

BAB I

PENDAHULUAN

Bab ini membahas mengenai latar belakang masalah, rumusan masalah, tujuan dan manfaat, ruang lingkup, serta sistematika penulisan pada skripsi berjudul *Automated Short Answer Grading Berbahasa Indonesia Menggunakan Matching Matrix dan Bidirectional Long-Short Term Memory (BiLSTM)*.

1.1 Latar Belakang

Tes atau ujian merupakan bagian penting dalam pendidikan pada jenjang dan instansi apa pun. Metode ini umum digunakan untuk mengevaluasi peserta didik dan mengukur ketercapaian target pembelajaran. Bentuk soal ujian dapat beraneka ragam, seperti pilihan ganda, benar-salah, esai, dan uraian jawaban singkat. Setiap bentuk soal memiliki kelebihan dan kekurangan masing-masing dalam menilai aspek tertentu.

Uraian jawaban singkat dinilai tepat untuk mengevaluasi kognitif (Salim dkk., 2022) serta menguji pemahaman dan pengetahuan (Lubis dkk., 2021; Wilianto & Girsang, 2023). Bentuk ini mengharuskan peserta didik untuk merangkai konsep dan menuliskannya sebagai jawaban dengan kalimat sendiri. Kemampuan mengingat, merancang, dan mengekspresikan dipadukan selama proses pengerjaan bentuk soal ini.

Namun, penilaian uraian jawaban singkat memiliki kesulitan, terutama dalam penilaian manual (Lubis dkk., 2021). Karena jawaban yang bervariasi, penilaian akan memakan waktu lama (Lubis dkk., 2021; Wilianto & Girsang, 2023), apalagi jika jumlah peserta didik atau mata ujian sangat banyak. Selain itu, masalah bias (Patil & Adhiya, 2023) dan inkonsistensi (Lubis dkk., 2021; Wilianto & Girsang, 2023) juga dapat muncul karena subjektivitas dan penurunan performa manusia. Untuk mengatasi masalah ini, penilaian jawaban singkat otomatis atau *automated short answer grading* (ASAG) dapat diterapkan sehingga waktu penilaian berkurang dan nilai ujian lebih akurat, objektif, dan konsisten.

ASAG merupakan tugas penilaian jawaban yang berupa satu frase hingga satu paragraf dengan teknik komputasi (Bonthu dkk., 2023). Tugas ini dapat dipandang sebagai regresi jika keluaran berupa nilai ataupun klasifikasi jika keluaran berupa kategori kognitif, seperti benar dan salah (Zhu dkk., 2022). Seperti penilaian jawaban singkat manual, ASAG berfokus pada semantik. Ketepatan dinilai berdasarkan makna, bukan berdasarkan tata bahasa atau gaya penulisan.

Pada ASAG, semantik yang digunakan untuk menilai tersebut tidak dapat hanya bersumber dari jawaban. Penilaian yang terbatas pada semantik jawaban, meskipun disertai dengan jawaban benar, tidak efektif menangani variasi representasi konseptual dalam konteks (Chen & Li, 2021). Hal ini dapat diatasi dengan penggunaan semantik soal. Integrasi soal dapat membantu ASAG memperoleh *prior knowledge* (Chen & Li, 2021), mengenali jawaban benar yang tidak biasa (Lv dkk., 2021), dan meningkatkan kualitas penilaian (Condor dkk., 2021).

Penelitian ASAG untuk bahasa Indonesia belum banyak dilakukan. Performa modelnya pun belum se bagus yang menggunakan data berbahasa Inggris. Rata-rata *error* absolut yang dihasilkan berada di antara 0,2–1,4 (Chamidah dkk., 2022; Haidir & Purwarianti, 2020; Lubis dkk., 2021; Salim dkk., 2022). Sebagai pembandingan, model ASAG untuk bahasa Inggris memiliki kisaran rata-rata *error* absolut 0,06–0,2 (Chen dkk., 2023; Patil & Adhiya, 2023; Wilianto & Girsang, 2023; Zhu dkk., 2022). Kondisi ini merupakan peluang untuk pengembangan ASAG berbahasa Indonesia.

