Penilaian Esai Pendek Otomatis Berdasarkan Similaritas Semantik dengan SBERT

Semantic Similarity-Based Short Essay Auto Scoring using SBERT

Nurul Chamidah[1], Mayanda Mega Santoni[2], Helena Nurramdhani Irmanda[3], Ria Astriratma[4], Yulnelly[5]

[1,2,3,4,5] Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jakarta

E-mail: [1]nurul.chamidah@upnvj.ac.id, [2]megasantoni@upnvj.ac.id, [3]helenairmanda@upnvj.ac.id, [4]astriratma@upnvj.ac.id, [5]yulnelly@upnvj.ac.id

Abstrak

Ujian dalam bentuk soal esai dianggap lebih baik dalam mengukur pemahaman dari pada soal berbentuk pilihan. Namun, jawaban esai memerlukan waktu dan tenaga lebih banyak untuk dievaluasi dan sering terjadi inkonsistensi. Maka dari itu, diperlukan suatu sistem penilaian esai otomatis yang dapat membantu evaluator dalam memberikan nilai dengan lebih cepat dan lebih konsisten. Penelitian ini bertujuan untuk mengevaluasi performa model penilaian esai otomatis dimana teks esai jawaban uji dan kunci jawaban dibandingkan secara semantik untuk mengetahui seberapa besar persamaan antara teks jawaban uji dan kunci jawaban. Semantik dari teks esai diperoleh dengan melakukan *word embeddings* dengan memanfaatkan model bahasa *pretrained* Siamese-BERT (SBERT) yang mentransformasi teks esai menjadi vektor sepanjang 512. Proses penilaian esai otomatis ini dimulai dengan praproses pada teks dengan menerapkan *case folding*, berikutnya *word embeddings* pada teks yang telah di praproses dengan SBERT. Vektor numerik dari kunci jawaban dan jawaban uji hasil *word embeddings* kemudian dibandingkan dengan *Cosine Similarity* untuk mendapatkan similaritas semantik sekaligus nilai esai yang merupakan output model. Evaluasi model penilaian esai otomatis ini dilakukan dengan membandingkan nilai dari model dengan nilai dari evaluator manusia. Pengukuran yang dipakai untuk mengukur performa penilaian esai otomatis ini adalah adalah dengan menghitung *Mean Absolute Error* (MAE) dan *Pearson Correlation*, dimana hasil penelitian ini menunjukan nilai rata-rata MAE sebesar 0.26 dan rata-rata korelasi sebesar 0.78.

Kata kunci: Penilaian Esai Otomatis, SBERT, Semantik, Esai.

Abstract

Essay questions are considered better in measuring students' understanding than multiple choice questions. However, essay answers require more time and effort to evaluate and there are often inconsistencies. Therefore, we need an automated essay scoring system that can assist evaluators in providing grades faster and more consistently. This study aims to evaluate the performance of the automatic essay scoring model where the students' answer essay and evaluator answer key are compared semantically to find out how similar is between the students' answer and evaluator's answer key. The semantics of the essay text are obtained by word embeddings using the Siamese-BERT (SBERT) pretrained language model which transforms the essay text into a 512-length vector. This automatic essay scoring process begins with preprocessing the text by applying case folding, then word embeddings on the text that has been preprocessed with SBERT. The numerical vectors of the answer keys and the test answers of the word embeddings are then compared with Cosine Similarity to obtain the semantic similarity as well as the essay skor which is the output of the model. The evaluation of this automatic essay scoring model is done by comparing model skor output with human evaluator skor. The metrics used to measure the performance of this automatic essay scoring is to calculate the Mean Absolute Error (MAE) and Pearson Correlation, where the results of this study show an average MAE value of 0.26 and an average correlation of 0.78.

Keywords: Automatic Essay Scoring, SBERT, Semantic, Essay.

1. PENDAHULUAN

Soal ujian pada umumnya dapat berbentuk pilihan seperti pilihan ganda, benar salah, dan mencocokkan dimana penjawab memilih jawaban yang paling tepat dari pilihan yang ada. Selain berbentuk pilihan, terdapat juga soal ujian berbentuk esai dimana penjawab harus menuliskan sendiri jawabannya. Soal pilihan jauh lebih mudah dinilai secara objektif dari pada soal esai. Namun, soal pilihan sulit digunakan untuk mengukur tingkat pemahaman yang memerlukan penjelasan [1].

