Analisis Komparatif dan Integrasi Model *Sentence Embedding* untuk Tugas *Semantic Textual Similarity*

1st Akhtar Muhammad Achsan
Prodi S1 Informatika
Fakultas Informatika
Bandung, Indonesia
akhtarmuha@student.telkomuniversity.ac.id

2nd Mochammad Arif Bijaksana
Prodi S1 Informatika
Fakultas Informatika
Bandung, Indonesia
arifbijaksana@telkomuniversity.ac.id

3rd Kemas Muslim Lhaksmana
Prodi S1 Informatika
Fakultas Informatika
Bandung, Indonesia
kemasmuslim@telkomuniversity.ac.id

***Abstrak* — *Semantic Textual Similarity* (STS) merupakan salah satu tugas penting dalam bidang *Natural Language Processing* (NLP) yang bertujuan mengukur tingkat kesamaan makna antara dua kalimat. Perkembangan model sentence embedding berbasis transformer telah meningkatkan kemampuan representasi semantik, namun setiap model memiliki karakteristik dan keunggulan yang berbeda sehingga belum terdapat satu model yang secara konsisten unggul pada seluruh skenario. Penelitian ini bertujuan melakukan analisis komparatif terhadap empat model *sentence embedding*, yaitu Sentence-BERT (SBERT), SimCSE, Language-agnostic BERT Sentence Embedding (LaBSE), dan Universal Sentence Encoder (USE), serta mengevaluasi pengaruh integrasi model melalui pendekatan *model fusion* berbasis *vector concatenation*. Eksperimen dilakukan menggunakan dataset STS Benchmark (STS-B) dan SICK dengan proses evaluasi berdasarkan metrik Pearson Correlation, Spearman Correlation, Mean Absolute Error (MAE), dan Mean Squared Error (MSE). Hasil penelitian menunjukkan bahwa SBERT menghasilkan performa terbaik pada kategori model tunggal di kedua dataset. Selain itu, pendekatan *model fusion* mampu meningkatkan kualitas representasi semantik dibandingkan sebagian besar model tunggal. Kombinasi SBERT dan SimCSE dengan normalisasi menghasilkan nilai Spearman Correlation tertinggi, yaitu 0,89384 pada STS-B dan 0,80028 pada SICK. Sementara itu, kombinasi SBERT dan USE menghasilkan nilai kesalahan prediksi terendah. Temuan ini menunjukkan bahwa integrasi embedding dari model yang memiliki karakteristik berbeda dapat menghasilkan representasi semantik yang lebih kaya, robust, dan efektif untuk tugas pengukuran kesamaan teks.**

***Kata kunci*— semantic textual similarity, sentence embedding, model fusion, SBERT, pemrosesan bahasa alami**

I. PENDAHULUAN

Perkembangan teknologi *Natural Language Processing* (NLP) dalam beberapa tahun terakhir telah mendorong peningkatan kemampuan sistem komputer dalam memahami dan merepresentasikan makna bahasa alami. Salah satu permasalahan fundamental dalam NLP adalah *Semantic Textual Similarity* (STS), yaitu tugas untuk mengukur tingkat kesamaan makna antara dua kalimat atau teks. STS memiliki peran penting dalam berbagai aplikasi, seperti semantic search, question answering, information retrieval, text clustering, dan sistem rekomendasi berbasis teks. Keberhasilan sistem STS sangat bergantung pada kualitas representasi semantik yang digunakan untuk merepresentasikan kalimat ke dalam bentuk numerik yang dapat diproses oleh mesin.

Perkembangan arsitektur transformer telah menghasilkan berbagai model *sentence embedding* yang mampu menghasilkan representasi semantik dengan kualitas yang semakin baik. Beberapa model yang banyak digunakan dalam penelitian STS antara lain Sentence-BERT (SBERT), SimCSE, Language-agnostic BERT Sentence Embedding (LaBSE), dan Universal Sentence Encoder (USE). SBERT memanfaatkan arsitektur Siamese Network untuk menghasilkan representasi kalimat yang efisien pada tugas kesamaan semantik, sedangkan SimCSE menerapkan pendekatan contrastive learning untuk meningkatkan kualitas representasi melalui pembelajaran hubungan antar kalimat. LaBSE dikembangkan untuk menghasilkan representasi semantik lintas bahasa dalam ruang embedding yang sama, sementara USE menawarkan pendekatan yang lebih ringan dengan tetap mempertahankan kemampuan representasi yang baik. Perbedaan arsitektur, strategi pelatihan, dan sumber data menyebabkan setiap model memiliki karakteristik yang berbeda dalam menangkap informasi semantik.

