Jurnal Sifo Mikroskil (JSM)
Volume 27, No 2, Oktober 2026 - Hal 337 - 352
DOI: https://doi.org/10.55601/jsm.v27i2.2092

e-ISSN: 2622-8130
ISSN: 1412-0100

Evaluasi Kinerja Model BiLSTM dan Additive Attention pada Pemodelan Semantic Textual Similarity Berbasis Pre-trained BERT

Chrisandy Wiranata[1], Muhammad Fahim Al-Mubarok[2], Sandy Agre Nicola[3], Ng Poi Wong[4], Sio Jurnalis Pipin[5]
[1,2,3,4,5]Universitas Mikroskil, Jl. Thamrin No. 124, 061-4573767
[1,2,3,4]Fakultas Informatika, Teknik Informatika, Universitas Mikroskil, Medan
[5]Fakultas Informatika, Teknologi Informasi, Universitas Mikroskil, Medan
e-mail: [1]221110252@students.mikroskil.ac.id, [2]221110196@students.mikroskil.ac.id, [3]221110040@students.mikroskil.ac.id, [4]poiwong@mikroskil.ac.id, [5]sio.pipin@mikroskil.ac.id

Dikirim: ***27-08-2026*** | *Diterima:* ***30-10-2026*** | *Diterbitkan:* ***31-10-2026***

Abstrak

Pengukuran kesamaan semantik antarteks (*Semantic Textual Similarity/STS*) merupakan tantangan utama dalam Pemrosesan Bahasa Alami (*NLP*) akibat variasi leksikal dan struktur kalimat. Penelitian ini bertujuan untuk menganalisis dan mengevaluasi kinerja integrasi arsitektur *Bidirectional Long Short-Term Memory* (*BiLSTM*) dan mekanisme *Additive Attention* yang berbasis representasi fitur *pre-trained BERT* dengan pendekatan *cross-encoder*. Evaluasi model dilakukan menggunakan benchmark dataset *Quora Question Pairs* (QQP) pada tingkat kalimat. Hasil pengujian pada data uji menunjukkan bahwa model mencapai nilai *loss* sebesar 0,3237, akurasi 0,8545, AUC 0,9307, serta *F1-score* sebesar 0,8837 untuk kelas *non*-duplikat dan 0,8059 untuk kelas duplikat. Meskipun kombinasi *BiLSTM* dan *Additive Attention* terbukti efektif dalam menangkap konteks dua arah dan pembobotan fitur relevan pada pasangan teks pendek, pengujian berbasis kasus pada tingkat dokumen menghasilkan skor kemiripan yang relatif rendah. Hal ini menunjukkan adanya keterbatasan generalisasi model yang dipengaruhi oleh domain dan karakteristik dataset QQP yang didominasi oleh kalimat pendek berstruktur sederhana.

Kata kunci: Semantic Textual Similarity, BiLSTM, Additive Attention, BERT, Quora Question Pair.

Abstract

Measuring Semantic Textual Similarity (STS) remains a fundamental challenge in Natural Language Processing (NLP) due to lexical variations and diverse sentence structures. This study aims to evaluate the performance of integrating Bidirectional Long Short-Term Memory (BiLSTM) architecture with an Additive Attention mechanism leveraging pre-trained BERT feature representations via a cross-encoder approach. The model was evaluated using the Quora Question Pairs (QQP) benchmark dataset at the sentence level. Experimental results on the test set demonstrate that the model achieved a loss of 0.3237, accuracy of 0.8545, AUC of 0.9307, and F1-scores of 0.8837 for the non-duplicate class and 0.8059 for the duplicate class. While the combination of BiLSTM and Additive Attention effectively captures bidirectional context and highlights relevant features in short text pairs, case-based evaluation at the document level yielded relatively low similarity scores. These findings reveal limitations in model generalization, heavily influenced by the domain and short-sentence characteristics of the QQP dataset.

Keywords: *Semantic Textual Similarity, BiLSTM, Additive Attention, BERT, Quora Question Pairs*

1. PENDAHULUAN

Pengukuran kesamaan semantik antarteks (*Semantic Textual Similarity*/STS) merupakan bagian dalam proses Pemprosesan Bahasa Alami (*Natural Language Processing*/NLP) yang bertujuan untuk mengkuantifikasi tingkat kesetaraan makna antara dua teks [1]. Dalam bidang NLP, perlu untuk membedakan konsep STS, *paraphrase detection*, dan *plagiarism detection* [2, 4]. STS berfokus pada pemetaan skor kontinu yang menggambarkan derajat kesamaan makna antarteks [4], sedangkan *paraphrase detection* umumnya merupakan tugas klasifikasi untuk menentukan apakah dua kalimat memiliki informasi yang identik melalui struktur leksikal yang berbeda [2]. Di sisi lain, *plagiarism detection* merupakan domain yang melibatkan identifikasi klaim atau gagasan tanpa atestasi [1], yang tidak hanya bergantung pada kemiripan semantik tetapi juga pada konteks atribusi dokumen [1, 2]. Pendekatan leksikal tradisional seperti *Cosine Similarity* berbasis TF-IDF maupun *string-matching* sering kali gagal menangkap kesetaraan makna pada teks yang mengalami perubahan struktur kalimat atau penggunaan sinonim (*paraphrastic plagiarism*) [2, 5].

Untuk mengatasi keterbatasan leksikal tersebut, perkembangan pemodelan bahasa mengarah pada penggunaan *contextual word embeddings* berbasis *Transformer*, seperti *Bidirectional Encoder Representations from Transformers* (BERT) [10, 15]. BERT terbukti unggul dalam menghasilkan representasi fitur kaya akan konteks melalui mekanisme *self-attention* dua arah [15, 20]. Dalam penerapan STS, arsitektur *cross-encoder* berbasis BERT memungkinkan token dari dua kalimat berinteraksi secara langsung, sehingga menghasilkan pemahaman relasi makna yang lebih mendalam dibandingkan pendekatan *bi-encoder* [14, 16]. Namun, melakukan *full fine-tuning* pada model BERT yang memiliki lebih dari 110 juta parameter membutuhkan sumber daya komputasi yang sangat besar [15]. Alternatifnya adalah pemanfaatan BERT sebagai *pretrained feature extractor* dengan parameter terbekukan (*frozen*) yang dikombinasikan dengan jaringan syaraf tiruan sekuensial seperti *Bidirectional Long Short-Term Memory* (BiLSTM) dan *Additive Attention* mulai banyak dieksplorasi [8, 9, 10]. BiLSTM berperan dalam menangkap urutan konteks jangka panjang secara maju dan mundur [8], sementara *Additive Attention* memberikan pembobotan adaptif pada fitur-fitur yang paling relevan [9, 20].