Soal ujian berbentuk esai dianggap lebih tepat dipakai untuk mengukur tingkat pemahaman dan pengetahuan, tapi lebih sulit dinilai karena harus dilakukan secara manual. Oleh karena itu, jawaban ujian esai memerlukan waktu yang lebih lama untuk dinilai daripada jawaban soal pilihan. Soal esai biasanya memiliki kunci jawaban atau rubrik penilaian yang digunakan sebagai standar acuan memberikan nilai. Namun, seringkali terjadi perbedaan atau inkonsistensi dalam memberikan nilai terhadap suatu jawaban karena penilaian manusia bersifat subjektif [2], [3]. Karena proses penilaian manusia membutuhkan banyak waktu, tenaga, dan tidak selalu objektif, maka diperlukan sistem penilaian esai otomatis yang dapat mengurangi biaya, waktu dan menentukan skor yang lebih akurat, objektif dan konsisten.

Banyak sistem penilaian esai otomatis telah dikembangkan, dimana sistem memanfaatkan *Natural Language Processing* (NLP) untuk menganalisis teks secara otomatis dan memberikan nilai pada teks esai. Biasanya, model penilaian esai otomatis memanfaatkan berbagai fitur linguistik baik berupa teks secara langsung missal dari kata yang terdapat pada teks, maupun fitur yang merupakan hasil training dari algoritma *Artificial Neural Network* (ANN) maupun *deep learning* [4]. Pendekatan deep learning menunjukkan bahwa pendekatan ANN pada sistem penilaian esai otomatis telah mencapai hasil yang baik [5], [6] dan menggunakan fitur yang dipelajari secara otomatis dari data.

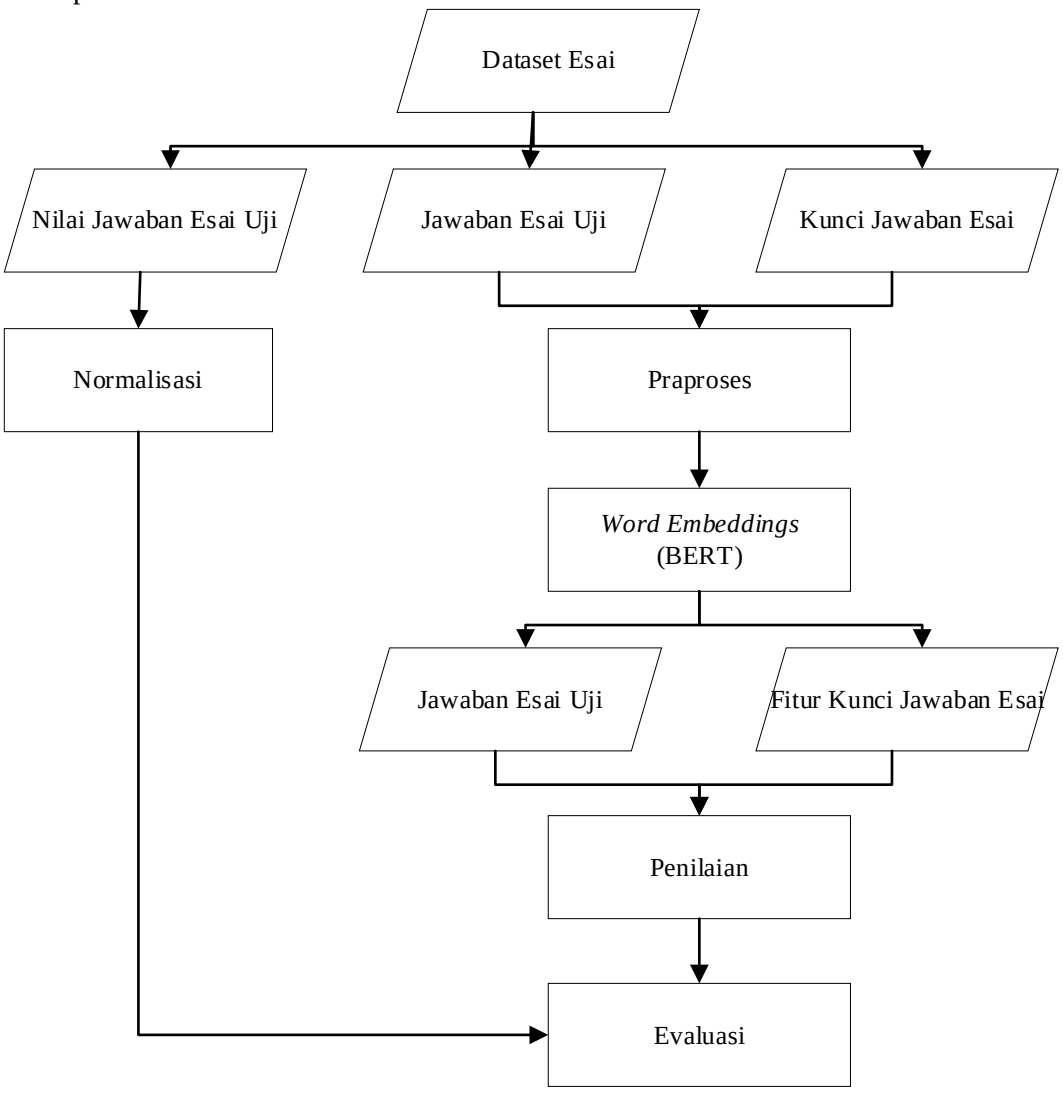
Model penilaian esai otomatis telah banyak dilakukan seperti dengan *Latent Semantic Analysis* (LSA) seperti pada penelitian [2], [7] dimana teks kunci jawaban dan jawaban uji dipetakan kedalam matriks. Penelitian [8] menggunakan *Vector Space Model*, dimana kata-kata dalam teks di bobotkan dengan teknik *Term Frequency-Inverse Document Frequency* (TF-IDF) kemudian nilai diberikan berdasarkan similaritas antara vektor kunci jawaban dengan vektor jawaban uji dengan *Cosine Similarity*. Selain *Cosine Similarity*, *Manhattan distance* dan *dice similarity* juga dipakai untuk menghitung similaritas antara teks esai uji dan kunci jawaban [9]. [10] memanfaatkan fitur NLP untuk penilaian esai otomatis. Penelitian-penelitian tersebut menggunakan fitur yang ada pada teks esai untuk penilaian esai otomatis.

