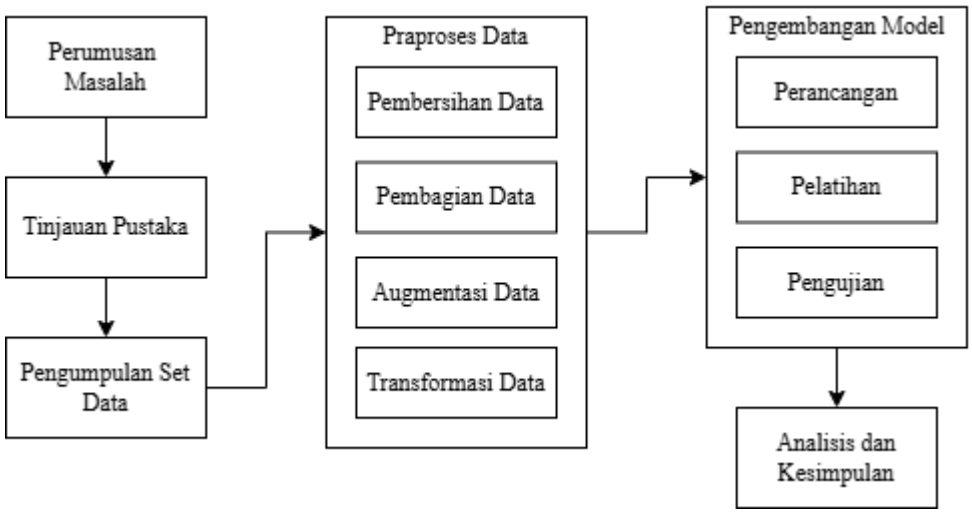
BAB III
METODE PENELITIAN

3.1 Desain Penelitian

Desain penelitian adalah tahapan, rencana, atau prosedur dalam melakukan penelitian. Desain penelitian yang digunakan mengacu pada metode (Santoso et al., 2020) yang dapat dilihat pada Gambar 3.1. Desain penelitian ini terdiri dari perumusan masalah, tinjauan pustaka, pengumpulan set data, pra proses data, pengembangan model, serta analisis dan kesimpulan. Penjelasan detail untuk setiap tahapan disampaikan sebagai berikut.



Gambar 3. 1 Desain Penelitian

3.1.1 Perumusan Masalah

Pada tahapan ini, peneliti mengidentifikasi masalah dengan mempelajari penelitian terdahulu terkait *Automatic Short Answer Scoring* (ASAS). Studi literatur dilakukan untuk menemukan celah penelitian dan mengisi celah tersebut dengan penelitian lanjutan. Dari perumusan masalah, ditemukan dasar penelitian seperti, latar belakang, masalah penelitian, tujuan penelitian, dan metode yang akan digunakan dalam penelitian.

3.1.2 Tinjauan Pustaka

Tahapan ini bertujuan untuk mengkaji berbagai teori, pendekatan, dan teknik yang telah digunakan dalam penelitian ASAS sebelumnya. Dengan melakukan ini, peneliti memperoleh pemahaman yang mendalam mengenai pendekatan dan model yang digunakan. Selain itu, tinjauan pustaka juga berisi penjelasan terkait variabel yang diteliti.

