Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.2023107338
Vol. 10, No. 6, Desember 2023, hlm. 1177-1184 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

AUTOMATED ESSAY SCORING MENGGUNAKAN *SEMANTIC TEXTUAL SIMILARITY* BERBASIS *TRANSFORMER* UNTUK PENILAIAN UJIAN ESAI

Kharisma Ayu Pradani[1], Lya Hulliyyatus Suadaa[*2]

[1,2]Politeknik Statistika STIS, Jakarta Timur
Email: [1]221910989@stis.ac.id, [2]lya@stis.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 05 Juni 2023, diterima untuk diterbitkan: 27 November 2023)

Abstrak

Ujian berbasis esai seringkali digunakan untuk menguji pemahaman siswa dalam menyelesaikan permasalahan. Tak terkecuali dalam pelaksanaan ujian di Politeknik Statistika STIS. Dalam melakukan penilaian pada jawaban tipe ini, dibutuhkan waktu serta tenaga yang besar, dan sering kali menimbulkan ketidakkonsistenan dalam penilaian. Hal ini dapat terjadi salah satunya karena perbedaan cara penilaian yang dilakukan oleh orang yang berbeda. Oleh karena itu diperlukan penyelesaian yang bisa mengefektifkan waktu, tenaga serta menjaga kekonsistenan aspek penilaian, diantaranya yaitu dengan *automated essay scoring* (AES). AES merupakan suatu model yang dilatih untuk menilai suatu esai secara otomatis berdasarkan kemiripan jawaban dengan kunci jawaban. Pada penelitian ini, metode yang diusulkan untuk menghitung kemiripan semantik teks berbahasa Indonesia antara jawaban esai dan kunci jawabannya yaitu model berbasis *Transformers* IndoBERT. Sebagai *baseline*, digunakan teknik ekstraksi fitur *Term Frequency - Inverse Document Frequency* (TF-IDF) dan penghitungan kemiripan fitur menggunakan *cosine similarity* dan *linear regression*. Selanjutnya nilai kemiripan tersebut dikonversi ke rentang nilai yang diinginkan sebagai prediksi nilai dari setiap esai. Berdasarkan hasil evaluasi, diperoleh bahwa model *fine-tuned* IndoBERT merupakan model terbaik dengan nilai MAE dan RMSE sebesar 0.1285 dan 0.2001.

Kata kunci: *Automated Essay Scoring (AES), Semantic textual similarity, Transformer, IndoBERT*

AUTOMATED ESSAY SCORING USING TRANSFORMER-BASED SEMANTIC TEXTUAL SIMILARITY FOR ESSAY ASSESSMENT

Abstract

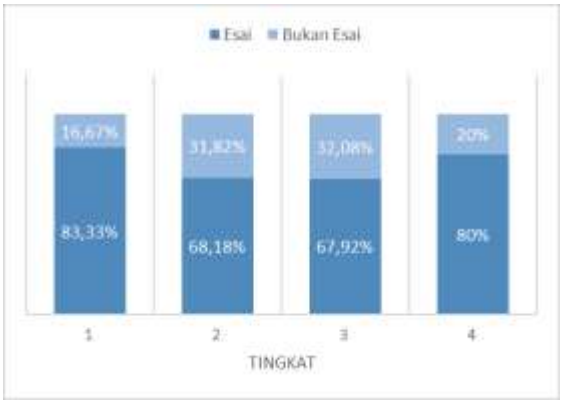
Essay-based exams are often used to test students' understanding of solving problems. However, assessing this type of answer takes a lot of time and effort and often results in inconsistencies. One of the reasons is the different ways between people while doing the assessment. Therefore, a solution is needed to streamline time, effort, and maintain consistency in aspects of assessment, including automated essay scoring (AES). AES is a model trained to assess an essay automatically based on the similarity of answers with the answer key. In this study, the method proposed to calculate the semantic similarity of Indonesian text between essay answers and answer keys is a model based on the Transformer BERT. As a baseline, the Term Frequency – Inverse Document Frequency (TF-IDF) feature extraction technique is used and calculating feature similarity using cosine similarity and linear regression. Then the similarity value is converted to the desired range of values as the predicted value of each essay. Based on the evaluation results, it was found that the fine-tuned IndoBERT model was the best model, with MAE and RMSE values of 0.1285 and 0.2001.

Keywords: *Automated Essay Scoring (AES), Semantic textual similarity, Transformer, IndoBERT*

1. PENDAHULUAN

Esai merupakan salah satu bentuk tulisan yang sering dipilih untuk menyatakan suatu hal yang membutuhkan penjelasan secara jelas dan rinci. Bentuk ini biasanya digunakan untuk menjawab soal atau memecahkan suatu permasalahan dari pertanyaan penalaran yang memerlukan alasan serta penjelasan. Sutoyo dalam (Kurniawati, & Pradnya, 2020) menyatakan bahwa soal esai merupakan soal yang digunakan untuk mengukur pencapaian hasil belajar dengan aspek yang kompleks. Tes dengan soal esai bertujuan mengukur kemampuan dari peserta dalam melakukan analisis, mengorganisasi dan mengekspresikan ide mengenai suatu hal.

