JURNAL TRIDI
Teknologi Informatika & Komputer
Vol. 2 No. 2 Bulan 2025
Universitas Bani Saleh FTID
P-ISSN: - , E-ISSN:

PENGUKURAN KEMIRIPAN KALIMAT BAHASA INDONESIA MENGGUNAKAN REPRESENTASI WORD EMBEDDING FASTTEXT

Belinda Eka Sarah Dewi

Universitas Bani Saleh, belinda@ubs.ac.id

ABSTRAK

Pengukuran kemiripan kalimat merupakan komponen penting dalam berbagai aplikasi pemrosesan bahasa alami (NLP), seperti pencarian semantik, sistem tanya jawab, dan deteksi plagiarisme. Penelitian ini bertujuan untuk mengevaluasi kemampuan model word embedding FastText dalam mengukur kemiripan semantik antar kalimat berbahasa Indonesia. Dataset yang digunakan adalah *Semantic Textual Similarity Benchmark* (STS-B) versi Bahasa Indonesia, yang memuat pasangan kalimat beserta skor kemiripan berdasarkan penilaian manusia. Setiap kalimat direpresentasikan sebagai rata-rata vektor dari kata-kata penyusunnya menggunakan model FastText pralatih untuk Bahasa Indonesia. Kemiripan antar kalimat dihitung menggunakan *cosine similarity*, dan hasilnya dibandingkan dengan skor referensi manusia menggunakan korelasi Pearson dan Spearman. Hasil evaluasi menunjukkan bahwa FastText mampu menangkap sebagian besar makna semantik antar kalimat, dengan nilai korelasi Pearson sebesar 0.3658 dan Spearman sebesar 0.4260. Meskipun demikian, hasil tersebut mengindikasikan bahwa pendekatan berbasis *word-level embedding* seperti FastText memiliki keterbatasan dalam memahami konteks yang lebih kompleks. Penelitian ini memberikan gambaran awal mengenai efektivitas FastText dalam tugas pengukuran *similarity* semantik untuk Bahasa Indonesia dan membuka peluang pengembangan metode representasi yang lebih kontekstual di masa depan.

Kata Kunci: FastText, Word Embedding, Semantic Similarity, NLP

ABSTRACT

Measuring sentence similarity is a crucial component in various natural language processing (NLP) applications, such as semantic search, question answering systems, and plagiarism detection. This study aims to evaluate the capability of the FastText word embedding model in capturing semantic similarity between Indonesian-language sentences. The dataset used is the Indonesian version of the Semantic Textual Similarity Benchmark (STS-B), which contains sentence pairs along with human-annotated similarity scores. Each sentence is represented by the average of its word vectors using a pre-trained FastText model for Indonesian. The similarity between sentence pairs is computed using cosine similarity and compared with the human reference scores through Pearson and Spearman correlation analyses. The evaluation results show that FastText is able to capture a substantial degree of semantic meaning between sentences, achieving a Pearson correlation of 0.3658 and a Spearman correlation of 0.4260. However, these results also indicate that word-level embeddings such as FastText have limitations in capturing deeper contextual relationships. This study provides an initial overview of FastText's effectiveness in sentence similarity tasks for Indonesian and highlights the need for more context-aware representation approaches in future research.

Keywords: *FastText, Word Embedding, Semantic Similarity, NLP*

JURNAL TRIDI
Teknologi Informatika & Komputer
Vol. 2 No. 2 Bulan 2025
Universitas Bani Saleh FTID
P-ISSN: - , E-ISSN:

Naskah diterima : 15 Juli 2025, Naskah dipublikasikan : 31 Juli 2025

PENDAHULUAN

Pemrosesan bahasa alami (*Natural Language Processing* atau NLP) merupakan bidang dalam kecerdasan buatan yang berfokus pada interaksi antara komputer dan bahasa manusia. Salah satu tantangan utama dalam NLP adalah memahami makna semantik dari teks, terutama dalam bentuk kalimat. Salah satu tugas penting dalam hal ini adalah *Semantic Textual Similarity* (STS), yaitu pengukuran tingkat kemiripan makna antara dua kalimat. STS memiliki banyak aplikasi praktis, seperti sistem pencarian informasi berbasis makna, sistem tanya jawab, pendeteksian plagiarisme, serta penyaringan konten otomatis (Agirre et al., 2012; Cer et al., 2017). Dalam konteks bahasa Indonesia, pengembangan dan evaluasi model STS masih cukup terbatas dibandingkan dengan penelitian berbahasa Inggris (Wilie et al., 2020).

