informasi yang akan dikumpulkan untuk chatbot ini meliputi kategori berikut sebagai kategori utama:

1. Jurusan dan Mata Kuliah: Informasi tentang jurusan-jurusan yang tersedia di Universitas Bina Nusantara beserta daftar mata kuliah yang ditawarkan dalam masing-masing jurusan.
2. Organisasi Kemahasiswaan: Data mengenai organisasi kemahasiswaan yang aktif di kampus, termasuk nama organisasi, deskripsi singkat, dan kegiatan yang diadakan oleh organisasi tersebut.
3. Biaya Pendidikan: Informasi tentang biaya pendidikan, termasuk biaya kuliah per semester, biaya pendaftaran, dan biaya lainnya yang perlu diperhatikan oleh calon mahasiswa.
4. Dosen: Daftar dosen yang mengajar di universitas, beserta informasi singkat mengenai latar belakang pendidikan dan penelitian mereka.
5. Alumni: Data mengenai alumni terbaik universitas, termasuk profil singkat, pekerjaan saat ini, dan prestasi yang telah diraih setelah lulus.
6. Enrichment dan Kegiatan Mahasiswa: Informasi tentang program-program enrichment dan kegiatan mahasiswa di kampus, seperti seminar, workshop, dan kegiatan ekstrakurikuler lainnya.
7. Fasilitas Kampus: Data mengenai fasilitas-fasilitas yang tersedia di kampus, seperti perpustakaan, laboratorium, dan fasilitas umum lainnya yang dapat digunakan oleh mahasiswa.

Contoh pelabelan data hierarki yang digunakan terdapat pada lampiran.

Pada tahap pengolahan data, setiap teks yang telah dikumpulkan akan diberi label dengan mengorganisasikan informasi secara hierarkis berdasarkan bab dan subbab. Setiap level dalam hierarki ini akan diberi sebuah level key yang merepresentasikan kategori tertentu.

0: text biasa, bukan merupakan level

1: subbab paling bawah seperti titel pendidikan, durasi

2: lokasi/kontak himpunan

3: jurusan/visi misi himpunan

4: fakultas/arti logo/kontak fasilitas

5: strata/syarat administrasi program/nama himpunan/jenis fasilitas

6: [placeholder]

7: kategori utama

**Tabel 3.1 Contoh List NER[/TODO this is wrong, will fix it later,will move to lampiran]**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Level 7 | Level 6 | Level 5 | Level 4 | Level 3 | Level 2 | Level 1 |
| Jurusan |  | S1 | Makna warna pada logo | CONTACT | Informasi Mengenai Keuangan dan Registrasi | LAYANAN KESEJAHTERAAN MAHASISWA |
| Jurusan |  | S2 | Makna bentuk pada logo | Apa saja prosedur pendaftaran? | Informasi mengenai Perkuliahan dan Ujian | 3. MENTORING |
| Organisasi Kemahasiswaan |  | Himpunan Mahasiswa Matematika | Mars HIMMAT | Apa saja dokumen yang dibutuhkan untuk registrasi? | Informasi Mengenai Studi Mahasiswa | 2. LAYANAN PEMANTAUAN AKADEMIK |
| Organisasi Kemahasiswaan |  | HIMPUNAN MAHASISWA STATISTIKA | Visi | Bagaimana cara melakukan pendaftaran? | BINUS @Bekasi | LAYANAN DISABILITAS |
| Organisasi Kemahasiswaan |  | HIMPUNAN MAHASISWA TEKNIK INFORMATIKA | Misi | TOEFL | BINUS @Bandung | DUKUNGAN SUMBER DAYA BELAJAR |
| Organisasi Kemahasiswaan |  | HIMPUNAN MAHASISWA SISTEM INFORMASI | Semboyan | Fast Track TOEFL Class (FTT) | BINUS @Malang | durasi |
| Organisasi Kemahasiswaan |  | HIMPUNAN MAHASISWA DKV | TAGLINE | Public Speaking and Performing | BINUS @Semarang | title |
| Organisasi Kemahasiswaan |  | HIMPUNAN MAHASISWA MANAGEMENT | PROFILE | Debate | BINUS ASO | S.Ak. |
| Organisasi Kemahasiswaan |  | HIMPUNAN MAHASISWA MARKETING COMMUNICATION | Journalism Club | Conversation Class | BINUS Online Learning | S.Ds. & BA (Hons) (Optional) |

Data yang telah di label ini kemudian akan diproses lagi, untuk digunakan dengan:

1. Tabel SQL

Data yang telah diberi label akan disimpan dalam database SQL dalam bentuk data per paragraf. Setiap paragraf akan memiliki beberapa level key yang mencerminkan struktur hirarkinya.

2. Pembuatan data untuk melatih model

Data berlabel ini akan digunakan untuk membuat kumpulan pertanyaan. Setiap pertanyaan akan diberi anotasi dengan teknik *sequence labeling* untuk menentukan level dari setiap kata di dalamnya.

3. Pembuatan data untuk indexing untuk tiap key

Data berlabel akan diproses untuk mengumpulkan level key yang unik pada setiap level hierarki.

Setelah proses pelabelan, data akan melalui proses indexing. Penelitian ini menggunakan library FAISS untuk melakukan indexing dan pencarian pada index. Teknik yang digunakan mencakup cosine similarity dan HNSW:

* + 1. Cosine Similarity

Teknik ini digunakan untuk menghitung tingkat kemiripan antara vektor suatu kata dengan kata-kata dalam basis data. Asumsi yang digunakan adalah bahwa pertanyaan pengguna yang telah diklasifikasikan oleh model NER (Named Entity Recognition) pada tiap level akan memiliki konteks yang mirip dengan kata kunci dalam index.

* + 1. HNSW

Teknik ini digunakan karena keunggulannya dalam efisiensi pencarian pada Semantic Similarity Search. Dengan HNSW, sistem dapat mencari representasi vektor pengetahuan yang relevan dengan pertanyaan pengguna secara lebih akurat, bahkan pada skala data yang besar. Teknik ini ideal untuk kasus di mana hasil pencarian menggunakan cosine similarity menunjukkan tingkat kemiripan yang rendah.

Melalui proses indexing ini, sistem dapat melakukan pencarian informasi secara cepat dan akurat, sehingga memungkinkan pengguna untuk mengakses pengetahuan yang relevan dengan lebih efisien.

Data *indexing* yang disimpan ke *vector database* meliputi:

1. Data Level Key Unik

Data berlabel diproses untuk mengumpulkan level key unik pada setiap level hierarki.