Berbagai penelitian terdahulu menunjukkan bahwa tidak terdapat satu model *sentence embedding* yang secara konsisten memberikan performa terbaik pada seluruh dataset dan domain teks. Sebagian model menunjukkan performa yang unggul pada dataset tertentu, namun mengalami penurunan performa ketika diterapkan pada karakteristik data yang berbeda. Kondisi ini menunjukkan bahwa setiap model cenderung menangkap aspek semantik yang berbeda sesuai dengan mekanisme pembelajaran yang digunakan. Oleh karena itu, pendekatan yang hanya mengandalkan satu model berpotensi membatasi kemampuan representasi semantik yang dihasilkan.

Salah satu pendekatan yang mulai banyak diteliti untuk mengatasi keterbatasan tersebut adalah *model fusion* atau *meta-embedding*, yaitu teknik yang menggabungkan representasi dari beberapa model *sentence embedding* menjadi satu representasi yang lebih kaya. Pendekatan ini didasarkan pada asumsi bahwa informasi semantik yang ditangkap oleh masing-masing model bersifat saling melengkapi sehingga kombinasi beberapa model berpotensi menghasilkan representasi yang lebih komprehensif dibandingkan model tunggal. Meskipun demikian, kajian mengenai efektivitas integrasi berbagai model *sentence embedding* modern, khususnya kombinasi SBERT, SimCSE, LaBSE, dan USE pada tugas STS, masih relatif terbatas dan belum banyak dianalisis secara komprehensif pada dataset standar yang berbeda karakteristik.

Berdasarkan permasalahan tersebut, penelitian ini melakukan analisis komparatif terhadap empat model *sentence embedding*, yaitu SBERT, SimCSE, LaBSE, dan USE, serta mengevaluasi efektivitas pendekatan *model fusion* berbasis *vector concatenation*. Evaluasi dilakukan menggunakan dua dataset benchmark, yaitu STS Benchmark (STS-B) dan SICK, dengan memanfaatkan metrik Pearson Correlation, Spearman Correlation, Mean Absolute Error (MAE), dan Mean Squared Error (MSE). Penelitian ini diharapkan dapat memberikan pemahaman yang lebih mendalam mengenai karakteristik masing-masing model, mengidentifikasi kombinasi model yang menghasilkan representasi semantik terbaik, serta memberikan kontribusi dalam pengembangan metode representasi kalimat untuk tugas *Semantic Textual Similarity*.

II. KAJIAN TEORI

A. Semantic Textual Similarity

Semantic Textual Similarity (STS) merupakan salah satu permasalahan penting dalam *Natural Language Processing* (NLP) yang bertujuan mengukur tingkat kesamaan makna antara dua kalimat atau teks [1]. Berbeda dengan pendekatan berbasis kesamaan leksikal, STS berfokus pada pemahaman makna semantik secara menyeluruh dengan mempertimbangkan konteks, sinonimi, serta hubungan makna antar kalimat. Dalam implementasinya, STS umumnya direpresentasikan sebagai skor numerik yang menunjukkan tingkat kesamaan semantik antara dua kalimat. Untuk menghitung kesamaan antar representasi kalimat, metode yang paling banyak digunakan adalah cosine similarity karena mampu mengukur kedekatan arah antar vektor tanpa dipengaruhi oleh panjang vektor [1].

B. Sentence Embedding

Sentence embedding merupakan teknik representasi teks yang mengubah kalimat menjadi vektor berdimensi tetap sehingga informasi semantik dapat direpresentasikan dalam bentuk numerik [2]. Perkembangan arsitektur transformer menghasilkan berbagai model *sentence embedding* yang memiliki karakteristik berbeda dalam menangkap makna kalimat.

Sentence-BERT (SBERT) dikembangkan dengan arsitektur Siamese Network untuk menghasilkan representasi kalimat yang dapat dibandingkan secara langsung menggunakan cosine similarity [3]. SimCSE memperkenalkan pendekatan contrastive learning yang menghasilkan embedding lebih diskriminatif melalui optimasi pasangan positif dan negatif selama proses pelatihan [4]. LaBSE dirancang untuk menghasilkan representasi semantik multibahasa sehingga memungkinkan pengukuran kesamaan lintas bahasa dalam satu ruang embedding yang sama [5]. Sementara itu, Universal Sentence Encoder (USE) menyediakan representasi kalimat yang efisien dan dapat digunakan pada berbagai tugas NLP, termasuk klasifikasi teks dan pengukuran kesamaan semantik [6].