Pada pelaksanaan ujian di Politeknik Statistika STIS, sering kali digunakan bentuk ujian esai, baik untuk ujian tengah semester (UTS) maupun ujian akhir semester (UAS) dimana hal tersebut ditujukan untuk menguji pemahaman mahasiswa dalam mata kuliah yang diujikan. Pada tahun akademik 2021/2022, persentase UTS dan UAS semester genap yang menggunakan esai sebagai format jawaban ada sebesar 73,58%. Adapun proporsi dari penggunaan esai dalam UTS dan UAS berdasarkan tingkatnya ditunjukkan pada gambar 1.



Gambar 1. Proporsi penggunaan esai pada UTS dan UAS tiap tingkat

Penilaian jawaban suatu esai, sering kali dilakukan secara manual sehingga membutuhkan waktu yang lama serta tenaga yang besar untuk menilai keseluruhan esai karena harus diperiksa satu per satu (Putri et al, 2022) (Muslicha, Putri, & Syadiyah, 2017). Guna mengatasi permasalahan tersebut, penyelesaian yang umum digunakan adalah dengan menambah tenaga penilai. Tetapi hal tersebut dapat menimbulkan permasalahan lain terkait kekonsistenan penilaian, dimana tiap penilai memiliki standar dan metode penilaian yang berbeda (Putri et al, 2022). Walaupun terdapat acuan kunci jawaban, dalam penilaian esai, kunci jawaban tersebut biasanya hanya berbentuk ide pokok dari jawaban yang benar untuk membantu penilai memahami secara garis besar jawaban seperti apa yang bernilai benar. Sedangkan esai menampilkan jawaban yang sangat variatif. Walau memiliki maksud yang sama, penulisan esai belum tentu sama (Muslicha, Putri, & Syadiyah, 2017). Perbedaan waktu, situasi, dan kondisi saat seseorang menilai esai juga dapat menghasilkan nilai yang tidak konsisten (Muslicha, Putri, & Syadiyah, 2017). Hal ini dikarenakan beberapa faktor, salah satunya adalah keterbatasan memori manusia dalam mengingat seluruh penilaian yang pernah dilakukan. Untuk jawaban yang sama, dua orang pengoreksi menghasilkan nilai yang berbeda (Putri et al, 2022) (Muslicha, Putri, & Syadiyah, 2017). Oleh karena itu, diperlukan suatu solusi yang bisa menyelesaikan masalah tentang banyaknya waktu dan tenaga, sekaligus ketidakkonsistenan proses penilaian esai.

Automated essay scoring (AES) merupakan salah satu penyelesaian yang ditawarkan untuk masalah tersebut. Definisi dari AES adalah sebuah tugas dari *machine learning* dalam NLP, dimana kita menciptakan suatu model yang bisa digunakan untuk memberikan nilai pada jawaban siswa yang berbentuk esai secara otomatis (Rajagede, 2021). Dalam penelitian (Rodriguez, Jafari, & Ormerod, 2019), penilaian AES menunjukkan kecenderungan lebih handal daripada manusia, menghemat waktu dan uang dalam melakukan penilaian esai dalam skala besar. Dalam hal ini AES melibatkan model statistik untuk mengekstraksi fitur penting yang terkandung dalam esai yang nantinya akan berguna untuk menetapkan nilai dalam suatu rentang angka (Beseiso, & Alzahrani, 2020).

Secara garis besar, cara kerja AES adalah memberikan skor pada esai berdasarkan pada fitur yang diekstrak dari teks esai. Proses dari penilaian terdiri dari dua fase (Rodriguez, Jafari, & Ormerod, 2019). Fase yang pertama adalah fase dimana terjadi pelatihan model untuk melakukan ekstraksi fitur yang akan menjadi kriteria penilaian esai. Sedangkan tahap yang kedua adalah membangun model mesin AES yang dilatih di data berlabel berdasarkan fitur dari kumpulan data yang telah terpilih (Beseiso, & Alzahrani, 2020).

Dalam penelitian ini, esai yang digunakan sebagai data dibatasi hanya untuk esai berbahasa Indonesia yang berbentuk teks. Soal yang ditanyakan juga dibatasi pada soal yang bersifat teoritis yang memerlukan jawaban dengan hanya menggunakan teks tanpa rumus dan angka. Adapun data yang digunakan terdiri atas jawaban esai, kunci jawaban, dan skor yang diperoleh dari dosen pada mata kuliah yang bersangkutan.

