BAB III
METODOLOGI PENELITIAN

3.1 Objek dan Subjek Penelitian

Subjek dalam penelitian ini adalah masalah yang dihadapi dalam model chatbot menggunakan arsitektur BERT. Objek dalam penelitian ini adalah model chatbot menggunakan arsitektur BERT dalam dataset terkait informasi perpustakaan Institut Teknologi Telkom Purwokerto dan MBKM.

3.2 Alat dan Bahan

Dalam penelitian ini alat dan bahan digunakan untuk menunjang keberhasilan penelitian. Alat dan bahan yang dimaksud adalah:

3.2.1. Alat

Perangkat keras yang digunakan yaitu :

a) Laptop Lenovo dengan spesifikasi:
    1) Processor : AMD A9-9425 RADEON R5, 5 COMPUTE CORES 2C+3G (2 CPUs), ~ 3.10 GHz
    2) Memory : 8 GB RAM DDR4
    3) Graphic Card : AMD Radeon RX Vega 5
    4) SSD : 512 GB

Perangkat Lunak yang digunakan yaitu :

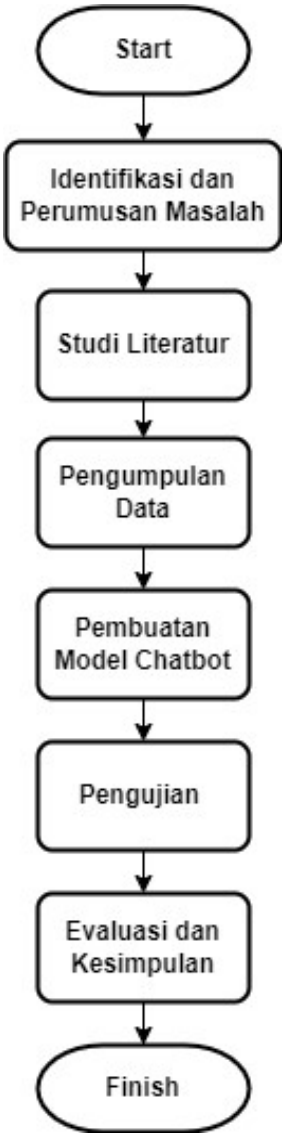
a) Google Colaboratory
b) Python 3.10

3.2.2. Bahan

Penelitian ini menggunakan data informasi mengenai perpustakaan Institut Teknologi Telkom Purwokerto dan MBKM untuk *dataset*, dimana dikombinasikan dengan *pre-trained* BERT saat *Fine-Tuning*.

3.3 Diagram Alir Penelitian

Ada beberapa tahapan dalam melakukan penelitian dalam penyusunan laporan penelitian ini. Berikut diagram alir penelitian yang dilakukan dalam penyusunan laporan ini yang dapat dilihat pada Gambar 3.1.



Gambar 3. 1 Diagram alir penelitian

3.2.1 Identifikasi Masalah

Proses mengidentifikasi dan mendefinisikan masalah yang akan dipecahkan atau diatasi dalam penelitian. Proses ini menggunakan pendekatan logika dan matematis. Tujuan dari identifikasi data atau informasi adalah untuk mengajukan pertanyaan mendasar terkait membuat model chatbot dengan arsitektur BERT sehingga penelitian tidak menghalangi diskusi.

3.2.2 Studi Literatur

Penelitian ini diperlukan referensi atau sumber data sebagai dasar dan pedoman pengembangan penelitian. Oleh karena itu penulis membaca, meneliti dan memahami konsep dan pembahasan terkait membangun *chatbot* menggunakan arsitektur BERT. Hasil yang diperoleh dijadikan sebagai dasar penulisan dan penelitian yang telah dilakukan.

3.2.3 Pengumpulan Data

Pengumpulan data menggunakan data mengenai informasi perpustakaan Institut Teknologi Telkom Purwokerto dan data mengenai MBKM diambil dari penelitian sebelumnya. Untuk data informasi perpustakaan dikumpulkan dengan mengajukan permohonan data kepada admin perpustakaan. Total dataset untuk penelitian ini berjumlah 220. Berikut sampel *dataset* yang dikumpulkan yang ada di Tabel 3.1.