Perbedaan arsitektur, strategi pelatihan, dan sumber data menyebabkan setiap model memiliki kemampuan yang berbeda dalam merepresentasikan informasi semantik. Oleh karena itu, performa model dapat bervariasi ketika diterapkan pada dataset dan domain yang berbeda [2].

C. Model Fusion

Keterbatasan model tunggal dalam menangkap seluruh aspek semantik mendorong berkembangnya pendekatan *model fusion* atau *meta-embedding*, yaitu teknik yang menggabungkan representasi dari beberapa model embedding menjadi satu representasi yang lebih kaya [7]. Pendekatan ini didasarkan pada asumsi bahwa setiap model mempelajari karakteristik semantik yang berbeda sehingga informasi yang dihasilkan dapat saling melengkapi.

Pada penelitian ini, metode *fusion* dilakukan menggunakan *vector concatenation*, yaitu menggabungkan seluruh dimensi embedding dari beberapa model menjadi satu vektor berdimensi lebih besar. Pendekatan ini mempertahankan informasi dari masing-masing model tanpa menghilangkan karakteristik representasi yang telah dipelajari sebelumnya [7]. Beberapa penelitian menunjukkan bahwa kombinasi embedding dari berbagai model mampu meningkatkan kualitas representasi semantik dibandingkan penggunaan model tunggal [8], [9].

D. Evaluasi Model

Evaluasi pada tugas STS dilakukan dengan membandingkan skor kesamaan yang dihasilkan model terhadap skor anotasi manusia sebagai ground truth [1]. Metrik utama yang digunakan dalam penelitian ini adalah Pearson Correlation dan Spearman Correlation karena keduanya telah menjadi standar evaluasi pada berbagai penelitian STS [1], [3], [4].

Pearson Correlation digunakan untuk mengukur hubungan linear antara skor prediksi dan skor referensi, sedangkan Spearman Correlation digunakan untuk mengukur konsistensi peringkat antar nilai. Selain metrik korelasi, penelitian ini juga menggunakan Mean Absolute Error (MAE) dan Mean Squared Error (MSE) untuk mengevaluasi tingkat kesalahan prediksi. Kombinasi metrik korelasi dan metrik error memberikan evaluasi yang lebih komprehensif terhadap kualitas representasi semantik yang dihasilkan oleh masing-masing model maupun hasil integrasi model..

III. METODE

Penelitian ini menggunakan pendekatan eksperimen kuantitatif untuk menganalisis performa model *sentence embedding* pada tugas Semantic Textual Similarity (STS) serta mengevaluasi efektivitas pendekatan model fusion dalam meningkatkan kualitas representasi semantik. Alur penelitian secara umum terdiri atas tahap pengumpulan

dataset, pra-pemrosesan data, ekstraksi *sentence embedding*, integrasi model menggunakan *vector concatenation*, perhitungan nilai kemiripan semantik, serta evaluasi performa model.

Dataset yang digunakan terdiri dari STS Benchmark (STS-B) dan SICK yang merupakan dua dataset standar yang banyak digunakan dalam penelitian STS [1], [18], [19]. Dataset STS-B memiliki rentang skor kesamaan 0–5 dan berasal dari berbagai domain seperti berita, forum diskusi, serta deskripsi gambar. Sementara itu, dataset SICK memiliki rentang skor 1–5 dan berfokus pada kalimat deskriptif yang berkaitan dengan aktivitas manusia dan objek sehari-hari. Penggunaan kedua dataset tersebut bertujuan untuk mengevaluasi performa model pada karakteristik data yang berbeda.

Pada tahap pra-pemrosesan, dilakukan penyeragaman format dataset sehingga setiap pasangan data memiliki struktur yang sama, yaitu sentence_a, sentence_b, dan score. Selain itu, dilakukan analisis karakteristik dataset yang meliputi distribusi skor kesamaan, panjang kalimat, serta ukuran kosakata untuk memperoleh gambaran mengenai kompleksitas data yang digunakan dalam eksperimen.

