Strategi RAG untuk Automasi Analisis Konten dalam Bahasa Indonesia dengan Model LLM

UJIAN KUALIFIKASI

Annisa Lyanzahra Utomo

99223117

PROGRAM DOKTOR TEKNOLOGI INFORMASI
UNIVERSITAS GUNADARMA
Agustus 2024

DAFTAR ISI

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi kecerdasan artifisal (AI) telah membawa perubahan signifikan dalam berbagai bidang, termasuk bidang psikologi. Dalam beberapa dekade terakhir, bidang psikologi telah berkembang pesat dengan munculnya pendekatan dan metode baru untuk memahami perilaku dan kondisi mental manusia (Kazdin, 2017). Konseling merupakan sebuah hubungan kolaboratif antara konselor profesional dengan individu, keluarga, atau kelompok. Tujuan utama konseling adalah memberdayakan klien untuk mencapai kesehatan mental, kesejahteraan, serta keberhasilan dalam pendidikan dan karir. Bagi mereka yang mengalami kesulitan psikologis atau interpersonal, konseling kesehatan mental menjadi intervensi utama yang membantu. Sesi konseling menerapkan pendekatan berpusat pada klien, menciptakan lingkungan yang aman dan suportif untuk membangun kepercayaan dan eksplorasi diri. Dalam sesi ini, klien didorong untuk menyelami pengalaman pribadinya, berbagi cerita intim, dan dibantu oleh terapis dalam menavigasi dialog untuk mencapai kesembuhan. Diskusi dalam sesi konseling mencakup berbagai topik, mulai dari peristiwa kehidupan terkini hingga introspeksi mendalam, yang semuanya berkontribusi pada perjalanan terapeutik (Kumar Adhikary et al., 2024).

Dokumentasi catatan konseling, yang merupakan ringkasan dari keseluruhan sesi, menjadi aspek penting dalam proses konseling. Catatan ini berfungsi merangkum pemicu stres klien dan prinsip-prinsip terapi yang diterapkan. Dokumentasi ini bermanfaat untuk berbagai keperluan, seperti membantu klien mengingat poin-poin penting dan kemajuan yang dicapai dalam sesi konseling, memfasilitasi komunikasi dan kolaborasi antar terapis, terutama dalam kasus transfer klien, memberikan bukti tertulis tentang proses konseling, yang dapat membantu melindungi konselor dan klien dalam situasi hukum, mempermudah pelacakan kemajuan klien dan membantu terapis dalam mengevaluasi efektivitas intervensi (Seligman, 2004).

Evaluasi sesi konseling melibatkan manusia sebagai penilai yang meringkas sesi dengan memberikan kode (pelabelan atau anotasi) untuk mengukur informasi yang diperoleh selama pertemuan konseling. Proses pemberian kode ini, disebut *observational coding*, menyediakan sistem organisasi berbasis teori yang memungkinkan data linguistik kompleks terstruktur untuk analisis lebih lanjut. Kode tersebut dapat mewakili topik pembicaraan (misalnya, obat-obatan), gejala yang diekspresikan (misalnya, depresi, kecemasan, kemarahan), dan perilaku verbal spesifik dalam pernyataan pasien (misalnya, memberi sinyal niat untuk mengubah atau mempertahankan perilaku) (Gaut et al., 2017). Namun, salah satu tantangan utama yang masih dihadapi oleh para konselor dan peneliti psikologi adalah menganalisis dan mengekstraksi informasi penting dari percakapan dengan klien. Percakapan ini seringkali panjang, kompleks, dan mengandung banyak informasi tersembunyi yang sulit untuk diidentifikasi secara manual dan proses analisis secara manual seringkali memakan waktu dan rentan terhadap bias subjektif. (Imel, Steyvers and Atkins, 2015).

Teknologi AI telah muncul sebagai alat yang berpotensi untuk membantu menganalisis teks dengan lebih baik. Salah satu pendekatan yang menjanjikan adalah penggunaan *Large Language Model* (LLM) yang dilatih pada korpus data yang besar untuk melakukan tugas-tugas seperti *natural language understanding*, *text generation*, dan *text summarization* (Brown et al., 2020). Penelitian sebelumnya oleh (Mullenbach et al., 2019) telah mengeksplorasi penggunaan LLM untuk menganalisis transkrip wawancara pasien, namun terbatas pada identifikasi gejala dan diagnosis. LLM telah menunjukkan kemampuan yang luar biasa dalam memahami dan mengolah teks, serta menghasilkan output yang akurat dan bermakna. Namun, sebagian besar penelitian sebelumnya berfokus pada domain yang lebih umum, seperti *question-aswering* (QA), *translation,* dan *text summarization* (Radford et al., 2019), (Rae et al., 2021).

Fine-tuning adalah teknik yang umum digunakan dalam memodifikasi model LLM yang telah dilatih sebelumnya agar dapat melakukan tugas-tugas spesifik, seperti klasifikasi teks atau analisis percakapan. *Fine-tuning* adalah teknik di mana sebuah model yang sebelumnya sudah dilatih pada dataset besar (*pre-training*)

kemudian disesuaikan lebih lanjut dengan dataset yang lebih kecil dan spesifik untuk tugas atau domain tertentu. Metode ini memungkinkan model untuk mempelajari rincian dari konteks baru sambil tetap mempertahankan pengetahuan yang diperoleh selama pelatihan awal. Meskipun *fine-tuning* dapat meningkatkan kemampuan model dalam skenario yang sangat spesifik, metode ini memiliki beberapa tantangan di mana proses ini membutuhkan sumber daya komputasi yang besar dan, jika dataset yang digunakan terbatas, dapat mengurangi kinerja model secara keseluruhan dan meningkatkan risiko *overfitting* (Ziegler et al., 2022).

Untuk mengatasi keterbatasan *fine-tuning*, *Retrieval-Augmented Generation* (RAG) telah diperkenalkan sebagai alternatif yang lebih adaptif. Teknik RAG menggabungkan kemampuan *text generation* dengan *information retrieval* dari sumber data eksternal. Dengan mengakses informasi dari korpus eksternal secara *real-time,* RAG dapat memberikan analisis yang relevan dan kontekstual, bahkan ketika menghadapi *input* yang tidak ada dalam data pelatihan. (Lewis et al., 2020). Hal ini membuat RAG sangat cocok untuk aplikasi yang membeutuhkan pemahaman mendalam dan respons yang informatif, dimaana variasi dan kompleksitas data sering tinggi. Oleh karena itu, penelitian ini berfokus pada eksplorasi dan pengembangan model analisis konten berbasis RAG untuk meningkatkan akurasi dan relevansi dalam pengkodean otomatis percakapan teks. Melalui penelitian ini, diharapkan model ini dapat memberikan analisis yang kontekstual dan adaptif terhadap variasi data, sehinnga menghasilkan interpretasi yang mendalam dan informatif.

1.2 Rumusan Masalah Penelitian

1. Bagaimana cara mengembangkan model LLM *Retrieval-Augmented Generation* (RAG) yang efektif untuk coding otomatis pada *client-counselor dialogue* (CCD) dalam bahasa Indonesia?
2. Bagaimana cara mengintegrasikan komponen *retrieval* dan *generation* dalam model RAG untuk menghasilkan *output* yang relevan dan kontekstual dari percakapan *client-counselor*?

3. Sejauh mana model RAG dapat memberikan *insight* yang berguna dalam konteks Konseling?
4. Bagaimana cara mengimplementasikan model RAG yang telah dikembangkan ke dalam sistem praktis untuk mendukung proses analisis konten dan penilaian dalam konteks konseling?

1.3 Batasan Masalah Penelitian

1. Penelitian ini hanya akan menggunakan data percakapan dalam bahasa Indonesia dari dialog konseling.
2. Analisis akan terfokus pada percakapan antara klien dan konselor tanpa mempertimbangkan data tambahan seperti catatan klinis atau hasil tes psikologis.
3. Penelitian ini fokus pada automasi berbasis teks dari percakapan konseling, mengabaikan modalitas lain seperti intonasi suara, ekspresi wajah, dan bahasa tubuh.