Tabel 3. 1 Sampel Dataset

No	Pertanyaan	Konteks	Jawaban
1	Berapa lama saya bisa	Perpustakaan ITTP memiliki ketentuan dalam peminjaman	Setiap buku dapat dipinjam selama 1 minggu.

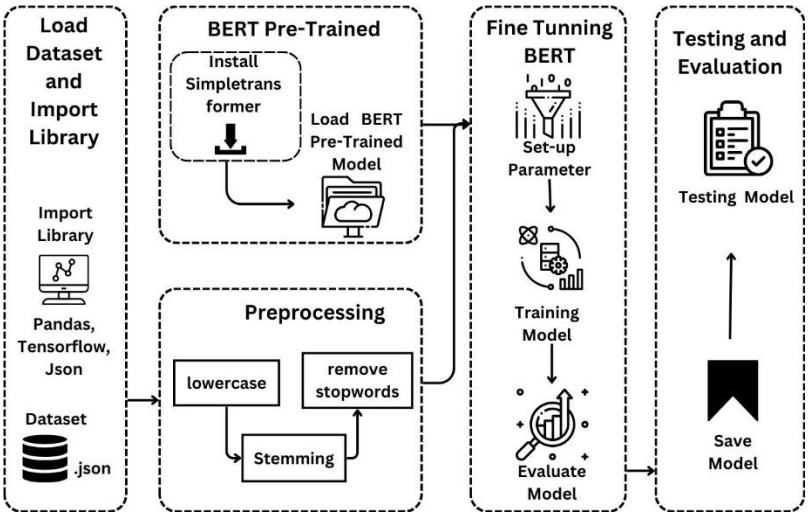
No	Pertanyaan	Konteks	Jawaban
	meminjam buku?	buku dimana setiap buku dapat dipinjam selama 1 minggu. Anda dapat memperpanjang masa peminjaman jika tidak ada peminjam lain yang menginginkannya.	
2	Apakah perpustakaan ini menyediakan layanan peminjaman buku secara daring?	Perpustakaan ITTP selain memiliki layanan luring juga menyediakan layanan peminjaman buku secara daring. Untuk peminjaman buku secara daring dapat dengan mengakses web dlibrary.ittelkom-pwt.ac.id.	Ya, perpustakaan ITTP menyediakan layanan peminjaman buku secara daring
3	Bagaimana cara register ke perpustakaan?	Perpustakaan ITTP memiliki ketentuan dalam kunjungan, dimana pengunjung harus reservasi di web dlibrary.ittelkom-	Silahkan akses reservasi di web dlibrary.ittelkom-pwt.ac.id pada bagian reservasi kunjungan.

No	Pertanyaan	Konteks	Jawaban
		pwt.ac.id pada bagian reservasi kunjungan.	
4	Apa itu Kampus Merdeka?	Kampus Merdeka adalah kebijakan yang dikeluarkan oleh Kemendikbudristek dengan memberikan hak kepada Mahasiswa untuk mengambil mata kuliah di luar program studi selama 1 semester dan berkegiatan di luar perguruan tinggi selama 2 semester. Perguruan tinggi diberikan kebebasan untuk menyediakan kegiatan Kampus Merdeka yang sesuai dengan kebutuhan dan minat mahasiswanya.	Kampus Merdeka adalah kebijakan yang dikeluarkan oleh Kemendikbudristek dengan memberikan hak kepada Mahasiswa untuk mengambil mata kuliah di luar program studi selama 1 semester dan berkegiatan di luar perguruan tinggi selama 2 semester.

No	Pertanyaan	Konteks	Jawaban
5	Apa tujuan dari kampus merdeka?	Tujuan kebijakan Kampus Merdeka adalah memberikan kesempatan kepada mahasiswa untuk memilih mata kuliah yang akan mereka tempuh berdasarkan keinginan sendiri serta mendorong mahasiswa meningkatkan soft skills serta hard skills agar siap bersaing dalam dunia global.	Memberikan kesempatan kepada mahasiswa untuk memilih mata kuliah yang akan mereka tempuh berdasarkan keinginan sendiri.

3.2.4 Pembuatan Chatbot

Tahapan berikutnya dilakukan pembuatan chatbot menggunakan arsitektur BERT. Berikut diagram alir dari proses pembuatan chatbot untuk penelitian ini yang dapat dilihat pada Gambar 3.2.