1. Data Teks Asli

Jika hasil pencarian pada index tidak menemukan jarak (similarity) yang lebih besar dari *threshold* (penelitian ini menggunakan n = 0.75), sistem akan melakukan pencarian ulang menggunakan data asli. Dalam proses penyimpanan index ini, data asli dipecah menjadi per kalimat. Setiap kalimat diolah menjadi index untuk memaksimalkan kemungkinan pencarian.

1. Database Design

Jika pencarian pada index tidak menemukan jarak (similarity) yang lebih kecil dari *threshold*, maka sistem akan melakukan query ke database. Database yang digunakan dalam penelitian ini dibagi menjadi tiga jenis:

1. *Hierarchical Database* digunakan untuk menyimpan level keys dan paragraf asli, teks kunci ini adalah data yang telah di label pada bagian data gathering dan di simpan di tabel sql, nantinya akan digunakan dalam query construction untuk mencari informasi lebih lanjut. Hierarchical Database akan diakses setelah menerima output cosine similarity search pada setiap level dari hasil model NER, untuk menentukan apakah perlu mengakses database lain dalam proses pencarian.

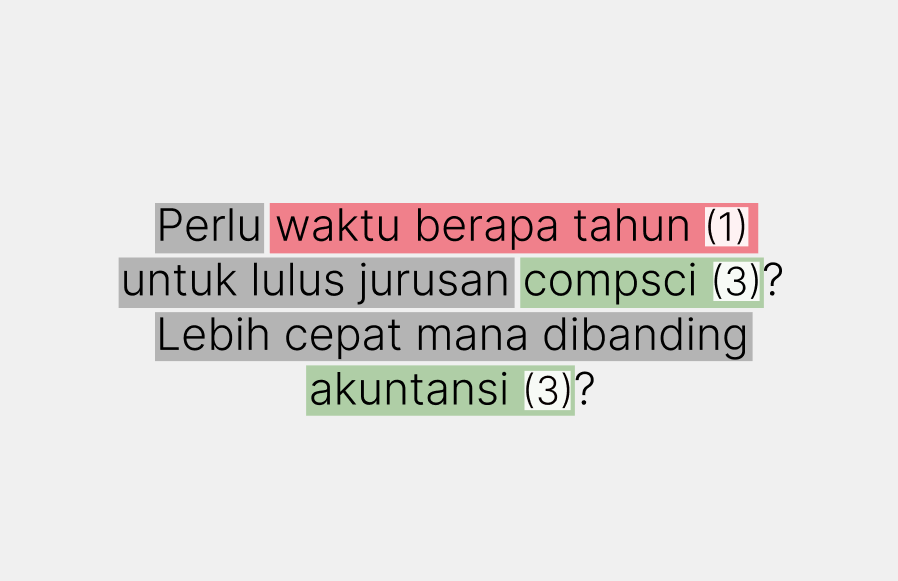
Tabel 3.2 Hierarchical Database Schema

|  |  |
| --- | --- |
| Kolom | Tipe Data |
| Text | VARCHAR(750) |
| level\_1 | VARCHAR(50) |
| level\_2 | VARCHAR(50) |
| level\_3 | VARCHAR(50) |
| level\_4 | VARCHAR(50) |
| level\_5 | VARCHAR(50) |
| level\_6 | VARCHAR(50) |
| level\_7 | VARCHAR(50) |

*Schema database* yang digunakan dipaparkan pada gambar 3.1, tujuan *schema* ini adalah untuk memungkinkan proses *query construction* atau proses formulasi query SQL dengan menambahkan kondisi tertentu pada tiap level. Dalam proses *query construction*, terdapat beberapa kondisi yang digunakan:

* + 1. **Penggunaan OR untuk Beberapa Level Sama**: Jika dalam satu pertanyaan terdapat beberapa level yang sama, maka kondisi pada query akan menggunakan operator OR.
    2. **Penggunaan AND untuk Beberapa Level Berbeda**: Jika dalam satu pertanyaan terdapat level yang berbeda, maka kondisi pada query akan menggunakan operator AND.

Sebagai contoh, pertanyaan di sini dibagi menjadi beberapa bagian:

 Gambar 3.2.2 Contoh NER Labeling

Model NER akan menghasilkan *query construction* seperti di atas, di mana teks dengan *highlight* abu-abu tidak memiliki level atau teks biasa, teks dengan *highlight* merah masuk ke level 1, dan teks dengan *highlight* hijau masuk ke level 3. Kemudian, hasil dari *query* akan dimasukkan ke QA model sebagai pengetahuan untuk menjawab pertanyaan pengguna:

**WHERE level\_1 = ‘DURASI’**  
**AND (iii) (level\_3 = ‘COMPUTER SCIENCE’ OR level\_3 = ‘AKUNTANSI’) (ii)**

Cosine similarity search dilakukan berdasarkan semua level key yang dihasilkan dari model NER, hasil dari cosine similarity ini kemudian akan digunakan untuk melakukan *query construction*.

1. *Vector Database* dalam penelitian ini dirancang untuk menyimpan representasi vektor dari data. Penelitian ini menggunakan 2 vector database yang menyimpan data yang telah diolah pada proses indexing seperti dijelaskan pada bab 3.2 no 1.
2. *Multimedia Database* dalam penelitian ini berfungsi sebagai tempat penyimpanan gambar atau media lainnya. Database ini hanya akan diakses apabila model Question Answering (QA) menentukan bahwa data multimedia diperlukan untuk menjawab pertanyaan pengguna. Pengambilan data multimedia dilakukan jika teks yang merepresentasikan data multimedia ada dalam jawaban model QA. Teks tersebut kemudian digunakan untuk mengambil data multimedia dari database.
3. Arsitektur model NER

Saat pengguna mengajukan pertanyaan, model bertanggung jawab untuk memberikan label pada setiap kata dalam pertanyaan tersebut sesuai dengan level yang telah ditentukan dalam tahap Data Gathering. Model ini dirancang untuk melakukan tugas NER yang bertujuan mengklasifikasikan setiap kata dalam pertanyaan pengguna ke dalam salah satu dari 8 level yang sudah ditentukan. Berbeda dengan NER konvensional, model ini mengklasifikasikan entitas secara hierarkis, mengikuti struktur level yang ada.

Hasil klasifikasi dari NER ini digunakan untuk menyusun kondisi pada query SQL, yang nantinya akan digunakan untuk mengambil informasi relevan dari basis data. Informasi yang diperoleh dari query SQL ini menjadi sumber pengetahuan bagi model question answering (QA) untuk memberikan jawaban yang tepat dan akurat terhadap pertanyaan pengguna.