Meskipun kombinasi model transformer dan jaringan sekuensial telah diterapkan pada berbagai tugas NLP [9, 11], masih terdapat *research gap* terkait evaluasi kinerja terukur dari arsitektur BiLSTM dan *Additive Attention* ketika ditempatkan di atas representasi *frozen* BERT *cross-encoder* [2, 3, 4]. Penelitian terdahulu berfokus pada *fine-tuning* langsung [15] atau penggunaan *pooling* sederhana [14] tanpa mengevaluasi seberapa efektif lapisan sekuensial dan pembobotan *attention* dalam mengekstrak pola semantik dari vektor *embedding* BERT yang dibekukan [2, 4]. Selain itu, terdapat kebutuhan untuk menganalisis sejauh mana model yang dilatih pada dataset pasangan kalimat pendek—seperti *Quora Question Pairs* (QQP) mampu melakukan generalisasi ketika diuji pada kasus skala dokumen [2, 4, 7].

Untuk menjawab celah penelitian tersebut, penelitian ini bertujuan untuk menganalisis dan mengevaluasi secara empiris kinerja integrasi model BiLSTM dan *Additive Attention* berbasis *Pre-trained* BERT *cross-encoder* dalam pemodelan *Semantic Textual Similarity* [2, 3, 4]. Kontribusi utama dari penelitian ini mencakup: (1) evaluasi komprehensif kinerja arsitektur BiLSTM-*Additive Attention* menggunakan benchmark dataset QQP pada tingkat kalimat [2, 4], (2) analisis efisiensi komputasi penggunaan *frozen* BERT dibandingkan potensi generalisasinya [2, 3], serta (3) analisis kritis keterbatasan generalisasi model saat dievaluasi pada pengujian berbasis kasus tingkat dokumen [2, 3, 4]. Sebagai bentuk implementasi praktis, model yang telah dievaluasi diintegrasikan ke dalam antarmuka aplikasi berbasis web guna memfasilitasi pengujian interaktif.

2. TINJAUAN PUSTAKA

2.1 *Natural Language Processing* (NLP)

Natural Language Processing (NLP) merupakan bidang dalam kecerdasan buatan yang berfokus pada bagaimana komputer dapat memahami, memproses, dan menghasilkan bahasa manusia secara alami. Tujuan utama dari NLP adalah penghubung komunikasi antara manusia dan mesin dengan memungkinkan sistem komputer menafsirkan teks atau ucapan layaknya manusia. Dalam penelitian ini, NLP berperan penting dalam mengubah bahasa alami menjadi representasi numerik yang dapat dipelajari oleh model deep learning [10].

Secara umum, NLP menggabungkan konsep linguistik dan pembelajaran mesin untuk memahami makna di balik bahasa. Bahasa manusia bersifat ambigu dan kontekstual, sehingga sistem NLP harus mampu menganalisis struktur sintaksis (tata bahasa) dan semantik secara bersamaan [11]. Untuk mencapai pemahaman tersebut, teks mentah harus melalui serangkaian proses pengolahan agar dapat dikenali oleh model. Tahapan utama dalam NLP meliputi preprocessing untuk membersihkan dan menormalkan data, tokenisasi untuk memecah kalimat menjadi unit-unit dasar, dan word embedding untuk merepresentasikan kata dalam bentuk vektor numerik [12].

2.2 *Semantic Textual Similarity* (STS)

Pendekatan Semantik bekerja dengan menilai kesamaan makna berdasarkan konteks dan relasi antar konsep dalam teks, bukan hanya kesamaan kata yang muncul. Dalam pendekatan ini, sistem berupaya memahami hubungan logis antara ide atau entitas dalam kalimat, seperti sinonimi, parafrase, atau keterkaitan sebab-akibat. Dengan memahami struktur semantik ini, model dapat mengenali bahwa dua kalimat berbeda secara leksikal dapat menyampaikan makna yang sama. Pendekatan ini banyak diterapkan dalam berbagai tugas NLP, termasuk semantic textual similarity (STS), karena mampu menangkap kesetaraan makna lintas variasi kalimat [13].

Dibandingkan metode leksikal, pendekatan semantik menawarkan pemahaman yang lebih dalam terhadap konteks bahasa. Model yang menggunakan analisis semantik mampu mengenali hubungan makna yang tidak eksplisit. Namun, kompleksitas pemrosesan dan kebutuhan data kontekstual yang besar membuat pendekatan ini lebih menuntut secara komputasi dibanding pendekatan leksikal. Meskipun demikian, hasil yang diperoleh jauh lebih representatif terhadap makna sebenarnya, menjadikannya dasar utama bagi sistem modern dalam mengukur kesamaan teks secara semantik [14].

2.3 BERT

BERT (*Bidirectional Encoder Representations from Transformers*) merupakan model berbasis arsitektur *Transformer* yang dirancang untuk menghasilkan representasi kontekstual suatu kata berdasarkan keseluruhan kalimat. BERT dibangun di atas arsitektur *Transformer encoder* yang menggunakan mekanisme *self-attention* untuk memahami hubungan antarkata secara dua arah dalam satu waktu. Keunggulan BERT terletak pada kemampuannya menangkap konteks kata berdasarkan posisi sebelum dan sesudahnya secara simultan [15]. Dalam sistem yang dikembangkan, BERT digunakan sebagai *pretrained feature extractor* dengan parameter yang dibekukan (*frozen*). Artinya, model tidak diperbarui selama proses pelatihan utama, melainkan dimanfaatkan untuk menghasilkan *embedding* yang telah kaya informasi kontekstual.

Untuk membandingkan dua kalimat, digunakan pendekatan *cross-encoder*. Dalam pendekatan ini, kedua kalimat digabungkan menjadi satu masukan sehingga setiap token dari kalimat pertama dan kedua dapat saling berinteraksi secara langsung melalui mekanisme *self-attention*. Strategi ini memungkinkan model memahami relasi makna secara lebih komprehensif dibandingkan pemrosesan terpisah [16].

2.4 *Bidirectional* LSTM (BiLSTM)

Long Short-Term Memory (LSTM), yaitu pengembangan dari *Recurrent Neural Network* (RNN) yang dirancang untuk mengatasi permasalahan *vanishing gradient* pada teks panjang [17]. LSTM

memiliki mekanisme gerbang *forget gate*, *input gate*, dan *output gate* yang berfungsi mengatur aliran informasi dalam memori jangka panjang. Dengan mekanisme tersebut, model mampu mempertahankan informasi penting dan membuang informasi yang tidak relevan selama proses pembelajaran. Untuk memperoleh pemahaman konteks yang lebih menyeluruh, penelitian ini menggunakan varian *Bidirectional* LSTM (BiLSTM). Berbeda dengan LSTM satu arah, BiLSTM memproses urutan teks dari dua arah, yaitu maju dan mundur. Pendekatan ini memungkinkan setiap kata dipahami berdasarkan konteks sebelum dan sesudahnya, sehingga menghasilkan representasi yang lebih komprehensif terhadap makna kalimat [18].