Gambar 3. 2 Diagram alir proses pembuatan *chatbot*

3.2.4.1 Load Dataset and Library

Tahap ini melakukan *import library* yang diperlukan seperti *Pandas*, *JSON*, *Sastrawi, Tensorflow*, dan *Pytorch*. serta mempersiapkan dataset yang telah dikumpulkan sebelumnya dimana dalam penelitian ini dataset diatur dalam format *JSON*.

3.2.4.2 Preprocessing

Tahap *pre-processing* dilakukan beberapa tahap untuk membuat dataset yang disiapkan dapat dipakai untuk model BERT-nya. Library yang dipakai untuk tahap ini yaitu *library Sastrawi*. Berikut tahapan *preprocessing* untuk datasetnya.

1. Lowercase

Proses pengubahan huruf kapital menjadi huruf biasa atau kecil, proses ini menggunakan fitur lowercase yang ada di python. Berikut contoh lowercase yang ada pada Tabel 3.2.

Tabel 3. 2 Lowercase

Sebelum	Sesudah
Apakah ada biaya yang harus dibayar selama mengikuti Studi Independen	apakah ada biaya yang harus dibayar selama mengikuti studi independen

2. Stemming

Proses isi proses mengubah kata-kata menjadi bentuk dasar dengan menghapus imbuhannya, proses ini menggunakan *library sastrawi* berupa *StemmerFactory*. Berikut contoh *stemming* yang ada pada Tabel 3.3.

Tabel 3. 3 *Stemming*

Sebelum	Sesudah
apakah ada biaya yang harus dibayar selama mengikuti studi independen	apakah ada biaya yang harus bayar lama ikut studi independen

3. Remove Stopword

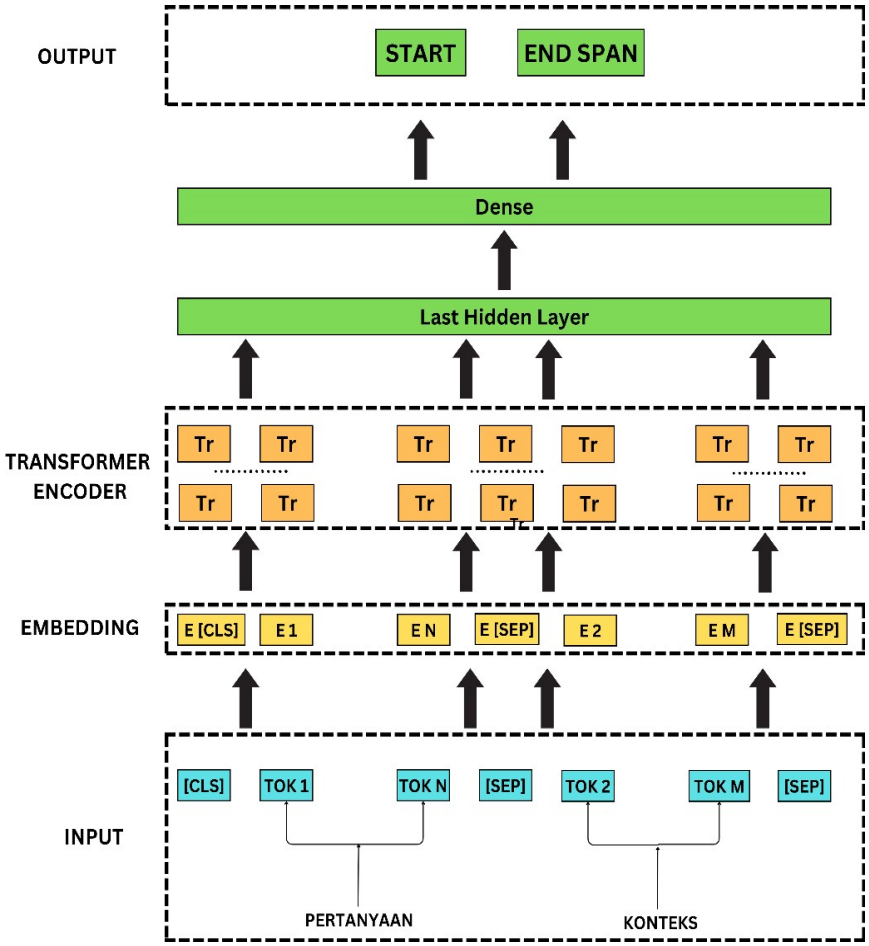
Proses isi menghapus kata yang tidak penting seperti kata sambung atau sejenisnya, proses ini menggunakan *library sastrawi* berupa *StopWordRemoverFactory*. Berikut contoh *remove stopword* yang ada pada Tabel 3.4.

