BAB V
KESIMPULAN DAN SARAN

Hasil penelitian ini membuktikan bahwa penggabungan arsitektur IndoBERT, BiLSTM, dan Attention mampu meningkatkan akurasi sistem *Automated Essay Scoring* pada teks berbahasa Indonesia secara signifikan. Berdasarkan evaluasi empiris, model yang diusulkan berhasil mencapai metrik *Quadratic Weighted Kappa* (QWK) sebesar 0,8508 dengan tingkat ketepatan (*Exact Accuracy*) menyentuh 63,27%. Kinerja tersebut secara konsisten mengungguli arsitektur dasar (*baseline*) BiLSTM tunggal dan model IndoBERT murni pada seluruh skenario pengujian. Implementasi mekanisme penyeimbangan kelas melalui modifikasi Weighted Huber Loss terbukti krusial dalam meningkatkan presisi prediksi untuk kelompok minoritas hingga 18,37%. Selain itu, stabilitas kinerja arsitektur ini telah divalidasi secara komprehensif melalui rentang *Bootstrap Confidence Intervals* 95% yang sangat sempit dan terpusat.

Pencapaian performa tersebut sejalan dengan tujuan utama penelitian, yakni mengembangkan model evaluasi otomatis yang dapat mereplikasi standar penilaian manusia secara akurat. Penggunaan *pretrained encoder* terbukti efektif dalam mengekstraksi representasi semantik leksikal secara mendalam dari esai berbahasa Indonesia. Selanjutnya, penambahan lapisan sekuensial berupa BiLSTM memfasilitasi pemodelan transisi argumen antarparagraf secara kronologis guna mempertahankan konteks struktural wacana. Mekanisme Attention juga beroperasi secara dinamis dengan memberikan bobot evaluasi pada klausa penentu yang paling memengaruhi kualitas teks secara keseluruhan. Sinergi dari ketiga komponen struktural tersebut berhasil mewujudkan sistem yang selaras dengan hierarki kognitif dari seorang penilai ahli.

Dari perspektif teoretis, temuan ini memberikan kontribusi esensial bagi literatur pemrosesan bahasa alami melalui validasi efektivitas pendekatan fusi struktural pada bahasa dengan sumber daya menengah. Studi ini menetapkan standar arsitektur baru dalam pengembangan *Automated Essay Scoring* khusus untuk bahasa Indonesia yang memiliki kompleksitas morfologis tinggi. Secara praktis, model ini menawarkan kerangka kerja evaluasi pendidikan yang objektif

serta konsisten untuk meminimalkan bias subjektivitas dari *human rater*. Lebih lanjut, optimasi ambang batas berbasis *coordinate-descent* pada model tersebut memunculkan wawasan psikometrik baru yang mengindikasikan bahwa jarak kognitif di antara skala penilaian ordinal manusia pada rubrik esai tidak bersifat ekuidistan.

Meskipun menghasilkan performa agregat yang tangguh, arsitektur usulan ini masih memiliki keterbatasan analitis terkait sumber data dan variasi genre wacana. Penggunaan korpus hasil terjemahan mesin dari *dataset* "PERSUADE 2.0" berisiko menghilangkan sebagian nuansa linguistik dan struktur retorika autentik khas penulis berbahasa Indonesia. Model ini juga menunjukkan penurunan presisi yang tajam ketika mengevaluasi esai dengan genre naratif fiksi akibat keterbatasan dalam memproses alur implisit. Analisis terhadap distribusi residu prediktif turut mengungkap keberadaan anomali *regression-to-the-mean* pada pemodelan ruang laten kontinu. Kondisi tersebut memicu algoritma untuk secara sistematis memberikan skor yang terlalu optimis pada esai berkualitas sangat rendah dan menurunkan skor aktual pada tulisan bermutu tinggi.

Untuk mengatasi batasan sistemik tersebut, penelitian mendatang perlu melakukan pengujian dan pelatihan ulang model menggunakan korpus esai autentik yang ditulis langsung oleh penutur asli bahasa Indonesia. Langkah ini sangat krusial guna memvalidasi ketahanan arsitektur terhadap kompleksitas morfologi serta keaslian gaya bahasa siswa tanpa adanya distorsi dari proses terjemahan. Modifikasi arsitektur tingkat lanjut dapat difokuskan pada perancangan *loss function* deterministik untuk menetralkan efek regresi terhadap nilai tengah populasi. Penggunaan pendekatan alternatif, seperti klasifikasi ordinal murni atau penerapan metode Label-Distribution-Aware Margin, sangat direkomendasikan agar ketegasan batas keputusan pada kelas skor ekstrem tetap terpelihara. Di samping itu, eksperimen lanjutan perlu mengkaji integrasi fitur stilistika guna mendongkrak sensitivitas model terhadap wacana bercorak penceritaan naratif.