Tahap berikutnya adalah ekstraksi *sentence embedding* menggunakan empat model yang memiliki karakteristik berbeda, yaitu Sentence-BERT (SBERT), SimCSE, Language-agnostic BERT Sentence Embedding (LaBSE), dan Universal Sentence Encoder (USE). Setiap model digunakan untuk menghasilkan representasi vektor dari masing-masing kalimat pada dataset. Embedding yang dihasilkan oleh model tunggal kemudian digunakan sebagai baseline dalam proses evaluasi.

Untuk meningkatkan kualitas representasi semantik, penelitian ini menerapkan pendekatan *weighted fusion* menggunakan metode *vector concatenation*. Metode ini dilakukan dengan menggabungkan embedding dari dua atau lebih model ke dalam satu vektor berdimensi lebih besar. Selain skenario tanpa normalisasi, dilakukan pula normalisasi embedding sebelum proses penggabungan untuk mengurangi perbedaan distribusi antar model dan meningkatkan stabilitas representasi hasil fusi. Eksperimen dilakukan pada berbagai kombinasi model untuk mengidentifikasi konfigurasi yang memberikan performa terbaik.

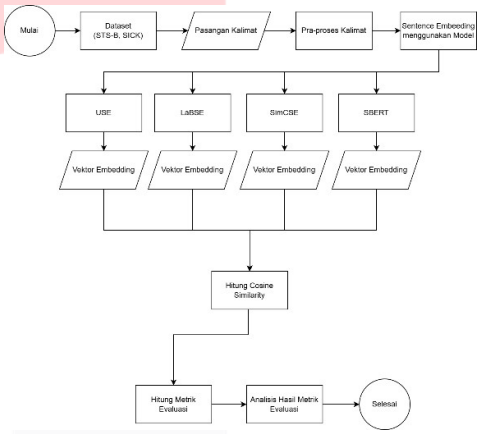
Nilai kesamaan semantik antar pasangan kalimat dihitung menggunakan cosine similarity karena metode ini mampu mengukur kedekatan arah antar vektor secara efektif. Hasil perhitungan cosine similarity kemudian dibandingkan dengan skor anotasi manusia yang terdapat pada dataset untuk mengevaluasi kualitas representasi semantik yang dihasilkan oleh masing-masing model maupun model hasil integrasi.

Evaluasi performa dilakukan menggunakan Pearson Correlation dan Spearman Correlation sebagai metrik utama karena keduanya merupakan standar evaluasi pada tugas STS [1]. Selain itu, digunakan Mean Absolute Error (MAE) dan Mean Squared Error (MSE) sebagai metrik tambahan untuk mengukur tingkat kesalahan prediksi terhadap nilai referensi. Penggunaan kombinasi metrik korelasi dan metrik error bertujuan memberikan evaluasi yang lebih komprehensif terhadap kemampuan model dalam merepresentasikan kesamaan semantik antar kalimat.

A. Alur Penelitian

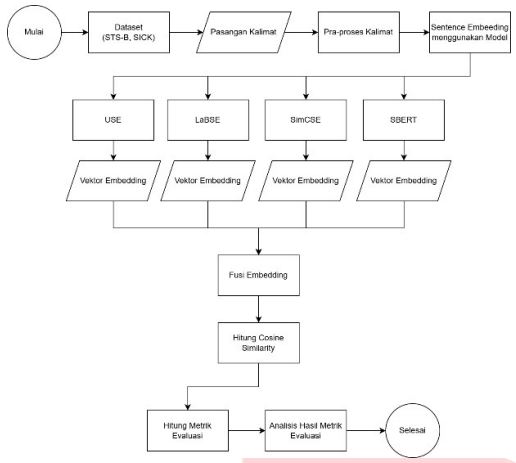
Penelitian ini terdiri atas dua skenario eksperimen, yaitu evaluasi model tunggal (single model) dan evaluasi model hasil integrasi (*multiple model fusion*). Pada skenario pertama, setiap model *sentence embedding* dievaluasi secara independen untuk mengukur kemampuan representasi semantik yang dihasilkan. Pada skenario kedua, embedding dari beberapa model digabungkan menggunakan metode *vector concatenation* untuk memperoleh representasi semantik yang lebih kaya dan komprehensif.