3.1.3 Pengumpulan Set Data

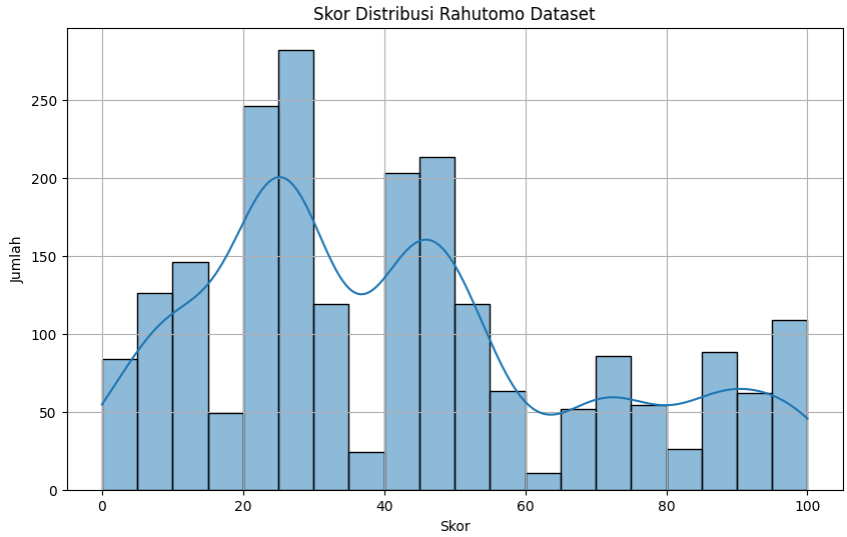
Pada tahapan ini, peneliti mengumpulkan data yang diperlukan untuk melatih dan menguji model ASAS. Data yang digunakan didapat dari penelitian (Rahutomo et al., 2018), yang pratinjau data dapat dilihat pada Tabel 3.1. Dataset ini terdiri dari 40 pertanyaan yang terbagi ke dalam empat kategori, yaitu politik, gaya hidup, olahraga, dan teknologi. Dimana masing-masing kategori memiliki 10 pertanyaan. Setiap pertanyaan disertai dengan jawaban referensi dan jawaban dari sekitar 50 mahasiswa. Total data yang tersedia sebanyak 2.162, dengan penilaian yang diberikan oleh tiga evaluator dengan skala 0-100, dimana 100 menunjukkan jawaban paling relevan dan 0 menunjukkan jawaban yang sepenuhnya salah. Dataset ini memenuhi karakteristik data yang dibutuhkan pada penelitian yaitu data memiliki kolom jawaban referensi, jawaban siswa, dan label skor dalam numerik. Gambar 3.2 menunjukkan distribusi skor pada dataset.

Tabel 3. 1 Pratinjau Dataset Rahutomo

Kategori	Pertanyaan	Jawaban Referensi	Jawaban Siswa	Skor 1	Skor 2	Skor 3
Lifestyle	Bagaimana hubungan berpikir positif terhadap pola hidup sehat?	Otak merupakan bagian terpenting dari tubuh manusia. Segala aktivitas akan dikoordinasikan dengan otak sebelum dijalankan oleh bagian tubuh. Pikiran positif akan menghindarkan anda dari stres,	karena pikiran bisa mempengaruhi kondisi tubuh kita	28	25	25

		meningkatkan rasa percaya diri serta menjaga kinerja organ tubuh lainnya tetap maksimal.				
Olahraga	Sebutkan 5 partai yang biasa dilombakan dalam bulu tangkis.	-Tunggal putra - Tunggal putri - Ganda putra - Ganda putri - Ganda campuran	Tunggal putra, tunggal putri, ganda putra, ganda putri dan ganda campuran	100	100	100
Politik	Sebutkan 5 butir sila Pancasila.	- Ketuhanan Yang Maha Esa, - kemanusiaan yang adil dan beradab, - persatuan Indonesia, - kerakyatan yang dipimpin oleh hikmat kebijaksanaan dalam permusyawaratan perwakilan, - keadilan sosial bagi seluruh rakyat Indonesia	1. Ketuhanan 2. Kemanusiaan 3. Persatuan 4. Kerakyatan 5. Keadilan	40	50	50

Teknologi	Apa kepanjangan dari LCD, CPU dan GPS?	LCD (Liquid Crystal Display), CPU (Central Processing Unit), GPS (Global Positioning System)	LCD = liquid crystal display CPU = central processing unit GPS = Global Positioning System	100	100	100



Gambar 3. 2 Distribusi Skor pada Dataset Rahutomo

3.1.4 Pra proses Data

Pada tahap ini, peneliti melakukan pra proses pada dataset yang mencakup beberapa langkah, yaitu pembersihan data, pembagian data, augmentasi data, dan transformasi data. Setiap tahapan dirancang untuk mengurangi noise, meningkatkan kualitas representasi data, dan menjaga distribusi data tetap konsisten dalam proses pelatihan model. Penjelasan rinci untuk masing-masing tahapan diberikan sebagai berikut.

3.1.4.1 Pembersihan Data

Pada tahap pembersihan data, dilakukan serangkaian proses untuk menghilangkan *noise* pada dataset. Pembersihan data yang dilakukan meliputi menghapus tanda baca, konversi teks menjadi huruf kecil, menghapus spasi

berlebih, dan menghapus data duplikat. Selain itu, peneliti juga menangani data yang tidak konsisten seperti skor yang berbeda terhadap jawaban siswa yang sama dan menghapus data yang hanya berisi string kosong. Dengan melakukan berbagai pembersihan data ini, peneliti memastikan data bersih dari noise dan duplikasi yang dapat mengganggu proses pembelajaran model.