Dari penelitian yang telah dilakukan sebelumnya, metode yang umum digunakan dalam memodelkan AES adalah dengan metode penghitungan *text similarity* menggunakan fitur TF-IDF (*Term Frequency - Inverse Document Frequency*). Seperti pada penelitian (Romadon & Lhaksamana, 2020), menyebutkan dalam penelitiannya bahwa TF-IDF bekerja lebih baik untuk mengurangi jumlah dimensi dari data. Pada penelitian (Ratna et al., 2019), dilakukan penilaian esai otomatis dengan menggunakan TF-IDF sebagai pembobot kata dari data esai. Penelitian lainnya, dilakukan (Septiandri & Winatmoko, 2020), dengan menggunakan data UKARA, model TF-IDF mendapatkan nilai *F1-Score* sebesar 0.812.

TF-IDF merupakan salah satu teknik yang sering digunakan untuk ekstraksi fitur teks dengan menghitung skor pembobotan kata (Lahitani, 2022). Pembobotan dilakukan berdasarkan jumlah frekuensi kata dalam sebuah dokumen dan ketersediaan kata pada keseluruhan dokumen yang ada. Dalam penelitian ini, metode TF-IDF digunakan sebagai *baseline* atau pembanding dengan metode yang lainnya.

Fitur yang dihasilkan TF-IDF diproses untuk menghasilkan skor yang diperoleh esai dengan menggunakan *cosine similarity* dan *linear regression*. Dalam penelitian sebelumnya, *cosine similarity* digunakan untuk mengukur tingkat kemiripan antara jawaban ahli dan jawaban siswa (Lahitani, 2022). Hasanah dalam (Lahitani, 2022) mengatakan bahwa *cosine similarity* lebih unggul dalam mengukur kemiripan dibanding model seperti *Jaccard.*

Model *linear regression* juga digunakan sebagai salah satu *learning model* dalam AES. Pada penelitian Manvi Mahana et al. (Mahana, Johns, & Apte, 2012), dilakukan pemodelan AES dengan menggunakan model *linear regression* yang diperoleh dari mempelajari fitur kata. Pada penelitinian ini, dikatakan bahwa dari referensi yang dirujuk mengindikasikan bahwa model ini bekerja dengan baik dalam AES sehingga metode ini dipilih sebagai *learning* model. Model final penelitian ini memperoleh nilai Kappa 0.73 pada seluruh delapan set esai. Model yang diperoleh bekerja dengan baik pada esai tanpa konten yang spesifik

Selain metode yang telah disebutkan, metode lain yang dapat dimanfaatkan dalam mengekstraksi fitur kalimat menjadi suatu representasi vektor adalah *Bidirectional Encoder Representations form Transformers* (BERT). Model dengan metode tersebut mendapatkan nilai *F1-score* lebih baik dibandingkan dengan model lainnya untuk dataset Ukara (Rajagede, 2021). Metode BERT didesain untuk melatih representasi dua arah secara mendalam dengan mengkondisikan bersama pada konteks yang ada di kiri dan kanan di semua lapisan, dimana hal ini berbeda dengan metode *embedding* sebelumnya yang melakukan pelatihan model dengan representasi searah, yaitu dari kiri ke kanan (Devlin, Chang, Lee, & Toutanova, 2018). BERT juga menghasilkan model yang lebih kaya fitur dengan model yang sederhana dan lebih cepat (Rajagede, 2021). Oleh karena itu, penelitian ini menggunakan model *pre-trained* IndoBERT yang merupakan model BERT yang telah dilatih dengan data berbahasa Indonesia yang besar (Bryan et al, 2020).

Output yang diharapkan dalam penelitian ini adalah model AES yang dapat menghasilkan skor berupa angka dalam rentang nilai yang telah ditentukan dan memiliki akurasi yang terbaik dalam memberikan skor atau nilai.

2. METODE PENELITIAN

2.1. Sumber Data

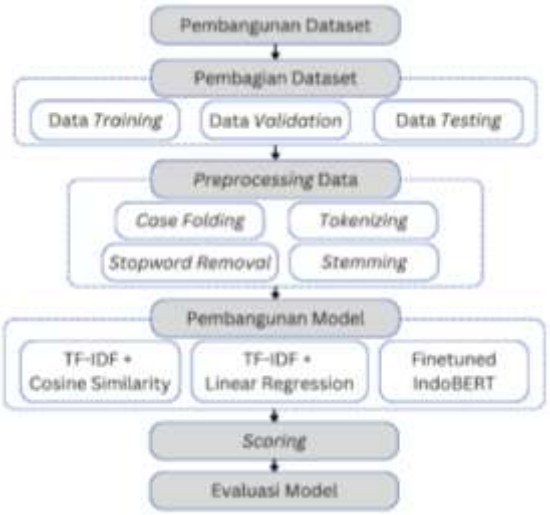
Dalam penelitian ini, data yang digunakan adalah data jawaban esai mata kuliah Information Retrieval di Politeknik Statistika STIS pada UAS Ganjil tahun 2022/2023. Setiap jawaban akan dihitung sebagai satu data. Esai yang digunakan dibatasi pada esai yang bersifat teoritis dan berbentuk teks berbahasa Indonesia, tanpa melibatkan hitungan ataupun rumus.

2.2. Tahapan Penelitian

Tahapan pada penelitian ini diilustrasikan pada gambar 2.