1.4 Tujuan Penelitian

1. Mengembangkan model LLM RAG yang dapat melakukan *coding* otomatis pada *client-counselor dialogue* (CCD) dalam bahasa Indonesia.
2. Mengintegrasikan komponen *retrieval* dan *generation* dalam model RAG untuk menghasilkan *output* yang relevan dan kontekstual dari *client-counselor dialogue* (CCD).
3. Mengevaluasi kegunaan dan efektivitas RAG dalam menghasilkan *insight* yang berguna dalam konteks Konseling.
4. Mengimplementasikan model yang dikembangkan dalam sistem praktis untuk mendukung proses analisis dan penilaian dalam konteks konseling.

1.5 Manfaat Penelitian

1. Dukungan dalam Pengambilan Keputusan: Model LLM yang dikembangkan dapat berfungsi sebagai alat bantu bagi konselor dalam proses pengambilan keputusan. Dengan memberikan analisis yang lebih cepat dan akurat dari dialog antara klien dan konselor, model ini dapat membantu konselor dalam

merancang intervensi yang lebih tepat dan informatif, serta sebagai pendukung dalam menentukan langkah-langkah berikutnya dengan lebih efektif.

2. Adaptasi untuk Bahasa dan Domain Spesifik: Penelitian ini berkontribusi pada adaptasi model LLM untuk bahasa Indonesia dan konteks konseling, yang dapat digunakan sebagai referensi untuk aplikasi teknologi serupa dalam bahasa dan domain lain.
3. Penerapan dan Evaluasi Model dalam Sistem Praktis: Penelitian ini menawarkan panduan tentang bagaimana model RAG dapat diimplementasikan dalam sistem konseling praktis, serta tantangan dan solusi dalam integrasi teknologi ke dalam praktik profesional.

BAB 2
TELAAH PUSTAKA

2.1 Natural Language Processing (NLP)

Natural Language Processing (NLP) adalah cabang dari ilmu komputer dan kecerdasan artifisial yang berfokus pada bagaimana komputer dapat memahami dan bekerja dengan bahasa manusia, baik lisan maupun tulisan. NLP mempelajari model matematika dan komputasi untuk mengembangkan berbagai sistem. NLP penting dalam ilmu komputer karena memungkinkan komputer untuk menangani kerumitan bahasa manusia. NLP mengeksplorasi bagaimana komputer dapat digunakan untuk memahami bahasa manusia untuk berbagai keperluan, seperti *text summarization, machine translation, speech recognition, spam detection, virtual assistant, chatbot, document classification* dan *sentiment analysis*. (Reshamwala, Mishra and Pawar, 2013).

Data preprocessing dalam NLP melibatkan serangkaian langkah untuk membersihkan, mengubah format, dan memperkaya data teks. Hal ini penting untuk memastikan bahwa data teks terstruktur dengan baik, bebas dari noise, dan konsisten dalam representasinya. Berikut beberapa teknik umum yang digunakan dalam preprocessing data pada NLP:

- *Tokenization*, proses memecah teks menjadi bagian-bagian yang lebih kecil, bisa berupa kata individual atau potongan kata. Hasilnya biasanya berupa indeks kata dan teks yang sudah ditokenisasi, dimana kata-kata tersebut mungkin akan direpresentasikan sebagai token numerik untuk digunakan dalam berbagai metode deep learning.
- *Case Folding*, proses mengubah semua huruf menjadi huruf kecil atau besar untuk menyamakan representasi kata.
- *Stopwords removal*, proses yang membuang kata yang dianggap tidak memiliki arti penting dari hasil *tokenizing*.
- *Stemming dan Lemmatization*, stemming adalah proses mengubah kata ke bentuk akarnya menggunakan aturan heuristik. Lemmatization adalah proses

untuk menemukan kata dasar dengan menganalisis morfologi kata menggunakan kamus.

- *Named-entity recognition (NER)*, proses mengidentifikasi dan mengklasifikasikan entitas yang memiliki nama, seperti orang, organisasi, atau lokasi.

Teknik ekstraksi fitur dalam NLP bertujuan untuk mengubah data teks mentah menjadi representasi numerik yang dapat dimengerti oleh algoritma machine learning. Representasi ini, yang disebut fitur, menangkap informasi penting dari teks, seperti makna kata, struktur kalimat, dan hubungan antar kata. Beberapa teknik umum yang digunakan dalam ekstraksi fitur NLP meliputi:

- *Bag-of-Words* (BoW), merupakan teknik yang memecah teks menjadi kata-kata individual, kemudian mewakili teks tersebut sebagai distribusi frekuensi kata-kata tersebut. Teknik ini berguna untuk mengekstrak wawasan bermakna dari data teks yang besar, seperti mengidentifikasi kata yang paling sering muncul, menganalisis sentimen, atau bahkan memprediksi tren kedepan. BoW dapat digunakan untuk berbagai macam aplikasi, mulai dari klasifikasi konten dan deteksi spam hingga analisis sentimen dan pengembangan chatbot.
- *Term Frequency-Inverse Document Frequency* (TF-IDF). *Term Frequency* merupakan jumlah frekuensi kemunculan sebuah kata dalam suatu dokumen, sedangkan *Inverse Document Frequency* merupakan perhitungaan di mana suatu kata tersebut tersebar dalam suatu dokumen. TF-IDF bertujuan untuk mengevaluasi seberapa relavan kata pada suatu dokumen dalam kumpulan dokumen. Proses ini digunakan untuk menghitung nilai bobot setiap kata pada dokumen, semakin besar nilai bobotnya maka peran kata tersebut di dalam dokumen menjadi penting. Teknik TF-IDF umum diguanakn untuk tugas *text classification*.
- *Word Embeddings*, merupakan teknik yang melatih model untuk mempresentasikan kata-kata sebagai vektor numerik yang menangkap makna dan hubungan semantic antar kata. Representasi tersebut dapat digunakan untuk berbagai tugas NLP selanjutnya, seperti analisis sentimen dan klasifikasi teks.

- *Part-of-Speech* (POS) tagging, teknik yang berfokus pada pemberian label kategori gramatikal kata-kata dalam kalimat, seperti kata benda, kata kerja, kata sifat, dll. POS tagging umumnya digunakan untuk berbagai tugas NLP seperti *text classification*, *sentiment analysis*, *machine translation*, dan *speech recognition*.
- *Latent Dirichlet Allocation* (LDA), model probabilistik generatif yang berasumsi setiap korpus dihasilkan dari campuran topik-topik tersembunyi. Masing-masing topik ini diwakili oleh probabilitas distribusi kata-kata. LDA bekerja dengan cara berulang kali menetapkan kata-kata dalam setiap dokumen ke topik dan menyesuaikan probabilitas kata-topik berdasarkan distribusi topik yang dihasilkan di seluruh dokumen. Hasil akhir dari LDA adalah serangkaian topik, yang masing-masing diwakili distribusi kata. LDA umum digunakan dalam aplikasi seperti *topic modeling*, *document clustering*, dan *information retrieval*.

Berbagai teknik telah dikembangkan untuk meningkatkan kemampuan model NLP dalam memahami dan menghasilkan bahasa alami. Berikut adalah beberapa teknik utama yang digunakan dalam model NLP:

- *Logistic regression*, teknik statistik yang digunakan untuk pemodelan prediktif, terutama dalam konteks klasifikasi biner. Dalam NLP, regresi logistik dapat digunakan untuk mengklasifikasikan teks ke dalam dua kategori, seperti sentimen positif atau negatif. Model ini memprediksi probabilitas suatu kejadian dengan menggunakan fungsi logit untuk memetakan input ke dalam ruang keluaran yang terbatas antara 0 dan 1.
- *Naïve Bayes*, metode klasifikasi berbasis probabilistik yang didasarkan pada Teorema Bayes. Teknik ini dianggap 'naïve' karena mengasumsikan bahwa semua fitur dalam dataset adalah independen. Dalam NLP, Naïve Bayes dapat digunakan untuk tugas klasifikasi teks seperti deteksi spam dan analisis sentimen.
- *Decision Tree*, teknik pemodelan prediktif yang menggunakan struktur pohon untuk membuat keputusan berdasarkan fitur input. Setiap simpul dalam pohon

mewakili fitur, dan setiap cabang mewakili keputusan berdasarkan fitur tersebut. Dalam NLP, Decision Tree dapat digunakan untuk tugas-tugas seperti klasifikasi teks dan analisis sentimen.