Alur eksperimen model tunggal ditunjukkan pada Gambar 1. Proses dimulai dengan pengambilan dataset STS Benchmark (STS-B) dan SICK yang kemudian melalui tahap pra-pemrosesan. Setiap pasangan kalimat direpresentasikan menggunakan model SBERT, SimCSE, LaBSE, dan USE secara terpisah. Hasil embedding digunakan untuk menghitung nilai kesamaan semantik menggunakan cosine similarity, kemudian dievaluasi menggunakan Pearson Correlation, Spearman Correlation, MAE, dan MSE.



GAMBAR 1

Pada skenario kedua, dilakukan integrasi embedding dari beberapa model menggunakan metode *vector concatenation* sebagaimana ditunjukkan pada Gambar 2. Embedding yang dihasilkan oleh masing-masing model digabungkan menjadi satu representasi vektor berdimensi lebih besar. Selain itu, dilakukan eksperimen dengan normalisasi embedding sebelum proses penggabungan untuk mengurangi perbedaan distribusi antar model. Representasi hasil integrasi kemudian digunakan untuk menghitung nilai kesamaan semantik dan dievaluasi menggunakan metrik yang sama dengan model tunggal.



GAMBAR 2

B. Ekstraksi dan Integrasi Sentence Embedding

Pada tahap ekstraksi fitur, setiap kalimat direpresentasikan ke dalam bentuk vektor menggunakan model SBERT, SimCSE, LaBSE, dan USE. Embedding yang dihasilkan kemudian digunakan baik sebagai representasi model tunggal maupun sebagai masukan pada proses integrasi model.

Metode integrasi yang digunakan dalam penelitian ini adalah *vector concatenation*. Teknik ini menggabungkan seluruh dimensi embedding dari beberapa model menjadi satu vektor representasi baru tanpa menghilangkan informasi dari masing-masing model. Secara matematis, proses integrasi embedding dapat dinyatakan sebagai:

$$E_{fusion} = [\, E_1 \oplus E_2 \oplus \cdots \oplus E_n] \qquad (1)$$

dengan E_{fusion} merupakan embedding hasil integrasi, E_i merupakan embedding yang dihasilkan oleh model ke-i, dan $\oplus$ menyatakan operasi konkatenasi vektor.

C. Perhitungan Similarity

Setelah representasi embedding diperoleh, kesamaan semantik antar pasangan kalimat dihitung menggunakan cosine similarity. Metode ini dipilih karena mampu mengukur kedekatan arah antar vektor tanpa dipengaruhi oleh panjang vektor sehingga sesuai untuk tugas *Semantic Textual Similarity*.

Nilai cosine similarity dihitung menggunakan Persamaan (2).

$$\text{cosine_similarity}\,(\mathrm{A},\mathrm{B}) = \frac{A \cdot B}{\|A\|\|B\|} \qquad (2)$$

dengan (A) dan (B) merupakan vektor embedding dari dua kalimat yang dibandingkan.

Nilai similarity yang diperoleh kemudian dibandingkan dengan skor anotasi manusia yang terdapat pada dataset STS-B dan SICK untuk mengukur kualitas representasi semantik yang dihasilkan oleh masing-masing model.

D. Evaluasi Performa Model

Evaluasi dilakukan menggunakan Pearson Correlation dan Spearman Correlation sebagai metrik utama. Pearson Correlation digunakan untuk mengukur hubungan linear antara skor prediksi dan skor referensi, sedangkan Spearman Correlation digunakan untuk mengukur kesesuaian peringkat antar nilai.

Pearson Correlation dihitung menggunakan Persamaan (3).

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \qquad (3)$$

Sedangkan Spearman Correlation dihitung menggunakan Persamaan (4).

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \qquad (4)$$

Selain metrik korelasi, digunakan Mean Absolute Error (MAE) dan Mean Squared Error (MSE) untuk mengukur tingkat kesalahan prediksi model. MAE dihitung menggunakan Persamaan (5).

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \qquad (5)$$

Sementara itu, MSE dihitung menggunakan Persamaan (6).

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (6)$$

Keempat metrik tersebut digunakan untuk memberikan evaluasi yang komprehensif terhadap performa model tunggal maupun model hasil integrasi dalam merepresentasikan kesamaan semantik antar kalimat.

IV. HASIL DAN PEMBAHASAN

Bagian ini membahas hasil eksperimen yang dilakukan terhadap model tunggal (single model) dan model hasil integrasi (*model fusion*) pada dataset STS Benchmark (STS-B) dan SICK. Analisis difokuskan pada pengaruh karakteristik masing-masing model terhadap performa STS serta efektivitas integrasi embedding dalam meningkatkan kualitas representasi semantik.