2.2.1. Pembangunan Dataset

Dataset dibangun dari lembar jawaban ujian yang berisi esai tulisan tangan, kunci jawaban, serta nilai untuk tiap esai dari dosen pengampu. Seluruh jawaban yang memenuhi batasan yang telah ditentukan, selanjutnya dilakukan perekaman jawaban secara manual. Dari data yang telah terekam, diperoleh data esai sebanyak 532 data yang berasal dari tujuh soal dan 76 siswa.



Gambar 2. Alur metode penelitian

2.2.2. Pembagian Dataset

Pada tahap ini, dataset yang didapatkan dibagi menjadi tiga kelompok data secara acak, yaitu data *training*, data *validation*, dan data *testing* dengan perbandingan 80:10:10.

2.2.3. Preprocessing Data

Pada tahapan ini, dilakukan beberapa tahapan mulai dari *case folding*, *tokenizing*, *stopword removal*, dan *stemming*. Penjelasan dari tahapan tersebut adalah sebagai berikut.

- *Case folding* dilakukan untuk menyamakan bentuk dari teks esai menjadi bentuk yang standar (menggunakan huruf kecil atau *lowercase*), serta menghilangkan karakter yang selain huruf.
- *Tokenizing* dilakukan dengan memotong string input berdasarkan pada kata yang menyusunnya dengan menggunakan spasi sebagai pemisah kata tersebut.
- *Stopword removal* dilakukan dengan cara membuang kata yang dianggap kurang penting, seperti kata ulang yang sering digunakan.
- *Stemming* dilakukan dengan mengubah kata-kata yang ada pada esai menjadi kata dasarnya.

Dalam teks berbahasa Indonesia, hal ini dilakukan dengan membuang imbuhan yang ada.

Untuk model *fine-tuned* IndoBERT, tahapan *preprocessing* yang diterapkan hanya *case folding* saja. Hal ini dikarenakan model tersebut menggunakan tokenisasi berdasarkan *subword* dan tidak memerlukan teknik *preprocessing* lainnya.

2.2.4. Pembangunan Model

Dalam pembangunan model, penelitian ini menggunakan ekstraksi fitur *Term Frequency - Inverse Document Frequency* (TF-IDF) dengan penghitungan skor menggunakan *cosine similarity* dan *linear regression* sebagai *baseline* dan mengusulkan *fine-tuning* model *pre-trained* IndoBERT. Penjelasan dari model yang digunakan dalam penelitian ini adalah sebagai berikut:

- TF-IDF dan *Cosine Similarity*

Metode pembobotan kalimat dengan menggunakan TF-IDF dilakukan dengan cara menghitung nilai *term frequency* (tf) yaitu jumlah kemunculan term (t) yang ada pada setiap dokumen (d) (Lahitani, 2022). *Inverse document frequency* (idf) adalah suatu perhitungan dari bagaimana term didistribusikan secara luas pada koleksi dokumen yang bersangkutan (Ahmad, Wardi, & Dewiani, 2018). Persamaan idf dijelaskan pada persamaan (1), dimana D merupakan jumlah semua dokumen dalam koleksi dan df_j merupakan jumlah dokumen yang mengandung term.

$$idf = \log(\frac{D}{df_j}) \quad (1)$$

TF-IDF merupakan gabungan dari kedua konsep tersebut, yaitu TF dan DF dan rumus penghitungannya dijelaskan pada persamaan (2) (Ahmad, Wardi, & Dewiani, 2018).

$$W_{ij} = tf \times idf \quad (2)$$

Bobot TF-IDF tersebut kemudian digunakan sebagai nilai vektor tiap esai dan kunci jawaban. Kemudian dilakukan penghitungan skor kemiripan antara vektor esai dan kunci jawaban dengan menggunakan *cosine similarity.*

Cosine similarity merupakan ukuran sudut antara vektor dokumen Da (titik (ax,bx)) dan Db (titik (ay,by)), dimana tiap vektor tersebut menggambarkan setiap kata yang ada dalam dokumen yang dibandingkan dan membentuk suatu segitiga, sehingga dapat diterapkan hukum *cosinus* (Ahmad, Wardi, & Dewiani, 2018). Adapun rumus penghitungannya dijelaskan pada persamaan (3).

$$\cos C = \frac{a_x b_x + a_y b_y}{\sqrt{a_x^2 + b_x^2} \times \sqrt{a_y^2 + b_y^2}} \quad (3)$$

Ketika dokumen memiliki kemiripan maka sudutnya akan bernilai nol dan kesamaannya adalah satu, begitu juga sebaliknya (Ahmad, Wardi, & Dewiani, 2018).