Selain menggunakan fitur yang ada pada teks esai, penelitian-penelitian yang berdasarkan ekstraksi fitur dimana fitur diperoleh dengan belajar dari data, atau disebut dengan fitur *pretrained*. Penelitian [11] memanfaatkan fitur pretrained Word2Vec sebagai *word embeddings* yang mentransformasi teks menjadi numerik. Penelitian [12] menggunakan model bahasa BERT yang merupakan hasil pelatihan model deep learning sebagai vektor untuk *word embeddings*. BERT merupakan model bahasa *pretrained* yang dirilis oleh tim Google AI Language, telah mencapai hasil yang sesuai dalam berbagai kasus NLP [13]. BERT telah diterapkan pada penilai esai otomatis sejak 2019 [14] dan memberikan akurasi yang baik. [15] Mengembangkan model bahasa BERT menjadi *Sentence*-BERT (SBERT), yakni fitur yang dibangun dengan Siemese LSTM. Vektor SBERT memiliki makna semantik dari teks dan telah diterapkan untuk penilaian esai berbahasa Inggris.

Berdasarkan penelitian-penelitian sebelumnya, maka dalam ini kami bertujuan mengevaluasi model penilaian esai otomatis dengan fitur yang memanfaatkan model *pretrained* SBERT sebagai ekstraksi fitur atau *word embeddings* untuk teks esai baik kunci jawaban maupun esai jawaban uji dalam Bahasa Indonesia. Kemudian nilai dari esai diperoleh dengan menghitung similaritas antara vektor jawaban uji dengan vektor kunci jawaban. Penelitian ini dilakukan dengan dataset berbahasa Indonesia pada jawaban dan kunci jawaban ujian mata kuliah Pengantar Basis Data.

2. METODE PENELITIAN

Gambar 1 menunjukkan metode penelitian yang digunakan dalam penelitian ini. Dataset terdiri dari teks jawaban esai uji yang merupakan jawaban, teks kunci jawaban dari soal esai, serta nilai yang terdiri dari nilai maksimum soal dan nilai dari jawaban esai yang merupakan hasil penilaian dari manusia. Data teks kunci jawaban serta jawaban esai uji di praproses dengan case folding serta penghilangan karakter non-alfabet. Teks yang telah dipraproses selanjutnya dilakukan *word embeddings* dengan fitur BERT yakni vektor sepanjang 512. Selanjutnya, fitur dari jawabn esai dan kunci jawaban dilihat similaritas vektornya dengan *Cosine Similarity* untuk menghasilkan nilai. Evaluasi dilakukan dengan membandingkan nilai hasil similaritas dengan nilai jawaban esai. Jika nilai similaritas semakin mendekati nilai jawaban, maka semakin baik performa dari penilaian esai otomatis.