A. Hasil Evaluasi Model Tunggal

TABEL 1

Model	Pearson STS-B	Spearman STS-B	Pearson SICK	Spearman SICK
SBERT	**0,881**	**0,881**	**0,839**	**0,805**
SimCSE	0,839	0,841	0,791	0,704
LaBSE	0,729	0,741	0,735	0,697
USE	0,808	0,809	0,771	0,693

Hasil pada Tabel 1 menunjukkan pola yang konsisten pada kedua dataset, yaitu SBERT menghasilkan performa terbaik dibandingkan model lainnya. Temuan ini mengindikasikan bahwa representasi semantik yang dihasilkan SBERT memiliki kesesuaian paling tinggi terhadap penilaian manusia. Performa tersebut berkaitan dengan arsitektur Siamese Network yang digunakan SBERT dan secara khusus dirancang untuk mengoptimalkan hubungan semantik antar kalimat. Dengan mekanisme pelatihan tersebut, ruang embedding yang dihasilkan menjadi lebih sesuai untuk pengukuran kesamaan menggunakan cosine similarity.

USE menunjukkan performa yang mendekati SBERT pada kedua dataset. Selisih performa yang relatif kecil menunjukkan bahwa USE memiliki kemampuan generalisasi yang baik dalam merepresentasikan makna kalimat meskipun menggunakan pendekatan yang lebih ringan dibandingkan model berbasis transformer modern. Sebaliknya, SimCSE dan LaBSE menghasilkan nilai korelasi yang lebih rendah. Pada SimCSE, pendekatan contrastive learning menghasilkan embedding yang sangat diskriminatif, namun sensitivitas terhadap karakteristik data menyebabkan performanya lebih fluktuatif. Sementara itu, LaBSE yang dirancang untuk representasi multibahasa menunjukkan performa yang lebih rendah karena sebagian kapasitas model digunakan untuk mempertahankan keselarasan representasi antar bahasa, bukan semata-mata untuk optimasi tugas STS dalam satu bahasa.

Selain itu, seluruh model memperoleh performa yang lebih tinggi pada STS-B dibandingkan SICK. Hal ini menunjukkan bahwa karakteristik dataset memiliki pengaruh terhadap kualitas representasi semantik yang dihasilkan. Struktur anotasi dan distribusi pasangan kalimat pada STS-B lebih sesuai dengan karakteristik pelatihan sebagian besar model embedding modern sehingga menghasilkan korelasi yang lebih tinggi.

B. Hasil Evaluasi Model Fusion

TABEL 2

Model	Pearson STS-B	Spearman STS-B	Pearson SICK	Spearman SICK
SBERT + SIMCSE	0,841	0,842	0,793	0,706
SBERT + SIMCSE (NORM)	**0,894**	**0,895**	**0,854**	**0,800**
SBERT + LABSE	0,869	0,870	0,832	0,794
SBERT + LABSE (NORM)	0,869	0,870	0,832	0,794
SBERT + USE	0,880	0,878	0,841	0,790
SBERT + USE (NORM)	0,880	0,878	0,841	0,790
SIMCSE + LABSE	0,840	0,841	0,792	0,704
SIMCSE + LABSE (NORM)	0,814	0,819	0,796	0,730
SIMCSE + USE	0,840	0,841	0,791	0,704
SIMCSE + USE (NORM)	0,844	0,844	0,797	0,709
LABSE + USE	0,793	0,798	0,776	0,718
LABSE + USE (NORM)	0,793	0,798	0,776	0,718
SBERT + SIMCSE + LABSE	0,841	0,842	0,793	0,707
SBERT + SIMCSE + LABSE (NORM)	0,877	0,878	0,842	0,790
SBERT + SIMCSE + USE	0,841	0,842	0,793	0,706
SBERT + SIMCSE + USE (NORM)	0,884	0,884	0,843	0,782
SBERT + LABSE + USE	0,863	0,863	0,830	0,781
SBERT + LABSE + USE (NORM)	0,863	0,863	0,830	0,781
SIMCSE + LABSE + USE	0,840	0,841	0,792	0,705
SIMCSE + LABSE + USE (NORM)	0,826	0,829	0,797	0,726
SBERT + SIMCSE + LABSE + USE	0,842	0,843	0,793	0,707
SBERT + SIMCSE + LABSE + USE (NORM)	0,870	0,870	0,834	0,777

Pola yang paling menonjol pada Tabel 2 adalah peningkatan performa yang signifikan pada kombinasi SBERT dan SimCSE setelah dilakukan normalisasi embedding. Sebelum normalisasi, kombinasi kedua model menghasilkan performa yang relatif lebih rendah dibandingkan SBERT tunggal. Namun setelah normalisasi diterapkan, nilai Pearson dan Spearman meningkat secara konsisten pada kedua dataset dan menjadi konfigurasi terbaik dalam penelitian ini.