- TF-IDF dan *Linear Regression*

Dilakukan pelatihan pemberian skor menggunakan *linear regression* dengan input berupa vektor TF-IDF dari esai dan kunci jawaban. *Simple linear regression* berasumsi bahwa terdapat hubungan linier antara variabel respon dan variabel penjelas. Dalam penelitian ini digunakan fungsi *LinearRegression* yang ada pada *library sklearn*, dimana metode ini mempelajari parameter dari model untuk *simple linear regression* dijelaskan pada persamaan (4) (Hackeling, 2017).

$$y = \alpha + \beta x \quad (4)$$

- *Fine-tuned* IndoBERT

IndoBERT merupakan model BERT yang telah dilatih dengan dataset berbahasa Indonesia sebanyak 250M kalimat yang berasal dari sumber yang tersedia di publik, seperti sosial media, blog, berita, dan website. Model pre-training yang digunakan dalam penelitian ini adalah indobenchmark/indobert-base-p1 yang telah tersedia pada penelitian (Bryan et al, 2020). Dari model tersebut dilakukan penyesuaian seperti modifikasi fungsi dan *layer* terakhir pada model sehingga *output* yang dihasilkan berupa skor. Untuk mendapatkan *hyperparameter* yang tepat untuk model yang dibangun, dilakukan juga *hyperparameter tuning* dengan cara membandingkan *loss* yang dihasilkan pada *train* dan *validation set*.

2.2.5. Scoring

Nilai prediksi yang diperoleh dari model dibulatkan menjadi satu desimal, dan disesuaikan ke dalam rentang 0 hingga 1. Proses penghitungan skor dari hasil yang diperoleh model dilanjutkan dengan mengkonversikan nilai ke rentang nilai yang diharapkan, dimana dalam penelitian ini adalah 0 hingga nilai maksimal. Adapun merujuk pada (Lahitani, 2022), rumus dari pengkonversian skor dijelaskan dalam persamaan (5).

$$skor = skor\ model \times nilai\ maks \quad (5)$$

2.2.6. Evaluasi Model

Evaluasi dilakukan dengan cara *cross validation* menggunakan 5-*fold*. Pada penelitian ini, matrik yang digunakan yaitu *mean absolute error* (MAE) dan *root mean squared error* (RSME). MAE adalah nilai rata-rata dari total eror mutlak. Perbedaan antara kedua nilai akan selalu bernilai positif atau mutlak, walaupun perbedaannya bernilai negatif (Arifin, Purnamasari, & Ratna, 2021). Cara menghitung MAE dijelaskan dalam persamaan (6).

$$MAE = \frac{\sum_{i=0}^{n} |y - \hat{y}|}{n} \quad (6)$$

RMSE adalah nilai akar kuadrat dari *mean square error* (MSE) dimana MSE adalah nilai rata rata dari jumlah error kuadrat.Nilai MSE akan selalu bernilai positif. Semakin kecil nilai MSE yang diperoleh, maka semakin baik model yang dievaluasi (Arifin, Purnamasari, & Ratna, 2021). Adapun untuk menghitung RMSE menggunakan rumus yang dijelaskan pada persamaan (7).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2} \quad (7)$$

3. HASIL DAN PEMBAHASAN

3.1. Pembangunan Dataset

Dalam penelitian ini, data yang digunakan adalah data jawaban esai mata kuliah Information Retrieval di Politeknik Statistika STIS pada UAS Ganjil tahun 2022/2023. Data tersebut berjumlah 532 esai, yang terdiri tujuh soal dengan jawaban 76 mahasiswa yang bersifat teoritis tanpa menggunakan rumus maupun hitungan. Dataset juga berisi kunci jawaban dan skor dari dosen.

Perekaman data dilakukan secara manual dengan merekam dari lembar jawaban mahasiswa. Semua soal memiliki nilai maksimal 10. Dalam penelitian ini, pelatihan dan penghitungan akurasi model dilakukan terhadap label yang berupa persentase nilai. Persentase nilai tersebut diperoleh dari pembagian skor dosen dengan nilai maksimal, sehingga diperoleh nilai antara 0 dan 1. Sebaliknya, untuk memperoleh nilai yang diharapkan, dilakukan dengan mengalikan nilai prediksi model dengan nilai maksimal soal. Contoh *dataset* yang disusun dapat dilihat pada tabel I. *Dataset* tersebut dibagi menjadi tiga bagian secara acak, yaitu data *training*, data *validation*, dan data *testing* dengan proporsi 80:10:10.