2.5 Mekanisme *Attention*

Mekanisme attention merupakan salah satu konsep penting dalam pemrosesan bahasa alami modern. Mekanisme attention memungkinkan suatu sistem untuk memfokuskan pemrosesan pada bagian teks yang paling bermakna terhadap tugas yang sedang dijalankan, seperti memahami makna kalimat, menerjemahkan teks, maupun membandingkan kesamaan semantik antarteks [19]. *Additive Attention* bekerja dengan memberikan bobot berbeda pada setiap keluaran BiLSTM, sehingga token yang lebih penting terhadap makna kalimat memperoleh perhatian lebih besar. Hasil dari proses ini adalah sebuah *context vector* yang merepresentasikan ringkasan informasi yang menyoroti bagian teks paling relevan, membantu model memusatkan perhatian pada kata-kata bermakna dan mengabaikan informasi yang kurang penting [20].

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif eksperimental yang terbagi ke dalam empat tahapan utama: prapemrosesan data, pembentukan arsitektur model, pelatihan dan *hyperparameter tuning*, serta evaluasi kinerja model.

3.1 Tahapan Penelitian

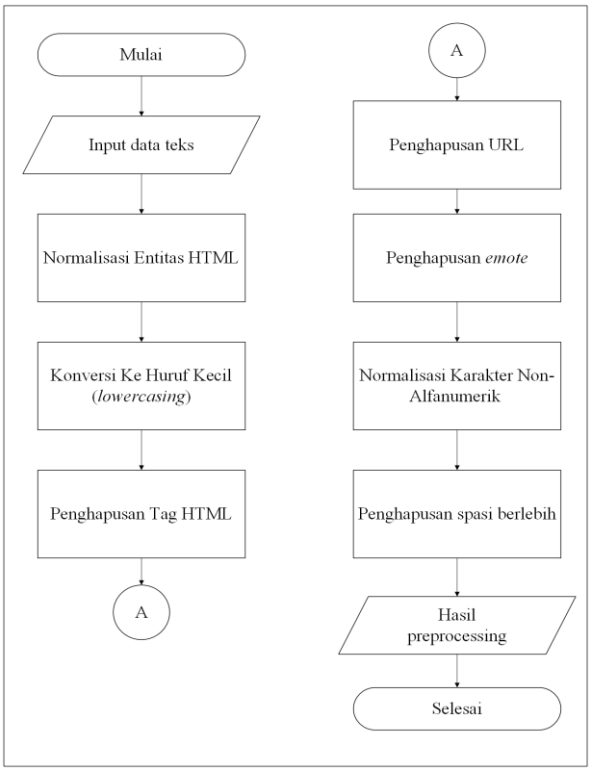
Penelitian ini menggunakan pendekatan kuantitatif dengan tahapan pelaksanaan peneilitian sebagai berikut:

1. Pengumpulan Data
   Tahap awal penelitian ini adalah penelusuran dan pengumpulan dataset *Quora Question Pairs* (QQP) dan literatur ilmiah melalui *Google Scholar* yang menjadi rujukan dalam perancangan, pelatihan, dan pengujian model.
2. Analisis kebutuhan
   a. Kebutuhan fungsional digambarkan menggunakan *use case* diagram.
   b. Kebutuhan non-fungsional dianalisis menggunakan metode PIECES.
3. Perancangan
   Tahap perancangan dilakukan dengan menggunakan alat bantu *wifeframe.cc* untuk menghasilkan antarmuka yang mudah digunakan, dan informatif.
4. Implementasi Sistem
   Tahap implementasi dilakukan dengan menggunakan *Python* dengan library *TensorFlow*, *Keras*, dan *Transformers* untuk STS, *FastAPI* sebagai *Backend* dan *Streamlit* sebagai *Frontend*.
5. Pengujian dan Evaluasi Tahap
   Pengujian menggunakan dataset *Quora Question Pairs* (QQP) dengan metrik F1 *score*, *precision*, dan *recall* untuk mengukur kemampuan model.

3.2 Analisa Proses

Alur penelitian ini terbagi menjadi dua tahapan utama, yaitu fase pelatihan (*training*) dan fase pengujian (*testing*). Pada kedua fase, data teks mentah terlebih dahulu melewati tahapan prapemrosesan (*preprocessing*) untuk membersihkan data. Tahapan prapemrosesan dapat dilihat pada Gambar 1.

Selanjutnya, teks dikonversi menjadi representasi vektor menggunakan BERT untuk tahap *tokenization* dan *embedding*.



Gambar 1. *Flowchart* Tahap Pra-Pemrosesan Data

3.2.1. Dataset dan Pembagian Data

Eksperimen dilakukan menggunakan *benchmark dataset Quora Question Pairs* (QQP) yang berisi pasangan pertanyaan untuk tugas *Semantic Textual Similarity* (STS) dan *paraphrase detection*. Dataset ini memiliki total 404.290 pasangan teks. Pembagian dataset dilakukan secara terpisah menjadi tiga bagian:

a. Data Pelatihan (*Training Set*): 80% (323.432 sampel) untuk melatih bobot model.
b. Data Validasi (*Validation Set*): 10% (40.429 sampel) untuk pemantauan *loss* dan *hyperparameter tuning*.
c. Data Pengujian (*Testing Set*): 10% (40.429 sampel) sebagai data independen untuk evaluasi akhir.

3.2.2. Prapemrosesan Data dan Ekstraksi Fitur

Pasangan teks (T_1, T_2) diproses menggunakan pendekatan *cross-encoder*. Kedua teks digabungkan menjadi satu format masukan: [CLS] T1 [SEP] T2 [SEP]. Prapemrosesan dilakukan menggunakan *tokenizer* dari *pre-trained* bert-base-uncased. Teks dikonversi menjadi token ID dan *attention mask* dengan batas panjang maksimum urutan (*maximum sequence length*) $N = 128$ token. Teks yang melebihi 128 token dipotong (*truncated*), sedangkan teks yang lebih pendek ditambahi *padding*.