Fenomena tersebut menunjukkan bahwa perbedaan distribusi dan skala embedding antar model dapat memengaruhi efektivitas proses integrasi. Ketika embedding digabungkan secara langsung, representasi dari salah satu model dapat mendominasi hasil akhir sehingga informasi komplementer tidak dimanfaatkan secara optimal. Normalisasi membantu menyeimbangkan kontribusi setiap model sehingga karakteristik semantik yang dipelajari oleh SBERT dan SimCSE dapat saling melengkapi.

Kombinasi SBERT dan SimCSE menghasilkan performa tertinggi karena kedua model dibangun menggunakan paradigma pelatihan yang berbeda. SBERT berfokus pada optimasi hubungan semantik antar kalimat, sedangkan SimCSE menekankan pembentukan struktur ruang embedding melalui contrastive learning. Integrasi kedua representasi tersebut menghasilkan embedding yang lebih kaya dibandingkan representasi model tunggal.

Menariknya, penambahan jumlah model tidak selalu menghasilkan peningkatan performa. Kombinasi tiga model tidak mampu melampaui performa SBERT+SimCSE yang telah dinormalisasi. Temuan ini menunjukkan bahwa kualitas integrasi lebih dipengaruhi oleh kompatibilitas karakteristik

model dibandingkan jumlah model yang digabungkan. Penambahan model yang memiliki informasi serupa atau redundan berpotensi meningkatkan dimensi embedding tanpa memberikan informasi semantik tambahan yang signifikan.

C. Analisis Error

Selain metrik korelasi, evaluasi juga dilakukan menggunakan MAE dan MSE untuk mengukur tingkat kesalahan prediksi model terhadap skor referensi. Hasil terbaik ditunjukkan pada Tabel III.

TABEL 3

Model	MAE STS-B	MSE STS-B	MAE SICK	MSE SICK
SBERT + USE	**1.569**	**3.369**	**0.787**	**1.024**
SBERT + USE (NORM)	**1.569**	**3.369**	**0.787**	**1.024**
SBERT + LABSE + USE	1.601	3.551	0.853	1.184
SBERT + LABSE + USE (NORM)	1.601	3.551	0.853	1.184
SIMCSE + LABSE	1.869	4.700	1.075	1.728

Berbeda dengan hasil korelasi, kombinasi SBERT dan USE menghasilkan nilai MAE dan MSE terendah pada kedua dataset. Hasil ini menunjukkan bahwa konfigurasi tersebut mampu menghasilkan prediksi yang secara numerik paling dekat dengan skor anotasi manusia. Kondisi ini mengindikasikan bahwa metrik korelasi dan metrik error mengevaluasi aspek performa yang berbeda. Korelasi mengukur kesesuaian pola hubungan antar data, sedangkan MAE dan MSE mengukur kedekatan nilai prediksi terhadap nilai referensi.

Perbedaan hasil antara metrik korelasi dan error menunjukkan bahwa tidak terdapat satu konfigurasi yang unggul pada seluruh aspek evaluasi. SBERT+SimCSE (Norm) lebih efektif dalam mempertahankan hubungan dan peringkat kesamaan antar kalimat, sedangkan SBERT+USE lebih efektif dalam meminimalkan deviasi nilai prediksi. Temuan ini menunjukkan bahwa pemilihan model terbaik perlu mempertimbangkan tujuan aplikasi yang akan dikembangkan.

Secara keseluruhan, hasil penelitian membuktikan bahwa integrasi embedding mampu meningkatkan kualitas representasi semantik apabila dilakukan pada kombinasi model yang tepat dan didukung oleh proses normalisasi yang sesuai. Selain itu, hasil eksperimen menunjukkan bahwa karakteristik pembelajaran masing-masing model memiliki pengaruh yang lebih besar terhadap performa dibandingkan jumlah model yang digunakan dalam proses integrasi.