Untuk dapat mengukur kemiripan antar kalimat, teks perlu direpresentasikan dalam bentuk numerik menggunakan teknik *text representation*. Salah satu pendekatan yang paling umum digunakan adalah *word embedding*, yaitu representasi vektor dari kata-kata berdasarkan distribusi atau konteks kemunculannya dalam korpus teks. Representasi ini memungkinkan model menghitung kedekatan semantik antar kata, yang kemudian digunakan untuk membentuk representasi kalimat. Beberapa teknik embedding populer yang telah banyak digunakan dalam tugas STS antara lain Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), dan FastText (Bojanowski et al., 2017). Masing-masing memiliki kelebihan dan kekurangan tergantung pada jenis data dan bahasa yang digunakan.

Model Word2Vec menghasilkan vektor kata berdasarkan konteks sekitar menggunakan arsitektur *skip-gram* atau *CBOW* (Continuous Bag-of-Words). Sementara itu, GloVe memanfaatkan statistik global dari korpus untuk membangun representasi vektor. FastText, yang dikembangkan oleh Facebook AI Research, merupakan pengembangan dari Word2Vec yang menyertakan *subword information*, yaitu model tidak hanya mempelajari kata secara keseluruhan, tetapi juga bagian-bagian katanya seperti *n-grams*. Hal ini membuat FastText lebih unggul dalam menangani kata-kata baru (*out-of-vocabulary*) dan cocok digunakan pada bahasa yang memiliki morfologi kompleks seperti Bahasa Indonesia (Grave et al., 2018). Selain metode klasik ini, pendekatan modern berbasis *transformer* seperti BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), dan XLM-R (Conneau et al., 2020) menawarkan representasi kontekstual yang lebih dalam, namun memerlukan sumber daya komputasi yang jauh lebih besar.

Penelitian ini secara khusus memilih FastText sebagai metode representasi karena model ini menyediakan keseimbangan antara kualitas representasi semantik dan efisiensi komputasi. FastText memiliki versi pralatih (*pretrained*) untuk Bahasa Indonesia dan dapat dengan mudah diterapkan tanpa pelatihan ulang yang memerlukan GPU (Jauhar et al., 2021). Selain itu, pendekatan berbasis FastText masih relevan dalam banyak sistem NLP skala ringan, terutama pada aplikasi real-time, perangkat bergerak, atau sistem

JURNAL TRIDI
Teknologi Informatika & Komputer
Vol. 2 No. 2 Bulan 2025
Universitas Bani Saleh FTID
P-ISSN: - , E-ISSN:

dengan sumber daya terbatas. Meskipun model kontekstual seperti IndoBERT memiliki akurasi yang tinggi dalam memahami konteks, kebutuhan akan pemrosesan yang lebih cepat dan ringan menjadikan FastText sebagai pilihan yang tepat untuk penelitian ini.

Melalui penelitian ini, dilakukan evaluasi terhadap performa FastText dalam mengukur kemiripan antar kalimat Bahasa Indonesia menggunakan dataset STS-B versi terjemahan (Lazarus & Wibowo, 2022). Representasi vektor kalimat dibentuk dengan menghitung rata-rata dari vektor kata, dan skor kemiripan dihitung menggunakan *cosine similarity*. Hasil dari model kemudian dibandingkan dengan skor penilaian manusia melalui analisis korelasi Pearson dan Spearman. Penelitian ini diharapkan dapat memberikan gambaran mengenai efektivitas FastText dalam menangkap makna semantik pada kalimat Bahasa Indonesia, serta menjadi pijakan awal bagi pengembangan sistem NLP yang lebih canggih dalam konteks bahasa lokal.

