Pembangunan Model Baseline Readability dari Buku Pelajaran Bahasa Indonesia Sekolah Dasar

Abstrak

Bahan bacaan adalah hal yang cukup penting bagi guru dalam proses untuk meningkatkan kemampuan siswa dalam memahami materi. Readability atau keterbacaan merupakan isu penting bagi praktisi dan akademisi di berbagai bidang dan minat, termasuk pendidikan, linguistik terapan, linguistik teks, ilmu perpustakaan, bisnis, kedokteran, dan komunikasi teknis. Terdapat beberapa penelitian sebelumnya yang melakukan penelitian terkait readability, readability model, maupun metode terkait, seperti pada artikel yang meneliti model keterbacaan dari bahasa cebuano yang menghasilkan metode gabungan dari metode Traditional dan pola suku kata (Syllables Pattern) sebagai yang terbaik. Penelitian ini membangun model readability dari buku pelajaran bahasa indonesia sekolah dasar menggunakan beberapa ekstraksi fitur, seperti prediktor tradisional atau prediktor surface-based, syllable pattern atau pola suku kata, dan IndoBERT dengan algoritma Random Forest dan Support Vector Machine, kemudian dilakukan hyperparameter tuning menggunakan Grid SearchCV. Hasil penelitian menunjukkan representasi saraf IndoBERT mencapai akurasi maksimum 72,1% saat menggunakan Support Vector Machine. Untuk mendukung penelitian selanjutnya, dataset dan kode penelitian ini telah dijadikan open-source.

Kata kunci: readability, surface-based, syllable patterns, Random Forest, Support Vector Machine.

1. Pendahuluan

Latar Belakang

Bahan bacaan adalah hal yang cukup penting bagi guru dalam proses untuk meningkatkan kemampuan siswa dalam memahami materi[1]. Readability atau keterbacaan merupakan isu penting bagi praktisi dan akademisi di berbagai bidang dan minat, termasuk pendidikan, linguistik terapan, linguistik teks, ilmu perpustakaan, bisnis, kedokteran, dan komunikasi teknis. Pada tingkat praktis, kriteria keterbacaan termasuk memilih bahan bacaan yang sesuai, mengkomunikasikan informasi teknis, medis, dan bisnis secara efektif kepada para profesional dan orang awam, mempersiapkan tes standar, menulis dan hal ini diperlukan untuk berbagai tugas, seperti mengajar keterampilan komunikasi. Secara teoritis tingkat, keterbacaan relevan dengan bidang-bidang seperti linguistik terapan, teori teks dan wacana, dan pemrosesan Bahasa alami [2].

Keterbacaan adalah kemudahan memahami atau pemahaman teks. Keterbacaan adalah seberapa mudah materi tertulis dapat dibaca dan dipahami. Namun, prediksi keterbacaan teks tidak selalu yang paling akurat dan tidak selalu mencerminkan kemudahan atau kesulitan teks. Dua faktor yang sering diteliti dengan cermat adalah kosa kata dan kalimat dalam teks [2].

Terdapat beberapa penelitian tentang readability, Penelitian terkait meneliti berbagai fitur linguistik mulai dari prediktor tradisional atau prediktor surface-based, fitur orthography-based menggunakan syllable pattern atau pola suku kata, dan representasi neural untuk mengembangkan model keterbacaan dasar untuk Bahasa Cebuano, penelitian tersebut menghasilkan model dengan kinerja terbaik dan kombinasi fitur untuk Cebuano mencapai sekitar 87,3% untuk semua metrik yang menggunakan kombinasi fitur tradisional dan pola suku kata dengan Random Forest [3]. Penelitian lain meneliti cara alternatif menggabungkan embeddings dari fitur BERT dengan fitur linguistik tradisional untuk model keterbacaan. Hasil dari eksperimen mereka menunjukkan bahwa metode tersebut mengungguli pendekatan klasik dalam penilaian keterbacaan menggunakan bahasa Inggris (OSE dan CCE) dan Filipina (Adarna), algoritma yang digunakan mencakup berbagai algoritma pembelajaran mesin, seperti Logistic Regression, Support Vector Machine, dan Random Forest. Peneliti juga menunjukkan bahwa BERT embeddings (informasi semantik dan sintaksis) dapat digunakan sebagai set fitur

pengganti lengkap untuk bahasa dengan sumber yang terbatas seperti bahasa Filipina dengan NLP tools yang terbatas untuk secara eksplisit mengekstraksi nilai fitur untuk menyelesaikan tugas [4].