- *Recurrent Neural Network* (RNN), RNN adalah jenis *neural network* yang dirancang untuk mengolah data sekuensial seperti teks. RNN memiliki memori internal sehingga dapat mempertimbangkan informasi sebelumnya dalam urutan saat memproses data. Dalam NLP, RNN sangat efektif untuk tugas-tugas seperti pemodelan bahasa, terjemahan mesin, dan analisis sentimen, karena kemampuannya untuk menangani data sekuensial seperti teks.
- *Autoencoders*, jenis *neural network* yang digunakan untuk pembelajaran representasi dengan cara mengkodekan *input* ke dalam representasi yang lebih kecil dan kemudian mendekodekannya kembali ke bentuk asli. Dalam NLP, *autoencoders* dapat digunakan untuk tugas-tugas seperti pengurangan dimensi dan deteksi anomali.
- *Sequence-to-sequence* (SEQ2SEQ), merupakan arsitektur *neural network* yang digunakan untuk mengubah satu urutan ke urutan lain. Model ini terdiri dari dua komponen utama: encoder dan decoder. Encoder mengubah input urutan menjadi representasi vektor, sementara decoder menghasilkan urutan output dari representasi tersebut. Dalam NLP, SEQ2SEQ dapat digunakan untuk tugas-tugas seperti terjemahan bahasa dan pembuatan teks.
- *Transformers*, merupakan arsitektur model yang telah merevolusi bidang NLP dengan kemampuannya untuk menangani dependensi jarak jauh dalam teks. Model ini menggunakan mekanisme perhatian (*attention mechanism*) yang memungkinkan pemrosesan paralel dan efisien. *Transformers* telah digunakan dalam berbagai aplikasi NLP, termasuk pemodelan bahasa, terjemahan mesin, dan pembuatan teks.

2.2 *Large Language Model* (LLM)

Large Language Model (LLM) adalah algoritma kecerdasan artifisial (AI) yang memanfaatkan *deep learning* dan kumpulan data yang sangat besar untuk memahami, meringkas, menghasilkan, dan memprediksi konten baru. LLM dilatih dengan menggunakan sejumlah besar teks melalui *self-supervised learning* dan

mampu menyelesaikan berbagai tugas dengan sangat baik. LLM dibangun menggunakan *neural network* yang terdiri dari berbagai parameter, yang biasanya mencakup miliaran bobot dan lebih banyak lagi. Model-model ini dilatih sebelumnya dengan jumlah data yang sangat besar untuk membantu mereka memahami kompleksitan dan hubungan dalam bahasa. Pada dasarnya, LLM adalah *neural network transformer* yang memprediksi teks yang kemungkinan akan muncul berikutnya (Marvin et al., 2024).

LLM terdiri dari *multiple layer neural network*, masing-masing dengan parameter yang dapat dilakukan *fine-tuning* dengan baik selama pelatihan, yang disempurnakan lebih lanjut oleh banyak lapisan yang dikenal dengan *attention mechanism*, yang memanggil bagian tertentu dari kumpulan data. Selama proses pelatihan, model-model ini belajar untuk memprediksi kata-kata berikutnya dalam sebuah kalimat berdasarkan konteks yang diberikan oleh kata-kata sebelumnya. Model ini melakukannya dengan mengaitkan skor probabilitas pada kemunculan kembali kata-kata yang telah dilakukan tokenisasi. Token-token ini kemudian diubah menjadi *embeddings*, yang merupakan representasi numerik dari konteks tersebut.

Untuk memastikan keakuratannya, proses ini melibatkan pelatihan LLM pada korpus teks yang sangat besar, hal ini memungkinkan model mempelajari tata bahasa, semantikm dan hubungan konseptual. Setelah dilatih, LLM dapat menghasilkan teks dengan memprediksi kata berikutnya secara mandiri berdasarkan input yang diterimanya, serta memanfaatkan pola dan pengetahuan yang telah diperolehnya. Hasilnya adalah bahasa yang koheren dan relevan secara kontekstual yang dapat dimanfaatkan untuk berbagai tugas NLU (*Natural Language Understanding*) dan NLG (*Natural Language Generation*). Performa model juga dapat ditingkatkan dengan teknik *prompt engineering*, *prompt-tuning*, *fine-tuning*, dan *Reinforcement Learning with Human Feedback* (RLHF) untuk menghilangkan bias, ujaran kebencian, dan jawaban yang secara factual tidak tepat atau yang dikenal sebagai "*hallucination*".

Fine-tuning adalah teknik di mana sebuah model yang sebelumnya sudah dilatih pada dataset besar (*pre-training*) kemudian disesuaikan lebih lanjut dengan

dataset yang lebih kecil dan spesifik untuk tugas atau domain tertentu. Metode ini memungkinkan model untuk mempelajari rincian dari konteks baru sambil tetap mempertahankan pengetahuan yang diperoleh selama pelatihan awal. Meskipun *fine-tuning* dapat meningkatkan kemampuan model dalam skenario yang sangat spesifik, metode ini memiliki beberapa tantangan di mana proses ini membutuhkan sumber daya komputasi yang besar dan, jika dataset yang digunakan terbatas, dapat mengurangi kinerja model secara keseluruhan dan meningkatkan risiko *overfitting*.

Untuk mengatasi keterbatasan *fine-tuning*, teknik *Retrieval-Augmented Generation* (RAG) dapat digunakan. RAG menggabungkan kemampuan *retrieval* informasi dari sumber eksternal dengan kemampuan model generatif, sehingga dapat menghasilkan *output* yang lebih relevan dan informatif dalam situasi di mana *fine-tuning* mungkin kurang optimal. Model dapat secara dinamis mencari dan memanfaatkan informasi tambahan yang relevan, memungkinkan analisis yang lebih mendalam dan respons yang lebih akurat.

2.3 Aplikasi dari LLM

LLM memiliki berbagai aplikasi yang luas dan bermanfaat dalam berbagai bidang, diantaranya sebagai berikut:

- *Chatbot* dan *Virtual Assistant*, LLM digunakan untuk mengembangkan chatbot dan asisten virtual yang mampu berinteraksi dengan pengguna secara alami dan responsif.
- *Language Translation*, LLM digunakan dalam sistem penerjemah bahasa otomatis yang akurat dan cepat, memfalisitasi komunikasi lintas bahasa.
- *Text Generation*, LLM dengan kemampuannya dalam memahami dan menghasilkan teks alami sehingga dapat digunakan dalam sistem *text generation* seperti untuk menulis artikel, membuat deskripsi produk, menciptakan karya fiksi, dst.
- *Sentiment Analysis*, LLM dimanfaatkan dalam analisis sentiment, yaitu mengidentifikasi emosi, sikap, dan opini dalam teks.
- *Content Summarization*, LLM dengan kemampuannya dalam memahami dan meringkas informasi penting dapat digunakan untuk membuat ringkasan dokumen, artikel berita, dan laporan yang efisien dan padat.