Gambar 1. Metode Penelitian

2.1. Dataset

Dataset diperoleh dari mata kuliah Pengantar Basis Data yang terdiri dari 4 soal esai pendek dikerjakan oleh 36 mahasiswa. Sehingga, total terdapat 144 jawaban esai uji dan nilai uji. Setiap soal memiliki kunci jawaban serta nilai maksimum jika jawaban benar. Gambar 2 menunjukkan contoh dataset dari salah satu jawaban esai uji. Dataset terdiri dari soal dengan nilai

maksimum tiap soal, kunci jawaban, jawaban uji yang merupakan jawaban dari siswa, serta nilai uji yang merupakan penilaian dari manusia terhadap jawaban uji.

Soal (Nilai =4)	Pada kunci relasi dikenal dua aturan yang berlaku yaitu integritas entity dan integritas referential. Jelaskan perbedaan kedua integritas tersebut!
Kunci Jawaban	Integritas entity mendefinisikan sebuah baris sebagai sebuah entitas yang unik untuk suatu tabel. Entity integritas memaksa integritas dari kolom atau primary key dari suatu tabel (melalui index, unique, constraint, primary key) tidak boleh null. Integritas referential memastikan bahwa seluruh nilai dari foreign key cocok dengan nilai primary key yang dihubungkan.
Jawaban Uji	1. INTEGRITAS ENTITY = Mengharuskan setiap entitas memiliki primary key, tabel harus memiliki kolom atau kumpulan kolom untuk menyediakan pengidentifikasian unik untuk baris dan tabel 2. INTEGRITAS REFERENTIAL = Dasar relasi antara foreign key dan primary key. Integritas referensial memastikan bahwa seluruh nilai dari foregin key cocok dengan nilai primary key yang dihubungkanya.
Nilai Uji	3

Gambar 2. Contoh Dataset

2.2. Praproses

Praproses pada teks jawaban uji dan kunci jawaban dilakukan dengan *case folding* untuk mengubah semua huruf alfabet menjadi huruf kecil dan menghilangkan karakter-karakter non-alfabet [16]. Contoh hasil praproses dari teks Gambar 2 dapat dilihat pada Gambar 3. Dapat dilihat pada Gambar 2 dan Gambar 3, praproses teks kunci jawaban antara lain mengubah huruf besar menjadi huruf kecil seperti kata “Integritas” menjadi “integritas”, dan “Entity” menjadi “Entity”. Sedangkan pada teks jawaban uji seperti kata “INTEGRITAS” menjadi “integritas”, “ENTITY” menjadi “entity”, “REFERENTIAL” menjadi “referential”, dst. Juga menghapus karakter non-alfabet seperti tanda titik (.), koma (,), sama dengan (=), dan lain-lain.

Kunci Jawaban	Jawaban Uji
integritas entity mendefinisikan sebuah baris sebagai sebuah entitas yang unik untuk suatu tabel entity integritas memaksa integritas dari kolom atau primary key dari suatu tabel melalui index unique constraint primary key tidak boleh null integritas referential memastikan bahwa seluruh nilai dari foreign key cocok dengan nilai primary key yang dihubungkan	integritas entity mengharuskan setiap entitas memiliki primary key tabel harus memiliki kolom atau kumpulan kolom untuk menyediakan pengidentifikasian unik untuk baris dan tabel integritas referential dasar relasi antara foreign key dan primary keyintegritas referensial memastikan bahwa seluruh nilai dari foregin key cocok dengan nilai primary key yang dihubungkanya

Gambar 3. Contoh Hasil Praproses pada Kunci Jawaban dan Jawaban Uji

2.3. Word Embeddings

Proses yang dipakai untuk mengubah teks menjadi numerik pada penelitian ini dilakukan dengan *word embedding* BERT. BERT didefinisikan sebagai jaringan transformator dua arah multilayer [17].Transformer adalah arsitektur jaringan saraf yang dirancang untuk menangani urutan data yang berurutan menggunakan mekanisme *attention*. Model BERT yang dipakai dalam penelitian ini merupakan arsitektur Siamese-BERT (SBERT) yang merupakan bagian dari *Hugging Face Transformers Library* [15] dimana perbedaan SBERT dengan BERT adalah SBERT memiliki makna semantik yang vektornya dapat dibandingkan menggunakan *Cosine Similarity*. Panjang vektor untuk fitur SBERT adalah 512. Contoh Word Embedding dengan SBERT dari hasil praproses Gambar 3 dapat dilihat pada Tabel 1.