Performa model ASAG dipengaruhi oleh metode yang digunakan pada setiap komponennya. Secara umum, terdapat tiga komponen model ASAG, yaitu praproses, representasi teks, dan asesmen jawaban (Haidir & Purwarianti, 2020). Metode pada komponen praproses tidak berpengaruh signifikan karena tidak berhubungan langsung dengan kalkulasi numerik. Untuk representasi teks, metode *embedding* atau mengubah teks menjadi vektor sering diterapkan belakangan ini. Metode tersebut melingkupi *word embedding* (Chen dkk., 2023; Lubis dkk., 2021; Patil & Adhiya, 2023), *sentence embedding* (Bonthu dkk., 2023; Chamidah dkk., 2022; Wilianto & Girsang, 2023), dan *contextual embedding* (Gomaa dkk., 2023; Salim dkk., 2022; Zhu dkk., 2022).

Word embedding mempunyai keterbatasan dalam menangkap makna kata yang sangat tergantung pada konteks (Lubis dkk., 2021). *Contextual embedding* mengatasi ini dengan kemampuan pemahaman kontekstual berkat model *transformer* (Gomaa dkk., 2023). Namun, pada dasarnya, hasil *contextual embedding* tetap representasi kata. Tidak ada representasi kalimat independen yang langsung dihitung (Reimers & Gurevych, 2019). Hal ini membuat penghitungan kesamaan kalimat sulit dilakukan.

Di lain sisi, *sentence embedding* menawarkan representasi kalimat secara utuh berukuran tetap. Pendekatan metode *sentence embedding* yang populer saat ini adalah penggunaan model Siamese-BERT. Model ini memiliki pemahaman konteks seperti pada *contextual embedding* karena dibangun dari arsitektur yang memuat *transformer* dan dilatih dengan jaringan *siamese* (Reimers & Gurevych, 2019). Model tersebut juga masih dapat ditingkatkan melalui *fine-tuning* atau pelatihan ulang model dengan data baru. Proses ini akan memperbarui perbendaharaan kata yang dipelajari oleh model.

Untuk asesmen jawaban, metode yang telah diterapkan antara lain BiLSTM (Gomaa dkk., 2023; Patil & Adhiya, 2023; Zhu dkk., 2022), *linear regression* (Haidir & Purwarianti, 2020; Salim dkk., 2022), dan *cosine similarity* (Bonthu dkk., 2023; Wilianto & Girsang, 2023). Metode *cosine similarity* dan *linear regression* memiliki kelemahan berupa tidak mampu menangkap ketergantungan berurutan dalam vektor representasi untuk menilai koherensi. Sebaliknya, BiLSTM mampu melakukan hal tersebut sehingga meningkatkan akurasi dan interpretasi jawaban (Gomaa dkk., 2023).

Namun, penggunaan BiLSTM dengan input model representasi berurutan dari dua teks tetap belum maksimal dalam menangkap informasi. Kedua representasi ini masih bersifat saling lepas, tanpa saling bertukar informasi. Ketiadaan interaksi yang jelas antara dua teks dalam proses pemetaan representasi akan menyebabkan banyak informasi tentang satu sama lain yang hilang (Feifei dkk., 2020).

Pertukaran informasi dua representasi teks dapat dilakukan dengan penambahan interaksi segera setelah representasi teks didapatkan. Model dengan arsitektur seperti

ini disebut sebagai *interaction model*. Pendekatan ini lebih dahulu sering digunakan untuk tugas seperti *text matching* dan *semantic text similarity* daripada tugas ASAG.

Terdapat berbagai metode untuk menghadirkan interaksi, salah satunya melalui pembentukan *matching matrix*. Ide dasar *matching matrix* adalah membuat matriks kecocokan, lalu mengekstrak fitur abstrak untuk menilai kecocokan kedua representasi teks. Elemen matriks ini dapat dihasilkan dengan mekanisme atensi, *cosine similarity*, *dot product*, atau *indicator function* (Huang & Cao, 2021). Penambahan interaksi akan membuat kedua teks menyintesis informasi yang cocok dan menangkap fitur interaktif sehingga penilaian kesamaan teks meningkat (Huang & Cao, 2021).

1.2 Rumusan Masalah

Berdasarkan latar belakang, masalah umum yang dirumuskan untuk penelitian ini adalah bagaimana penerapan *Matching Matrix* dan *Bidirectional Long Short-Term Memory* (BiLSTM) untuk tugas *Automated Short Answer Grading* (ASAG) pada uraian jawaban singkat berbahasa Indonesia. Masalah khusus penelitian ini antara lain:

1. Bagaimana kombinasi *Matching Matrix* dan BiLSTM dalam arsitektur model ASAG dengan kinerja terbaik pada uraian jawaban singkat berbahasa Indonesia?
2. Bagaimana pengaruh BiLSTM terhadap kinerja model ASAG pada uraian jawaban singkat berbahasa Indonesia?
3. Bagaimana pengaruh *Matching Matrix* terhadap kinerja model ASAG pada uraian jawaban singkat berbahasa Indonesia?