2.3.1 Model LLM

a) BERT (Bidirectional Encoder Representation from Transformers)

BERT, yang diperkenalkan oleh Google pada tahun 2018, adalah model pra-pelatihan yang menggunakan arsitektur *Transformer* untuk memahami konteks dari kedua arah, baik kiri maupun kanan dari sebuah token dalam urutan teks. Pendekatan bidirectional ini memungkinkan BERT untuk menangkap hubungan yang lebih kompleks antara kata-kata dalam sebuah kalimat, sehingga meningkatkan akurasi dalam berbagai tugas NLP seperti klasifikasi teks, penerjemahan bahasa, dan ekstraksi informasi.

b) RoBERTa (*Robustly Optimized* BERT *Pretraining Approach*)

RoBERTa dirilis oleh Facebook AI pada tahun 2019, adalah pengembangan lebih lanjut dari BERT dengan optimasi dalam proses pra-pelatihan. RoBERTa melibatkan peningkatan jumlah data pelatihan dan ukuran batch, penggunaan pendekatan dinamis dalam masking selama pra-pelatihan, serta penghapusan tugas NSP (*Next Sentence Prediction*). Dengan optimasi ini, RoBERTa berhasil meningkatkan performa di berbagai benchmark NLP dibandingkan dengan BERT asli.

c) GPT

GPT adalah model generatif yang dirancang oleh OpenAI dan pertama kali diperkenalkan pada tahun 2018. Model ini menggunakan arsitektur Transformer dan berfokus pada prediksi teks secara autoregressive. GPT dilatih untuk memprediksi kata berikutnya dalam sebuah urutan teks berdasarkan konteks sebelumnya, membuatnya sangat efektif dalam tugas-tugas seperti pembuatan teks, dialog interaktif, dan penerjemahan bahasa. Versi-versi selanjutnya adalah GPT-2 (2019) dan GPT-3 (2020).

d) DistilBERT (*Distilled* BERT)

DistilBERT diperkenalkan oleh Hugging Face pada tahun 2019, adalah versi yang lebih ringan dan efisien dari BERT yang dikembangkan melalui proses distilasi pengetahuan. Distilasi ini melibatkan penyederhanaan model BERT dengan tetap mempertahankan sebagian besar kemampuan aslinya. DistilBERT dirancang untuk mengurangi waktu inferensi dan penggunaan

memori, menjadikannya pilihan ideal untuk aplikasi dengan keterbatasan sumber daya komputasi.

e) Longformer (*Long-range dependence transformer*)

Longformer diperkenalkan oleh Allen Institute for AI pada tahun 2020, adalah model Transformer yang dirancang untuk menangani dependensi jarak jauh dalam teks. Longformer memperkenalkan mekanisme perhatian linier yang memungkinkan pemrosesan sekuens teks yang lebih panjang dibandingkan dengan Transformer tradisional. Hal ini menjadikannya sangat efektif dalam tugas-tugas yang memerlukan analisis teks panjang, seperti analisis dokumen dan ekstraksi informasi dari artikel panjang.

f) ELECTRA (*Efficient Lifelong End-to-End Text Recognition with Attention*)

ELECTRA dirilis oleh Google Research pada tahun 2020, adalah model pra-pelatihan yang memperkenalkan pendekatan baru dalam pembelajaran representasi teks. ELECTRA melibatkan pembelajaran untuk membedakan antara token asli dan token yang telah digantikan oleh generator. Pendekatan ini memungkinkan ELECTRA untuk belajar dengan efisien dan menghasilkan representasi teks yang akurat, serta menunjukkan kinerja yang baik dalam berbagai tugas NLP dengan biaya komputasi yang rendah.

g) T5 (*Text-to-Text Transfer Transformer*)

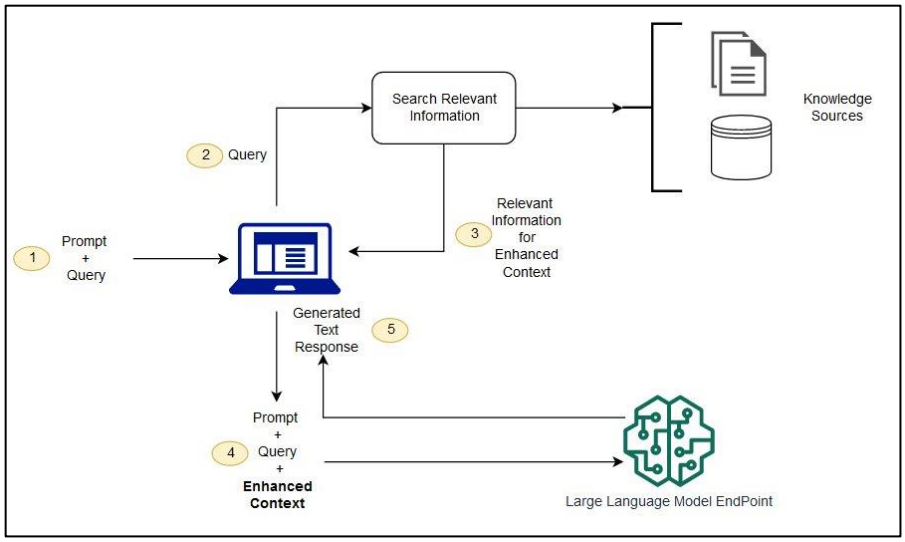
T5 dikembangkan oleh Google Ai pada tahun 2019. T5 dilatih pada kumpulan data teks dan kode yang sangat besar menggunakan kerangka text-to-text. Model T5 mampu melakukan tugas-tuags berbasis teks yang telah dilakukan *pre-training*. T5 juga dapat disesuaikan (*fne-tuned*) untuk melakukan tugas-tugas lain. T5 telah diterapkan dalam berbagai aplikasi, termasuk *chatbot, machine translation,* t*ext summarization, code generation,* dan robotika.

2.4 *Retrieval-Augmented Generation* (RAG)

Retrieval Augmented Generation (RAG) merupakan sebuah paradigma baru dalam pengembangan *Large Language Models* (LLM) yang bertujuan untuk meningkatkan kemampuan generasi teks dengan mengintegrasikan retrieval dari sumber-sumber eksternal. Pendekatan ini diperkenalkan oleh (Lewis et al., 2020)

sebagai upaya untuk mengatasi keterbatasan inheran dalam LLMs konvensional, yaitu kecenderungan untuk menghasilkan output yang tidak akurat atau tidak koheren ketika membahas topik-topik yang membutuhkan pengetahuan faktual yang spesifik.

RAG mengombinasikan kekuatan LLM dalam membangkitkan teks yang alami dan masuk akal dengan kemampuan *retrieval* informasi dari korpus eksternal yang besar, seperti Wikipedia atau sumber-sumber lainnya. Dalam arsitektur RAG, LLM dilatih untuk menghasilkan teks dengan mempertimbangkan konteks masukan serta informasi yang diperoleh dari komponen retrieval. Komponen retrieval ini bertanggung jawab untuk mengidentifikasi dan mengambil potongan teks yang relevan dari corpus eksternal berdasarkan masukan yang diberikan. Potongan teks yang diambil kemudian disediakan kepada LLM sebagai konteks tambahan untuk membangkitkan keluaran akhir yang lebih akurat dan informatif.



Gambar 2. 1 Aliran konseptual menggunakan RAG dengan LLM.

Paradigma RAG berkembang dalam tiga tahap utama: *Naive* RAG, *Advanced* RAG, dan Modular RAG. *Naive* RAG merupakan tahap awal dalam evolusi RAG yang muncul bersamaan dengan adopsi luas dari arsitektur *Transformer*. Pada tahap ini, RAG mengandalkan proses yang sederhana dan linier, mencakup tiga langkah utama: *indexing*, *retrieval*, dan *generation*. Pada tahap *indexing*, data mentah diubah menjadi representasi vektor yang dapat disimpan dalam basis data vektor.

Kemudian, pada tahap retrieval, sistem mengambil potongan dokumen yang paling relevan berdasarkan kemiripan semantik dengan *query* pengguna. Akhirnya, tahap *generation* melibatkan model bahasa yang menghasilkan jawaban dengan memanfaatkan informasi yang diambil. Meskipun metodologi ini cukup efisien, *Naive* RAG memiliki beberapa keterbatasan, seperti tantangan dalam presisi dan *recall* saat retrieval, serta risiko "halusinasi" selama *generation*.

Advanced RAG memperkenalkan optimasi yang lebih canggih untuk mengatasi kelemahan dari Naive RAG. Perbaikan ini mencakup strategi pra-retrieval dan pasca-retrieval yang dirancang untuk meningkatkan kualitas dan relevansi informasi yang diambil. Pada tahap *pre-retrieval*, fokus utamanya adalah mengoptimalkan struktur indexing dan query asli pengguna melalui teknik seperti *query rewriting* dan *query expansion.* Sementara itu, tahap pasca-*retrieval* melibatkan proses seperti reranking dan kompresi konteks untuk memastikan bahwa informasi yang diambil relevan dan terfokus pada pertanyaan pengguna. Dengan strategi ini, *Advanced* RAG berhasil meningkatkan ketepatan dan mengurangi redundansi dalam informasi yang diambil, sehingga meningkatkan kualitas jawaban yang dihasilkan.