Tabel 1. Contoh Hasil Word Embedding dengan SBERT

	1	2	3	4	…	510	511	512
Vektor kunci jawaban	-0.055	1680395.000	-584515.000	-1830183.000	…	-3926777.000	-3243494.000	-0.051
Vektor jawaban uji	-0.029	4938433.690	-1910702.330	-670217.490	…	-3356042.500	-2905505.340	-0.049

Dapat dilihat pada Tabel 1, masing-masing teks jawaban uji dan kunci jawaban ditransformasikan kedalam vektor sepanjang 512 dimana vektor ini mengandung makna semantik yang dapat dibandingkan. Kedekatan semantik dua vektor ini akan dibahas pada proses penilaian.

2.4. Penilaian

Sebagaimana pada penelitian [15] bahwa vektor SBERT memiliki makna semantik yang dapat dibandingkan. Pada penelitian ini dilakukan perhitungan similaritas antara vektor jawaban uji dengan kunci jawaban untuk mengetahui seberapa besar persamaan semantik antara dua vektor tersebut. Semakin sama dua vektor, berarti semakin besar pula similaritas semantiknya yang artinya semakin sama antara kunci jawaban dengan jawaban uji dan berpengaruh pada nilai yang semakin besar. Sebaliknya, bila similaritasnya semakin kecil, berarti jawaban uji dan kunci jawaban semakin berbeda dan nilai akan semakin kecil. Similaritas antara vektor jawaban ujian kunci jawaban pada penelitian ini menggunakan *Cosine Similarity*. *Cosine Similarity* menghitung L2- Norm dot product dari dua vektor [18]. Jika x dan y adalah barus vektor, maka *Cosine Similarity cossim* dapat dilihat pada persamaan 1.

$$cossim(x,y) = \frac{(xy^{\top})}{(\|x\|\|y\|)} \quad (1)$$

Dimana x adalah vektor jawaban uji dan y adalah vektor kunci jawaban. Semakin mirip dua vektor, maka semakin besar nilai similaritasnya. Nilai *Cosine Similarity* pada vektor jawaban uji dan kunci jawaban pada Tabel 1 adalah 0.846534 yang artinya memiliki similaritas yang cukup tinggi, atau dapat diartikan dengan model penilaian esai otomatis, jawaban tersebut memiliki nilai 84.6% dari nilai maksimal.

2.5. Normalisasi

Normalisasi dilakukan pada data nilai uji. Normalisasi dilakukan karena nilai maksimum untuk setiap soal berbeda-beda sehingga diperlukan normalisasi agar range nilai sama. Normalisasi dilakukan dengan membandingkan nilai uji suatu jawaban dengan nilai maksimum dari jawaban soal tersebut. Sehingga, setelah normalisasi range nilai uji adalah antara nol hingga satu. Normalisasi *min-max* [19] digunakan untuk menormalisasi nilai uji. Teknik *min-max* digunakan dalam penelitian ini untuk menyamakan range nilai uji. Yakni, nilai 0 jika jawaban salah total dan nilai maksimum berupa nilai soal. Dengan range nilai yang pasti ini, maka setelah normalisasi nilai uji minimum adalah 0 dan nilai maksimum adalah 1. Normalisasi *min-max* untuk nilai uji dapat dilihat pada persamaan 2.

$$n_{baru} = \frac{n_{uji}}{n_{soal}} \quad (2)$$

Dimana n_{baru} merupakan nilai uji hasil normalisasi, n_{uji} adalah nilai uji dari jawaban yang diberikan oleh evaluator manusia, dan n_{soal} merupakan nilai maksimum soal jika jawaban benar. Nilai n_{baru} hasil normalisasi ini digunakan dalam evaluasi performa sebagai nilai pembanding dari nilai model. Semakin kecil selisih antara nilai baru dengan nilai model, maka semakin baik performa model. Dapat dilihat pada contoh Gambar 2, nilai soal $n_{soal} = 4$, dan nilai uji $n_{uji} = 3$, maka nilai baru n_{baru} hasil normalisasi adalah 3/4 = 0.75.

2.6. Evaluasi

Evaluasi dilakukan untuk mengetahui seberapa baik performa model penilaian esai otomatis. Performa ini diukur dengan *Mean Absolute Error* (MAE) dan *Pearson Correlation.* MAE digunakan untuk mengukur rata-rata selisih nilai yang dihasilkan model penilaian esai jika dibandingkan dengan nilai yang diberikan oleh evaluator manusia. Sedangkan *Pearson Correlation* untuk mengetahui seberapa sesuai nilai dari model dengan nilai dari evaluator.