V. KESIMPULAN

Penelitian ini telah melakukan analisis komparatif terhadap empat model *sentence embedding*, yaitu Sentence-BERT (SBERT), SimCSE, Language-agnostic BERT Sentence Embedding (LaBSE), dan Universal Sentence Encoder (USE), serta mengevaluasi efektivitas integrasi model menggunakan metode *vector concatenation* pada tugas *Semantic Textual Similarity* (STS). Hasil eksperimen pada dataset STS Benchmark (STS-B) dan SICK menunjukkan bahwa SBERT merupakan model tunggal dengan performa terbaik berdasarkan metrik Pearson Correlation dan Spearman Correlation. Temuan ini menunjukkan bahwa representasi semantik yang dihasilkan SBERT memiliki tingkat kesesuaian yang tinggi terhadap penilaian manusia dalam mengukur kesamaan makna antar kalimat.

Hasil penelitian juga menunjukkan bahwa pendekatan *model fusion* mampu meningkatkan kualitas representasi semantik apabila diterapkan pada kombinasi model yang tepat. Kombinasi SBERT dan SimCSE yang disertai proses normalisasi menghasilkan performa terbaik dengan nilai Spearman Correlation sebesar 0,89514 pada dataset STS-B dan 0,80028 pada dataset SICK. Peningkatan tersebut mengindikasikan bahwa integrasi embedding dari model yang memiliki karakteristik pembelajaran berbeda dapat menghasilkan representasi yang lebih kaya dan komplementer dibandingkan penggunaan model tunggal. Selain itu, hasil eksperimen menunjukkan bahwa proses normalisasi memiliki peran penting dalam meningkatkan efektivitas integrasi dengan mengurangi perbedaan distribusi antar embedding.

Di sisi lain, kombinasi SBERT dan USE menghasilkan nilai MAE dan MSE terendah pada kedua dataset, yang menunjukkan kemampuan terbaik dalam meminimalkan deviasi prediksi terhadap skor referensi. Temuan ini mengindikasikan bahwa model dengan performa korelasi terbaik tidak selalu menghasilkan tingkat kesalahan prediksi terendah, sehingga pemilihan konfigurasi model perlu mempertimbangkan tujuan evaluasi yang digunakan.

Secara keseluruhan, penelitian ini membuktikan bahwa integrasi *sentence embedding* melalui pendekatan *vector concatenation* dapat meningkatkan performa tugas *Semantic Textual Similarity*. Untuk penelitian selanjutnya, pendekatan integrasi yang lebih kompleks, seperti *weighted fusion*, *attention-based fusion*, atau *meta-embedding*, dapat dieksplorasi lebih lanjut guna memperoleh representasi semantik yang lebih efektif dan robust pada berbagai domain serta karakteristik data.

REFERENSI

[1] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, “SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation,” in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer, and D. Jurgens, Eds., Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 1–14. doi: 10.18653/v1/S17-2001.

[2] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, “MTEB: Massive Text Embedding Benchmark,” in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds.,

Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2014–2037. doi: 10.18653/v1/2023.eacl-main.148.

[3] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. doi: 10.18653/v1/D19-1410.

[4] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6894–6910. doi: 10.18653/v1/2021.emnlp-main.552.

[5] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT Sentence Embedding," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 878–891. doi: 10.18653/v1/2022.acl-long.62.

[6] D. Cer *et al.*, "Universal Sentence Encoder for English," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, E. Blanco and W. Lu, Eds., Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 169–174. doi: 10.18653/v1/D18-2029.

[7] D. Kiela, C. Wang, and K. Cho, "Dynamic Meta-Embeddings for Improved Sentence Representations," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds., Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 1466–1477. doi: 10.18653/v1/D18-1176.

[8] H. Tsukagoshi, R. Sasano, and K. Takeda, "Comparison and Combination of Sentence Embeddings Derived from Different Supervision Signals," in *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, V. Nastase, E. Pavlick, M. T. Pilehvar, J. Camacho-Collados, and A. Raganato, Eds., Seattle, Washington: Association for Computational Linguistics, Jul. 2022, pp. 139–150. doi: 10.18653/v1/2022.starsem-1.12.

[9] N. Poerner, U. Waltinger, and H. Schütze, "Sentence Meta-Embeddings for Unsupervised Semantic Textual Similarity," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 7027–7034. doi: 10.18653/v1/2020.acl-main.628.