1. *Question Answering* model

Question Answering (QA) model adalah jenis model NLP yang dirancang untuk menjawab pertanyaan pengguna. Penelitian ini menggunakan arsitektur RAG yang memungkinkan model QA untuk menjawab berdasarkan informasi yang telah didapatkan setelah *database query*. Model QA yang digunakan pada penelitian ini adalah model BERT yang telah di fine tune dengan dataset SQuAD (Stanford Question Answering Dataset) IndoBERT-SQuAD. Penggunaan model BERT yang telah di fine-tune dengan dataset SQuAD memungkinkan model untuk memahami pertanyaan dengan lebih baik dan memberikan jawaban yang akurat berdasarkan konteks pertanyaan. Selain menggunakan IndoBERT-SQuAD, penelitian ini juga memanfaatkan Gemini AI sebagai model tambahan untuk memperkaya representasi bahasa dan meningkatkan kemampuan model dalam memahami nuansa yang lebih kompleks dalam pertanyaan, serta sebagai pembanding kemampuan menjawab pertanyaan antara indoBERT dengan gemini AI. Gemini AI adalah sebuah platform kecerdasan buatan (AI) yang dikembangkan oleh Google DeepMind. Diluncurkan pertama kali pada Maret 2023, Gemini AI adalah sebuah model bahasa besar (large language model/LLM) yang dikembangkan oleh Google DeepMind.

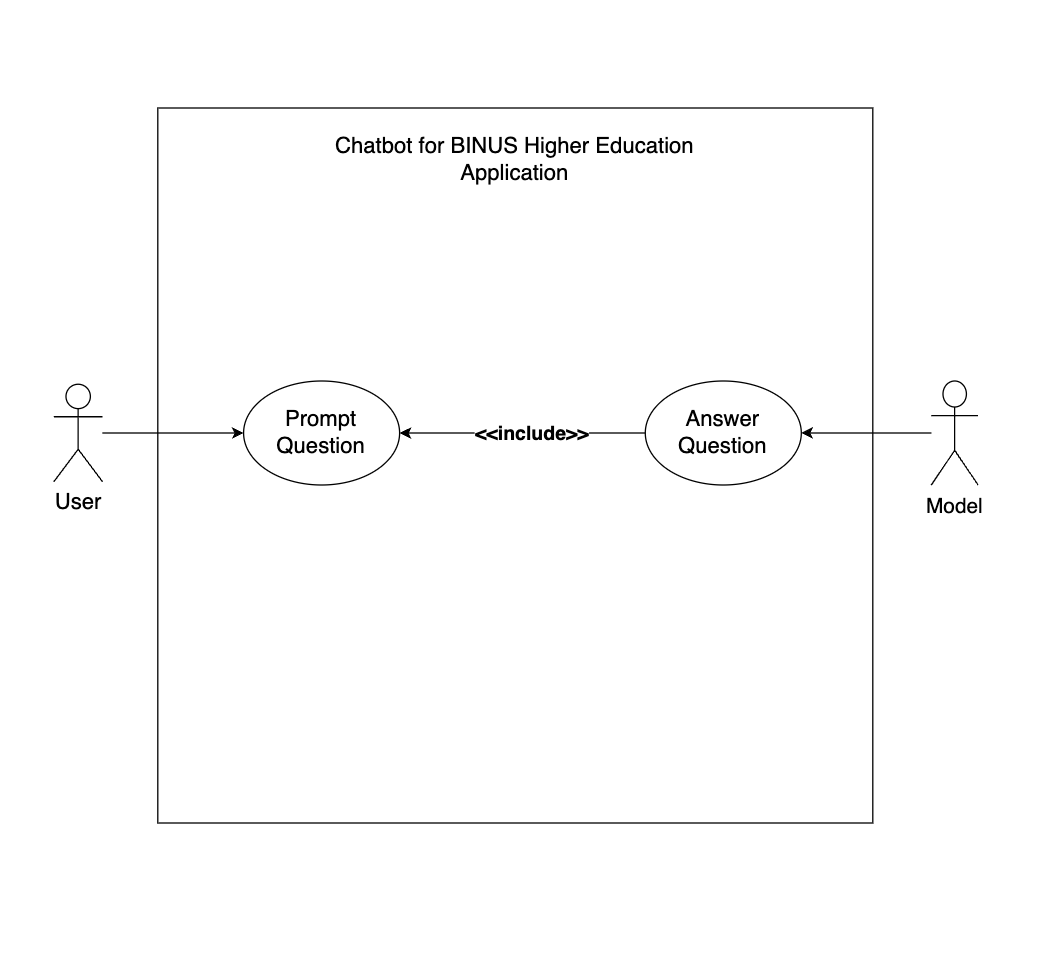
1. Penerapan model

Model yang dilatih akan diimplementasikan dalam sebuah aplikasi web yang dibangun dengan menggunakan kerangka kerja FastAPI sebagai backend dan Streamlit sebagai frontend. FastAPI akan berperan sebagai server API, mengelola HTTP *request* dari pengguna, menjalankan inferensi model, dan berinteraksi dengan database untuk mendapatkan informasi yang relevan. Streamlit akan menyediakan antarmuka pengguna yang interaktif, memungkinkan pengguna untuk memasukkan pertanyaan teks dan menerima jawaban secara real-time. Aplikasi ini akan difokuskan pada penyediaan informasi umum tentang Universitas Bina Nusantara. Selain itu, aplikasi ini juga akan dilengkapi dengan fitur umpan balik pengguna, yang memungkinkan pengguna untuk memberikan masukan terkait kualitas jawaban model. Jika model tidak dapat menjawab pertanyaan pengguna, sistem akan mengarahkan pengguna ke layanan bantuan yang relevan.

## 3.3 Evaluasi

**/TODO**

## 3.4 Perancangan UML

Gambar 3.5 Use Case Diagram

### 3.4.1 Prompt Question

Di halaman utama, akan ditampilkan kotak teks sebagai tempat pengguna menuliskan pertanyaan kepada chatbot. Chatbot akan menjawab pertanyaan dari pengguna tersebut. Selain itu, pengguna juga akan diberikan beberapa rekomendasi *prompt*.