Modular RAG merupakan tahap paling maju dalam evolusi RAG, menawarkan fleksibilitas dan modularitas yang lebih besar dalam sistemnya. *Modular* RAG memperkenalkan berbagai modul spesifik yang dapat diintegrasikan atau diganti sesuai dengan kebutuhan tugas tertentu. Modul baru seperti *search module* untuk pencarian kemiripan, *memory module* untuk pengingat berkelanjutan, dan *predict module* untuk mengurangi redundansi, memberikan kemampuan adaptasi yang lebih tinggi pada sistem RAG. Pendekatan ini memungkinkan proses *retrieval* dan *generation* yang tidak lagi terbatas pada urutan linier, tetapi dapat dilakukan secara iteratif dan adaptif. *Modular* RAG juga memanfaatkan teknologi seperti *reinforcement learning* untuk terus meningkatkan kinerja dan relevansi jawaban.

2.4.1 Evaluasi RAG

Evaluasi RAG berfokus pada dua aspek utama: kualitas *retrieval* dan kualitas *generation*. Kualitas *retrieval* dinilai berdasarkan relevansi konteks yang diambil

dari basis data, sementara kualitas *generation* diukur melalui ketepatan dan relevansi jawaban yang dihasilkan. Evaluasi ini melibatkan penggunaan metrik seperti Hit Rate, MRR, dan NDCG untuk menilai *retrieval*, serta BLEU dan ROUGE untuk menilai *generation*. Evaluasi RAG juga mencakup kemampuan model untuk menangani informasi yang tidak relevan, mengintegrasikan informasi dari berbagai sumber, dan menghindari kesalahan dalam menyajikan informasi yang keliru.

2.4.2 Alat dan Implementasi RAG

Beberapa alat dan teknologi mendukung implementasi RAG, termasuk framework seperti LangChain, LLamaIndex, dan HayStack. Alat-alat ini menyediakan API dan integrasi yang memungkinkan pengguna untuk mengadopsi dan menyesuaikan RAG sesuai dengan kebutuhan mereka. Beberapa tools dan perpustakaan yang populer untuk mengimplementasikan pendekatan RAG, antara lain:

1. HuggingFace RAG: Perpustakaan ini menyediakan implementasi end-to-end dari RAG, dengan dukungan untuk beberapa arsitektur retriever dan generator yang berbeda.
2. FAISS: Perpustakaan ini digunakan untuk pencarian vektor efisien dan mendukung pendekatan retrieval berbasis dense.
3. Pyserini: Perpustakaan ini digunakan untuk pencarian teks dan retrieval berbasis sparse, serta mendukung indeks terbalik dan pencarian Boolean.
4. Haystack: Ini adalah kerangka kerja sumber terbuka untuk membangun pipeline pencarian dan pertanyaan-jawaban, termasuk komponen RAG.

2.5 Client-Counselor Dialogue dalam Konseling

Client-Counselor dialogue (CCD) atau dialog antara klien dan konselor merupakan elemen penting dalam proses psikoterapi. Komunikasi yang efektif antara kedua belah pihak memiliki pengaruh signifikan terhadap keberhasilan psikoterapi. Salah satu aspek krusial dalam dialog klien-konselor adalah kemampuan konselor untuk menciptakan lingkungan yang aman dan terbuka bagi klien. Tiga kondisi utama yang harus hadir dalam hubungan psikoterapi adalah

kehangatan (*warmth*), pemahaman empatik (*empathic understanding*), dan penerimaan tanpa syarat (*unconditional positive regard*). Ketika klien merasakan bahwa konselor benar-benar memahami dan menerima mereka tanpa menghakimi, mereka cenderung lebih terbuka dan berkomunikasi secara lebih mendalam. Konselor yang terampil dalam *active listening*, merefleksikan perasaan, dan memvalidasi pengalaman klien dapat memfasilitasi pembangunan kepercayaan dan mengembangkan aliansi psikoterapi yang kuat (Norcross & Wampold, 2011).

Pendekatan yang berpusat pada klien (*client-centered approach*) menekankan bahwa klien adalah ahli dalam kehidupannya sendiri, dan konselor berperan sebagai fasilitator yang membantu klien menemukan solusi mereka sendiri. Dialog yang kolaboratif dan saling menghormati, di mana kedua belah pihak berkontribusi secara aktif, dapat meningkatkan keterlibatan klien dan memperdalam pemahaman mereka terhadap masalah yang dihadapi (Bohart & Tallman, 2010).

Selain itu, beberapa studi telah meneliti pengaruh gaya komunikasi dan penggunaan bahasa dalam dialog klien-konselor. Penggunaan bahasa yang mudah dimengerti, jelas, dan bebas dari jargon teknis dapat meningkatkan pemahaman klien dan membangun hubungan yang lebih kuat. Konselor yang mampu menyesuaikan gaya komunikasi mereka dengan preferensi dan karakteristik unik setiap klien cenderung lebih efektif dalam memfasilitasi perubahan psikoterapi (Simmons et al., 2008).

2.6 Rangkuman Hasil Penelitian Terkait

Penelitian "*Automated Coding of Implicit Motives: A Machine-Learning Approach*" oleh Joyce S. Pang dan Hiram Ring (2020) mengeksplorasi penggunaan model *machine learning* untuk mengotomatisasi pengkodean motif implisit dalam teks. Penelitian ini menggunakan arsitektur CNN, CNN2D, dan TCN untuk memprediksi skor motif seperti *achievement* (pencapaian), *affiliation* (afiliasi), dan *power* (kekuasaan) pada beberapa dataset yang belum pernah dilihat sebelumnya. Hasil penelitian menunjukkan bahwa prediksi motif oleh model pembelajaran mesin memiliki tingkat korelasi moderat hingga tinggi dengan skor motif yang dikodekan oleh manusia, dengan nilai korelasi (r) berkisar antara 0,30 hingga 0,86. Hasil tersebut mengindikasikan potensi metode ini dalam mengotomatisasi proses

yang sebelumnya memerlukan banyak waktu dan tenaga ahli. Meski demikian, penelitian ini memiliki kesulitan dalam memahami bahasa figuratif, ironi, atau sarkasme, yang dapat mengurangi akurasi prediksi.

Penelitian "*Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*" oleh Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela (2020) memperkenalkan konsep Retrieval Augmented Generation (RAG) yang mengombinasikan memori parametrik (pre-trained seq2seq model) dan memori non-parametrik. Model RAG dievaluasi pada berbagai tugas NLP yang memerlukan pengetahuan intensif, termasuk beberapa tugas QA *open-domain* seperti Natural Questions dan TriviaQA. Penelitian ini menunjukkan bahwa RAG memberikan hasil yang lebih spesifik, beragam, dan faktual dibandingkan dengan model parametric-only. Model RAG menunjukkan peningkatan kinerja yang signifikan pada tugas-tugas QA open-domain, mengungguli model parametric-only dan arsitektur retrieve-and-extract untuk tugas yang spesifik. Model RAG menghasilkan bahasa yang lebih faktual dan spesifik dibandingkan model baseline parametric-only, dan Retrieval index dapat diperbarui tanpa perlu melatih ulang model.

Penelitian "*Topic Modeling Enchancement using Word Embeddings*" oleh Siriwat Limwattana dan Santitham Prom-on (2021) mengusulkan Deep Word-Topic Latent Dirichlet Allocation (DWT-LDA), sebuah model LDA yang dikembangkan dengan menggunakan teknik word embedding. Word embedding memungkinkan model untuk mempelajari. Penelitian ini menggunakan dataset dari forum diskusi online Pantip.com untuk bahasa Thailand dan dataset ulasan produk dari Amazon.com untuk bahasa Inggris. DWT-LDA mampu membuat kata kunci yang lebih spesifik sehingga menghasilkan anotasi topik yang lebih jelas dibandingan dengan LDA biasa.