1.3 Tujuan dan Manfaat

Tujuan umum penelitian ini adalah menerapkan *Bidirectional Long Short-Term Memory* (BiLSTM) dan *Matching Matrix* untuk tugas *Automated Short Answer Grading* (ASAG) pada uraian jawaban singkat berbahasa Indonesia. Tujuan khusus penelitian ini antara lain:

1. Menemukan kombinasi *Matching Matrix* dan BiLSTM dalam arsitektur model ASAG dengan kinerja terbaik pada uraian jawaban singkat berbahasa Indonesia

2. Mengetahui pengaruh *Matching Matrix* terhadap kinerja model ASAG pada uraian jawaban singkat berbahasa Indonesia.
3. Mengetahui pengaruh BiLSTM terhadap kinerja model ASAG pada uraian jawaban singkat berbahasa Indonesia.

Manfaat penelitian ini adalah penilaian ujian uraian jawaban singkat berbahasa Indonesia dapat dilakukan tanpa memakan banyak waktu dengan hasil nilai ujian yang lebih akurat, objektif, dan konsisten.

1.4 Ruang Lingkup

Ruang lingkup dalam menerapkan *Bidirectional Long Short-Term Memory* (BiLSTM) dan *Matching Matrix* untuk tugas *automated short answer grading* pada penelitian ini adalah sebagai berikut:

1. Terdapat empat set data dengan domain dan jenis soal masing-masing yang akan diproses secara terpisah pada penelitian ini.
2. Setiap set data akan menghasilkan model terbaiknya masing-masing.
3. Terdapat dua bentuk input data model, yakni matriks dan vektor, yang dibentuk dari fitur soal dan jawaban serta satu input label model dari fitur rata-rata.
4. Penilaian pada ASAG bertipe regresi dengan keluaran model berupa nilai *float*.
5. Model *sentence embedding* Siamese-BERT yang digunakan ialah IndoSBERT yang di-*fine-tuning* secara terpisah pada keempat set data berbahasa Indonesia yang digunakan.

1.5 Sistematika Penulisan

Laporan penelitian skripsi ini disusun dalam sistematika terstruktur sehingga mudah dipahami dan dipelajari. Sistematika yang dimaksud adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini membahas mengenai latar belakang masalah, rumusan masalah, tujuan dan manfaat, ruang lingkup, serta sistematika penulisan skripsi tentang *automated short answer grading* (ASAG) menggunakan *matching matrix* dan BiLSTM untuk uraian jawaban singkat berbahasa Indonesia.

BAB II LANDASAN TEORI

Bab ini membahas mengenai *state of the art* dan teori-teori untuk landasan penelitian skripsi, meliputi *automated short answer grading* (ASAG), prapemrosesan data, *sentence embedding*, model IndoSBERT, metode *fine-tuning* SimCSE, *cosine similarity*, *matching matrix*, *Bidirectional Long Short-Term Memory* (BiLSTM), *dropout*, *Symmetric Mean Absolute Percentage Error* (SMAPE), dan *Mean Absolute Error* (MAE).

BAB III METODOLOGI PENELITIAN

Bab ini membahas mengenai metodologi penelitian skripsi, meliputi pengumpulan data, prapemrosesan data, *sentence embedding*, pembagian data, pengembangan model ASAG, dan pengujian model ASAG.

BAB IV HASIL DAN PEMBAHASAN

Bab ini membahas mengenai lingkungan implementasi, skema pengujian, serta hasil dan analisis untuk skenario 1 (arsitektur dengan *matching matrix* dan BiLSTM), skenario 2 (arsitektur tanpa *matching matrix* dengan BiLSTM), skenario 3 (arsitektur dengan *matching matrix* tanpa BiLSTM), skenario 4 (arsitektur tanpa *matching matrix* dan BiLSTM), skenario 5 (perbandingan kinerja model), dan skenario 6 (pengujian model terbaik).

BAB V KESIMPULAN DAN SARAN

Bab ini membahas mengenai kesimpulan dari uraian yang telah dijelaskan pada bab-bab sebelumnya dan saran pengembangan penelitian lebih lanjut.