TINJAUAN PUSTAKA

1. Semantic Textual Similarity (STS)

Tugas *Semantic Textual Similarity* (STS) bertujuan untuk mengukur seberapa mirip dua potongan teks secara semantik, biasanya dalam bentuk kalimat. STS telah menjadi salah satu topik penting dalam pengembangan sistem NLP sejak diperkenalkan dalam kompetisi SemEval (Semantic Evaluation) oleh Agirre et al. (2012), yang menjadi tonggak awal benchmarking model-model dalam memahami kedekatan makna antar teks. Dalam STS Benchmark (Cer et al., 2017), setiap pasangan kalimat diberi skor kemiripan dari 0 hingga 5 berdasarkan penilaian manusia, dan dijadikan standar evaluasi bagi model-model pembelajaran mesin maupun representasi teks.STS tidak hanya mengukur kesamaan kata (*lexical similarity*), tetapi juga mempertimbangkan struktur, sinonimi, dan pemahaman makna yang lebih dalam, sehingga diperlukan metode representasi yang mampu menangkap nuansa semantik dengan baik.

2. Word Embedding dan Representasi Semantik

Word embedding merupakan teknik untuk memetakan kata ke dalam ruang vektor berdimensi tetap yang menyimpan informasi semantik. Word2Vec, diperkenalkan oleh Mikolov et al. (2013), menggunakan pendekatan prediktif berbasis konteks lokal dengan arsitektur *skip-gram* atau *CBOW*. Sementara itu, GloVe (Pennington et al., 2014) menggunakan pendekatan statistik global untuk menghasilkan representasi kata. Kelemahan utama kedua metode ini adalah ketidakmampuannya dalam menangani kata baru atau morfologi kompleks karena setiap kata dianggap sebagai unit tunggal.

FastText (Bojanowski et al., 2017) hadir sebagai solusi dengan memasukkan informasi *subword* dalam proses pelatihan. Dengan menggunakan karakter *n-gram*, FastText dapat membentuk vektor untuk kata-kata yang sebelumnya tidak ditemukan (*out-of-vocabulary*/OOV), yang membuatnya lebih unggul terutama pada bahasa seperti Bahasa Indonesia yang kaya morfologi. Grave et al. (2018) memperluas model ini dengan melatih FastText untuk lebih dari 150 bahasa, termasuk Bahasa Indonesia, dan menyediakan pretrained vector yang dapat digunakan langsung untuk berbagai tugas NLP.

3. Representasi Kalimat dan Similarity Measurement

Dalam tugas STS, kalimat biasanya direpresentasikan dengan berbagai pendekatan, mulai dari metode sederhana seperti *bag-of-words*, hingga metode yang lebih canggih

JURNAL TRIDI
Teknologi Informatika & Komputer
Vol. 2 No. 2 Bulan 2025
Universitas Bani Saleh FTID
P-ISSN: - , E-ISSN:

seperti *sentence embedding*. Salah satu pendekatan yang umum dilakukan adalah merata-ratakan (*average pooling*) vektor kata dalam suatu kalimat untuk menghasilkan satu vektor representasi kalimat (Kenter & de Rijke, 2015). Meski sederhana, pendekatan ini masih banyak digunakan karena efisien dan dapat memberikan hasil yang cukup baik jika word embedding yang digunakan mengandung informasi semantik yang kaya. Kemiripan antar kalimat kemudian dihitung menggunakan metrik seperti *cosine similarity*, yang mengukur seberapa dekat arah vektor antar dua representasi kalimat. Nilai ini kemudian dibandingkan dengan skor manusia menggunakan korelasi Pearson dan Spearman (Cer et al., 2017) sebagai indikator performa model.

METODE PENELITIAN

1. Jenis Penelitian

Penelitian ini merupakan penelitian kuantitatif eksperimental dengan pendekatan evaluasi kinerja model NLP. Fokus utamanya adalah menganalisis seberapa baik representasi word embedding FastText dapat menangkap kemiripan semantik antar kalimat dalam Bahasa Indonesia berdasarkan pengukuran kesesuaian terhadap skor penilaian manusia.