Persamaan *Mean Absolute Error* (MAE) dapat dilihat pada persamaan 3 [20]. MAE didefinisikan sebagai selisih antara nilai output model dengan nilai uji yang diberikan oleh evaluator manusia.

$$MAE = \frac{\sum|n_{sistem} - n_{uji}|}{Jum} \quad (3)$$

Dengan n_{model} berupa nilai output dari model, n_{uji} berupa nilai uji yang merupakan hasil penilaian dari evaluator, dan Jum merupakan jumlah total jawaban yang dievaluasi.

Korelasi nilai output dari model dengan nilai uji dihitung dengan *Pearson Correlation*. Korelasi ini dipakai untuk mendapatkan tingkat kesesuaian atau kesepakatan antara nilai uji yang diberikan oleh evaluator manusia dengan nilai output dari model penilaian esai otomatis. Persamaan *Pearson Correlation* dapat dilihat pada persamaan 4 korelasi *Pearson* didefinisikan sebagai perbandingan antara kovarian dengan perkalian standar deviasi dari nilai uji dan nilai output model penilaian esai otomatis.

$$corr(n_{model}, n_{uji}) = \frac{cov(n_{model},\ n_{uji})}{stdev(n_{model}).stdev(n_{uji})} \quad (4)$$

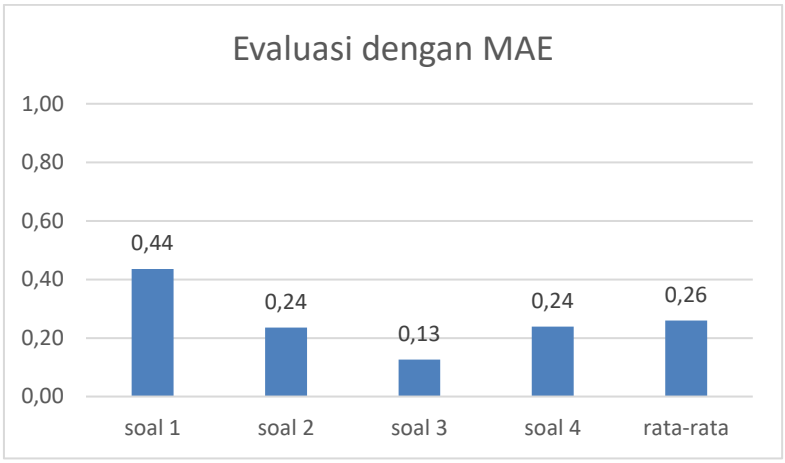
Dengan $corr$ adalah nilai korelasi dan cov adalah kovarian antara nilai output model penilaian esai otomatis dengan nilai uji, dan $stdev$ adalah standar deviasi untuk masing-masing nilai, baik nilai dari output model dan nilai uji. Kriteria kesuksesan model dibagi menjadi kurang, baik, dan sangat baik. kriteria korelasi tergolong kurang jika korelasi < 0.4, dikatakan baik jika nilai korelasi antara 0.4 hingga 0.75, dan sangat baik jika nilai korelasi > 0.75 [18].

3. HASIL DAN PEMBAHASAN

Penelitian dilakukan terhadap dataset yang terdiri dari 4 soal dan 4 kunci jawaban dengan jawaban uji merupakan hasil ujian oleh 36 orang mahasiswa. Praproses untuk data teks jawaban dan kunci jawaban dilakukan dengan *case folding* serta menghilangkan karakter-karakter selain huruf. Teks hasil praproses selanjutnya di embedding dengan vektor SBERT yang memiliki makna semantik dimana proses ini mengubah teks jawaban dan kunci jawaban menjadi vektor sepanjang 512. Vektor hasil *word embeddings* antara jawaban dan kunci jawaban selanjutnya di bandingkan dengan meilihat similaritasnya, yakni dengan *Cosine Similarity*. Semakin besar similaritas antara vektor jawaban dan kunci jawabannya, berarti jawaban semakin mendekati atau semakin sama dengan kunci jawaban secara semantik. Nilai similaritas ini merupakan nilai yang diberikan oleh model penilaian otomatis dalam persentase.