Tabel 3. 4 *Remove Stopword*

Sebelum	Sesudah
apakah ada biaya yang harus bayar lama ikut studi independen	biaya bayar lama ikut studi independen

3.2.4.3 Arsitektur BERT

Tahap ini melakukan instalasi *simpletransformer* terlebih yang kemudian dilanjutkan dengan memanggil pre-trained model BERT yang akan digunakan. penelitian ini menggunakan beberapa model BERT seperti *bert-base-uncased, bert-base-multilingual* dan *indobert-base-uncased* untuk mendapatkan model terbaik. Penelitian ini menggunakan task *question answering* dalam membuat model chatbot pada ketiga model BERT tersebut. Untuk Arsitektur BERT dengan *task question answering* dapat dilihat pada Gambar 3.3.



Gambar 3. 3 Arsitektur BERT *Question Answering*

Arsitektur BERT dengan *question answering task* memiliki alir seperti Gambar 3.3, cara kerja BERT diawali dengan menginputkan pertanyaan dan konteks dimana sebelum dilakukan *embedding*, pertanyaan dan konteks dalam *question answering task* akan dimasukkan ke dalam model BERT dengan diberi token CLS di awal dan token SEP sebagai pemisah antara pertanyaan dan konteks.

Sesudah tahap *embedding*, token-token ini akan diubah menjadi representasi vektor yang kaya secara semantik dan digunakan untuk melatih model dalam tahap fine-tuning untuk menyelesaikan *Question Answering Task*. Di layer *transformer encoder*, representasi vektor tersebut akan mengalami serangkaian transformasi berulang untuk memperoleh representasi yang lebih kompleks dan kaya secara semantik seperti *self attention mechanism* yang memberikan nilai perhatian pada kata yang penting secara random sehingga menghasilkan inputan konseptual.

Sesudah *layer encoder* akan masuk ke *last hidden layer* dimana representasi vektor dari token CLS *(Classification Token)* diambil sebagai representasi akhir dari pertanyaan dan konteks. Representasi vektor di token CLS mengandung informasi tentang keseluruhan pertanyaan dan konteks, dan akan digunakan untuk memprediksi jawaban yang tepat. Dari hasil di *last hidden layer* akan masuk ke dense dengan aktivasi softmax yang menghasilkan probabilitas yang disesuaikan untuk setiap token sebelum ke output layer. Output layer yang dihasilkan berupa probabilitas untuk setiap token dalam konteks, yang menunjukkan sejauh mana token tersebut mungkin merupakan awal (*start span*) atau akhir (*end span*) dari jawaban yang relevan. Berikut contoh sederhana dalam proses di arsitektur BERT.

Pertama dengan inputan untuk arsitektur BERT terdapat question dan context.

Question :

“dimana lokasi perpustakaan ITTP?”

Context :

“Perpustakaan ITTP berlokasi di Gedung Rektorat Lantai 1.”

Proses selanjutnya dilakukan tokenisasi dari *question* dan *context* , sehingga menjadi seperti berikut.

'[CLS]', 'dimana', 'lokasi', 'perp', '##su', '##tak', '##aan', '[UNK]', '?', '[SEP]', '[UNK]', '[UNK]', 'berlokasi', 'di', 'gedung', 'rektorat', 'lantai', '1', '.', '[SEP]'

Hasil dari tokenisasi tersebut ada token [UNK], dimana token tersebut dipakai untuk istilah asing atau singkatan yang tidak ada didalam *corpus* BERT. Setelah tokenisasi dilanjutkan dengan Embedding dengan hasil seperti berikut.