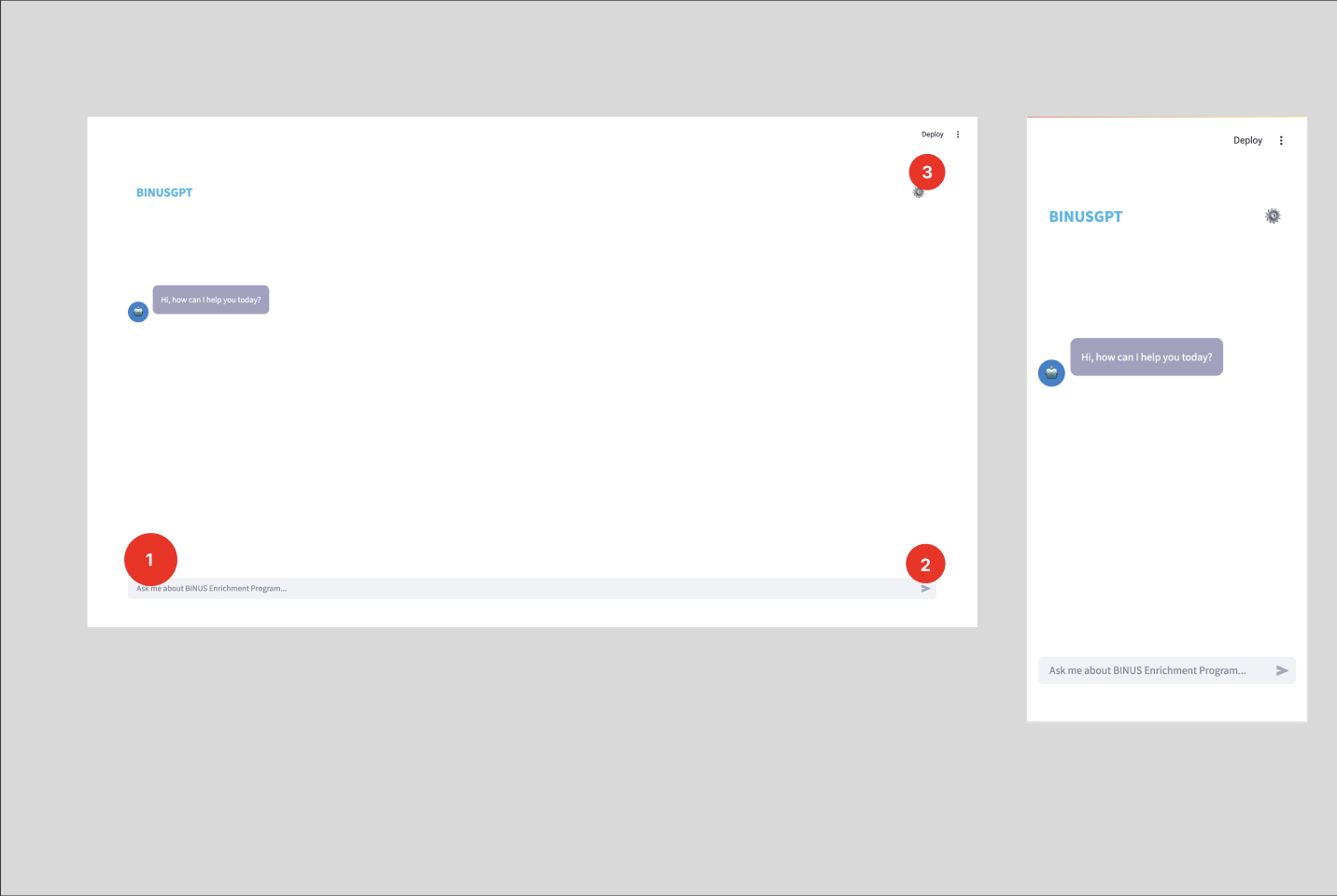
### 3.4.2 Give Feedback

Jika pengguna ingin memberi masukan, melaporkan masalah, ataupun ulasan terkait pengalaman menggunakan chatbot, pengguna dapat memberi *feedback*.

### 3.4.3 Get Help

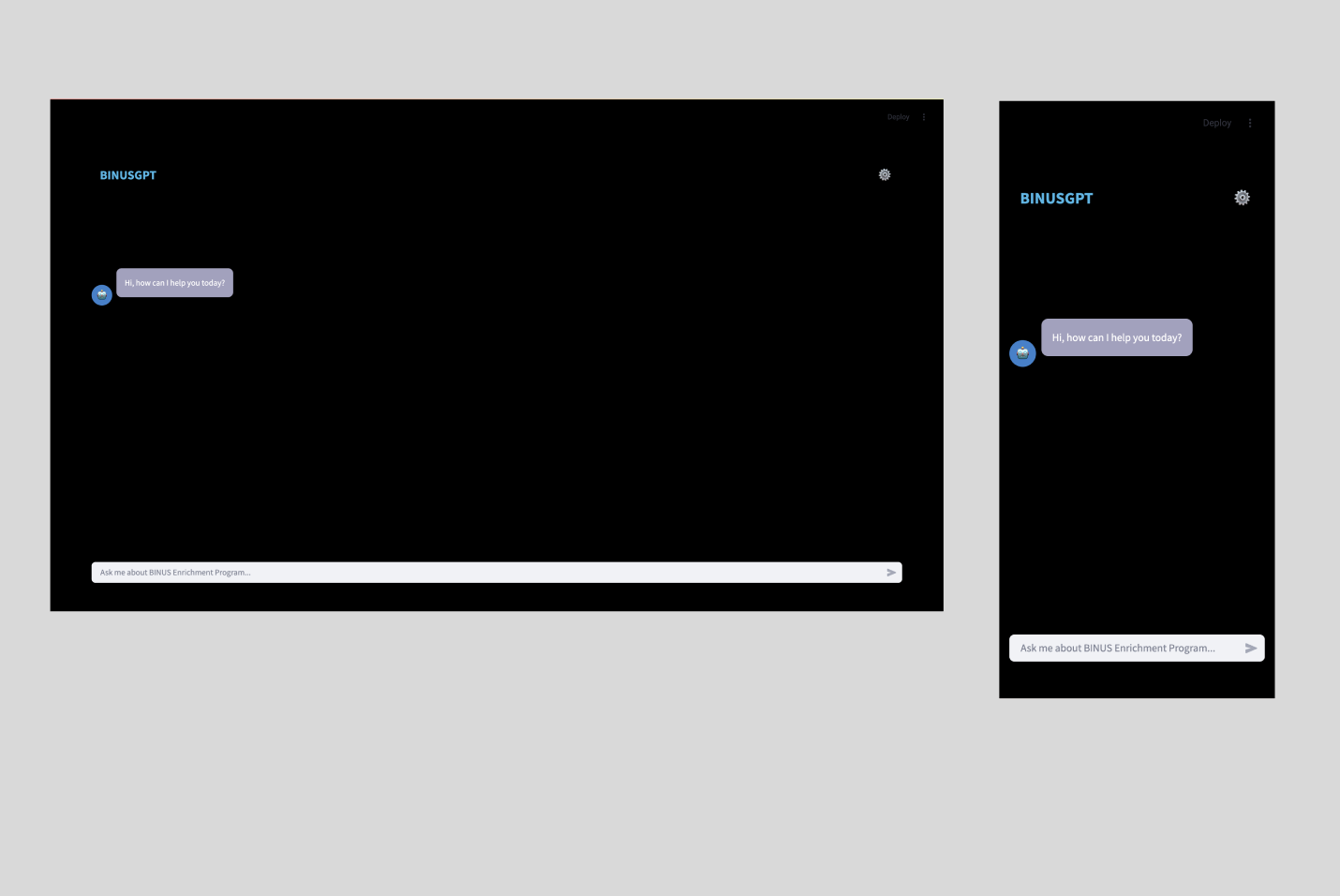
Jika jawaban yang diberikan AI kurang memuaskan pengguna ataupun pengguna kurang memahami jawaban, pengguna dapat meminta kontak dengan service center, yang akan membantu menjawab pertanyaan pengguna.