3.1.4.2 Pembagian Data

Penelitian ini menggunakan dua skenario pembagian data untuk menguji performa model pada berbagai aspek evaluasi, yaitu *specific-prompt* dan *cross-prompt*. Berikut adalah penjelasan mendetail tentang kedua skenario pembagian data tersebut.

1. Skenario *Specific-Prompt*

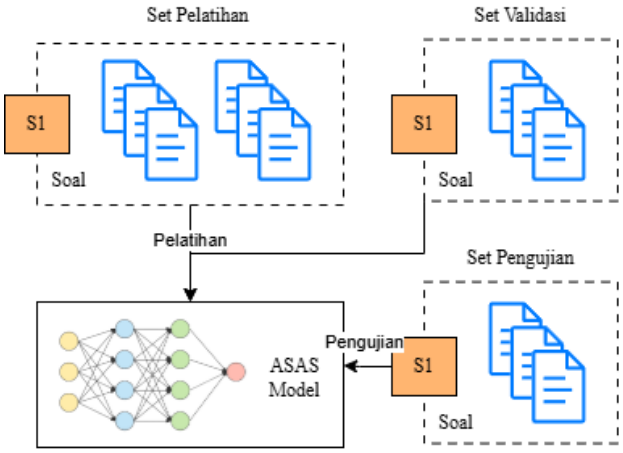
Pada skenario *specific-prompt*, data dibagi berdasarkan jawaban dari setiap siswa untuk tiap pertanyaan. Skenario ini digunakan untuk mengevaluasi performa model ketika memprediksi skor jawaban berdasarkan pertanyaan yang sudah dilatih sebelumnya. Visualisasi skenario ini dapat dilihat pada Gambar 3.3. Sebagai contoh, jika dalam dataset terdapat 10 pertanyaan, dan setiap pertanyaan memiliki 10 jawaban dari siswa, dengan rasio pembagian 8:1:1, maka pembagiannya akan dilakukan sebagai berikut.

- Data Pelatihan (80%): 8 jawaban dari setiap pertanyaan digunakan untuk data pelatihan.
- Data Validasi (10%): 1 jawaban dari setiap pertanyaan digunakan untuk data validasi.
- Data Pengujian (10%): 1 jawaban dari setiap pertanyaan digunakan untuk data pengujian.

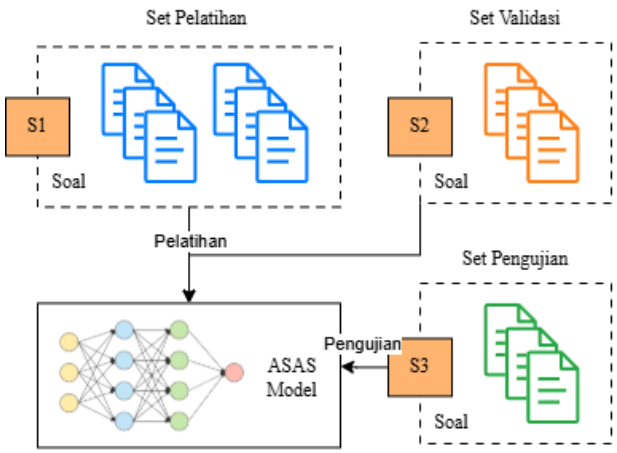
2. Skenario *Cross-Prompt*

Pada skenario *cross-prompt*, data dibagi berdasarkan pertanyaan, bukan berdasarkan jawaban siswa. Skenario ini digunakan untuk mengevaluasi performa model ketika memprediksi skor jawaban untuk pertanyaan yang belum dilihat sebelumnya. Visualisasi skenario ini dapat dilihat pada Gambar 3.4. Sebagai contoh, jika dalam dataset terdapat 10 pertanyaan, dan setiap pertanyaan memiliki 10 jawaban dari siswa, dengan rasio pembagian 8:1:1, maka pembagiannya akan dilakukan sebagai berikut.

- Data Pelatihan (80%): 8 pertanyaan beserta 10 jawaban dari tiap pertanyaan digunakan untuk data pelatihan.
- Data Validasi (10%): 1 pertanyaan beserta 10 jawaban dari pertanyaan digunakan untuk data validasi.
- Data Pengujian (10%): 1 pertanyaan beserta 10 jawaban dari pertanyaan digunakan untuk data pengujian.