3.2.3. Arsitektur Model BiLSTM-Additive Attention

Model terdiri dari empat komponen utama yang disusun secara sekuensial:

a. **BERT Feature Extractor (Frozen):** Vektor *embedding* diekstrak dari luaran lapisan terakhir bert-base-uncased (128×768). Seluruh parameter BERT dibekukan (*frozen*) untuk efisiensi komputasi.
b. **Bidirectional LSTM Layer:** Vektor *embedding* diteruskan ke lapisan BiLSTM dengan 128 unit tersembunyi (*hidden units*). BiLSTM memproses urutan dari arah maju ($\overrightarrow{h_t}$) dan mundur ($\overleftarrow{h_t}$), menghasilkan representasi konteks $h_t = \left[\overrightarrow{h_t}\,;\,\overleftarrow{h_t}\right]$ berdimensi 256.
c. **Additive Attention Layer:** Mekanisme *Additive Attention* menghitung pembobotan a_t untuk setiap *hidden state* h_t menggunakan persamaan:

$$u_t = \tanh(W_a h_t + b_a)$$
$$\alpha_t = \frac{\exp(u_t^T v_a)}{\sum_{j=1}^{N} \exp(u_j^T v_a)}$$
$$c = \sum_{t=1}^{N} \alpha_t \, h_t$$

di mana W_a, b_a, dan v_a adalah parameter perhatian yang dilatih, dan c merupakan *context vector* ringkasan semantik.
d. Klasifikasi dan Konversi Skor Kemiripan: *Context vector* diteruskan ke *Dense layer* dengan fungsi aktivasi *Sigmoid* untuk menghasilkan probabilitas kemiripan $p \in [0,1]$. Skor kemiripan kontinu S_{sim} dalam rentang $0 - 100\%$ untuk tampilan antarmuka web dihitung dengan memetakan nilai probabilitas:

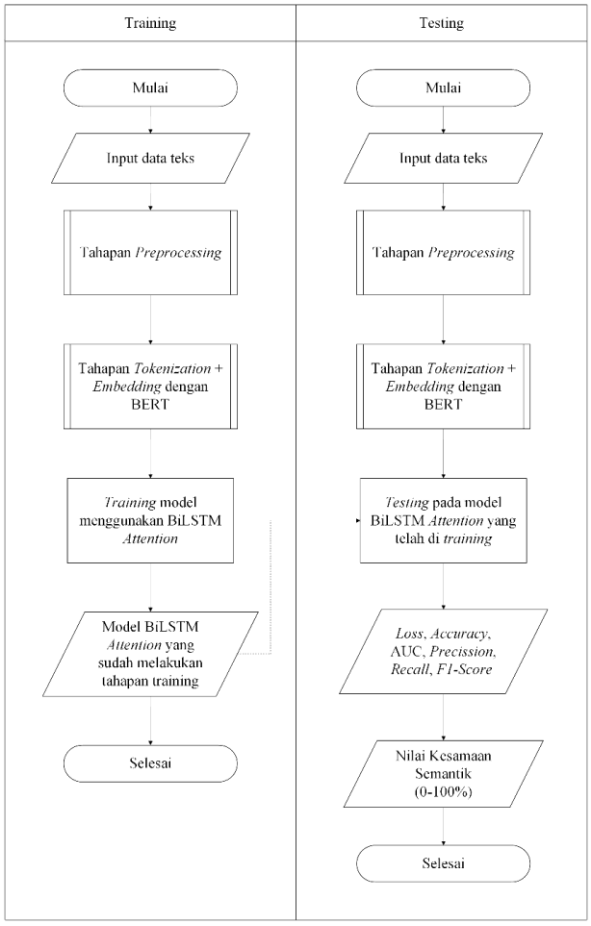
$$S_{sim} = p \times 100\%$$

3.2.4. Pengaturan Pelatihan dan Hyperparameter Tuning

Pelatihan dilakukan menggunakan *optimizer* AdamW dengan *learning rate* 2×10^{-4}, *weight decay* 0,01, *batch size* 64, dan *loss function Binary Cross-Entropy*. *Hyperparameter tuning* dilakukan menggunakan algoritma *Hyperband* melalui pustaka Optuna sebanyak 10 *trial* pada *validation set* untuk menentukan kombinasi unit BiLSTM dan tingkat *dropout* (0,3) optimal.

3.2.5. Metrik Evaluasi dan Lingkungan Eksperimen

Evaluasi model dilakukan secara menyeluruh menggunakan metrik *Binary Cross-Entropy Loss*, *Accuracy*, *Area Under Curve* (AUC), *Precision*, *Recall*, dan *F1-Score*. Eksperimen dijalankan pada *environment* Python 3.10 dengan *framework* PyTorch 2.1, HuggingFace Transformers, dan NVIDIA RTX 3060 GPU (12 GB VRAM).



Gambar 2. *Flowchart* Proses Pengukuran *Semantic Text Similarity*

Pada fase pelatihan seperti pada gambar 2, model BiLSTM dengan mekanisme *Attention* dilatih untuk mempelajari pola fitur dari data. Model yang telah optimal kemudian diimplementasikan pada fase pengujian. Performa model dievaluasi menggunakan metrik *Loss*, *Accuracy*, AUC, *Precision*, *Recall*, dan *F1-Score*. Output akhir dari sistem ini adalah nilai kesamaan semantik dalam rentang 0-100%.

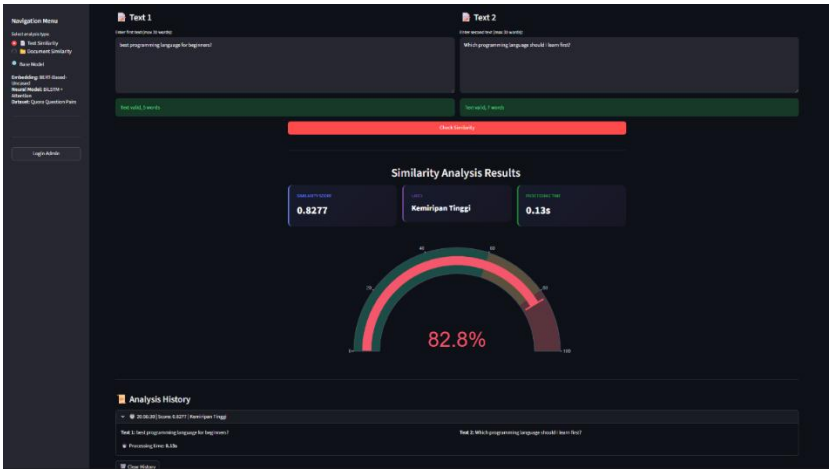
4. HASIL DAN PEMBAHASAN

4.1 Hasil

Hasil akhir dari aplikasi *Semantic Textual Similarity* diimplementasikan dalam bentuk aplikasi berbasis *website*.