## 3.5 Perancangan UI

Gambar 3.6 Rancangan UI Mode Terang

Pada gambar 3.6 terdapat beberapa angka dalam lingkaran merah. Penjelasan dari masing-masing angka tersebut antara lain:

1. Nomor 1 merupakan *text box* yang akan menerima input prompt dari user.
2. Nomor 2 merupakan ke tombol *send* untuk mengirim input prompt dari user yang sudah diketik oleh user di teks prompt.
3. Nomor 3 merupakan ke tombol pengaturan, tombol pengaturan akan mengalihkan user ke halaman pengaturan, dimana halaman ini menyediakan pilihan seperti pengaturan umum, bahasa, dan keamanan.



Gambar 3.7 Rancangan UI Mode Gelap

Dalam perancangan *User Interface* aplikasi Chatbot penyedia informasi Universitas Bina Nusantara, peneliti merancang 2 jenis mode, yakni mode terang dan mode gelap. Dari kedua mode tersebut, peneliti juga merancang 2 jenis desain untuk jenis perangkat berbeda yakni berbasis Web dan juga berbasis Smartphone.

Dalam perancangan UI chatbot mode terang, peneliti memilih template warna dengan kode warna antara lain #3cb7e8, #A3A1BF, dan #ffffff. Sedangkan, dalam perancangan UI chatbot mode gelap, peneliti memilih template warna dengan kode warna antara lain #3cb7e8, #A3A1BF, #000000.

Adapun juga fungsi dari masing-masing mode, yakni:

**Fungsi Mode Terang:**

1. **Visibilitas yang lebih baik di lingkungan terang:**Di bawah sinar matahari atau di ruangan yang terang, mode terang lebih mudah dibaca karena memiliki kontras tinggi antara teks yang gelap dan latar belakang yang terang.
2. **Menjadi standar untuk sebagian besar antarmuka:**Mode terang telah menjadi pilihan umum untuk tampilan di berbagai aplikasi, sehingga pengguna merasa lebih familiar dan terbiasa dengan pengaturan ini.
3. **Kejelasan visual yang lebih baik:**Latar belakang yang cerah memudahkan pengguna untuk melihat detail gambar, grafik, atau elemen visual lainnya, sehingga fokus pada konten jadi lebih mudah.

**Fungsi Mode Gelap:**

1. **Mengurangi ketegangan mata di kondisi minim cahaya:**Saat digunakan di tempat yang kurang pencahayaan, mode gelap membuat layar tidak terlalu menyilaukan, sehingga mata lebih nyaman dan tidak cepat lelah.
2. **Mengurangi efek silau:**Mode gelap juga mengurangi intensitas cahaya dari layar, sehingga mengurangi risiko silau, terutama saat digunakan di tempat yang cahayanya rendah.
3. **Membantu meningkatkan fokus:**Banyak pengguna merasa mode gelap membantu mereka lebih fokus pada isi layar tanpa terganggu oleh pencahayaan yang terlalu terang di sekitarnya.

Selain dari perancangan mode, peneliti memilih tema dan konsep desain aplikasi ini didasari konsep minimalis untuk memberikan kesan bersih, rapi, dan juga menyoroti fungsi aplikasi BINUSGPT itu sendiri.

# BAB 4 HASIL DAN PEMBAHASAN

Bab ini menyajikan deskripsi cara penggunaan aplikasi chatbot serta hasil dan pembahasan dari pengujian dua model chatbot, yaitu IndoBERT-SQuAD dan Gemini AI dalam penyediaan informasi mengenai Universitas Bina Nusantara. Tujuan pengujian ini adalah untuk membandingkan performa kedua model dalam menjawab 30 test cases yang telah dirancang, serta mengevaluasi keakuratan dan relevansi jawaban yang diberikan oleh chatbot.

## 4.1 [Process bangun model] -> placeholder name

**/TODO**

## 4.2 Pengujian Model Chatbot

### 4.2.1 Test Case

Pengujian chatbot dilakukan menggunakan 30 test cases yang mewakili variasi pertanyaan yang umum diajukan seputar informasi di Universitas Bina Nusantara. Test case ini berfungsi sebagai alat evaluasi untuk mengukur performa dua model chatbot, yaitu IndoBERT-SQuAD dan Gemini AI, dalam menjawab pertanyaan berbasis pengetahuan Universitas Bina Nusantara. Test case dibagi menjadi tiga kategori berdasarkan jumlah kata dalam pertanyaan:

1. **Kategori Short (3-6 kata)**:  
   Berisi pertanyaan yang terdiri dari 3 hingga 6 kata, seperti "Jurusan di Binus?" atau "Biaya kuliah Binus." Pertanyaan pendek ini dirancang untuk mengevaluasi kemampuan chatbot dalam mengenali dan memberikan respons cepat terhadap permintaan yang singkat dan langsung. Kategori ini sering muncul dalam penggunaan sehari-hari, di mana pengguna ingin mendapatkan jawaban singkat tanpa detail berlebihan.
2. **Kategori Medium (7-10 kata)**:  
   Berisi pertanyaan dengan 7 hingga 10 kata, contohnya "Apa saja fasilitas kampus yang tersedia di Binus?" atau "Bagaimana cara bergabung dengan organisasi mahasiswa?" Kategori medium ini menguji sejauh mana model dapat memahami pertanyaan yang lebih spesifik.
3. **Kategori Long (≥11 kata)**:  
   Berisi pertanyaan yang lebih panjang, 11 kata atau lebih, seperti "Bisakah Anda memberi informasi mengenai proses pendaftaran untuk mahasiswa internasional di Binus?" atau "Apa persyaratan untuk mendaftar di program studi teknologi informasi di Binus?" Pertanyaan dalam kategori ini bertujuan untuk mengevaluasi pemahaman chatbot terhadap permintaan yang rumit dan rinci.