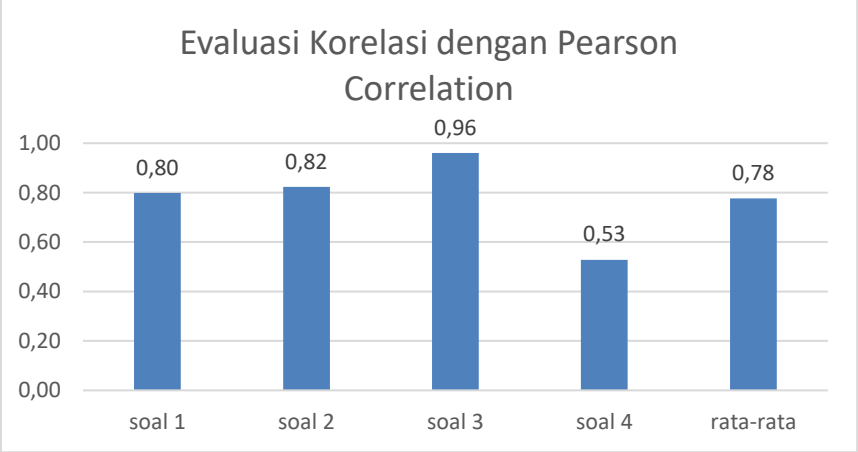
Nilai dari output model selanjutnya dievaluasi dengan *Mean Absolute Error* (MAE). Nilai uji dari dataset dinormalisasi kedalam range 0 -1 agar nilai uji dan nilai output model berada pada rentang yang sama. Hasil evaluasi dengan MAE pada nilai uji dengan nilai output model dapat dilihat pada Gambar 4.

Gambar 4 menunjukkan hasil evaluasi MAE antara nilai uji dengan nilai output model. Dari hasil MAE tersebut dapat dilihat bahwa nilai yang berasal dari output model yang memanfaatkan SBERT sebagai *word embeddings* yang memiliki makna semantik untuk soal 1 memiliki MAE 0.44, soal 2 sebesar 0.24, soal 3 sebesar 0.13, dan soal 4 memiliki MAE 0.24 dengan rata-rata untuk seluruh soal adalah 0.26.



Gambar 4. Evaluasi dengan Mean Absolute Error (MAE)

Evaluasi korelasi dengan *Pearson Correlation* antara nilai uji dan nilai dari output model dapat dilihat pada Gambar 5. Dari Gambar 5 dapat disimpulkan bahwa nilai output model memiliki korelasi yang sangat baik dengan nilai uji untuk soal 1 hingga 3. Sedangkan soal 4 memiliki korelasi yang baik dan rata-rata untuk seluruh penilaian memiliki korelasi yang sangat baik. Hal ini menunjukkan korelasi antara penilaian dengan model yang diusulkan dan nilai uji yang merupakan hasil penilaian manusia memiliki korelasi sangat baik.



Gambar 5. Evaluasi Korelasi dengan *Pearson Correlation*

4. KESIMPULAN DAN SARAN

Similaritas semantik pada penilaian esai otomatis dalam penelitian ini memanfaatkan vektor dari *word embeddings SBERT* yang dibandingkan dengan *Cosine Similarity*. Pemrosesan teks jawaban uji dan kunci jawaban dilakukan dengan melakukan case folding, kemudian menerapkan *word embeddings* SBERT yang mentransformasi teks menjadi vektor numerik sepanjang 512. Vektor hasil embeddings dari teks kunci jawaban dan jawaban uji selanjutnya dibandingkan dengan *Cosine Similarity* untuk melihat seberapa besar similaritas antara dua vektor tersebut. Similaritas yang semakin besar menunjukkan semakin mirip pula teks kunci jawaban dengan jawaban uji secara semantik.

Evaluasi nilai hasil output model penilaian esai pendek otomatis dengan *Pearson Correlation* menunjukkan bahwa secara keseluruhan penggunaan SBERT sebagai *word embeddings* pada penilaian esai otomatis memiliki korelasi yang sangat baik antara nilai output sistem dengan nilai dari evaluator manusia. Namun, *error* atau kesalahan sistem penilaian masih

cukup besar yakni dengan rata-rata keseluruhan mencapai 26% yang artinya terdapat rata-rata selisih nilai model dan nilai manusia sebesar 26%.

Pengembangan penialain esai pendek otomatis berdasarkan similaritas semantik masih diperlukan, dengan melakukan pelatihan ulang pada model SBERT dengan dataset yang lebih besar atau menggunakan materi dari esai yang diberikan dengan harapan vector yang dihasilkan oleh model SBERT lebih mendekati konteks materi esai yang diberikan.

DAFTAR PUSTAKA

[1] H. Rababah and A. T. Al-Taani, "An automated scoring approach for Arabic short answers essay questions," in *ICIT 2017 - 8th International Conference on Information Technology, Proceedings*, Oct. 2017, pp. 697–702. doi: 10.1109/ICITECH.2017.8079930.