Gambar 3. 3 Skenario Pembagian Data *Specific-Prompt*



Gambar 3. 4 Skenario Pembagian Data *Cross-Prompt*

3.1.4.3 Augmentasi Data

Setelah proses pembagian data, peneliti melakukan tahap augmentasi pada data pelatihan untuk meningkatkan keragaman data serta menyeimbangkan distribusi skor. Pada penelitian ini, metode augmentasi yang digunakan adalah teknik *prompt engineering* dengan pendekatan substitusi sinonim pada tingkat kata,

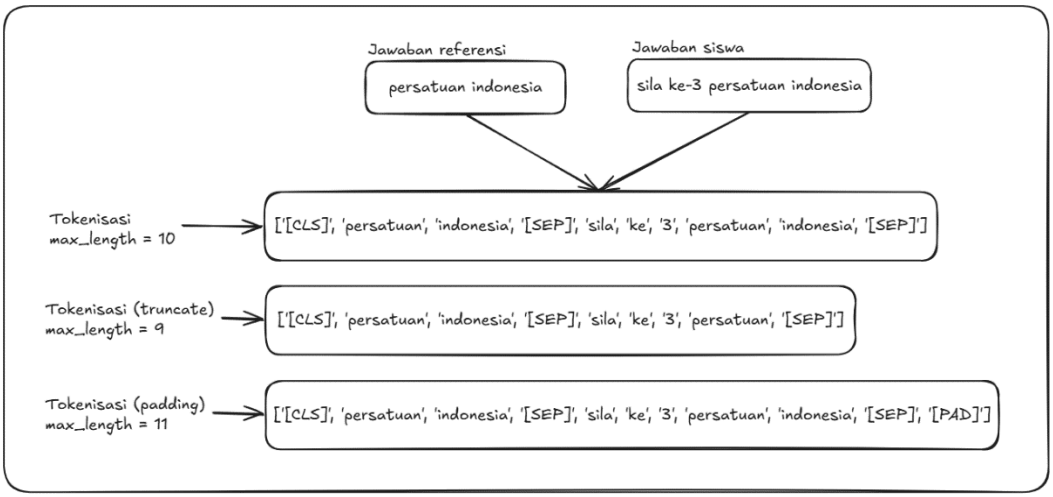
yang juga digunakan dalam penelitian (Wijanto & Yong, 2024). Proses penggantian sinonim ini dilakukan menggunakan model *Large Language Model* (LLM) *open-source*, yaitu Gemma 3 varian 4B parameter.

3.1.4.4 Transformasi Data

Peneliti melakukan transformasi data pada kolom jawaban dan label pada dataset untuk mempersiapkan input yang sesuai dengan kebutuhan model. Transformasi ini meliputi normalisasi skor, pembentukan format input teks, tokenisasi, dan encoding. Normalisasi dilakukan pada kolom label atau skor dengan mengubah rentang nilai dari 0–100 menjadi 0–1. Selanjutnya, peneliti menyesuaikan format input teks dengan pendekatan yang digunakan. Setelah format input dibuat, dilakukan proses tokenisasi dan encoding untuk mengubah data teks menjadi representasi numerik yang dapat diproses oleh model. Penjelasan lebih lanjut mengenai proses format input, tokenisasi, dan encoding pada masing-masing pendekatan disajikan sebagai berikut.