2. Data Penelitian

Data yang digunakan dalam penelitian ini adalah dataset STS Benchmark Bahasa Indonesia yang diadaptasi dari *Semantic Textual Similarity Benchmark* (STS-B) dan telah diterjemahkan oleh Lazarus dan Wibowo (2022). Dataset ini berisi pasangan kalimat berbahasa Indonesia beserta skor penilaian kemiripan semantik yang diberikan oleh manusia, dalam skala 0 (sangat tidak mirip) hingga 5 (sangat mirip). Struktur dataset terdiri atas beberapa kolom seperti ditunjukkan pada Gambar 1, namun hanya kolom text_1, text_2, dan correlation (skor referensi manusia) yang digunakan dalam penelitian ini.

	domain	data	type	score	correlation	text_1	text_2
0	main-captions	MSRvid	2012test	0024	2.500	Seorang gadis sedang menata rambutnya.	Seorang gadis sedang menyisir rambutnya.
1	main-captions	MSRvid	2012test	0033	3.600	Sekelompok pria bermain sepak bola di pantai.	Sekelompok anak laki-laki sedang bermain sepak...
2	main-captions	MSRvid	2012test	0045	5.000	Seorang wanita mengukur pergelangan kaki wanit...	Seorang wanita mengukur pergelangan kaki wanit...
3	main-captions	MSRvid	2012test	0063	4.200	Seorang pria sedang memotong mentimun.	Seorang pria sedang mengiris mentimun.
4	main-captions	MSRvid	2012test	0066	1.500	Seorang pria sedang memainkan harpa.	Seorang pria sedang memainkan keyboard.

Gambar 1. Dataset STS Benchmark Bahasa Indonesia

3. Tahapan Penelitian

Proses penelitian dilakukan melalui beberapa tahapan sebagai berikut:

a. Pra-pemrosesan Data
   - Mengunduh dataset dari repositori Hugging Face.
   - Memastikan dataset bersih dari nilai kosong dan duplikat.

b. Pemuatan Model Word Embedding
   - Menggunakan FastText pretrained model Bahasa Indonesia yang telah dilatih oleh Facebook AI Research.
   - Model diunduh dalam format .vec, kemudian dimuat ke dalam memori menggunakan pustaka gensim.

c. Representasi Kalimat

JURNAL TRIDI
Teknologi Informatika & Komputer
Vol. 2 No. 2 Bulan 2025
Universitas Bani Saleh FTID
P-ISSN: - , E-ISSN:

- Masing-masing kalimat pada pasangan text_1 dan text_2 direpresentasikan sebagai rata-rata dari vektor kata penyusunnya (*average pooling*).
- Bila suatu kata tidak ditemukan dalam model, maka diabaikan dari perhitungan rata-rata.

d. Penghitungan Similarity
- Kemiripan antar pasangan kalimat dihitung menggunakan cosine similarity antara vektor representasi text_1 dan text_2. *Cosine similarity* mengukur kedekatan arah dua vektor dalam ruang berdimensi tinggi tanpa mempertimbangkan magnitudonya, sehingga sangat sesuai digunakan dalam model word embedding. Secara matematis, *cosine similarity* antara dua vektor kalimat A dan kalimat B didefinisikan sebagai berikut:

$$CosineSimilarity(A,B) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

Dengan:

$$A \cdot B = \sum_{i=1}^{n} A_i \cdot B_i$$

$$\|A\| = \sqrt{\sum_{i=1}^{n} A_i^2}$$

$$\|B\| = \sqrt{\sum_{i=1}^{n} B_i^2}$$

$\|A\|$ dan $\|B\|$ merupakan magnitudo (Panjang) dari vector A dan vector B.

e. Evaluasi Kinerja
- Skor *similarity* dari model dibandingkan dengan skor penilaian manusia menggunakan dua metrik korelasi:
  - *Pearson correlation coefficient*: untuk mengukur hubungan linier antara dua skor.
  - *Spearman rank correlation*: untuk mengukur kesesuaian urutan atau ranking antara dua skor.

f. Visualisasi Hasil
- Visualisasi scatterplot antara skor manusia dan skor model menggunakan pustaka matplotlib dan seaborn.