Tabel 1. Contoh Dataset

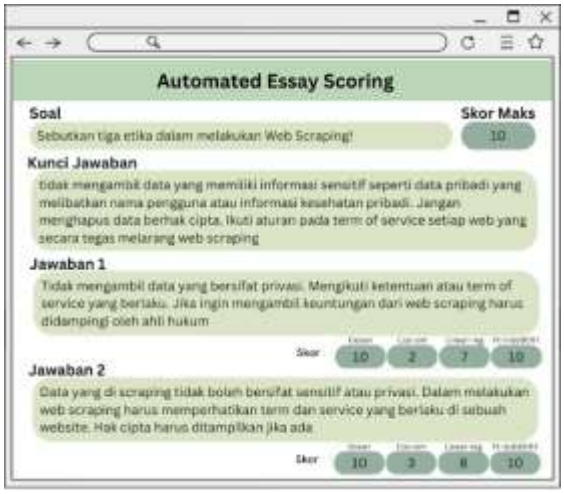
No	Esai	Kunci Jawaban	Skor Dosen	$\frac{Skor}{Nilai\ Maks}$
1	Pada tahap training dilakukan pemrosesan yang menghasilkan fungsi gamma yang nantinya fungsi tersebut akan diterapkan pada tahap testing untuk pengklasifikasian teks	Pada tahap training, klasifikasi teks berarti mempelajari satu classifier gamma yang dapat memetakan dokumen ke kelas kelas. Pada tahap testing klasifikasi teks berarti menentukan kelas yang paling tepat untuk dokumen d menggunakan classifier gamma	8	0.8
2	Memperhatikan web yang akan diambil datanya, apakah data mengandung privasi dari pengguna. Memerlukan izin jika terdapat hak cipta. Tidak menggunakan data untuk kepentingan di luar dari kepentingan yang diperbolehkan	Tidak mengambil data yang memiliki informasi sensitif seperti data pribadi yang melibatkan nama pengguna atau informasi kesehatan pribadi. Jangan menghapus data berhak cipta. Ikuti aturan pada term of service setiap web yang secara tegas melarang web scraping	10	1

3.2. Penilaian Esai Otomatis

Untuk membersihkan data yang akan menjadi input dari model, terlebih dahulu dilakukan *preprocessing,* yaitu *case folding*, *tokenizing*, *stopword removal*, dan *stemming*. *Stopword removal* dilakukan dengan menggunakan *library* nltk untuk bahasa Indonesia yang telah tersedia dan siap digunakan. Sedangkan untuk proses *stemming* dilakukan dengan menggunakan *library* Sastrawi. Sastrawi merupakan *library* yang sering digunakan untuk *stemming* bahasa Indonesia (Saputra, Adji, & Permanasari, 2015).

Selanjutnya, dilakukan ekstraksi fitur dengan TF-IDF. TF-IDF merupakan metode untuk mengekstraksi fitur dari sekumpulan dokumen dengan memperhatikan jumlah kemunculan *term* pada setiap dokumen dan distribusi *term* tersebut pada seluruh dokumen yang ada. Pada tahap ekstraksi fitur, data hasil *preprocessing* diproses lebih lanjut untuk membangun *dictionary* yang merupakan kumpulan kata yang digunakan pada seluruh dokumen. Untuk tiap kata pada *dictionary* tersebut diberi bobot dengan nilai TF-IDF. Bobot inilah yang digunakan untuk membentuk vektor jawaban dan vektor kunci jawaban dari tiap esai. Vektor diproses untuk menghasilkan nilai dengan *cosine similarity* dan *linear regression*.

Pada model pertama, nilai prediksi diperoleh dengan menghitung kemiripan vektor esai dan vektor kunci jawaban menggunakan *cosine similarity*, sedangkan model kedua menggunakan *linear regression* untuk menghasilkan nilai prediksi skor esai dengan input berupa gabungan vektor esai dan kunci jawaban. Pelatihan model *linear regression* dilakukan dengan menggunakan fungsi yang ada pada *library* Sklearn. Nilai prediksi dari dua metode tersebut dibulatkan menjadi hanya satu angka dibelakang koma dan dimasukkan ke dalam rentang 0 hingga 1, yang selanjutnya nilai tersebut dikalikan dengan nilai maksimal tiap soal. Pada penelitian ini, semua soal memiliki nilai maksimal 10. Dari proses *scoring* tersebut didapatkan nilai yang diharapkan. Beberapa contoh hasil prediksi nilai dari model pertama dan kedua ditampilkan pada gambar 3.



Gambar 3. Contoh hasil prediksi nilai

Selanjutnya dilakukan pembangunan model ketiga dengan melakukan penyesuaian (*fine-tuning*) terhadap *pre-trained* model IndoBERT. IndoBERT menawarkan banyak fungsi, salah satunya adalah menentukan apakah dua kalimat sama atau berbeda. Penelitian ini menggunakan model *pretrained indobert-base-p1* (Bryan et al, 2020). Model tersebut merupakan model BERT yang telah dilatih dengan menggunakan *dataset* berbahasa Indonesia. Tahapan *preprocessing* yang dilakukan adalah *case folding*. Selanjutnya dilakukan *fine-tuning* model IndoBERT dengan *input* berupa teks gabungan dari jawaban esai dan kunci jawaban, serta *output* berupa nilai esai.

Pelatihan model dilakukan dengan cara 5-*fold cross validation*. Proses *hyperparameter tuning* dilakukan untuk menentukan nilai *epoch* terbaik yang akan digunakan pada tiap *fold*. Proses ini dilakukan dengan membandingkan nilai *loss* pada data *train* dan data *validation* untuk setiap *epoch*. Proses ini dilakukan pada rentang *epoch* 1 hingga 20. Grafik *loss* pada data *train* dan data *validation* untuk *fold* pertama ditunjukkan pada gambar 4.