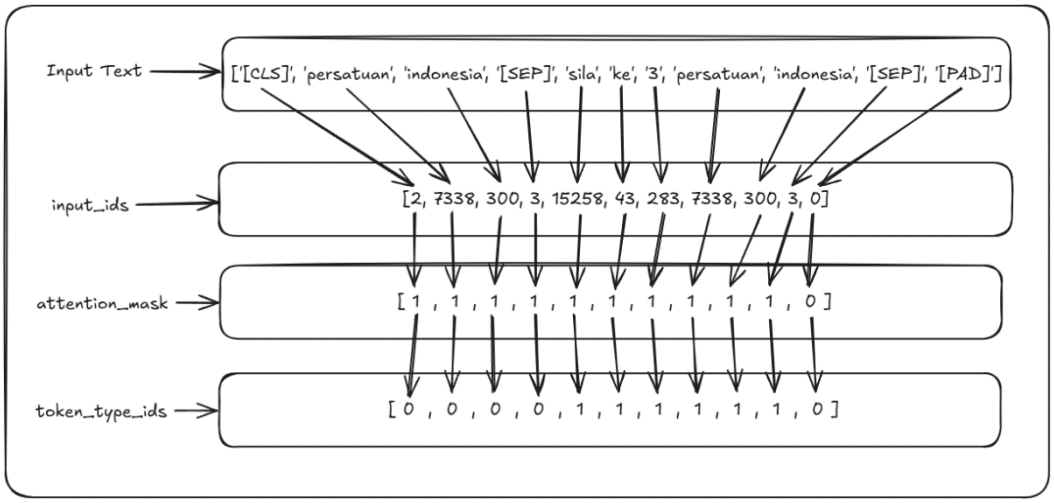
1. *Direct Scoring*

Pada pendekatan ini, jawaban referensi dan jawaban siswa digabungkan ke dalam satu input teks dengan format. "[CLS] jawaban referensi [SEP] jawaban siswa [SEP]". Format ini digunakan karena pendekatan *direct scoring* hanya menerima satu input teks, sehingga penggabungan ini memungkinkan model mempertimbangkan kedua bagian teks secara bersamaan. Input teks ini kemudian melalui proses tokenisasi menggunakan tokenizer dari model pralatih IndoBERT, yaitu BertTokenizer. Untuk memastikan panjang input token konsisten pada seluruh data, diterapkan parameter *max_length* sebesar 512. Jika jumlah token pada input melebihi batas ini, maka akan dilakukan *truncation* atau pemotongan terhadap token yang berlebih. Sebaliknya, jika jumlah token kurang dari 512, maka akan dilakukan *padding* dengan menambahkan token khusus [PAD] hingga panjang input mencapai 512 token. Mekanisme ini penting untuk menjaga kesesuaian struktur input dengan arsitektur model. Contoh proses tokenisasi ditampilkan pada Gambar 3.5.



Gambar 3. 5 Contoh Proses Tokenisasi

Transformasi selanjutnya adalah melakukan encoding pada token-token tersebut. Encoding dilakukan untuk mengubah token menjadi tiga komponen fitur berbentuk numerik, yaitu *input_ids*, *attention_mask*, dan *token_type_ids*. Komponen input_ids berisi representasi indeks token dalam *vocabulary*, *attention_mask* berisi penanda 0 atau 1 untuk membedakan token asli dengan token padding ([PAD]), dan token_type_ids berisi penanda 0 atau 1 untuk membedakan antara segmen jawaban referensi dan jawaban siswa. Contoh proses encoding ditampilkan pada Gambar 3.6.



Gambar 3. 6 Contoh Proses Encoding

2. Similarity-Based Scoring

Pada pendekan ini, jawaban referensi dan jawaban siswa tidak digabungkan menjadi satu teks, tetapi diolah secara terpisah. Masing-masing teks dikemas dalam

format input. "[CLS] jawaban [SEP]". Format ini dipilih karena arsitektur model yang digunakan menerima dua teks sebagai input terpisah. Tokenisasi dilakukan secara terpisah untuk masing-masing teks menggunakan AutoTokenizer dari model Sentence BERT. Parameter tokenisasi yang digunakan serupa dengan pendekatan sebelumnya, yaitu *max_length* sebesar 512, serta menerapkan *truncation* dan *padding* sesuai kebutuhan.

Transformasi selanjutnya adalah proses encoding terhadap token-token hasil tokenisasi. Berbeda dengan pendekatan sebelumnya yang menghasilkan tiga komponen input, pada pendekatan ini encoding hanya menghasilkan dua komponen input numerik, yaitu *input_ids* dan *attention_mask*. Kedua komponen ini memiliki fungsi yang sama seperti pada pendekatan sebelumnya, dimana *input_ids* merepresentasikan indeks token dalam *vocabulary*, dan *attention_mask* menandai token asli dan token hasil padding. Pendekatan ini tidak menggunakan komponen *token_type_ids*, karena jawaban siswa dan jawaban referensi tidak digabungkan dalam satu input teks, tetapi diproses secara terpisah. Oleh karena itu, informasi mengenai perbedaan antar segmen tidak diperlukan dalam struktur input model.

3.1.5 Pengembangan Model

Tahap ini merupakan bagian inti dari penelitian, di mana peneliti merancang, melatih, dan menguji dua pendekatan untuk tugas ASAS, yaitu *direct scoring* dan *similarity-based scoring*. Kedua pendekatan ini menggunakan arsitektur dasar berbasis model Transformer, namun memiliki perbedaan signifikan dalam alur pemrosesan input dan mekanisme prediksi skor.