1. Halaman *Text Similarity Analysis*

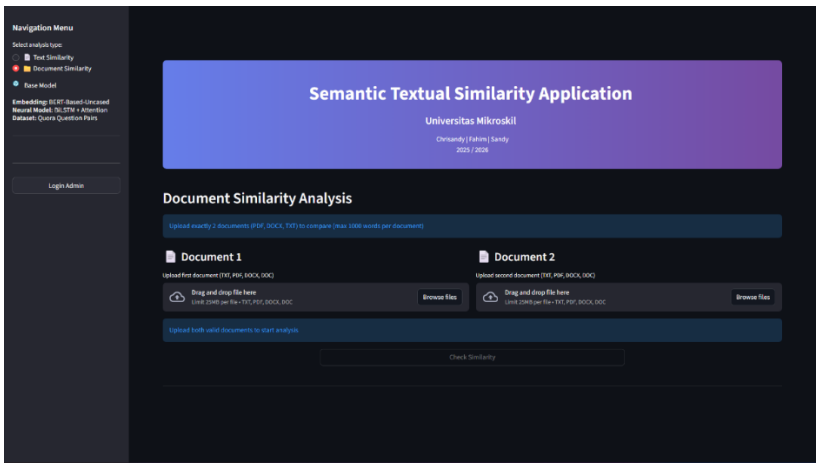
   Gambar 5 memperlihatkan tampilan pengujian *Text Similarity*, pengguna dapat memasukkan dua buah teks dengan batas maksimal jumlah kata yang telah ditentukan oleh sistem. Sistem kemudian memproses kedua teks dan menampilkan hasil analisis secara visual dalam bentuk *similarity score*, label tingkat kemiripan, waktu pemrosesan, dan *gauge chart.*



Gambar 3. Tampilan Hasil *Text Similarity Analysis*

2. Halaman *Document Similarity Analysis*

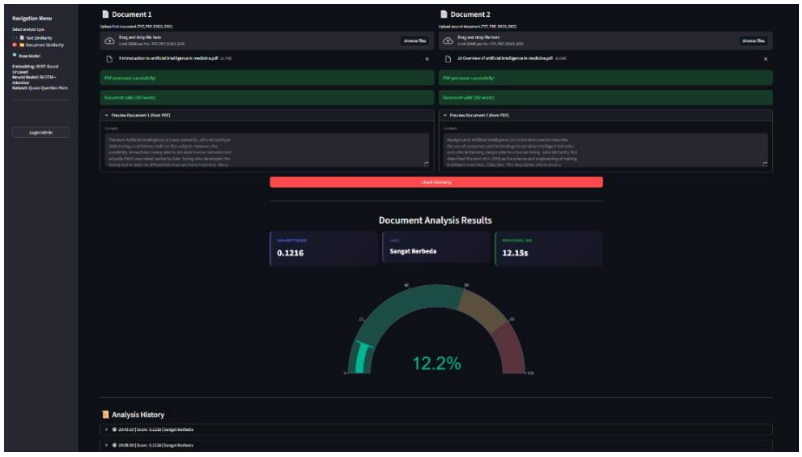
Gambar 4 memperlihatkan tampilan *Document Similarity Analysis*, pengguna dapat mengunggah dua dokumen untuk dibandingkan dan pengguna dapat melihat *preview* isi dokumen.



Gambar 4. Tampilan Halaman *Document Similarity Analysis*

Gambar 5 memperlihatkan tampilan pengujian *Document Similarity.* Proses analisis dilakukan dengan membagi dokumen menjadi bagian-bagian teks, kemudian menghitung tingkat kesamaan semantik antar dokumen secara keseluruhan. Hasil analisis ditampilkan dalam bentuk *similarity score*, label tingkat kemiripan, waktu pemrosesan, dan *gauge chart*

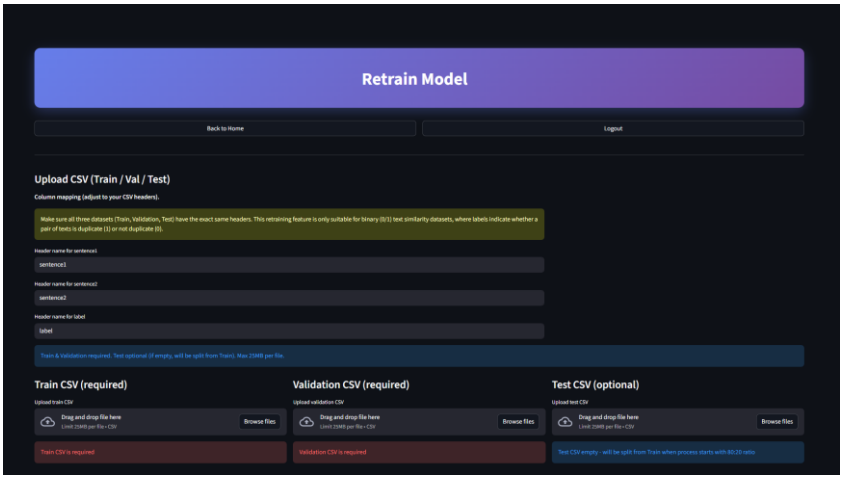
2622-8130 Jurnal Sifo Mikroskil (JSM) - Volume 27, No 2, Oktober 2026
DOI: https://doi.org/10.55601/jsm.v27i2.2092



Gambar 5. Tampilan Hasil *Document Similarity Analysis*

3. Halaman Pelatihan Ulang Model

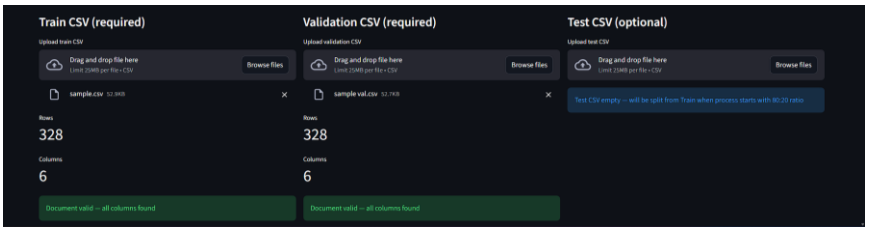
Gambar 6 dan Gambar 7 menunjukkan tampilan *Retrain Model* yang berfungsi untuk melakukan proses pelatihan ulang model, administrator dapat mengunggah *dataset* dalam format CSV dan administrator dapat mendefinisikan secara manual kolom yang merepresentasikan data *sentence1*, *sentence2*, dan *label*.



