BAB I
PENDAHULUAN

1.1 Latar Belakang

Dalam era pendidikan digital, kebutuhan akan sistem penilaian otomatis semakin meningkat, terutama untuk menilai jawaban teks pendek dalam skala besar secara efisien, objektif, dan konsisten. *Automatic Short Answer Scoring* (ASAS) bertujuan untuk mengotomasi proses penilaian tersebut secara andal. Pendekatan yang umum digunakan adalah dengan membandingkan jawaban siswa dengan jawaban referensi yang telah ditentukan sebelumnya. Namun, keragaman struktur, kosakata, dan gaya bahasa dalam jawaban siswa menyebabkan proses pembandingan semantik menjadi kompleks (Chamidah et al., 2023).

Untuk mengatasi tantangan tersebut, berbagai teknik *Natural Language Processing* (NLP) telah diterapkan guna mengekstrak fitur linguistik dari teks jawaban, yang selanjutnya digunakan sebagai input dalam model *machine learning*. Seiring perkembangan teknologi NLP, peneliti kini banyak melakukan *fine-tuning* pada model pra-latih dengan arsitektur transformer, seperti BERT (Mahmood & Abdulsamad, 2024). Model ini mampu menghasilkan representasi vektor yang kontekstual dan bermakna, bahkan ketika teks memiliki variasi dalam struktur, kosakata, dan gaya bahasa.

Dalam pengembangan model penilaian otomatis untuk jawaban singkat, terdapat dua pendekatan yang umum digunakan, yaitu *direct scoring* dan *similarity-based scoring*. Pendekatan *direct scoring* menggunakan representasi vektor mentah dari model secara langsung untuk memprediksi skor, dengan asumsi bahwa vektor tersebut mengandung fitur linguistik yang relevan untuk tugas prediksi. Beberapa studi, seperti (Salim et al., 2022), telah menguji performa berbagai varian IndoBERT untuk memprediksi skor jawaban pada dataset berbahasa Indonesia. Sementara itu, studi lain seperti (Kaya & Cicekli, 2024) menggabungkan beberapa arsitektur, seperti BERT, LSTM, dan CNN, untuk meningkatkan kualitas representasi fitur.

Di sisi lain, pendekatan *similarity-based scoring* menggunakan nilai kesamaan semantik antara representasi vektor jawaban siswa dan jawaban referensi untuk memprediksi skor, dengan asumsi bahwa tingkat kemiripan semantik

merupakan indikator utama dalam penilaian oleh model penilaian. Pendekatan ini banyak mengandalkan arsitektur siamese seperti Siamese LSTM atau Sentence-BERT (SBERT), yang memungkinkan pemrosesan dua input teks secara paralel. Pendekatan ini banyak digunakan dalam penelitian ASAS karena sesuai dengan karakteristik ASAS yang melibatkan perbandingan dua teks.

Beberapa studi telah mengimplementasikan pendekatan ini dengan variasi arsitektur. Penelitian (Haidir & Purwarianti, 2020) menggunakan model SBERT untuk ekstraksi fitur dan *logistic regression* untuk model prediktif, dengan fitur utama berupa *cosine similarity* serta beberapa kombinasi vektor. Penelitian (Chamidah et al., 2023) menggunakan model SBERT untuk ekstraksi fitur dan LSTM untuk model prediktif, dengan fitur utama berupa *manhattan distance*. Namun, belum ditemukan penelitian yang secara eksplisit membandingkan kedua pendekatan tersebut dalam lingkungan eksperimen yang sama, seperti penggunaan dataset dan pembagian data yang seragam. Padahal, perbandingan ini penting untuk memperoleh pemahaman yang lebih objektif dan menyeluruh terkait kelebihan dan keterbatasan masing-masing pendekatan.

Selain pendekatan, skenario pembagian dataset juga penting dalam penelitian ASAS. Umumnya, dataset terdiri dari sejumlah pertanyaan, dimana setiap pertanyaan disertai satu jawaban referensi dan diikuti oleh banyak jawaban siswa. Berdasarkan tinjauan literatur, terdapat dua skenario dalam pembagian dataset, yaitu *specific-prompt* dan *cross-prompt*. Skenario *specific-prompt* membagi data berdasarkan jawaban siswa untuk setiap pertanyaan. Skenario ini mengevaluasi kemampuan model dalam menilai jawaban siswa yang baru dari pertanyaan yang telah dilatih sebelumnya (*unseen student answers*). Pendekatan ini banyak digunakan dalam penelitian sebelumnya, seperti pada (Salim et al., 2022).