4. Alat dan Perangkat

Penelitian ini dilakukan menggunakan Google Colaboratory (Google Colab) sebagai lingkungan pemrograman. Bahasa pemrograman yang digunakan adalah Python 3. Perpustakaan (*library*) utama yang digunakan antara lain:

- gensim untuk pemrosesan model FastText
- pandas dan numpy untuk manipulasi data
- scipy.stats untuk analisis korelasi

JURNAL TRIDI
Teknologi Informatika & Komputer
Vol. 2 No. 2 Bulan 2025
Universitas Bani Saleh FTID
P-ISSN: - , E-ISSN:

- matplotlib dan seaborn untuk visualisasi hasil
- datasets dari Hugging Face untuk pemuatan dataset

5. Teknik Analisis Data

Analisis data dilakukan secara kuantitatif dengan menghitung nilai korelasi Pearson dan Spearman antara skor cosine similarity model FastText dengan skor kemiripan dari manusia. Interpretasi hasil mengikuti pedoman umum:

- Nilai korelasi mendekati 1.0 berarti prediksi model sangat mirip dengan penilaian manusia.
- Nilai korelasi mendekati 0.0 menunjukkan bahwa model gagal merepresentasikan kemiripan semantik.

Interpretasi disertai analisis visual distribusi data dan pembahasan atas keterbatasan metode.

HASIL DAN PEMBAHASAN

1. Analisis Hasil Evaluasi

Penelitian ini bertujuan untuk mengevaluasi kemampuan model word embedding FastText dalam mengukur kemiripan semantik antar kalimat berbahasa Indonesia. Setelah dilakukan proses representasi kalimat menggunakan rata-rata vektor kata dari FastText *pretrained* model Bahasa Indonesia, diperoleh skor cosine similarity untuk setiap pasangan kalimat. Skor tersebut kemudian dibandingkan dengan skor penilaian manusia dari dataset STS-B Indonesia dengan menggunakan metrik korelasi Pearson dan Spearman.

```
[ ] from scipy.stats import spearmanr, pearsonr
    import pandas as pd
    import numpy as np

    # Convert the 'correlation' column to numeric, coercing errors
    ground_truth_scores = pd.to_numeric(ds['test']['correlation'], errors='coerce')

    # Remove any rows where the conversion resulted in NaN
    valid_indices = ~np.isnan(ground_truth_scores)
    ground_truth_scores = ground_truth_scores[valid_indices]
    model_scores_filtered = np.array(model_scores)[valid_indices]


    pearson = pearsonr(model_scores_filtered, ground_truth_scores)[0]
    spearman = spearmanr(model_scores_filtered, ground_truth_scores)[0]

    print(f"Pearson Correlation: {pearson:.4f}")
    print(f"Spearman Correlation: {spearman:.4f}")

    Pearson Correlation: 0.3658
    Spearman Correlation: 0.4260
```

Gambar 2. Nilai Korelasi Pearson dan Korelasi Spearman

Hasil pengujian menunjukkan bahwa model FastText menghasilkan nilai korelasi Pearson sebesar 0.3658 dan nilai Spearman sebesar 0.4260 terhadap skor penilaian