Gambar 6. Tampilan Halaman Retrain Model Bagian Upload Dataset



Gambar 7. Tampilan Halaman Retrain Model Bagian Dataset Summary

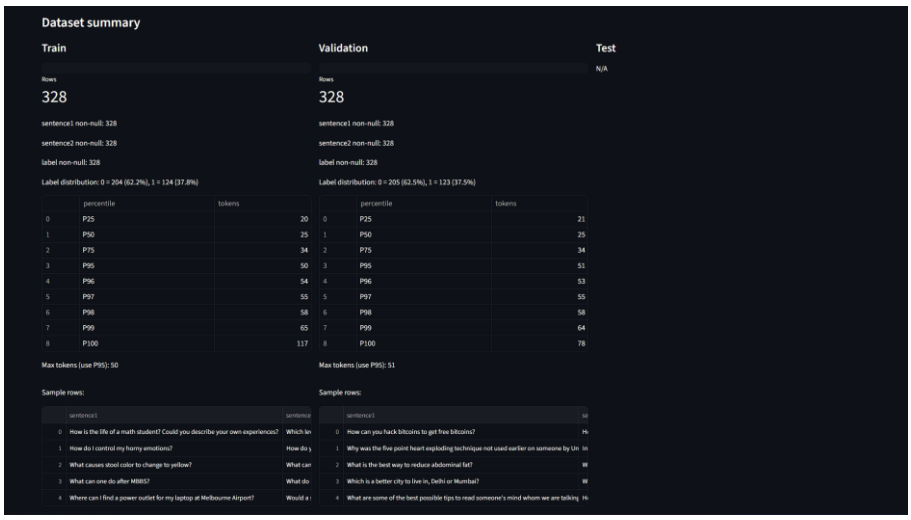


Gambar 8. Tampilan Hasil Upload Dataset

Setelah *dataset* berhasil diunggah, sistem akan menyajikan analisis statistik deskriptif pada panel *Dataset Summary* (Gambar 9). Informasi yang disajikan meliputi jumlah data, distribusi kelas label,

2622-8130 Jurnal Sifo Mikroskil (JSM) - Volume 27, No 2, Oktober 2026
DOI: https://doi.org/10.55601/jsm.v27i2.2092

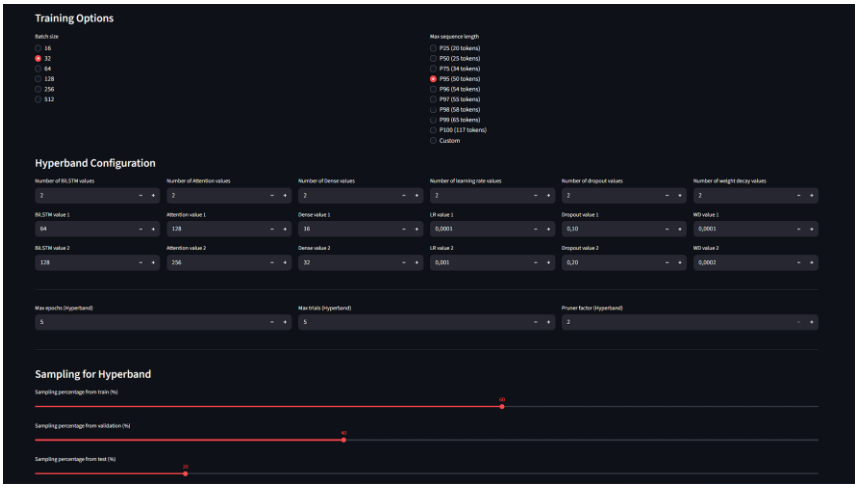
statistik token kalimat, serta contoh data dari *dataset* yang berfungsi sebagai acuan dalam memverifikasi kualitas dan keseimbangan data sebelum proses pelatihan dimulai.



Gambar 9. Tampilan Hasil Dataset Summary

Gambar 10 menunjukkan tampilan konfigurasi pelatihan model yang terintegrasi dengan algoritma *Hyperband* untuk penentuan *hyperparameter* yang optimal. Administrator memiliki fleksibilitas untuk menentukan parameter pelatihan yang meliputi:

a. *General Training Options* berupa pengaturan ukuran *batch* dan pengaturan panjang maksimum urutan.
b. *Hyperband Configuration* berupa pengaturan jumlah unit pada lapisan BiLSTM, lapisan *Attention*, lapisan *Dense*, nilai *Learning Rate*, tingkat *Dropout*, dan *Weight Decay*.
c. *Hyperband Training Constraints* berupa pengaturan jumlah *epoch*, jumlah percobaan, dan *pruner factor*.
d. *Data Sampling* berupa pengaturan persentase penggunaan dataset.
e. *Retrain Full Training Data* berupa pengaturan pelatihan model dengan menggunakan *hyperparameter* terbaik dengan melatih menggunakan seluruh dataset dengan jumlah *epoch* yang ditentukan (Gambar 14).
f. *Model Name* berupa nama model hasil proses *training* yang akan disimpan.



Gambar 10. Tampilan Halaman Retrain Model Bagian Konfigurasi

4.2 Pembahasan

Model *semantic similarity* dilatih dan dievaluasi menggunakan dataset *Quora Question Pairs* (QQP) sebagai benchmark kesamaan makna pada level kalimat. Dalam penelitian ini, BERT digunakan dalam kondisi *frozen* sebagai penyedia *contextual embedding* tanpa *fine-tuning*, sementara adaptasi terhadap tugas dilakukan melalui penambahan BiLSTM dan *additive attention*. Pendekatan ini dipilih untuk menjaga keseimbangan antara performa dan efisiensi komputasi, mengingat *fine-tuning* penuh BERT melibatkan pembaruan sekitar 110 juta parameter, sedangkan BiLSTM dan *attention* hanya menambahkan parameter dalam skala ratusan ribu hingga kurang dari satu juta.

Proses pelatihan dilakukan dalam dua tahap. Tahap pertama berupa *hyperparameter tuning* menggunakan metode *Hyperband* melalui *Optuna* dengan 10 *trial*, meliputi *tuning learning rate*, ukuran hidden layer BiLSTM, *dropout*, dan *weight decay*. Konfigurasi terbaik dipilih berdasarkan nilai *validation loss* terendah dan digunakan pada tahap kedua, yaitu pelatihan penuh selama 20 *epoch* menggunakan seluruh data pelatihan dan validasi.

Hasil pelatihan menunjukkan konvergensi yang stabil, dengan performa akhir pada data pengujian sebesar *loss* 0.3237, *accuracy* 0.8545, dan AUC 0.9307 (Tabel 1). Konsistensi nilai pada data pelatihan, validasi, dan pengujian menunjukkan bahwa model memiliki kemampuan generalisasi yang baik. Analisis *confusion matrix* memperlihatkan distribusi kesalahan yang relatif seimbang pada kedua kelas (Tabel 2). Nilai *precision*, *recall*, dan *F1-score* juga menunjukkan performa yang proporsional, meskipun model sedikit lebih kuat dalam mengidentifikasi pasangan non-duplikat dibandingkan pasangan duplikat (Tabel 3).