Gambar 4. *Loss* model IndoBERT *Fold* 1

Pada gambar 4, terlihat bahwa nilai *loss* terendah berada pada *epoch* 16, sedangkan untuk *epoch* selanjutnya, terlihat bahwa nilai dari *loss* yang diperoleh baik pada data *training* maupun data *validation* tidak berubah signifikan, sehingga pada *fold* pertama, parameter *epoch* yang digunakan adalah 16. Mekanisme tersebut dilakukan untuk kelima *fold*. Contoh hasil prediksi nilai yang diperoleh dengan model ketiga ditampilkan pada gambar 3.

3.3. Evaluasi

Metrik yang digunakan untuk mengevaluasi tiap model adalah nilai MAE dan RMSE. Penghitungan MAE dan RMSE dilakukan dengan menggunakan *library* Sklearn dengan membandingkan skor dari model dan skor dosen pada data *testing*. Hasil evaluasi model tersebut ditampilkan pada tabel 2.

Tabel 2. Hasil Evaluasi Model AES pada Data Testing

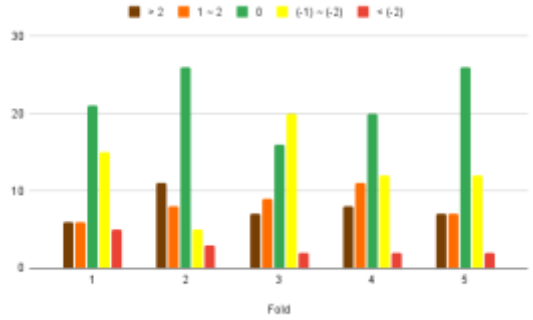
Model	MAE	RMSE
TF-IDF + Cosine Similarity	0.4375	0.5255
TF-IDF + Linear Regression	0,2087	0,2948
IndoBERT	**0.1285**	**0.2001**

Berdasarkan evaluasi yang ditampilkan pada Tabel II, terlihat bahwa model *fine-tuned* IndoBERT memiliki nilai MAE dan RMSE yang paling kecil dengan nilai secara berturut-turut adalah 0,1285 dan 0,2001. Semakin kecil nilai MAE ataupun RMSE menunjukkan bahwa tingkat kesalahan nilai prediksi akan semakin kecil, sehingga dari ketiga model tersebut terbukti bahwa model *fine-tuned* IndoBERT lebih baik dalam memprediksi nilai esai dibandingkan dengan model lainnya.

Hal ini dapat terjadi dikarenakan dalam melakukan penilaian esai, penting untuk melakukan penilaian dengan membandingkan makna dari esai dengan kunci jawaban, bukan hanya kemiripan pemilihan kata. Model *fine-tuned* IndoBERT menggunakan arsitektur *Transformers* dengan mekanisme *attention*, dimana model dapat memahami teks secara makna. Walaupun kata yang dipilih berbeda, jika memiliki makna yang sama, model akan dapat mengidentifikasi dua teks tersebut sebagai teks dengan kemiripan yang tinggi. Di lain sisi, metode ekstraksi fitur dengan TF-IDF menilai kepentingan dari sebuah kata dengan melihat jumlah dari kata itu pada suatu esai dan bagaimana kata itu didistribusikan pada seluruh esai. Dengan kata lain, TF-IDF hanya menilai dari seberapa sering kata tersebut digunakan. Model dengan metode ekstraksi fitur tersebut tidak mempertimbangkan makna dari kata dalam kalimat esai, sehingga kata yang sama akan dinilai sama walaupun memiliki makna yang berbeda, begitu pula kebalikannya.

Gambar 5 menampilkan perbandingan nilai hasil prediksi model terbaik yang dihasilkan, yaitu *fine-tuned* IndoBERT, dengan nilai dari dosen pada data *testing* pada tiap *fold* nya. Untuk yang berwarna coklat menggambarkan jumlah nilai prediksi yang lebih besar dari nilai dosen sebanyak lebih dari 2 poin. Warna jingga menunjukkan jumlah nilai prediksi lebih besar dari nilai dosen dengan selisih 1 - 2 poin. Warna hijaumenunjukkan jumlah nilai prediksi yang bernilai sama dengan nilai dosen. Warna kuning menunjukkan jumlah nilai prediksi yang mendapat nilai lebih kecil

dari nilai dosen dengan selisih 1 - 2 poin, sedangkan warna merah menunjukkan jumlah nilai prediksi yang lebih kecil dari nilai dosen lebih dari 2 poin.



Gambar 5. Prediksi model IndoBERT dibandingkan nilai dosen pada data test

Jika dari gambar tersebut, dapat dilihat bahwa secara umum model *fine-tuned* IndoBERT menghasilkan distribusi hasil prediksi dengan cukup baik, tidak ada kecenderungan untuk memberikan nilai lebih besar atau lebih kecil dari nilai dosen. Bahkan model tersebut menghasilkan nilai sama dengan nilai dosen paling banyak dibandingkan kelompok lain.

4. KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa *dataset* telah berhasil dibangun dengan melakukan proses perekaman esai, kunci jawaban, dan nilai pada lembar jawaban mahasiswa. Kemudian telah dibangun model berbasis *Transformers* menggunakan *fine-tuned* IndoBERT untuk menghitung *semantic textual similarity* dan model pembanding dengan TF-IDF yang dikombinasikan dengan *cosine similarity* dan *linear regression*. Dari ketiga model tersebut, diperoleh bahwa model *fine-tuned* IndoBERT merupakan model terbaik yang dapat diaplikasikan untuk menilai esai secara otomatis pada ujian esai di Politeknik Statistika STIS.

Adapun saran untuk penelitian selanjutnya, yang pertama adalah dapat mengintegrasikan model AES dengan o*ptical character recognition* (OCR) untuk mengenali tulisan tangan mahasiswa dan mengubahnya menjadi format yang dapat diproses oleh model. Saran kedua, dengan memperluas batasan bentuk esai yang dapat diproses sehingga tidak terbatas pada esai berbahasa Indonesia tanpa melibatkan rumus atau hitungan.

DAFTAR PUSTAKA

AHMAD, R., WARDI, dan DEWIANI, 2018. E-learning Automated Essay Scoring System Menggunakan Metode Searching Text Similarity Matching Text. Jurnal KPE, 22(1).

ARIFIN, A.R., PURNAMASARI, P.D., dan RATNA, A.A.P., 2021. Automatic Essay Scoring for Indonesian Short Answers using Siamese Manhattan Long Short-Term Memory. Proc. of the 3rd International Conference on Electrical, Communications and Computer Engineering (ICECCE).

BESEISO, M., dan ALZAHRANI, S., 2020. An Empirical Analysis of BERT Embedding for Automated Essay Scoring. International Journal of Advanced Computer Science and Applications, 11(10), p.204-210.

BRYAN, W., KARISSA, V., GENTA, I.W., SAMUEL, C., XIAOHONG, L., ZHI, Y.L., SIDIK, S. et al, 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. arXiv preprint arXiv:2009.05387.

DEVLIN, J., CHANG, M.W., LEE, K., dan TOUTANOVA, K., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

HACKELING, G., 2017. Mastering Machine Learning with scikit-learn. Packt Publishing Ltd.

KURNIAWATI, F.E., dan PRADNYA, W.M., 2020. Implementasi Algoritma Winnowing Pada Sistem Penilaian Otomatis Jawaban Esai Pada Ujian Online Berbasis Web. Jurnal Khatulistiwa Informatika, 6(2), pp.169-175.

LAHITANI, A.R., 2022. Automated Essay Scoring Menggunakan Cosine Similarity pada Penilaian Esai Multi Soal. Jurnal Kajian Ilmiah, 22(2), p.107-118.

MAHANA, M., JOHNS, M., dan APTE, A.,2012 Automated Essay Grading Using Machine Learning. CS229 Machine Learning - Autumn 2012 Stanford University.

MUSLICHA, C.C., PUTRI, P.C., SYADINAH, S., 2017. Tes Essay. Universitas Mulawarman.

PUTRI, H., SUSIANI, D., WANDANI, N.S., dan PUTRI, F.A., 2022. Instrumen Penilaian Hasil Pembelajaran Kognitif pada Tes Uraian dan Tes Objektif. Jurnal Papeda: Jurnal Publikasi Pendidikan Dasar, 4(2), p.139-148.

RAJAGEDE, R.A., 2021. Improving automatic essay scoring for Indonesian language using simpler model and richer feature. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, p.11-18.

RATNA, A. A. P., KHAIRUNISSA, H., KALTSUM, A., IBRAHIM, I., dan PURNAMASARI, P. D., 2019. Automatic essay grading for Bahasa Indonesia with support vector machine and latent semantic analysis. In *2019 International Conference on*

Electrical Engineering and Computer Science (ICECOS) (pp. 363-367). IEEE.

RODRIGUEZ, P.U., JAFARI, A., dan ORMEROD, C.M., 2019. Language model and automated essay scoring. arXiv preprint arXiv:1909.09482.

ROMADON, A. W., LHAKSMANA, K. M., KURNIAWAN, I., dan RICHASDY, D., 2020. Analyzing TF-IDF and word embedding for implementing automation in job interview grading. In *2020 8th International Conference on Information and Communication Technology (ICoICT)* (pp. 1-4). IEEE.

SAPUTRA, N., ADJI, T.B., dan PERMANASARI, A.E., 2015. Analisis sentimen data presiden Jokowi dengan preprocessing normalisasi dan stemming menggunakan metode naive bayes dan SVM. Jurnal Dinamika Informatika, 5(1).

SEPTIANDRI, A. A., dan WINATMOKO, Y. A., 2020. Ukara 1.0 challenge track 1: automatic short-answer scoring in bahasa indonesia. arXiv preprint arXiv:2002.12540.