Penelitian "*Towards Coding Social Science Datasets with Language Models*" oleh Christopher Michael Rytting, Taylor Sorensen, Lisa Argyle, Ethan Busby, Nancy Fulda, Josh Gubler, dan David Wingate (2023) mengeksplorasi penggunaan Large Language Models (LLM) dalam pengkodean data dalam ilmu *social science*.

Fokus utama dari penelitian ini adalah untuk menguji kemampuan LLM, seperti GPT-3, dalam melakukan analisis teks yang kompleks dan memberikan hasil yang relevan dengan konteks sosial tertentu. Hasil penelitian menunjukkan model ini mampu menangkap nuansa dan konteks dari teks yang dianalisis, meskipun ada keterbatasn seperti model kesulitan dalam menangani teks yang sangat ambigu, yang mungkin memerlukan penilaian manusia untuk interpretasi yang lebih akurat. Rata-rata akurasi yang dicapai oleh model dalam penelitian ini adalah sekitar 85%.

Penelitian "*Can Large Language Models Reasons about Medical Questions*?" oleh Valentin Lievin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, dan Ole Winther (2024) mengevaluasi kemampuan LLM, seperti GPT-3.5 dan Llama 2, dalam menjawab dan memberikan alasan mengenai pertanyaan medis yang kompleks. Penelitian ini menggunakan tiga dataset utama: MedQA-USMLE, MedMCQA, dan PubMedQA. Penelitian menemukan bahwa model GPT-3.5 mampu mencapai skor yang melewati batas kelulusan pada ketiga dataset medis: MedQA-USMLE (60.2%), MedMCQA (62.7%), dan PubMedQA (78.2%). Model open-source Llama 2 70B juga menunjukkan kinerja yang baik dengan akurasi 62.5% pada MedQA-USMLE. Beberapa keterbatasan dalam penelitian ini adalah model cenderung dapat menghasilkan fakta yang salah (*hallucination*), dan ketergantungan pada anotasi ahli untuk evaluasi.

Penelitian "*Towards Understanding Counseling Conversations: Domain Knowledge and Large Language Models*" oleh Younghun Lee, Dan Goldwasser, dan Laura Schwab Reese (2024) meneliti pemahaman percakapan konseling dengan model BERT-based dan pengetahuan domain khusus tentang konseling. Tujuan utama penelitian ini adalah untuk meningkatkan pemahaman terhadap percakapan antara konselor dan klien, serta memprediksi hasil percakapan, yaitu apakah klien merasa lebih positif setelahnya. Data yang digunakan berasal dari percakapan antara klien dan konselor dari Childhelp National Child Abuse Hotline. Integrasi pengetahuan domain khusus tentang konseling dengan fitur yang dihasilkan LLM menunjukkan peningkatan kinerja model 15% lebih baik. Hal ini menunjukkan bahwa pengetahuan domain spesifik yang relevan dapat membantu

model LLM dalam memahami konteks percakapan konseling dengan lebih baik dan membuat prediksi yang lebih akurat.

No	Nama	Judul	Tujuan Penelitian	Metode	Hasil
1	Joyce S. Pang dan Hiram Ring	Automated Coding of Implicit Motives: A Machine-Learning Approach	Mengotomatisasi pengkodean motif implisit dalam teks.	Menggunakan arsitektur CNN, CNN2D, dan TCN	Prediksi motif memiliki tingkat korelasi moderat hingga tinggi dengan skor motif yang dikodekan oleh manusia, dengan nilai korelasi (r) berkisar antara 0,30 hingga 0,86.
2	Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin,	Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks	Memperkenalkan konsep Retrieval Augmented Generation (RAG	Mengombinasikan memori parametrik (pre-trained seq2seq model) dan memori non-parametrik.	RAG menunjukkan peningkatan kinerja yang signifikan pada tugas-tugas QA open-domain, mengungguli

	Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela (2020)				model parametric-only dan arsitektur retrieve-and-extract untuk tugas yang spesifik.
3	Siriwat Limwattana dan Santitham Prom-on (2021)	Topic Modeling Enchancement using Word Embeddings	Mengusulkan Deep Word-Topic Latent Dirichlet Allocation (DWT-LDA)	Mengembangkan model LDA dengan menggunakan teknik word embedding	DWT-LDA mampu membuat kata kunci yang lebih spesifik sehingga menghasilkan anotasi topik yang lebih jelas dibandingan dengan LDA biasa.
4	Christopher Michael Rytting, Taylor	Towards Coding Social Science Datasets with	Mengeksplorasi penggunaan Large Language Models (LLM)		Rata-rata akurasi yang dicapai oleh model dalam

	Sorensen, Lisa Argyle, Ethan Busby, Nancy Fulda, Josh Gubler, dan David Wingate (2023)	Language Models	dalam pengkodean data dalam ilmu *social science*		penelitian ini adalah sekitar 85%.
5	Valentin Lievin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, dan Ole Winther (2024)	Can Large Language Models Reasons about Medical Questions?	Mengevaluasi kemampuan LLM, seperti GPT-3.5 dan Llama 2, dalam menjawab dan memberikan alasan mengenai pertanyaan medis yang kompleks		Model GPT-3.5 mampu mencapai skor yang melewati batas kelulusan pada ketiga dataset medis: MedQA-USMLE (60.2%), MedMCQA (62.7%), dan PubMedQA (78.2%). Model open-source Llama 2 70B juga menunjukkan

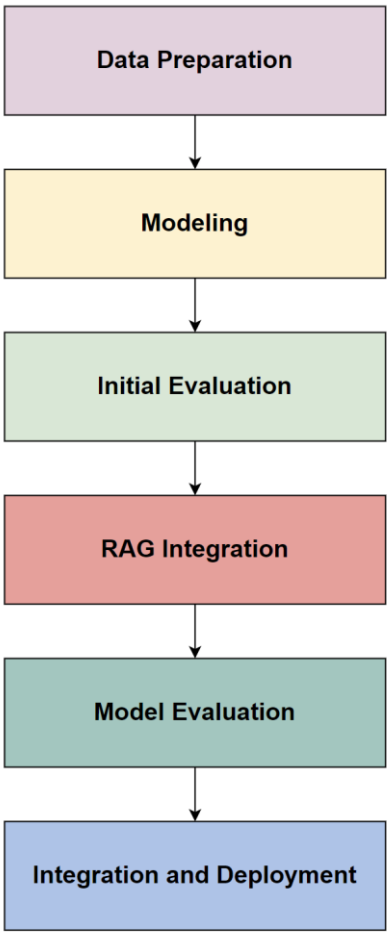
					kinerja yang baik dengan akurasi 62.5% pada MedQA-USMLE

BAB 3
METODE PENELITIAN

3.1 Gambaran Umum Penelitian

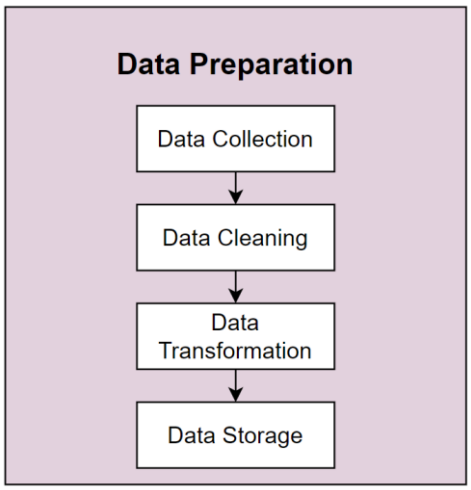
Penelitian ini berfokus pada pengembangan strategi *Retrieval-Augmented Generation* (RAG) untuk automasi analisis konten teks konseling dalam bahasa Indonesia. Tujuan utama dari penelitian ini adalah untuk membangun model LLM yang mampu melakukan *coding* otomatis terhadap dialog antara klien dan konselor, dengan memanfaatkan teknik RAG. Dalam prosesnya, penelitian ini akan menggunakan dua jenis data utama, yaitu Data *Client-Counselor Dialogue* (CCD) dan Data Korpus Referensi, yang akan diolah dan diintegrasikan melalui model RAG. Sebelum digunakan, data akan diperiksa untuk menghilangkan informasi pribadi yang dapat mengidentifikasi individu. Selain itu, data korpus referensi akan dikumpulkan dari sumber-sumber eksternal yang relevan untuk mendukung komponen *retrieval*, memungkinkan model untuk mengakses informasi tambahan yang berguna dalam proses analisis dan pengodean. Setiap langkah dalam pengembangan model ini akan melibatkan pemodelan RAG, serta evaluasi performa model menggunakan metrik retrieval dan generation.