### 4.2.2 Performa Model

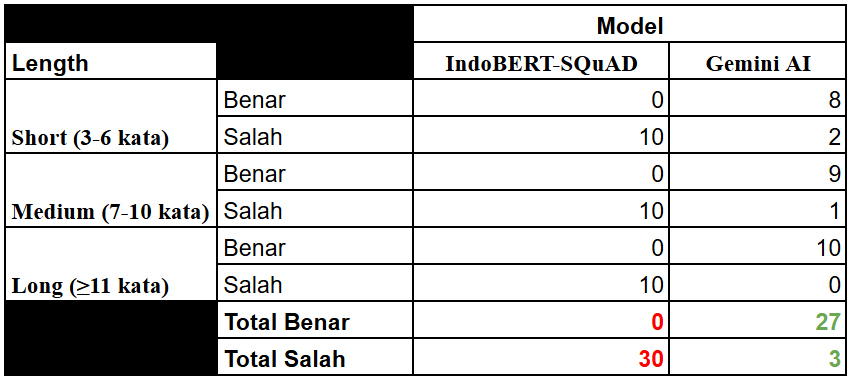
Untuk menilai performa dari masing-masing model, IndoBERT-SQuAD dan Gemini AI, terhadap 30 test cases yang disiapkan, digunakan metode voting yang melibatkan tiga evaluator. Setiap evaluator akan memberikan penilaian independen terhadap jawaban yang dihasilkan oleh model, menilai apakah output tersebut dianggap benar atau salah berdasarkan relevansi dan akurasi jawaban.

Setiap jawaban dari model akan ditentukan benar atau salah melalui proses voting dengan langkah-langkah berikut:

1. **Pemberian Vote oleh Evaluator**:  
   Ketiga evaluator memeriksa jawaban yang diberikan oleh model untuk setiap test case dan memberikan vote, yaitu “benar” atau “salah”, tergantung pada seberapa baik jawaban tersebut sesuai dengan konteks dan informasi yang diminta.
2. **Penentuan Mayoritas**:  
   Setelah semua evaluator memberikan vote, hasilnya dianalisis berdasarkan mayoritas. Jika dua dari tiga evaluator memilih “benar” dan satu memilih “salah,” maka hasil mayoritas (2 suara) akan diambil, dan jawaban model tersebut akan dianggap benar. Sebaliknya, jika mayoritas memilih “salah,” maka jawaban dianggap salah.
3. **Hasil Akhir Berdasarkan Voting**:  
   Hasil voting ini kemudian digunakan untuk menentukan performa model secara keseluruhan. Dengan metode ini, model dapat dinilai secara objektif karena setiap jawaban melewati proses penilaian kolektif dari tiga evaluator.

Melalui metode voting ini, penilaian hasil model menjadi lebih reliabel, mengurangi potensi bias, dan memberikan gambaran yang lebih adil mengenai performa masing-masing model. Dengan begitu, hasil perbandingan performa model dapat diukur dengan lebih akurat dan konsisten. Berikut adalah tabel hasil benar atau salahnya output kedua model berdasarkan proses voting.

Tabel 4.1 Hasil Voting Output Model



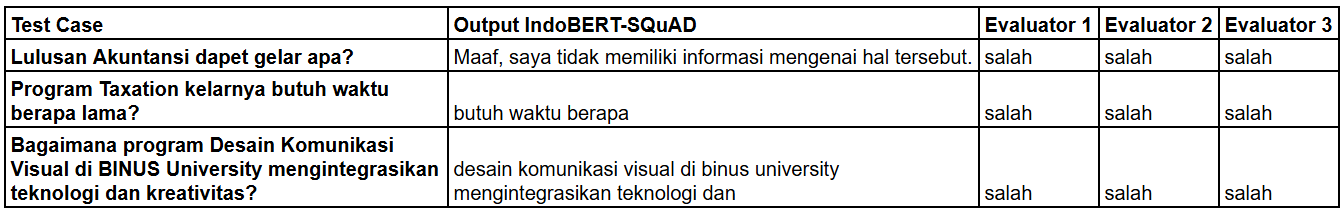
Hasil output model pada 30 test cases menunjukkan bahwa tidak ada output yang dianggap benar oleh mayoritas evaluator untuk model IndoBERT-SQuAD. Sedangkan untuk model Gemini AI, sebagian besar dari output yang dihasilkan dianggap benar oleh evaluator, dengan hasil 27 output benar dari 30 total test cases dan mendapatkan akurasi 90%. Hal ini menunjukkan bahwa model IndoBERT-SQuAD memiliki keterbatasan dalam memberikan output yang relevan terhadap konteks pertanyaan, atau memberikan output yang tidak lengkap.

Sementara itu, model Gemini AI menunjukkan kemampuan yang jauh lebih unggul dalam memahami konteks pertanyaan dan memberikan output yang relevan. Keunggulan Gemini AI ini dapat disebabkan oleh beberapa faktor, seperti arsitektur model yang mendukung pemrosesan dan pemahaman bahasa alami secara lebih mendalam. Performa Gemini AI yang konsisten pada ketiga jenis pertanyaan (short, medium, dan long) menunjukkan bahwa model ini lebih mampu menangani kompleksitas dan variasi dalam input pertanyaan dibandingkan IndoBERT-SQuAD.

### 4.2.3 Contoh Output

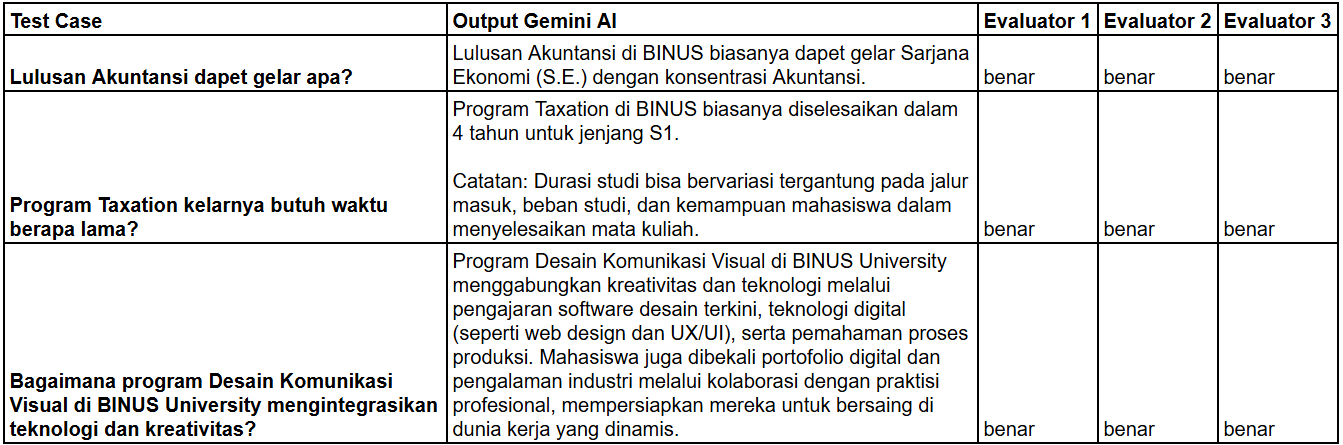
Bagian ini menyajikan contoh output dari kedua model chatbot, IndoBERT-SQuAD dan Gemini AI, untuk masing-masing kategori test case: short, medium, dan long. Ditampilkan satu test case dari setiap kategori sebagai sampel yang merepresentasikan berbagai tipe pertanyaan yang diujikan. Tabel 4.2 dan tabel 4.3 menunjukkan contoh pertanyaan yang digunakan sebagai test case serta membandingkan jawaban yang diberikan oleh kedua model untuk memahami perbedaan dalam akurasi dan kualitas respons antara model IndoBERT-SQuAD dan Gemini AI.