Pada tahap perancangan, peneliti merancang arsitektur untuk masing-masing pendekatan. Setelah perancangan selesai, dilakukan proses pelatihan model menggunakan data pelatihan dan validasi yang berasal dari Rahutomo Dataset. Proses ini menghasilkan model dengan bobot parameter terbaik untuk masing-masing pendekatan berdasarkan performa validasi. Selanjutnya, model yang telah dilatih dievaluasi menggunakan data pengujian untuk mengukur kemampuan generalisasi terhadap data yang belum pernah dilihat sebelumnya. Penjelasan lebih rinci mengenai arsitektur dan alur kerja masing-masing pendekatan disampaikan pada bagian selanjutnya.

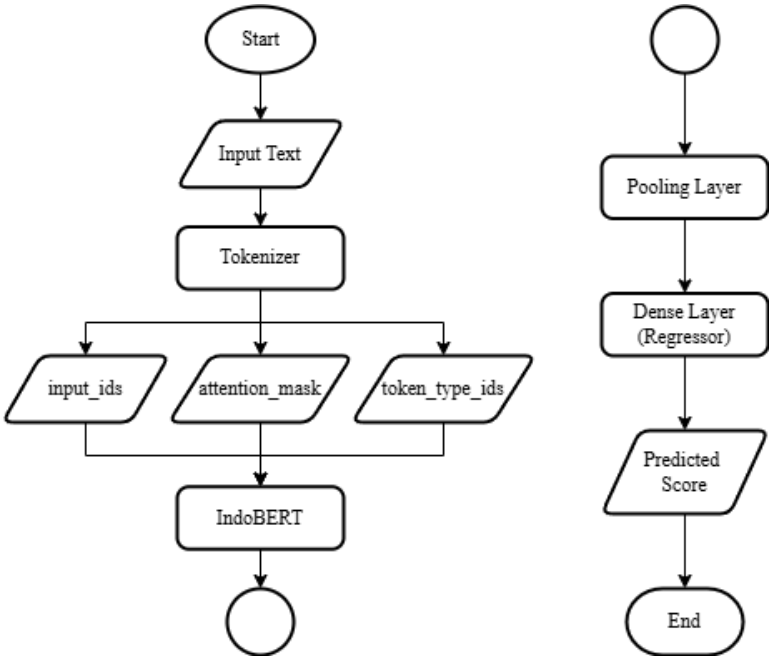
3.1.5.1 *Direct scoring*

Pada pendekatan *direct scoring*, model yang digunakan adalah IndoBERT dengan varian *indobert-lite-base-p2*, yaitu model BERT yang telah dipra-latih secara khusus pada korpus berbahasa Indonesia. Hal ini memungkinkan model untuk memahami konteks semantik dalam Bahasa Indonesia dengan lebih baik. Selain IndoBERT, agar perbandingan dengan varian model Sentence-BERT yang merupakan multilingual model. Oleh karena itu, pendekatan *direct scoring* menggunakan model tambahan yaitu multilingual BERT (mBERT) yang juga multilingual model, dengan Bahasa Indonesia yang sudah termasuk ke dalamnya. Varian mBERT yang digunakan adalah *google-bert/bert-base-multilingual-cased*. Detail konfigurasi dari kedua model yang digunakan dapat dilihat pada Tabel 3.2 berikut.

Tabel 3. 2 Detail Konfigurasi Model Pendekatan Direct Scoring

Model	*# Params*	*# Layers*	*# Heads*	*Emb. Size*	*Hidden Size*	*FFN Size*
IndoBERT	11.7M	12	12	128	768	3072
mBERT	177.8M	12	12	768	768	3072

Input teks yang telah diformat dan melalui proses tokenisasi serta encoding sebagaimana dijelaskan pada Sub Bab 3.1.4.4, akan menghasilkan tiga komponen fitur, yaitu *input_ids*, *attention_mask*, dan *token_type_ids*. Ketiga komponen ini diproses oleh model IndoBERT. Output embedding dari model kemudian diproses oleh *layer pooling* untuk menghasilkan embedding dari keseluruhan input yang kemudian diteruskan ke *dense layer* (regressor) untuk memetakan representasi vektor menjadi skor prediksi dalam rentang kontinu [0, 1]. Pendekatan ini memungkinkan pelatihan secara *end-to-end*, di mana proses ekstraksi fitur oleh *encoder* dan prediktif oleh *regressor* dilakukan secara bersamaan dalam satu arsitektur. Arsitektur lengkap dari pendekatan *direct scoring* ditampilkan pada Gambar 3.7.