Tabel 1. Hasil Evaluasi Performa pada Data Pelatihan, Validasi, dan Pengujian

Dataset	***Loss***	***Accuracy***	**AUC**
Training	0.3401	0.8454	0.9221
Validation	0.3235	0.8551	0.9310
Testing	0.3237	0.8545	0.9307

Tabel 2. *Confusion Matrix* pada Data Pengujian

	Prediksi 0	**Prediksi 1**
Aktual 0	40202 (TN)	5692 (FP)
Aktual 1	4895 (FN)	21981 (TP)

Tabel 3. Nilai *Precision*, *Recall*, dan *F1-Score* untuk Setiap Label pada Data Pengujian

Label	***Precision***	***Recall***	***F1-Score***
0	0.8915	0.8760	0.8837
1	0.7943	0.8179	0.8059

Untuk membuktikan kontribusi empiris dari komponen BiLSTM dan *Additive Attention* di atas representasi *frozen* BERT, dilakukan eksperimen pembanding (*Ablation Study*) menggunakan konfigurasi pelatihan yang identik. Hasil perbandingan disajikan pada Tabel 2.

Tabel 4. Hasil *Ablation Study* Arsitektur Model pada Data Uji

Arsitektur Model	***Loss***	***Accuracy***	**AUC**	***F1-Score (Macro)***
Frozen BERT + Dense (*Baseline*)	0,3812	0,8120	0,8850	0,7980
Frozen BERT + BiLSTM + Dense	0,3450	0,8380	0,9120	0,8260
** *Frozen* BERT + BiLSTM + Additive Attention (Usulan)**	0,3237	0,8545	0,9307	0,8448

Hasil *Ablation Study* membuktikan bahwa penambahan lapisan BiLSTM meningkatkan akurasi sebesar 2,6% dari *baseline*, dan integrasi *Additive Attention* memberikan kontribusi tambahan sebesar 1,65% pada akurasi serta meningkatkan AUC hingga 0,9307. Hal ini mengonfirmasi bahwa *Additive Attention* secara efektif memfokuskan bobot pada fitur-fitur token yang memuat informasi semantik penting.

Secara arsitektural, keberadaan BiLSTM berkontribusi dalam menangkap konteks dua arah, sementara mekanisme *attention* membantu model memfokuskan representasi pada bagian teks yang paling relevan terhadap pengukuran kemiripan. Kombinasi ini terbukti efektif untuk tugas kesamaan semantik pada teks pendek seperti QQP.

Pada pengujian *document similarity*, model diterapkan pada 10 abstrak dokumen ilmiah (Tabel 4) dengan panjang 150–250 kata. Karena model dilatih pada teks pendek, dokumen dipecah menjadi segmen maksimal 30 kata dan skor *similarity* dihitung berdasarkan rata-rata tiga pasangan segmen dengan nilai tertinggi. Hasil menunjukkan bahwa skor *similarity* antar dokumen cenderung rendah, bahkan pada dokumen dengan topik serupa (Tabel 5).

Tabel 5. Detail 10 Dokumen Yang Digunakan Untuk Pembandingan Dokumen

No	Judul	Tahun	Tautan
1	*The regulation of artificial intelligence*	2024	https://link.springer.com/article/10.1007/s00146-023-01650-z
2	*Artificial Intelligence: from ethics to law*	2020	https://www.sciencedirect.com/science/article/abs/pii/S030859612030029X
3	*Artificial Intelligence in Education: a review*	2020	https://ieeexplore.ieee.org/abstract/document/9069875
4	*Artificial Intelligence Education Ethical Problems and Solutions*	2018	https://ieeexplore.ieee.org/abstract/document/8468773
5	*Artificial Intelligence as a New Category of Civil Law*	2020	https://www.ceeol.com/search/article-detail?id=914573
6	*Role of AI in education*	2023	https://www.researchgate.net/publication/369545925_Role_of_AI_in_Education
7	*State of the art and practice in AI in education*	2022	https://www.researchgate.net/publication/364994764_State_of_the_art_and_practice_in_AI_in_education
8	*Education for AI, not AI for Education: The Role of Education*	2022	https://link.springer.com/article/10.1007/s40593-021-00270-2
9	*Introduction to artificial intelligence in medicine*	2019	https://pubmed.ncbi.nlm.nih.gov/30810430/
10	*Overview of artificial intelligence in medicine*	2019	https://pmc.ncbi.nlm.nih.gov/articles/PMC6691444

Tabel 6. Hasil Pengujian Terhadap 10 Abstrak dari Dokumen

Rentang Nilai Similarity	Jumlah Pasangan Dokumen	Interpretasi Umum
0% - < 1%	13	Kesamaan semantik sangat rendah, dominan pada perbandingan lintas domain
1% - < 3%	14	Kesamaan rendah dengan sedikit irisan topik

3% - < 7%	7	Kesamaan rendah hingga menengah, umumnya pada domain yang berdekatan
7% - < 12%	7	Kesamaan relatif lebih tinggi pada dokumen dengan topik serupa
≥12%	4	Kesamaan tertinggi, dokumen sangat beririsan secara domain dan terminologi

Pengujian tingkat dokumen dilakukan sebagai studi kasus eksploratif (*case study*) menggunakan 10 pasang dokumen abstrak artikel ilmiah untuk melihat perilaku model saat menerima masukan teks panjang. Dokumen terlebih dahulu diurai menjadi segmen-segmen kalimat sebelum dihitung kemiripan rata-ratanya. Hasil pengujian menunjukkan nilai kemiripan yang relatif rendah (rata-rata skor $< 45\%$). Hal ini tidak menunjukkan kegagalan model, melainkan mengonfirmasi keterbatasan batas generalisasi (*domain bias*). Model yang dilatih khusus pada pasangan kalimat pendek berstruktur sederhana (QQP) mengalami penurunan responsivitas terhadap konteks wacana kompleks dan paragraf panjang. Oleh karena itu, klaim efektivitas arsitektur BiLSTM-*Additive Attention* dalam penelitian ini dibatasi secara ketat pada ranah pemodelan kesamaan semantik tingkat kalimat (*short-text STS*).

Rendahnya skor ini mengindikasikan keterbatasan model dalam merepresentasikan hubungan semantik global pada teks panjang. Hal ini diduga dipengaruhi oleh karakteristik dataset pelatihan QQP yang didominasi pertanyaan pendek dan bahasa non-teknis. Dengan demikian, model yang dilatih pada domain spesifik belum dapat digeneralisasi secara optimal ke dokumen ilmiah tanpa adaptasi data atau pelatihan lanjutan berbasis domain.