Berdasarkan uraian di atas, penelitian ini telah menghasilkan klasifikasi multi kelas yang dibagi menjadi K1 (anak umur 6-7), K2 (anak umur 7-8), K3 (anak umur 8-9) mengenai pembangunan model readability dari buku pelajaran Bahasa Indonesia sekolah dasar menggunakan prediktor tradisional atau prediktor surface based, fitur orthography-based menggunakan syllable pattern atau pola suku kata, dan representasi neural. Algoritma yang dipilih adalah Random Forest dan Support Vector Machine. Random Forest dipilih karena salah satu yang terbaik di antara algoritma klasifikasi mampu mengklasifikasikan sejumlah besar data dengan akurat [5]. Support Vector Machine dipilih karena menjadi salah satu classifier yang paling kuat mampu menangani vektor fitur dimensi tak terbatas [5].

Topik dan Batasannya

Topik dan Batasan dalam penelitian ini adalah mengetahui hasil dan performansi model readability yang dibangun berdasarkan fitur ekstraksi yang dilakukan pada penelitian ini. Batasan masalah pada penelitian ini adalah melakukan pembangunan model readability dari buku pelajaran Bahasa Indonesia sekolah dasar menggunakan prediktor tradisional atau prediktor surface-based, fitur orthography-based menggunakan syllable pattern atau pola suku kata, dan representasi neural. Algoritma yang dipilih adalah Random Forest dan Support Vector Machine. Data yang digunakan dalam penelitian ini merupakan buku pelajaran Bahasa Indonesia sekolah dasar kelas 1, 2, dan 4 yang didapatkan dari website [7].

Tujuan

Tujuan dalam penelitian tugas akhir ini adalah membangun model readability dari buku pelajaran Bahasa Indonesia sekolah dasar menggunakan prediktor tradisional atau prediktor surface-based, fitur orthography-based menggunakan syllable pattern atau pola suku kata, dan representasi neural. Algoritma yang dipilih adalah Random Forest dan Support Vector Machine serta melakukan analisis atas hasil dari pembangunan model.

Organisasi Tulisan

Pada bab 2 membahas mengenai studi yang berkaitan dengan penelitian ini. Bab 3 pada penelitian ini membahas alur sistem yang dibangun dan penjelasan singkat dan setiap alur pada sistem tersebut. Bab 4 pada penelitian ini membahas mengenai hasil pengujian dan evaluasi dari hasil pengujian tersebut. Bab 5 membahas mengenai kesimpulan dari penelitian ini yang didapatkan dari analisis hasil pengujian.

2. Studi Terkait

Terdapat beberapa penelitian sebelumnya yang melakukan penelitian terkait readability, readability model, maupun metode terkait, seperti Lloyd Lois Antonie Reyes, Michael Antonio Ibañez, Ranz Sapinit, Mohammed Hussien, dan Joseph Marvin Imperial yang meneliti model keterbacaan dari Bahasa Cebuano [3]. Selanjutnya ada Joseph Marvin Imperial yang meneliti keterbacaan menggunakan BERT [4]. Selanjutnya Tovly Deutsch, Masoud Jasbi, dan Stuart Shieber mencoba beberapa fitur linguistik untuk meningkatkatkan kinerja model keterbacaan [6]. Terdapat beberapa model, teori dan hasil yang didapatkan pada penelitian tersebut. Berikut merupakan beberapa penelitian yang terkait dengan topik dan metode dari penelitian ini.

Pada penelitian pertama, peneliti meneliti berbagai fitur linguistik mulai dari prediktor tradisional atau prediktor surface-based, fitur orthography-based menggunakan syllable pattern atau pola suku kata, dan representasi neural untuk mengembangkan model baseline readability dasar untuk Bahasa Cebuano, Penelitian tersebut menghasilkan model dengan kinerja terbaik dan kombinasi fitur untuk Cebuano mencapai sekitar 87,3% untuk semua metrik yang menggunakan kombinasi fitur tradisional dan pola suku kata dengan Random Forest[3].

Pada penelitian kedua, peneliti meneliti cara alternatif menggabungkan embeddings dari information-rich BERT dengan fitur linguistik tradisional untuk readability. Hasil dari eksperimen mereka menunjukkan bahwa metode tersebut mengungguli pendekatan klasik dan vanilla dalam penilaian readability menggunakan bahasa Inggris (OSE dan CCE) dan Filipina (Adarna) dataset dalam berbagai algoritma pembelajaran mesin, seperti Logistic Regression, Support Vector Machine, dan Random Forest. Peneliti juga menunjukkan bahwa BERT embeddings (informasi semantik dan sintaksis) dapat digunakan sebagai set fitur pengganti lengkap untuk