Tahapan penelitian dimulai dengan pengumpulan dan persiapan data, di mana Data Dialog Klien-Konselor dan Data Korpus Referensi dikumpulkan, dibersihkan, dan ditransformasikan ke dalam bentuk yang dapat digunakan oleh sistem. Langkah selanjutnya melibatkan pemodelan RAG, yang mencakup pemodelan agent RAG, proses retrieval, dan generation. Setelah pemodelan selesai, evaluasi awal dilakukan untuk menilai kinerja sistem. Selanjutnya, sistem RAG diintegrasikan secara penuh dan dievaluasi kembali dengan metrik retrieval dan generation serta melalui penilaian pakar. Akhirnya, model diintegrasikan ke dalam aplikasi yang ramah pengguna untuk diterapkan dalam sesi konseling oleh praktisi konseling dan peneliti. Penelitian ini akan menjunjung tinggi prinsip etika dengan menganonimkan data dan tidak melibatkan subjek manusia secara langsung.



Gambar 3.1 Alur Penelitian

3.2 *Data Preparation*



Gambar 3.2 Tahap Data Preparation

Persiapan data melibatkan beberapa sub-tahapan yang bertujuan untuk memastikan kualitas dan format data sesuai dengan kebutuhan model.

3.2.1 *Data Collection*

Pada tahap ini, data yang diperlukan untuk membangun model RAG dikumpulkan dari berbagai sumber. Pada tahap ini, dua jenis data dikumpulkan:

- Data Dialog Klien-Konselor: Data ini berisi transkrip dialog antara klien dan konselor (*client-counselor dialogue* (CCD)), yang akan digunakan sebagai *input* utama untuk model. Data ini bisa diperoleh dari transkrip CCD. Sebelum digunakan, data akan diperiksa untuk menghilangkan informasi pribadi yang dapat mengidentifikasi individu.
- Data Korpus Referensi: Korpus referensi adalah koleksi teks yang berisi informasi teoretis, kajian literatur, atau sumber lain yang relevan dengan domain Konseling. Korpus ini digunakan sebagai sumber untuk membangun bagian *retrieval model*.

3.2.2 *Data Cleaning*

Tahap ini bertujuan untuk membersihkan data dari elemen-elemen yang tidak relevan atau tidak konsisten. Proses ini melibatkan penghapusan noise, duplikasi, dan inkonsistensi yang mungkin terdapat dalam data. Tujuan dari pembersihan data adalah untuk meningkatkan kualitas data dan memastikan bahwa data yang digunakan bebas dari kesalahan yang dapat mempengaruhi hasil akhir.

3.2.3 *Data Transformation*

Pada tahap data *transformation* data mentah diubah menjadi format yang siap digunakan dalam sistem Retrieval-Augmented Generation (RAG).

- *Chunking*: Memecah teks menjadi potongan-potongan kecil seperti kalimat atau paragraf. Contoh: Teks "Saya merasa cemas akhir-akhir ini. Apakah ada cara untuk mengatasi ini?" dipecah menjadi "Saya merasa cemas akhir-akhir ini." dan "Apakah ada cara untuk mengatasi ini?"
- *Vectorization*: Mengubah potongan teks menjadi representasi numerik (vektor) yang bisa diproses oleh model machine learning. Contoh: Kalimat

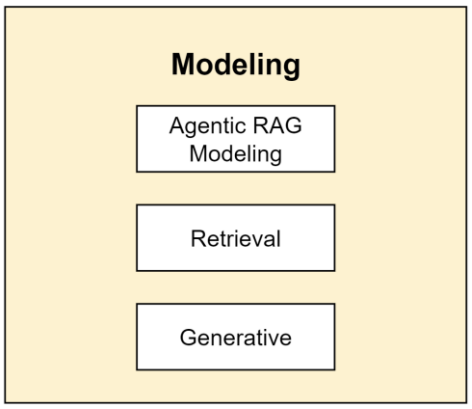
"Saya merasa cemas" diubah menjadi vektor seperti [0.25, -0.77, 0.56, ..., -0.12] menggunakan teknik *embedding.*

- *Embeddings:* Representasi numerik dari kata atau kalimat yang menangkap makna semantik dan hubungan antar kata. Contoh: Kata "kucing" diubah menjadi vektor embedding dengan model seperti BERT, yang bisa menjadi [0.12, -0.85, 0.45, ..., 0.21].
- *Indexing:* Menyimpan dan mengorganisir vektor dalam struktur data untuk pencarian cepat dan efisien. Contoh: Vektor dari berbagai kalimat diindeks menggunakan Faiss untuk memungkinkan pencarian berdasarkan kemiripan vektor.

3.2.4 *Data Storage*

Setelah data diproses, data tersebut disimpan dalam sistem penyimpanan yang memungkinkan akses cepat dan efisien selama proses modeling. Penyimpanan data harus dirancang sedemikian rupa sehingga data dapat diakses dengan mudah oleh model retrieval dan generatif.

3.3 *Modeling*



Gambar 3. 3 Tahap Modeling

Tahap ini mencakup pengembangan dan penyesuaian model RAG, yang terdiri dari model *retrieval* dan model generatif. Pemodelan dilakukan dengan memilih dan menyusun agent RAG yang tepat, serta mengoptimalkan model *retrieval* dan generatif agar dapat bekerja secara efektif bersama-sama.

- Agent RAG Modeling: Pada tahap ini, agent RAG dipilih dan disesuaikan untuk memastikan bahwa model *retrieval* dan generatif dapat bekerja secara harmonis. Pemodelan agent RAG melibatkan penentuan parameter dan arsitektur yang sesuai dengan tujuan sistem, serta integrasi antara komponen retrieval dan generatif.
    - *Retrieval Model*: Model ini bertugas untuk mencari dan mengambil informasi yang relevan dari korpus referensi berdasarkan *input* yang diberikan oleh pengguna.
    - *Generative Model*: Model ini bertugas untuk menghasilkan teks baru berdasarkan *input* yang diberikan. Model generatif ini biasanya dilatih menggunakan data dialog klien-konselor.

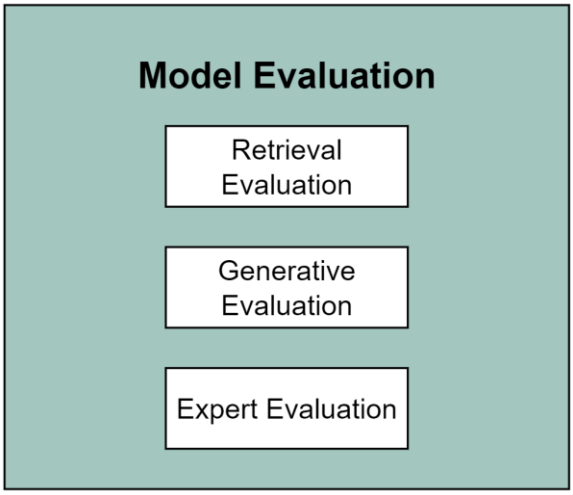
3.4 *Initial Evaluation*

Initial evaluation atau evaluasi awal dilakukan setelah integrasi awal antara model *retrieval* dan model generatif. Evaluasi ini bertujuan untuk mengukur kinerja sistem RAG secara keseluruhan sebelum diintegrasikan sepenuhnya ke dalam sistem yang lebih besar.

3.5 RAG *Integration*

Tahap ini melibatkan integrasi penuh antara model retrieval dan model generatif untuk membentuk sistem RAG yang lengkap. Integrasi dilakukan dengan memastikan bahwa kedua model dapat berfungsi secara kohesif dan saling mendukung dalam proses retrieval dan generation.

3.6 *Model Evaluation*



Gambar 3.5 Tahap Model Evaluation

Model yang telah dilatih kemudian dievaluasi untuk memastikan performanya sesuai dengan yang diharapkan.