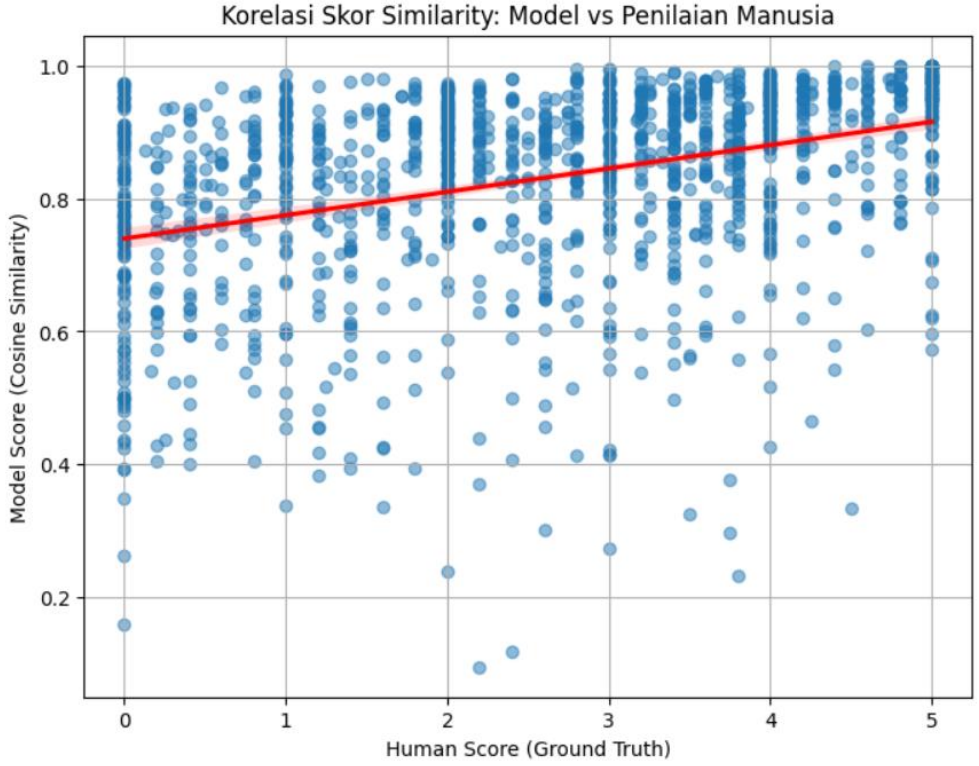
JURNAL TRIDI
Teknologi Informatika & Komputer
Vol. 2 No. 2 Bulan 2025
Universitas Bani Saleh FTID
P-ISSN: - , E-ISSN:

manusia seperti ditunjukkan pada Gambar 2. Nilai Pearson menunjukkan kekuatan hubungan linier antara hasil model dan penilaian manusia, sementara Spearman menunjukkan seberapa konsisten ranking kemiripan yang dihasilkan model dengan ranking dari penilaian manusia. Dengan kata lain, meskipun tidak menunjukkan hubungan yang sangat kuat, hasil ini mengindikasikan adanya kemampuan FastText dalam menangkap sebagian makna semantik antar kalimat Bahasa Indonesia.

2. Analisis Visualisasi Korelasi Kemiripan Kalimat

Gambar 3 menampilkan scatter plot yang menggambarkan hubungan antara skor kemiripan yang diberikan oleh model FastText (*cosine similarity*) dengan skor penilaian manusia *(ground truth*) dari dataset STS Bahasa Indonesia. Sumbu horizontal menunjukkan skor penilaian manusia yang berkisar antara 0 hingga 5, sedangkan sumbu vertikal menunjukkan skor *cosine similarity* dari model yang berkisar antara 0 hingga 1.

Dari visualisasi tersebut, dapat dilihat bahwa sebagian besar titik berakumulasi di rentang skor *cosine similarity* antara 0.7 hingga 1.0. Hal ini menunjukkan bahwa model FastText cenderung memberikan skor kemiripan yang tinggi, bahkan terhadap pasangan kalimat yang dinilai kurang mirip oleh manusia (misalnya skor 0–2). Distribusi ini mengindikasikan bahwa FastText memiliki keterbatasan dalam membedakan pasangan kalimat yang secara semantik berbeda tetapi memiliki kata-kata dengan kemiripan bentuk (morfologi) atau leksikal yang tinggi.



Gambar 3. Plot Korelasi Skor Similarity Model vs Penilaian Manusia

Gambar 3 menampilkan *scatter plot* yang menggambarkan hubungan antara skor kemiripan yang diberikan oleh model FastText (*cosine similarity*) dengan skor penilaian manusia (*ground truth*) dari dataset STS Bahasa Indonesia. Sumbu horizontal

JURNAL TRIDI
Teknologi Informatika & Komputer
Vol. 2 No. 2 Bulan 2025
Universitas Bani Saleh FTID
P-ISSN: - , E-ISSN:

menunjukkan skor penilaian manusia yang berkisar antara 0 hingga 5, sedangkan sumbu vertikal menunjukkan skor *cosine similarity* dari model yang berkisar antara 0 hingga 1. Dari visualisasi tersebut, dapat dilihat bahwa sebagian besar titik berakumulasi di rentang skor cosine similarity antara 0.7 hingga 1.0. Hal ini menunjukkan bahwa model FastText cenderung memberikan skor kemiripan yang tinggi, bahkan terhadap pasangan kalimat yang dinilai kurang mirip oleh manusia (misalnya skor 0–2). Distribusi ini mengindikasikan bahwa FastText memiliki keterbatasan dalam membedakan pasangan kalimat yang secara semantik berbeda tetapi memiliki kata-kata dengan kemiripan bentuk (morfologi) atau leksikal yang tinggi.