Sebaliknya, skenario *cross-prompt* membagi data berdasarkan pertanyaan sehingga model dievaluasi untuk kemampuannya dalam menilai jawaban siswa dari pertanyaan yang belum pernah dilatih sebelumnya (*unseen questions*). Skenario ini menuntut model untuk memiliki kemampuan generalisasi yang lebih tinggi. Namun, skenario ini jarang digunakan dan kurang dieksplor, penelitian yang menggunakan skenario ini adalah (Haidir & Purwarianti, 2020).

Penelitian terbaru oleh (Ferreira Mello et al., 2025), mulai menerapkan kedua skenario pembagian dataset ini dengan mengeksplorasi berbagai teknik *prompt engineering* pada model GPT-4. Namun, belum ada penelitian berbasis BERT yang membandingkan performa model pada kedua skenario ini dalam lingkungan eksperimen yang seragam. Kebanyakan penelitian yang menggunakan BERT hanya fokus pada salah satu skenario. Padahal, evaluasi pada kedua skenario tersebut penting untuk memberikan gambaran menyeluruh terkait performa model dalam kondisi yang berbeda.

Oleh karena itu, penelitian ini bertujuan untuk membandingkan pendekatan *direct scoring* dan *similarity-based scoring* pada skenario *specific-prompt* dan *cross-prompt*, guna memberikan pemahaman yang lebih komprehensif terkait kemampuan masing-masing pendekatan dalam dua kondisi evaluasi yang berbeda. Dalam melakukan analisis, penelitian ini tidak hanya menggunakan metrik evaluasi seperti RMSE dan Pearson Correlation, tetapi juga mengevaluasi stabilitas model melalui analisis residual untuk mengidentifikasi outlier, serta performa model pada level soal, yang masih minim dieksplorasi pada penelitian sebelumnya. Dengan analisis yang sistematis dan mendalam, penelitian ini berkontribusi dalam memberikan wawasan terkait efektivitas kedua pendekatan pada dua skenario berbeda berdasarkan metrik evaluasi, stabilitas prediksi, dan performa pada level soal. Hasil dari penelitian ini diharapkan dapat menjadi dasar bagi penelitian lanjutan yang ingin mengoptimalkan pendekatan yang paling efektif, serta bagi pengembang sistem yang ingin mengintegrasikan model ASAS ke platform pembelajaran.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijabarkan, berikut beberapa rumusan masalah yang akan dijawab dalam penelitian ini.

1. Apa konfigurasi terbaik untuk masing-masing pendekatan (*direct* dan *similarity-based*) pada skenario *specific-prompt* dan *cross-prompt* berdasarkan metrik evaluasi dan jumlah outlier?
2. Bagaimana perbandingan performa antara pendekatan *direct scoring* dan *similarity-based scoring* pada skenario *specific-prompt* dan *cross-prompt* ditinjau dari metrik evaluasi, jumlah outlier, serta performa pada level soal?

1.3 Tujuan Penelitian

Berdasarkan perumusan masalah sebelumnya, tujuan dari penelitian ini dapat dirangkum dalam poin-poin berikut ini.

1. Menentukan konfigurasi terbaik untuk masing-masing pendekatan, yaitu *direct scoring* dan *similarity-based scoring*, pada skenario *specific-prompt* dan *cross-prompt* berdasarkan metrik evaluasi dan jumlah outlier.
2. Membandingkan performa antara pendekatan *direct scoring* dan *similarity-based scoring* pada skenario *specific-prompt* dan *cross-prompt* dengan mempertimbangkan metrik evaluasi, jumlah outlier, serta performa pada level soal.

1.4 Manfaat Penelitian

Berdasarkan tujuan yang telah dijabarkan, penelitian ini diharapkan memberikan berbagai manfaat yang dapat dirinci sebagai berikut.