Gambar 3. 7 Arsitektur Model Pendekatan *Direct Scoring*

3.1.5.2 Similarity-Based Scoring

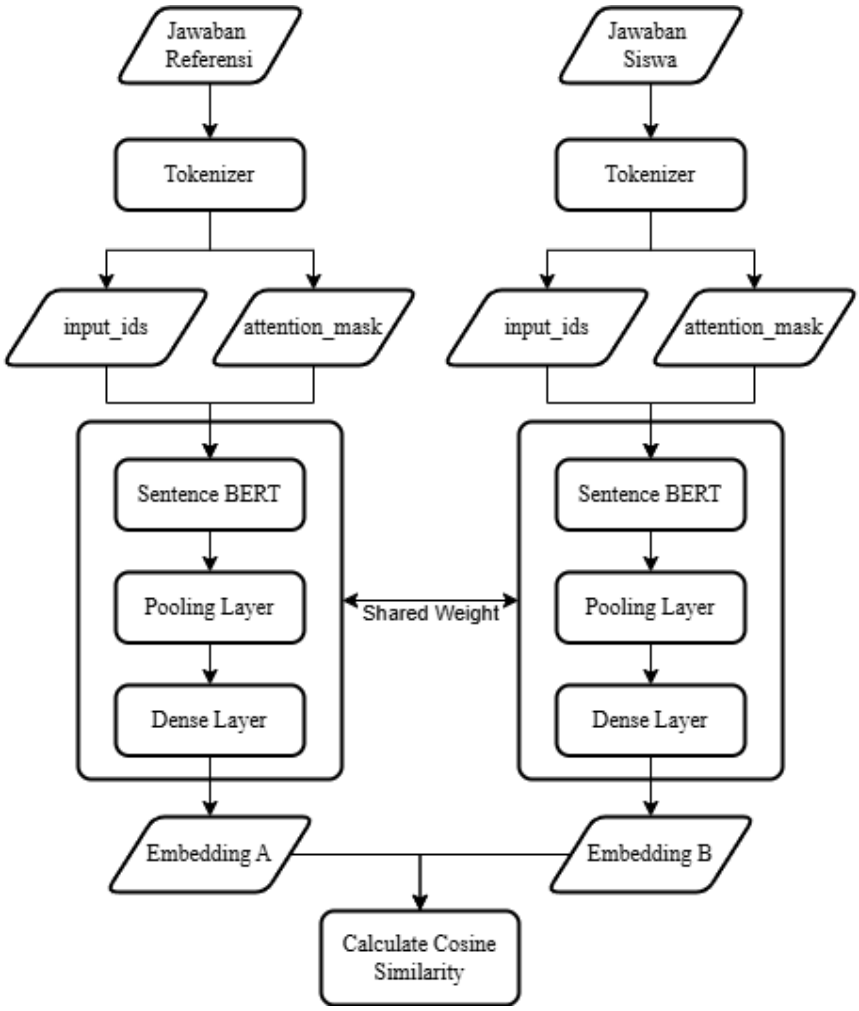
Pada pendekatan *similarity-based scoring*, model yang digunakan adalah Sentence-BERT dengan varian *sentence-transformer/distiluse-base-multilingual-cased-v2*. Model ini merupakan hasil modifikasi dari BERT yang dioptimalkan untuk menghasilkan embedding semantik pada level kalimat, sehingga cocok untuk tugas yang melibatkan perbandingan antar teks. Detail konfigurasi model Sentence-BERT dapat dilihat pada Tabel 3.3 berikut.

Tabel 3. 3 Detail Konfigurasi Model Baseline Sentence BERT

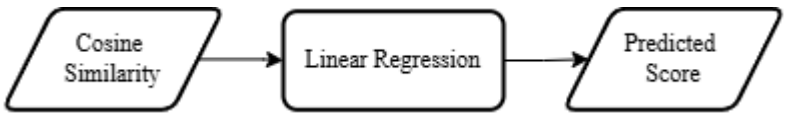
# Params	*# Layers*	*# Heads*	*Emb. Size*	*Hidden Size*	*FFN Size*
135 M	6	12	768	768	3072

Input terdiri dari dua teks terpisah, yaitu jawaban referensi dan jawaban siswa. Kedua input diproses secara paralel menggunakan *encoder* yang sama (*shared weights*), menghasilkan dua vektor representasi tetap melalui *layer pooling*. Embedding dari kedua teks kemudian dibandingkan menggunakan *cosine similarity*, yang merepresentasikan tingkat kemiripan semantik antara jawaban siswa dan referensi. Arsitektur Model untuk tahapan ini ditampilkan pada Gambar 3.8. Nilai kesamaan ini digunakan sebagai fitur dalam model regresi linear untuk

memprediksi skor akhir, seperti yang ditampilkan pada Gambar 3.9. Proses pelatihan terdiri dari dua tahap, yaitu fine-tuning model Sentence-BERT terhadap skor manual dan pelatihan model regresi linear menggunakan fitur *cosine similarity*.