Meskipun demikian, terdapat tren positif yang tampak melalui garis regresi linear berwarna merah, yang menunjukkan kecenderungan meningkatnya skor *cosine similarity* seiring dengan naiknya skor penilaian manusia. Hal ini diperkuat oleh hasil evaluasi kuantitatif yang menghasilkan nilai korelasi Pearson sebesar 0.3658 dan korelasi Spearman sebesar 0.4260. Korelasi ini termasuk dalam kategori sedang, menandakan bahwa meskipun FastText belum mampu sepenuhnya menangkap makna semantik secara kompleks, model ini tetap memiliki kemampuan dasar untuk mengenali kemiripan kata dan frasa dalam Bahasa Indonesia. Secara keseluruhan, pendekatan ini masih bermanfaat dalam aplikasi NLP sederhana yang tidak membutuhkan pemahaman konteks secara mendalam, tetapi kurang efektif dalam menangani variasi sintaksis dan nuansa makna kalimat. Hal ini membuka ruang untuk penggunaan metode representasi kalimat yang lebih kontekstual, seperti BERT atau model transformer lainnya, dalam penelitian lanjutan.

3. Interpretasi Hasil

Nilai korelasi yang diperoleh berada pada kisaran sedang (*moderate*), yang berarti FastText cukup efektif dalam mengukur similarity semantik, meskipun belum ideal untuk tugas-tugas yang memerlukan pemahaman konteks yang lebih dalam. Model ini mampu mengenali kata-kata yang memiliki kemiripan morfologis atau bentuk turunan kata, berkat penggunaan *subword embedding*. Misalnya, kalimat yang menggunakan kata "menyisir" dan "menata rambut" dapat tetap memiliki representasi vektor yang saling berdekatan.

Namun, pendekatan ini memiliki keterbatasan dalam memahami konteks kalimat secara menyeluruh. Karena FastText menghitung representasi kalimat sebagai rata-rata dari kata-kata, model ini mengabaikan urutan kata dan hubungan sintaksis. Oleh karena itu, kalimat yang berbeda secara struktur tetapi memiliki kosakata serupa dapat dinilai terlalu mirip oleh model, meskipun secara semantik maknanya berbeda. Selain itu, FastText tidak mampu memahami makna kalimat dalam konteks pragmatis, ironi, atau implikatur, yang sering kali diperlukan dalam pemrosesan bahasa alami tingkat lanjut.

4. Implikasi dan Keterbatasan

Temuan dalam penelitian ini memberikan gambaran bahwa model word embedding tradisional seperti FastText masih dapat digunakan untuk tugas pengukuran kemiripan kalimat dalam Bahasa Indonesia, terutama pada sistem dengan keterbatasan sumber daya komputasi. Namun, pendekatan ini tidak cukup kuat untuk menangani kalimat yang mengandung ambiguitas makna atau konteks yang kompleks. Kelemahan lain dari metode ini adalah ketergantungan terhadap kelengkapan kosa kata dalam model pretrained yang digunakan.

JURNAL TRIDI
Teknologi Informatika & Komputer
Vol. 2 No. 2 Bulan 2025
Universitas Bani Saleh FTID
P-ISSN: - , E-ISSN:

Selain itu, pendekatan rata-rata embedding kata (*average pooling*) juga memiliki keterbatasan, karena menganggap semua kata dalam kalimat memiliki kontribusi yang sama terhadap makna keseluruhan kalimat.