1. Bagi peneliti

Dapat memberikan pemahaman secara mendalam yang didukung oleh data mengenai akurasi dan stabilitas model antara pendekatan *direct scoring* dan *similarity-based scoring* pada skenario *specific-prompt* dan *cross-prompt* untuk penelitian selanjutnya.

2. Bagi praktisi

Melalui analisis performa model, penelitian ini diharapkan dapat memberikan rekomendasi praktis mengenai pendekatan yang paling efisien pada masing-masing skenario pembagian data untuk keperluan integrasi model ke dalam sistem.

1.5 Batasan Penelitian

Penelitian ini memiliki beberapa batasan untuk memfokuskan ruang lingkup penelitian, yaitu sebagai berikut.

1. Penelitian ini hanya menggunakan set data Rahutomo yang berbahasa Indonesia, sehingga hasil penelitian terbatas pada konteks bahasa Indonesia dan mungkin tidak dapat digeneralisasi untuk bahasa lain.
2. Model pra-latih yang digunakan pada penelitian ini terbatas pada *encoder-only* model, seperti BERT dan variannya.

3. Fokus penelitian adalah penilaian jawaban singkat yang memiliki kunci jawaban dan tidak mencakup penilaian esai panjang.
4. Metrik evaluasi model hanya menggunakan *Pearson Correlation* dan *Root Mean Squared Error* (RMSE)

1.6 Sistematika Penulisan

Sistematika penulisan dalam penelitian ini disusun secara terstruktur dengan tujuan memberikan gambaran yang komprehensif dan jelas mengenai isi dan alur penelitian. Sistematika ini dirancang untuk memastikan bahwa setiap aspek dari penelitian dapat dipahami dengan mudah oleh pembaca dan memberikan informasi yang diperlukan untuk mendalami topik penelitian. Berikut adalah sistematika penulisan yang diterapkan dalam penelitian ini.

BAB I. PENDAHULUAN

Bab ini membahas hal yang melatarbelakangi penelitian mengenai penilaian uraian otomatis dan manfaatnya, penelitian sebelumnya, dan celah penelitian pada bagian latar belakang. Berdasarkan latar belakang ini, dirumuskan beberapa masalah yang akan dijawab melalui penelitian pada bagian rumusan masalah. Menjelaskan apa saja yang ingin dicapai dari penelitian pada bagian tujuan penelitian. Memberikan kontribusi bagi berbagai pihak dari penelitian pada bagian manfaat penelitian. Mengidentifikasi ruang lingkup dan batasan yang ditetapkan pada penelitian untuk memfokuskan penelitian pada bagian batasan penelitian. Memberikan panduan mengenai isi dokumen untuk memudahkan pembaca dalam mengikuti alur penelitian pada bagian sistematika penulisan.

BAB II. KAJIAN PUSTAKA

Bab ini menyediakan tinjauan pustaka yang berisi penelitian terdahulu dan landasan teori dari topik penelitian. Landasan teori ini mencakup teori terkait penilaian uraian secara umum, jenisnya, tantangannya, solusinya melalui penilaian esai otomatis, metode yang digunakan yaitu NLP, machine learning, deep learning, arsitektur dan model yang digunakan. Fokus utama kajian pustaka adalah memberikan pemahaman mendalam terkait teori dan teknologi yang mendasari penelitian ini serta mendukung pengembangan konsep dan metodologi penelitian.

BAB III. METODOLOGI PENELITIAN

Bab ini menguraikan secara rinci rencana dan alur penelitian untuk mengembangkan dan mengevaluasi model penilaian esai otomatis. Penelitian dimulai dari perumusan masalah, tinjauan pustaka, pengumpulan data, praproses data, pengembangan model, serta analisis dan kesimpulan.

BAB IV. HASIL DAN PEMBAHASAN

Bab ini menyediakan hasil dan pembahasan dari awal hingga akhir eksperimen. Terdiri dari pengolahan data, skenario pemodelan, hasil dan pembahasan setiap eksperimen secara detail.

BAB V. PENUTUP

Bab ini menyajikan kesimpulan penelitian yang berisi jawaban dari pertanyaan pada rumusan masalah dan mengukur pencapaian tujuan penelitian. Selain itu, diberikan juga rekomendasi dan saran berupa langkah lanjutan yang dapat digunakan untuk memperbaiki atau memperluas penelitian di masa depan.