

1. Pendahuluan

Latar Belakang

Pendidikan memegang peranan penting dalam mengembangkan sumber daya manusia yang berkualitas. Melalui pendidikan, manusia mendapatkan ilmu yang berguna untuk mengembangkan berbagai hal. Pendidikan berkualitas mengacu pada proses pembelajaran pada semua jenjang pendidikan, salah satunya adalah evaluasi belajar[1]. Evaluasi belajar dilakukan dengan mengerjakan soal ujian untuk menguji tingkat kemampuan siswa. Proses merancang soal yang sesuai dengan tujuan pembelajaran perlu memperhatikan tingkat kesulitan soal, agar dapat membedakan tingkat kemampuan siswa yang berbeda-beda. Pentingnya perancangan soal juga memudahkan untuk mengidentifikasi tingkat kesulitan soal yaitu mudah, sedang dan sulit.

Konsep Natural Language Processing(NLP) dapat diterapkan dalam bidang pendidikan melalui klasifikasi teks untuk mengidentifikasi tingkat kesulitan soal. Klasifikasi adalah proses mengklasifikasikan dokumen ke dalam satu atau lebih kategori yang telah ditentukan sebelumnya atau ke dalam kelas-kelas dokumen yang serupa [2]. Klasifikasi dalam NLP dapat secara langsung menganalisis teks dan membentuk kategori sesuai dengan konten yang diberikan. Pendekatan klasifikasi teks membantu dalam mengidentifikasi tingkat kesulitan soal dengan memprediksi nilai-nilai variabel yang tidak diketahui berdasarkan nilai variabel lainnya.

Penelitian tentang prediksi tingkat kesulitan soal telah dilakukan untuk soal *High-stakes Medical Exam*. Hasil menunjukkan bahwa embeddings memiliki kekuatan prediksi tertinggi, diikuti oleh fitur linguistik dengan menggunakan algoritma Random Forest[3]. Pada penelitian lain memprediksi kesulitan soal menggunakan *Item Response Theory*(IRT) yang berhasil memprediksi pertanyaan ya atau tidak. Kesalahan prediksi untuk pertanyaan ya atau tidak memiliki rata-rata prediksi terkecil yaitu 0,68[4]. Penelitian tersebut mencoba menggunakan sinonim atau mengubah pertanyaan untuk meningkatkan kesulitan soal menjadi 0,53 dan 1,02.

Pada Tugas Akhir ini, dibangun model untuk identifikasi tingkat kesulitan soal pada mata pelajaran Bahasa Indonesia dan Matematika. Model yang dibangun menggunakan metode klasifikasi Random Forest, SVM, Gaussian, dan Dense NN seperti metode yang telah digunakan pada penelitian sebelumnya[3] tetapi kami mencoba menambahkan metode lain untuk eksperimen yaitu Logistic Regression. Dengan menggunakan metode *machine learning* dan pendekatan *deep learning* ini bisa mengklasifikasikan dengan baik. Penelitian ini memprediksi tingkat kesulitan soal yang dikembangkan dengan tiga ekstraksi fitur yaitu fitur embeddings menggunakan Word2Vec, fitur leksikal menggunakan TF-IDF dan fitur sintaksis menggunakan metode POS tagging. *Labeling* pada penelitian ini dilakukan dengan survey kepada guru Bahasa Indonesia dan Matematika tingkat SD dan SMP. Label dibagi menjadi 3 kategori yaitu mudah, sedang, dan sulit berdasarkan soal yang menjadi dataset. Sebelum melakukan pemodelan, terdapat pra-pemrosesan data yang dilakukan untuk mempersiapkan teks agar lebih sesuai dengan kebutuhan saat pemodelan.

Topik dan Batasannya

Topik pada penelitian ini adalah memprediksi tingkat kesulitan soal yang menggunakan klasifikasi teks. Klasifikasi menggunakan model *machine learning* Random Forest, Linear Regression, SVM, Gaussian dan *deep learning* Dense NN, klasifikasi juga dikembangkan dengan menggunakan ekstraksi fitur. Batasan dari penelitian ini yaitu: (1) Dataset yang digunakan adalah soal Bahasa Indonesia dan Matematika tingkat Sekolah Dasar dan Sekolah Menengah Pertama; (2) Data berasal dari Bank Soal sebanyak 548 soal; (3) Model klasifikasi yang digunakan adalah Random Forest, Linear Regression, SVM, Gaussian dan Dense NN; (4) Ekstraksi fitur yang digunakan adalah fitur embeddings dengan metode Word2Vec dan fitur linguistic dengan metode TF-IDF dan Stanza.

Tujuan

Tujuan dari penelitian ini adalah mengaplikasikan metode klasifikasi teks soal pilihan ganda Bahasa Indonesia dan Matematika untuk identifikasi tingkat kesulitan soal serta menganalisis hasilnya. Tingkat kesulitan soal berupa kategori yaitu mudah, sedang dan sulit. Klasifikasi teks menghasilkan hasil prediksi tingkat kesulitan soal yang akan dievaluasi dan dianalisis.

Organisasi Tulisan

Penelitian ini terdiri dari beberapa bagian. Bagian pertama berisi pendahuluan yang menjelaskan topik dan tujuan pada penelitian ini. Bagian kedua menjelaskan studi terkait yang mendukung topik penelitian. Bagian ketiga berisi system yang dibangun untuk menjelaskan metode atau model yang digunakan. Bagian keempat adalah Evaluasi yang menjelaskan hasil pengujian dan analisis hasil pengujian. Bagian kelima adalah kesimpulan dan saran berdasarkan penelitian yang sudah dilakukan.