Tabel 4.2 Performa Model IndoBERT-SQuAD



Pada Tabel 4.2, model IndoBERT-SQuAD memberikan jawaban yang tidak sesuai. Untuk test case "Lulusan Akuntansi dapet gelar apa?", model hanya menjawab "Maaf, saya tidak memiliki informasi mengenai hal tersebut". Untuk test case kategori medium "Program Taxation kelarnya butuh waktu berapa lama?", jawaban model sangat singkat dan tidak spesifik, yaitu "butuh waktu berapa". Sementara itu, untuk kategori long "Bagaimana program Desain Komunikasi Visual di BINUS University mengintegrasikan teknologi dan kreativitas?", model hanya mengulang sebagian kata kunci dari pertanyaan "desain komunikasi visual di binus university mengintegrasikan teknologi dan" tanpa memberikan jawaban yang berarti. Ketiga jawaban ini dinilai salah oleh semua evaluator.

Tabel 4.3 Performa Model Gemini AI



Sebaliknya, pada Tabel 4.3 terlihat bahwa model Gemini AI memberikan jawaban yang lebih detail dan relevan. Pada test case "Lulusan Akuntansi dapet gelar apa", model menjawab dengan jelas bahwa "Lulusan Akuntansi di BINUS mendapat gelar Sarjana Ekonomi (S.E.) dengan konsentrasi Akuntansi". Untuk kategori medium "Program Taxation kelarnya butuh waktu berapa lama?", model memberikan jawaban yang komprehensif, menjelaskan bahwa "Program Taxation diselesaikan dalam 4 tahun untuk jenjang S1", disertai catatan tambahan tentang variasi durasi studi. Pada kategori long "Bagaimana program Desain Komunikasi Visual di BINUS University mengintegrasikan teknologi dan kreativitas?", jawaban model menjelaskan secara rinci bagaimana program Desain Komunikasi Visual di BINUS University mengintegrasikan teknologi terkini, seperti desain 3D dan UX/UI, serta pentingnya kolaborasi dengan praktisi profesional. Ketiga jawaban ini dinilai benar oleh semua evaluator.

## 4.3 Diskusi

Bagian ini bertujuan untuk membahas hasil dari evaluasi performa kedua model chatbot, IndoBERT-SQuAD dan Gemini AI, berdasarkan 30 test case yang telah dilakukan. Pembahasan akan mencakup kelebihan model Gemini AI, penyebab dari perbedaan akurasi antara kedua model, dan juga membahas lebih dalam mengenai beberapa jawaban yang salah dari Gemini AI.

Hasil evaluasi menunjukkan bahwa model Gemini AI memiliki keunggulan yang signifikan dibandingkan IndoBERT-SQuAD dalam memberikan jawaban yang relevan, akurat, dan sesuai konteks pertanyaan. Gemini AI mencapai tingkat akurasi sebesar 90% pada 30 test case yang diuji, sementara IndoBERT-SQuAD tidak memberikan jawaban yang dianggap benar oleh mayoritas evaluator. Keunggulan Gemini AI disebabkan oleh kemampuannya dalam memahami konteks pertanyaan secara lebih mendalam sehingga dapat memberikan jawaban yang lebih terstruktur dan informatif.

Namun, ada 3 jawaban dari Gemini AI dinilai salah oleh evaluator, ketiganya karena tidak sesuai dengan data knowledge yang tersedia di dalam database.

Tabel 4.4 Jawaban Salah Model Gemini AI

 Untuk test case "Program Business Creation di lokasi mana?”, model memberikan output "Program Business Creation ada di BINUS University." Jawaban ini tidak spesifik, berdasarkan data knowledge yang tersedia, Program Business Creation terletak di kampus tertentu di BINUS University. Jawaban ini dinilai salah karena model tidak memberikan informasi lokasi yang spesifik sesuai data, yang menunjukkan kemungkinan keterbatasan akses atau interpretasi terhadap detail lokasi dalam database.

Untuk test case “Apa akreditasi program Fashion?”, model memberikan output “Maaf, saya tidak memiliki informasi tentang akreditasi program Fashion di BINUS.” dan memberikan saran untuk mengunjungi website resmi BINUS untuk mendapatkan informasi yang detail. Jawaban ini menunjukkan bahwa model tidak mengambil jawaban yang sesuai dengan data knowledge yang tersedia dalam database mengenai akreditasi program Fashion, yaitu program Fashion memiliki akreditasi “A”. Meskipun model memberikan respons yang sopan dan menyarankan tindakan alternatif, jawaban ini tetap dinilai salah karena tidak memberikan informasi yang spesifik dan sesuai kebutuhan pengguna.

Untuk test case “Akreditasi Entrepreneurship Business Creation di BINUS apa?”, model memberikan output “Akreditasi Entrepreneurship Business Creation di BINUS adalah A.”. Jawaban ini salah karena data yang tersedia di knowledge base menyebutkan bahwa akreditasi program ini adalah "Unggul" Kesalahan ini menunjukkan bahwa model mungkin telah mengakses data yang salah, atau terjadi kesalahan dalam penyimpanan atau interpretasi data oleh model.