[2] R. Adhitia and A. Purwarianti, "PENILAIAN ESAI JAWABAN BAHASA INDONESIA MENGGUNAKAN METODE SVM - LSA DENGAN FITUR GENERIK," *Jurnal Sistem Informasi*, vol. 5, no. 1, p. 33, Jul. 2012, doi: 10.21609/jsi.v5i1.260.

[3] X. Peng, D. Ke, and B. Xu, "Automated Essay Scoring Based on Finite State Transducer: towards ASR Transcription of Oral English Speech," pp. 8–14, 2012.

[4] Y. Farag, H. Yannakoudakis, and T. Briscoe, "Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input," *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 263–271, Apr. 2018, doi: 10.48550/arxiv.1804.06898.

[5] E. B. Page, "Computer grading of student prose, using modern concepts and software," *The Journal of experimental education*, vol. 62, no. 2, pp. 127–142, 1994.

[6] S. Valenti, S. Valenti, F. Neri, and A. Cucchiarelli, "An Overview of Current Research on Automated Essay Grading," *Journal of Information Technology Education: Research*, vol. 2, no. 1, pp. 319–330, 2003.

[7] R. B. Aji, Z. A. Baisal, and Y. Firdaus, "Automatic Essay Grading System Menggunakan Metode Latent Semantic Analysis E-78 E-79," *Seminar Nasional Aplikasi Teknologi Informasi*, vol. 2011, no. Snati, pp. 1–9, 2011.

[8] J. Zeniarja, A. Salam, and I. Achsanu, "Sistem Koreksi Jawaban Esai Otomatis (E-Valuation) dengan Vector Space Model pada Computer Based Test (CBT)," *Seri Prosiding Seminar Nasional Dinamika Informatika*, vol. 4, no. 1, Apr. 2020, Accessed: Aug. 04, 2020. [Online]. Available: http://prosiding.senadi.upy.ac.id/index.php/senadi/article/view/134

[9] F. Rahutomo, Y. P. Putra, and M. H. Ali, "Implementasi Manhattan Distance dan Dice Similarity pada Ujian Esai Daring Berbahasa Indonesia," *Seminar Informatika Aplikatif Polinema*, pp. 171–174, 2019, Accessed: Aug. 04, 2020. [Online]. Available: http://jurnalti.polinema.ac.id/index.php/SIAP/article/view/373

[10] N. Chamidah, M. Mega Santoni, H. Nurramdhani Irmanda, R. Astriratma, F. Ilmu Komputer, and U. Pembangunan Nasional Veteran Jakarta, "Penilaian Esai Pendek Otomatis dengan Pencocokan Kata Kunci Frasa Nomina," *Techno.Com*, vol. 20, no. 4, pp. 489–498, Nov. 2021, doi: 10.33633/tc.v20i4.5043.

[11] F. Ginter and J. Kanerva, "Fast training of word2vec representations using n-gram corpora," 2014.

[12] M. Beseiso and S. Alzahrani, "An empirical analysis of BERT embedding for automated essay scoring," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 10, 2020.

[13] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, pp. 4171–4186, Oct. 2018, doi: 10.48550/arxiv.1810.04805.

[14] P. U. Rodriguez, A. Jafari, and C. M. Ormerod, "Language models and Automated Essay Scoring," 2019.

[15] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 3982–3992, Aug. 2019, doi: 10.48550/arxiv.1908.10084.

[16] F. Pratama, "Rancang Bangun Aplikasi Peringkas Tkes Otomatis Artikel Berbahasa Indonesia Mengunakan Metode Term Frequency Inverse Document Frequency (TF-IDF) dan K-mean Clustering," *Fakultas Sains dan Teknologi*, Apr. 2014.

[17] A. Vaswani *et al.*, "Attention is All you Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[18] T. F. de C. Marshall and J. L. Fleiss, "Statistical Methods for Rates and Proportions.," *The Statistician*, vol. 25, no. 1, p. 70, 1976, doi: 10.2307/2988144.

[19] Jiawei. Han, Micheline. Kamber, and Jian. Pei, *Data Mining: Concepts and Techniques*. Elsevier Science, 2012. doi: 10.1016/C2009-0-61819-5.

[20] G. Brassington, "Mean absolute error and root mean square error: which is the better metric for assessing model performance?," *Geophysical Research Abstracts*, vol. 19, pp. 2017–3574, 2017.