- Model *Retrieval*: Model retrieval dievaluasi kembali untuk memastikan konsistensi dan kinerja setelah integrasi. Metrik seperti Hit Rate, MRR, dan NDCG digunakan untuk menilai efektivitas retrieval dalam skenario yang lebih kompleks. Contoh: MRR digunakan untuk mengukur peringkat rata-rata dari hasil yang diambil oleh model retrieval, yang menunjukkan seberapa cepat model menemukan informasi yang relevan.
- Model Generatif: Model generatif dievaluasi kembali menggunakan metrik seperti ROUGE dan BLEU untuk memastikan bahwa teks yang dihasilkan tetap akurat dan relevan setelah integrasi. Contoh: ROUGE digunakan untuk menilai seberapa baik output model generatif sesuai dengan teks referensi setelah model diintegrasikan dengan retrieval.
- *Expert Evaluation*: Evaluasi model oleh pakar di bidang psikologi untuk memastikan bahwa *output* yang dihasilkan model sesuai dengan standar profesional dan relevan untuk aplikasi di dunia nyata. Contoh: Meminta pakar psikologi untuk menilai apakah respon yang dihasilkan model terhadap kasus-kasus tertentu sudah tepat dalam konteks konseling.

Contoh gambaran *input* dan *output* model:

- *Input*

Transkrip Dialog Klien-Konselor: Ini adalah teks transkrip dari sesi konseling yang berisi percakapan antara klien dan konselor. Contoh teks:

```
Klien: Saya merasa cemas setiap kali harus berbicara di depan umum.
Konselor: Apa yang Anda pikirkan saat merasa cemas?
Klien: Saya merasa takut jika orang lain akan menilai saya.
```

- *Output*

Output berupa *Labeled Transcripts*, yaitu transkrip yang telah diberi label atau kode berdasarkan kategori psikologi yang relevan. Contoh *Output*:

```
Klien: Saya merasa cemas setiap kali harus berbicara di depan umum.
Kode: [Kecemasan Sosial]
Konselor: Apa yang Anda pikirkan saat merasa cemas?
Kode: [Pemikiran Negatif]
Klien: Saya merasa takut jika orang lain akan menilai saya.
Kode: [Ketakutan akan Penilaian]
```

3.7 Integration & Deployment

Setelah sistem dievaluasi dan disempurnakan, tahap akhir adalah mengintegrasikan sistem RAG ke dalam lingkungan kerja yang sesungguhnya dan menerapkannya untuk digunakan secara praktis. Penerapan ini melibatkan pengujian akhir dan pemantauan sistem untuk memastikan bahwa sistem berfungsi dengan baik dalam lingkungan nyata.

3.8 Jadwal Estimasi Penelitian

Jadwal Estimasi Penelitian menjelaskan mengenai rancangan kegiatan yang dilakukan selama penelitian beserta estimasi waktu tiap kegiatan yang dilakukan.

Tabel 3.1 Jadwal Penelitian

No	Uraian Kegiatan	Tahun 1		Tahun 2		Tahun 3	
		Sem1	Sem2	Sem1	Sem2	Sem1	Sem2
1	Penyusunan Proposal						
1	Ujian Kualifikasi						
2	Persiapan data penelitian						
3	Pengolahan penelitian						
4	*Progress report* 1						
5	*Progress report* 2						
6	Publikasi						
7	Sidang Tertutup						
8	Sidang Terbuka						

DAFTAR PUSTAKA

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., Mccandlish, S., Radford, A., Sutskever, I. and Amodei, D., n.d. *Language Models are Few-Shot Learners*. [online] Available at: <https://commoncrawl.org/the-data/>.

Cao, J., Tanana, M., Imel, Z.E., Poitras, E., Atkins, D.C. and Srikumar, V., 2019. *Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes*. [online] Available at: <https://www.>.

Gaut, G., Steyvers, M., Imel, Z.E., Atkins, D.C. and Smyth, P., 2017. Content Coding of Psychotherapy Transcripts Using Labeled Topic Models. *IEEE Journal of Biomedical and Health Informatics*, 21(2), pp.476–487. https://doi.org/10.1109/JBHI.2015.2503985.

Han, G., Liu, W., Huang, X. and Borsari, B., 2024. Chain-of-Interaction: Enhancing Large Language Models for Psychiatric Behavior Understanding by Dyadic Contexts. [online] Available at: <http://arxiv.org/abs/2403.13786>.

Imel, Z.E., Steyvers, M. and Atkins, D.C., 2015. Computational psychotherapy research: Scaling up the evaluation of patient-provider interactions. *Psychotherapy*, 52(1), pp.19–30. https://doi.org/10.1037/a0036841.

Kazdin, A.E., 2017. *Addressing the treatment gap: A key challenge for extending evidence-based psychosocial interventions*. *Behaviour Research and Therapy*, https://doi.org/10.1016/j.brat.2016.06.004.

Kumar Adhikary, P., Srivastava, A., Kumar, S., Singh, S.M., Manuja, P., Gopinath, J.K., Krishnan, V., Kedia, S., Deb, K.S. and Chakraborty, T., 2024. *Exploring the Efficacy of Large Language Models in Summarizing Mental Health Counseling Sessions: A Benchmark Study*. https://doi.org/10.48550/arXiv.2402.19052.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S. and Kiela, D., n.d. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. [online] Available at: <https://github.com/huggingface/transformers/blob/master/>.

Marvin, G., Hellen, N., Jjingo, D. and Nakatumba-Nabende, J., 2024. Prompt Engineering in Large Language Models. pp.387–402. https://doi.org/10.1007/978-981-99-7962-2_30.

Radeva, I., Popchev, I., Doukovska, L. and Dimitrova, M., 2024. Web Application for Retrieval-Augmented Generation: Implementation and Testing. *Electronics (Switzerland)*, 13(7). https://doi.org/10.3390/electronics13071361.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., n.d. *Language Models are Unsupervised Multitask Learners*. [online] Available at: <https://github.com/codelucas/newspaper>.

Rae, J.W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick, J., Cassirer, A., Powell, R., Driessche, G. van den, Hendricks, L.A., Rauh, M., Huang, P.-S., Glaese, A., Welbl, J., Dathathri, S., Huang, S., Uesato, J., Mellor, J., Higgins, I., Creswell, A., McAleese, N., Wu, A., Elsen, E., Jayakumar, S., Buchatskaya, E., Budden, D., Sutherland, E., Simonyan, K., Paganini, M., Sifre, L., Martens, L., Li, X.L., Kuncoro, A., Nematzadeh, A., Gribovskaya, E., Donato, D., Lazaridou, A., Mensch, A., Lespiau, J.-B., Tsimpoukelli, M., Grigorev, N., Fritz, D., Sottiaux, T., Pajarskas, M., Pohlen, T., Gong, Z., Toyama, D., d'Autume, C. de M., Li, Y., Terzi, T., Mikulik, V., Babuschkin, I., Clark, A., Casas, D. de Las, Guy, A., Jones, C., Bradbury, J., Johnson, M., Hechtman, B., Weidinger, L., Gabriel, I., Isaac, W., Lockhart, E., Osindero, S., Rimell, L., Dyer, C., Vinyals, O., Ayoub, K., Stanway, J., Bennett, L., Hassabis, D., Kavukcuoglu, K. and Irving, G., 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. [online] Available at: <http://arxiv.org/abs/2112.11446>.

Reshamwala, A., Mishra, D. and Pawar, P., 2013. *REVIEW ON NATURAL LANGUAGE PROCESSING*. [online] *An International Journal (ESTIJ*, Available at: <https://www.researchgate.net/publication/235788362>.

Seligman, L., 2004. *Documentation, Report Writing, and Record Keeping in Counseling*. https://doi.org/10.1007/978-1-4419-8927-7_11.

Tran, T., Yin, Y., Tavabi, L., Delacruz, J., Borsari, B., Woolley, J.D., Scherer, S. and Soleymani, M., 2023. Multimodal Analysis and Assessment of Therapist Empathy in Motivational Interviews. In: *ACM International Conference Proceeding Series*. Association for Computing Machinery. pp.406–415. https://doi.org/10.1145/3577190.3614105.