Gambar 3. 8 Arsitektur *Encoder* Sentence-BERT pada Pendekatan *Similarity-Based Scoring*



Gambar 3. 9 Model Prediksi Skor Menggunakan *Cosine Similarity* dan Regresi Linear

3.1.5.3 Eksperimen Konfigurasi Model

Pada tahap ini, dilakukan eksperimen terhadap beberapa komponen penting dalam pipeline model untuk memperoleh konfigurasi optimal dari masing-masing pendekatan. Pada pendekatan *direct scoring*, eksperimen dilakukan terhadap

kombinasi tiga variabel utama, yaitu augmentasi digunakan untuk melihat pengaruh data dengan dan tanpa augmentasi, strategi pooling (CLS, Mean, Max, Attention Pooling) digunakan untuk melihat metode ekstraksi representasi yang paling relevan, dan nilai dropout (0.1, 0.3, 0.5) digunakan untuk melihat dampak regularisasi. Sementara itu, pada pendekatan *similarity-based scoring*, eksperimen dilakukan terhadap kombinasi dua variabel utama, yaitu pengaruh augmentasi dan strategi pooling.

3.1.6 Analisis dan Kesimpulan

Pada tahap ini, dilakukan analisis terhadap performa dua pendekatan, yaitu *direct scoring* dan *similarity-based scoring*, pada dua skenario pembagian data, yaitu *specific-prompt* dan *cross-prompt*. Tujuan analisis ini adalah untuk mengevaluasi efektivitas masing-masing pendekatan dalam memprediksi skor pada data pengujian.

Sebelum dilakukan perbandingan antar pendekatan pada masing-masing skenario, terlebih dahulu dilakukan analisis terhadap eksperimen internal untuk menentukan konfigurasi model terbaik untuk setiap kombinasi pendekatan dan skenario. Setelah mendapatkan konfigurasi model terbaik, model tersebut digunakan untuk evaluasi antar pendekatan pada setiap skenario.

Evaluasi dilakukan dengan beberapa metode, baik secara kuantitatif maupun lanjutan. Secara kuantitatif, digunakan dua metrik evaluasi, yaitu *Root Mean Squared Error* (RMSE) untuk mengukur tingkat kesalahan absolut dan *Pearson Correlation* untuk menilai hubungan linier antara skor prediksi dan skor referensi. Penjelasan lengkap mengenai metrik ini dapat dilihat pada Sub Bab 2.11.

Selain itu, dilakukan analisis residual untuk mengidentifikasi outlier berdasarkan ambang batas tetap. Penjelasan lengkap mengenai analisis residual dan outlier dapat dilihat pada Sub Bab 2.12. Ambang batas tetap digunakan dalam penelitian ini untuk memastikan proses identifikasi outlier konsisten dan adil ketika membandingkan kedua pendekatan. Penggunaan ambang batas yang sama penting agar tidak terjadi bias karena perbedaan distribusi residual masing-masing model. Nilai ambang batas ditentukan berdasarkan analisis awal terhadap distribusi residual dari kedua pendekatan menggunakan metode interquartile range (IQR). Hasil analisis menunjukkan bahwa nilai Q3 dari residual sebagian besar berada di

sekitar 0.1, sementara upper bound IQR berada di sekitar 0.2. Oleh karena itu, dua nilai threshold tetap yang digunakan dalam evaluasi adalah 0.1 dan 0.2.

Evaluasi lanjutan mencakup analisis pada tingkat soal dengan mengamati nilai RMSE dan jumlah outlier per soal, serta analisis kualitatif dengan membandingkan hasil prediksi dari kedua pendekatan untuk data atau soal tertentu. Hasil analisis ini digunakan untuk menarik kesimpulan terkait efektivitas masing-masing pendekatan pada setiap skenario. Dengan melakukan analisis ini, diharapkan dapat memberikan gambaran menyeluruh terkait performa model dari berbagai sisi, sekaligus memberikan rekomendasi untuk penelitian selanjutnya.