## 4.4 Penggunaan Chatbot

Berikut adalah deskripsi singkat mengenai cara penggunaan aplikasi chatbot yang dikembangkan:

1. **User Interface:** Aplikasi chatbot ini memiliki antarmuka berbasis web yang dirancang sederhana dan mudah diakses, memungkinkan pengguna langsung mengakses kolom chat utama setelah membuka aplikasi. Pada halaman utama, terdapat kolom teks untuk memasukkan pertanyaan serta tombol *send* untuk memulai percakapan dengan chatbot. Selain itu, aplikasi ini menyediakan beberapa menu bantuan dan contoh pertanyaan yang sering diajukan, sehingga pengguna dapat mengetahui pertanyaan-pertanyaan umum yang dapat dijawab oleh chatbot.
2. **Cara Bertanya:** Pengguna dapat mengetik pertanyaan secara langsung di kolom chat dengan berbagai variasi kata dan frasa, baik dalam bahasa formal maupun informal. Aplikasi ini juga dilengkapi dengan fitur koreksi kesalahan ketik (typo correction) yang memungkinkan chatbot untuk mengidentifikasi dan memahami pertanyaan meskipun terdapat kesalahan pengetikan.
3. **Proses Interaksi:** Setelah pengguna mengirimkan pertanyaan, chatbot akan memproses input tersebut melalui model yang telah dilatih dengan informasi terkait Universitas Bina Nusantara. Chatbot memberikan jawaban berdasarkan informasi relevan yang terdapat dalam basis data, dengan tujuan memberikan jawaban yang akurat dan lengkap.
4. **Contoh Penggunaan:** Sebagai contoh, pengguna yang bertanya "Akreditasi Jurusan Film apa?" akan mendapatkan jawaban “Akreditasi Jurusan Film di BINUS adalah "Unggul"

## 4.4 Tampilan Chatbot

**/TODO**

# BAB 5 KESIMPULAN DAN SARAN

## 5.1 Kesimpulan

Pembuatan chatbot dengan data knowledge yang spesifik tentang BINUS University memberikan kemudahan dalam mencari informasi terkait Universitas Bina Nusantara. Chatbot ini berpotensi menjadi alat yang sangat bermanfaat, terutama bagi calon mahasiswa, mahasiswa aktif, maupun masyarakat umum yang membutuhkan informasi seputar universitas secara cepat dan efisien.

Dari hasil pengujian, dapat disimpulkan bahwa model Gemini AI memiliki performa yang lebih unggul dibandingkan dengan IndoBERT-SQuAD dalam menjawab test case yang diberikan. Hal ini terlihat dari tingkat akurasi yang tinggi pada Gemini AI, yaitu sebesar 90%, dibandingkan dengan IndoBERT-SQuAD yang tidak mampu memberikan jawaban yang dianggap benar oleh mayoritas evaluator. Keunggulan Gemini AI mencakup kemampuan memberikan jawaban yang lebih relevan, informatif, dan sesuai dengan konteks pertanyaan.

Namun demikian, meskipun menunjukkan performa yang lebih baik, Gemini AI masih dapat memberikan jawaban yang salah. Terdapat beberapa test case di mana output yang dihasilkan salah karena ketidaksesuaian dengan data knowledge yang tersedia. Hal ini menunjukkan bahwa meskipun Gemini AI lebih unggul, masih ada ruang untuk perbaikan, terutama dalam memastikan keakuratan data dan relevansi jawaban yang diberikan.

## 5.2 Saran

Berdasarkan kesimpulan di atas, terdapat beberapa rekomendasi untuk pengembangan lebih lanjut:

1. Peningkatan basis data knowledge adalah salah satu prioritas utama. Pembaruan dan validasi rutin terhadap data knowledge perlu dilakukan untuk memastikan informasi yang dimiliki chatbot selalu lengkap, relevan, dan akurat. Hal ini penting agar chatbot dapat memberikan jawaban yang lebih informatif dan minim kesalahan.
2. Pengembangan model perlu difokuskan untuk meningkatkan kemampuan chatbot dalam menangani pertanyaan yang spesifik terhadap domain BINUS University. Dengan melakukan fine-tuning pada dataset yang relevan, performa model dapat dioptimalkan sehingga lebih cocok dengan kebutuhan pengguna dan konteks universitas.
3. Uji coba yang lebih luas perlu dilakukan dengan menggunakan test case yang lebih variatif, baik dari segi panjang, kompleksitas, maupun cakupan pertanyaan. Uji coba ini dapat dilakukan setelah penambahan data knowledge untuk mengevaluasi sejauh mana pembaruan dan pengembangan yang dilakukan dapat meningkatkan performa model secara keseluruhan.

# DAFTAR PUSTAKA

Atmauswan, P. S., & Abdullahi, A. M. (2 ent Chatbot For University Information System Using Natural Language Approach. Albukhary Social Business Journal, 3(2), 59–64.

Nguyen, T. T., Le, A. D., Hoang, H. T., & Nguyen, T. (2021). NEU-chatbot: Chatbot for admission of National Economics University. Computers and Education: Artificial Intelligence, 2, 100036.

Copeland, B. (2024, January 9). Artificial Intelligence (AI) | Definition, Examples, Types, Applications, Companies, & Facts. Encyclopedia Britannica. https://www.britannica.com/technology/artificial-intelligence

Devlin, J. (2018, October 11). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.org. https://arxiv.org/abs/1810.04805

Kuhlman, D. (2012). A Python Book: Beginning Python, Advanced Python, and Python Exercises. Platypus Global Media.

Tank, U. (2022, September 27). Top 7 In-Demand Programming Languages. Retrieved from Stackify: https://stackify.com/top-7-in-demand-programming-languages/

Hiremath, G., Hajare, A., Bhosale, P., Nanaware, R., & Wagh, K. S. (2018). Chatbot for education system. *International Journal of Advance Research, Ideas and Innovations in Technology*, *4*(3), 37-43.

Bhat, V. (2023). Retrieval Augmented Generation (RAG) based Restaurant Chatbot with AI Testability. *ResearchGate*, 41-45.

Windiatmoko, Y., Rahmadi, R., & Hidayatullah, A. F. (2021, February). Developing facebook chatbot based on deep learning using rasa framework for university enquiries. In *IOP conference series: materials science and engineering* (Vol. 1077, No. 1, p. 012060). IOP Publishing.

Ahmad, N. A., Hamid, M. H. C., Zainal, A., & Baharum, Z. (2020). UNISEL Bot: Designing Simple Chatbot System for University FAQs. *International Journal of Innovative Technology and Exploring Engineering*, *9*(2), 4689-4693.

Le Hoanh Su, T. D. H., Thi-Yen-Linh, T., Thi-Duyen-Ngoc, N., Bao-Tuyen, L., & Ha-Phuong-Truc, N. (2020). Development of an AI Chatbot to support admissions and career guidance for universities. *International Journal of Emerging Multidisciplinary Research*, *4*(2), 11-17.

Muhyidin, A., & Setiawan, M. A. F. (2021). Developing UNYSA chatbot as information services about yogyakarta state university. In *Journal of Physics: Conference Series* (Vol. 1737, No. 1, p. 012038). IOP Publishing.

Shchegoleva, L., & Burdin, G. (2021). Chatbot for applicants on university admission issues. In *Conference of Open Innovations Association, FRUCT* (No. 29, pp. 491-494). FRUCT Oy.

# LAMPIRAN