'input_ids': [3, 3357, 3197, 3641, 3321, 12137, 1603, 1, 35, 4, 1, 1, 9474, 1485, 3183, 22757, 5844, 21, 18, 4], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

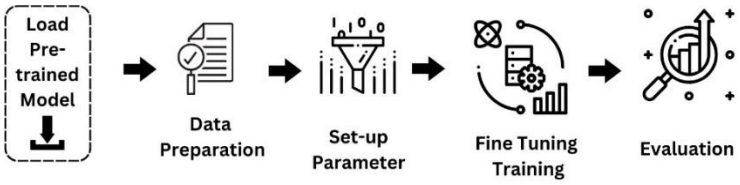
Hasil embedding tersebut akan diproses pada *encoder transformers* hingga *hidden layer* terakhir untuk menghasilkan *start span* dan *end span*. Dan output yang dihasilkan seperti berikut.

start_span = tensor(10)

end_span = tensor(18)

Setelah digabungkan dari start_span dan end_span menghasilkan jawaban “berlokasi di gedung rektorat lantai 1”.

3.2.4.4 Fine-tuning BERT



Gambar 3. 4 Diagram Alir *Fine Tuning* BERT

Tahap *fine tuning* dapat dilihat pada Gambar 3.4 untuk detailnya, dimana terdapat *load* beberapa model BERT seperti *bert-base-uncased, bert-base-multilingual* dan *indobert-base-uncased*. Data preparation dari dataset yang telah di preprocessing, *Set-up Parameter* melakukan konfigurasi parameter untuk tiap model dari menggunakan GPU tidak saat training datanya, penyimpanan hasil output training, epochs, batch size, dan lain sebagainya.

Tahapan berikutnya dilanjutkan dengan *Fine Tuning Training* dimana proses ini melakukan training tiap model yang ada untuk mencari model terbaiknya. Model BERT yang telah di *Fine tuning* akan dievaluasi dalam kinerjanya seperti dalam memprediksi jawabannya menggunakan *F1-Score* dan *Exact Match* (EM). Setelah selesai semua proses, semua model yang telah di training dibandingkan dengan satu sama lainya untuk hasil evaluasinya. Dan model terbaik didapatkan dari hasil evaluasi yang nilainya lebih tinggi dari semua model lainnya.

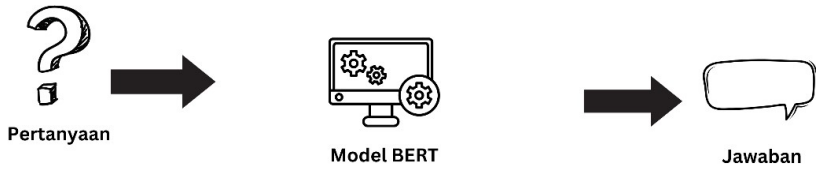
3.2.4.5 Evaluasi Model BERT

Tahap testing dan evaluasi menggunakan model BERT yang telah dilakukan *fine-tuning* dan disimpan modelnya. Untuk mendapatkan hasil yang baik pada evaluasi menggunakan data pertanyaan yang berhubungan dengan dataset sebelumnya namun tidak termasuk ke dalam dataset yang di *fine-tuning* sebelumnya.

Metric yang dipakai dalam evaluasi ini yaitu *F1-score* dan *Exact Match* (EM).

3.2.5 Pengujian

Skenario pengujian model *chatbot* bahasa Indonesia menggunakan metode BERT *Question-Answering* dapat dilihat pada Gambar 3.5.



Gambar 3. 5 Skema Pengujian Model *Chatbot*

Pengujian model dilakukan dengan skenario data di luar konteks dataset dan data di dalam konteks dataset. Pengujian dilakukan dua tahap dimana tahap pertama dilakukan dengan menggunakan *model.predict* yang disediakan oleh library. Tahap pertama ini dilakukan untuk menguji keakuratan model dalam memberikan jawaban dengan inputan pertanyaan dan konteks. Tahap kedua pengujian menggunakan *Haystack library*, dimana dari pertanyaan yang dimasukkan ke dalam model BERT untuk diproses mencari *probabilitas* jawabannya. Dalam model BERT ditentukan 5 *probabilitas* jawaban dan 5 *probabilitas* konteks untuk menghasilkan jawaban, dan dicari nilai tertinggi dari *probabilitas* jawaban. Pada *Haystack* library melakukan pengecekan kesesuaian konteks dan jawaban dari korpus yang tersedia dari model BERT. Sehingga nilai *score probabilitas* jawaban yang paling tinggi akan menjadi luaran dari jawaban yang sesuai.