5. KESIMPULAN

Berdasarkan hasil eksperimen dan evaluasi yang telah dilakukan, dapat disimpulkan bahwa:

1. Integrasi arsitektur *Bidirectional Long Short-Term Memory* (BiLSTM) dan mekanisme *Additive Attention* berbasis representasi *frozen pre-trained* BERT *cross-encoder* terbukti efektif dalam memodelkan *Semantic Textual Similarity* (STS) pada tingkat kalimat (*short-text*). Evaluasi pada *benchmark dataset Quora Question Pairs* (QQP) menghasilkan nilai *loss* sebesar 0,3237, *accuracy* 0,8545, AUC 0,9307, serta *F1-score* 0,8837 untuk kelas non-duplikat dan 0,8059 untuk kelas duplikat.
2. Hasil eksperimen pembanding mengonfirmasi bahwa penambahan lapisan sekuensial BiLSTM dan pembobotan *Additive Attention* memberikan kontribusi peningkatan akurasi sebesar 4,25% dibandingkan penggunaan *baseline frozen* BERT murni.
3. Pengujian berbasis kasus pada skala dokumen (*document-level*) menunjukkan keterbatasan generalisasi model dengan skor kemiripan yang relatif rendah. Hal ini mengonfirmasi bahwa efektivitas arsitektur ini dibatasi oleh karakteristik *domain* data pelatihan QQP yang didominasi oleh pasangan kalimat pendek berstruktur sederhana.

6. SARAN

Berdasarkan hasil penelitian dan keterbatasan yang telah diidentifikasi, beberapa saran untuk pengembangan penelitian selanjutnya adalah sebagai berikut:

1. Melakukan perbandingan antara pendekatan *cross-encoder* dan *bi-encoder* untuk menganalisis *trade-off* antara performa, efisiensi komputasi, serta penerapannya pada teks dengan konteks yang lebih panjang, seperti dokumen.
2. Menggunakan *dataset* yang berbeda dan lebih representatif serta lebih fokus terhadap *long-text* dibanding *short-text*.

DAFTAR PUSTAKA

[1] T. Foltýnek, N. Meuschke, dan B. Gipp, "Academic plagiarism detection: A systematic literature review," *ACM Comput. Surv.*, vol. 52, no. 6, 2020, doi: 10.1145/3345317.

[2] A. Amirzhanov, C. Turan, dan A. Makhmutova, "Plagiarism types and detection methods: a systematic survey of algorithms in text analysis," *Front. Comput. Sci.*, vol. 7, no. Ml, 2025, doi: 10.3389/fcomp.2025.1504725.

[3] "Global Plagiarism Report 2018–2024: COVID & AI Impact." Diakses: 5 Januari 2026. [Daring]. Tersedia pada: https://plagiarismsearch.com/global-plagiarism-trends

[4] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, dan L. Specia, "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, hlm. 1–14, 2017, doi: 10.18653/v1/s17-2001.

[5] V. Thompson, "Methods for Detecting Paraphrase Plagiarism," hlm. 1–21, 2017, [Daring]. Tersedia pada: http://arxiv.org/abs/1712.10309

[6] M. Han, X. Zhang, X. Yuan, J. Jiang, W. Yun, dan C. Gao, "A survey on the techniques, applications, and performance of short text semantic similarity," *Concurr. Comput.*, vol. 33, no. 5, hlm. 1–17, 2021, doi: 10.1002/cpe.5971.

[7] Z. H. Amur, Y. Kwang Hooi, H. Bhanbhro, K. Dahri, dan G. M. Soomro, "Short-Text Semantic Similarity (STSS): Techniques, Challenges and Future Perspectives," *Applied Sciences (Switzerland)*, vol. 13, no. 6, 2023, doi: 10.3390/app13063911.

[8] A. Kamineni, M. Shrivastava, H. Yenala, dan M. Chinnakotla, "Siamese LSTM with Convolutional Similarity for Similar Question Retrieval," *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing, iSAI-NLP 2018 - Proceedings*, hlm. 1–7, 2018, doi: 10.1109/iSAI-NLP.2018.8692937.

[9] J. Chen, T. Zhang, Z. Yan, Z. Zheng, W. Zhang, dan J. Zhang, "Attention-based BiLSTM with positional embeddings for fake review detection," *J. Big Data*, vol. 12, no. 1, 2025, doi: 10.1186/s40537-025-01130-9.

[10] R. Patil, S. Boit, V. Gudivada, dan J. Nandigam, "A Survey of Text Representation and Embedding Techniques in NLP," *IEEE Access*, vol. 11, hlm. 36120–36146, 2023, doi: 10.1109/ACCESS.2023.3266377.

[11] D. Khurana, A. Koli, K. Khatter, dan S. Singh, "Natural language processing: state of the art, current trends and challenges | SpringerLink," hlm. 3713–3744, 2021, [Daring]. Tersedia pada: https://link.springer.com/article/10.1007/s11042-022-13428-4

[12] C. P. Chai, "Comparison of text preprocessing methods," *Nat. Lang. Eng.*, vol. 29, no. 3, hlm. 509–553, 2023, doi: 10.1017/S1351324922000213.

[13] S. Boudaa, T. Boudaa, dan A. El Haddadi, "Semantic Textual Similarity: Overview and Comparative Study between Arabic and English," *Computacion y Sistemas*, vol. 28, no. 3, hlm. 1209–1228, 2024, doi: 10.13053/CyS-28-3-5035.

[14] N. Reimers dan 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*, hlm. 3982–3992, 2019, doi: 10.18653/v1/d19-1410.

[15] J. Devlin, M. W. Chang, K. Lee, dan 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, no. Mlm, hlm. 4171–4186, 2019.

[16] S. Humeau, K. Shuster, M. A. Lachaux, dan J. Weston, "Poly-encoders: architectures and pre-training strategies for fast and accurate multi-sentence scoring," *8th International Conference on Learning Representations, ICLR 2020*, hlm. 1–14, 2020.

[17] I. D. Mienye, T. G. Swart, dan G. Obaido, "Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications," *Information (Switzerland)*, vol. 15, no. 9, hlm. 1–34, 2024, doi: 10.3390/info15090517.

[18] M. Krichen dan A. Mihoub, "Long Short-Term Memory Networks: A Comprehensive Survey," *AI (Switzerland)*, vol. 6, no. 9, hlm. 1–21, 2025, doi: 10.3390/ai6090215.

[19] S. Chaudhari, V. Mithal, G. Polatkan, dan R. Ramanath, "An Attentive Survey of Attention Models," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 5, hlm. 1–33, 2021, doi: 10.1145/3465055.

[20] Sneha, Soily Ghosh, et al., "BiLSTM-LIME: integrating NLP and advanced machine learning models for fake news detection" *Discover Artificial Intelligence*, vol. 6, 125 (2026), doi: 10.1007/s44163-026-00852-w

2622-8130
Jurnal Sifo Mikroskil (JSM) - Volume 27, No 2, Oktober 2026
DOI: https://doi.org/10.55601/jsm.v27i2.2092