PENUTUP

Simpulan

Penelitian ini bertujuan untuk mengevaluasi kemampuan model word embedding FastText dalam mengukur kemiripan semantik antar kalimat berbahasa Indonesia. Dengan menggunakan pendekatan rata-rata vektor kata (*average pooling*) dan pengukuran cosine similarity, representasi kalimat yang dihasilkan oleh FastText dibandingkan dengan skor penilaian manusia dari dataset STS-Bahasa Indonesia. Hasil evaluasi menunjukkan bahwa nilai korelasi Pearson sebesar 0.3658 dan Spearman sebesar 0.4260, yang mengindikasikan adanya hubungan positif sedang antara prediksi model dan penilaian manusia. Visualisasi scatter plot memperlihatkan bahwa model cenderung memberikan skor similarity tinggi secara umum, bahkan terhadap kalimat yang tidak terlalu mirip, sehingga mengindikasikan bahwa FastText belum mampu membedakan makna secara kontekstual secara menyeluruh.

Meskipun demikian, model FastText masih memiliki keunggulan dari sisi efisiensi dan kemampuannya menangani morfologi bahasa yang kaya, seperti Bahasa Indonesia. Oleh karena itu, pendekatan ini tetap relevan untuk aplikasi NLP ringan dan berbasis resource terbatas.

Saran

Berdasarkan hasil dan keterbatasan yang ditemukan dalam penelitian ini, beberapa saran dapat diajukan untuk pengembangan penelitian di masa mendatang:

1. Eksplorasi model yang lebih kontekstual seperti IndoBERT atau XLM-RoBERTa untuk melihat sejauh mana representasi berbasis konteks dapat meningkatkan korelasi dengan penilaian manusia.
2. Menggabungkan FastText dengan teknik representasi kalimat yang lebih kompleks, seperti attention mechanism, weighted pooling, atau sentence embedding, untuk meningkatkan akurasi dalam memahami makna kalimat.
3. Menambahkan analisis berbasis kategori kalimat, seperti kalimat naratif vs perintah, atau sinonim vs parafrase, agar diperoleh pemahaman yang lebih rinci terhadap kinerja model dalam skenario berbeda.
4. Mengembangkan dataset STS Bahasa Indonesia yang lebih besar dan bervariasi untuk memperkuat proses evaluasi model word embedding pada Bahasa Indonesia.

Daftar Pustaka

Agirre, E., Cer, D., Diab, M., & Gonzalez-Agirre, A. (2012). *SemEval-2012 Task 6: A pilot on semantic textual similarity*. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics* (pp. 385–393).

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). *Enriching Word Vectors with Subword Information*. *Transactions of the Association for Computational Linguistics*, 5, 135–146.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). *SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation*. *Proceedings of SemEval*, 1–14.

JURNAL TRIDI
Teknologi Informatika & Komputer
Vol. 2 No. 2 Bulan 2025
Universitas Bani Saleh FTID
P-ISSN: - , E-ISSN:

Conneau, A., et al. (2020). *Unsupervised cross-lingual representation learning at scale*. In *ACL*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In *NAACL-HLT*.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). *Learning word vectors for 157 languages*. In *LREC*.

Jauhar, S. K., Firdaus, A., & Pratama, A. (2021). *Evaluating Word Embeddings for Indonesian Textual Similarity*. In *International Conference on Data and Software Engineering (ICoDSE)*.

Kenter, T., & de Rijke, M. (2015). *Short text similarity with word embeddings*. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (pp. 1411–1420). ACM.

Lazarus, M. A., & Wibowo, R. A. (2022). *STS Benchmark for Indonesian: A Translated and Preprocessed Version*. Hugging Face Datasets. https://huggingface.co/datasets/LazarusNLP/stsb_mt_id

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. *arXiv preprint arXiv:1907.11692*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. *arXiv preprint arXiv:1301.3781*.

Pennington, J., Socher, R., & Manning, C. D. (2014). *GloVe: Global Vectors for Word Representation*. In *EMNLP*.

Wilie, B., Vincentio, K. A., Winata, G. I., & Cahyawijaya, S. (2020). *IndoNLU: Benchmark and resources for evaluating Indonesian Natural Language Understanding*. In *Findings of EMNLP*.