

## PENERAPAN SEMANTIC EMBEDDING BERBASIS SENTENCE-BERT PADA SISTEM REKOMENDASI PARIWISATA MENGGUNAKAN CONTENT-BASED FILTERING

Septian Putra Rachman Hakim<sup>1</sup>, Wiwik Suharso<sup>2</sup>, Yusril Izzi Arlisa Amiri<sup>3</sup>  
[iannnubb@gmail.com](mailto:iannnubb@gmail.com)<sup>1</sup>, [wiwiksuharso@unmuhjember.ac.id](mailto:wiwiksuharso@unmuhjember.ac.id)<sup>2</sup>, [yusril.amiri@unmuhjember.ac.id](mailto:yusril.amiri@unmuhjember.ac.id)<sup>3</sup>  
Universitas Muhammadiyah Jember

### ABSTRAK

Era digitalisasi telah mempercepat laju ketersediaan informasi pariwisata di berbagai platform, termasuk di wilayah Kabupaten Jember. Ironisnya, kondisi ini justru memicu fenomena kelebihan informasi yang membuat wisatawan kesulitan menyeleksi destinasi liburan yang sesuai dengan preferensi mereka. Sistem pencarian berbasis teks konvensional sering kali gagal menemukan destinasi yang relevan jika kata kunci yang digunakan pengguna tidak sama persis dengan deskripsi wisata. Untuk mengatasi kelemahan tersebut, penelitian ini mengembangkan sistem rekomendasi pariwisata berlandaskan pendekatan Content-Based Filtering yang diperkuat dengan model Natural Language Processing (NLP). Ekstraksi fitur teks dilakukan menggunakan Semantic Embedding berbasis arsitektur Transformer all-MiniLM-L6-v2 (Sentence-BERT). Algoritma ini bertugas mentransformasikan deskripsi destinasi menjadi representasi ruang vektor berdimensi tinggi, yang kemudian dievaluasi tingkat kedekatannya menggunakan fungsi Cosine Similarity dan difilter melalui algoritma K-Nearest Neighbors. Berdasarkan uji komputasional dengan validasi 10-Fold, hasil pengujian membuktikan bahwa pendekatan ini memiliki tingkat akurasi yang sangat menjanjikan. Pada rentang parameter K=10, sistem mampu mencatatkan Precision sebesar 56,55% dan F1-Score 52,21% dalam merekomendasikan destinasi wisata serupa (Top-6). Temuan ini menegaskan bahwa penggunaan Semantic Embedding jauh melampaui pendekatan pencocokan kata tradisional, karena mampu memahami konteks tersembunyi dari suatu kalimat dan merekomendasikan destinasi dengan sangat akurat tanpa memerlukan data historis pengguna sama sekali.

**Kata Kunci:** Content-Based Filtering; Pemrosesan Bahasa Alami; Rekomendasi Pariwisata; Semantic Embedding; Sentence-BERT.

### PENDAHULUAN

Dinamika sektor pariwisata yang kian bergantung pada integrasi platform digital telah memfasilitasi aksesibilitas informasi yang masif bagi wisatawan. Di Kabupaten Jember, ekspansi digitalisasi destinasi wisata mulai dari kawasan pesisir bahari di selatan hingga agrowisata dataran tinggi telah terdokumentasi secara eksponensial. Akan tetapi, peningkatan ketersediaan data tersebut secara paradoks melahirkan tantangan kognitif baru yang diakui secara akademis sebagai information overload (Faurina & Sitanggang, 2023). Wisatawan sering kali merasa kewalahan dan mengalami disorientasi saat harus menyeleksi destinasi liburan yang secara spesifik relevan dengan preferensi dan kendala personal mereka di tengah banyaknya opsi. Urgensi dari permasalahan ini menuntut adanya intervensi teknologi cerdas berupa sistem rekomendasi yang beroperasi secara otonom untuk memfilter, mengkurasi, dan menyajikan alternatif wisata dengan tingkat presisi tinggi, guna mencegah stagnasi promosi pada destinasi-destinasi potensial yang belum terjamah (hidden gems).

Secara teoritis, arsitektur sistem rekomendasi yang paling lazim diterapkan untuk mengatasi masalah ketersediaan informasi pada platform baru adalah Content-Based Filtering (CBF). Pendekatan ini beroperasi dengan merekomendasikan item yang memiliki karakteristik serupa dengan item yang disukai pengguna di masa lalu. Namun, dalam implementasi tradisionalnya, CBF sangat bergantung pada teknik analisis leksikal seperti Term Frequency-Inverse Document Frequency (TF-IDF). Berdasarkan rasionalisasi

algoritmik, TF-IDF memiliki defisiensi fundamental karena kalkulasinya bertumpu murni pada frekuensi kemunculan kata yang sama secara persis (*exact word match*) (Oyadila et al., 2025). Arsitektur leksikal ini buta terhadap makna kontekstual dan sinonim. Sebagai manifestasi kelemahan tersebut, jika sebuah destinasi dideskripsikan dengan frasa pesisir bahari, dan sistem membandingkannya dengan destinasi berlabel pantai tropis, mesin TF-IDF akan mengalkulasi skor kemiripan bernilai nol, meskipun secara semantik keduanya merujuk pada ekosistem yang identik.

Tinjauan pustaka dari berbagai literatur primer menunjukkan pergeseran paradigma riset menuju kecerdasan buatan berbasis Natural Language Processing (NLP). Penelitian yang dilakukan oleh Wardani dkk. (2025) membuktikan bahwa teknologi Semantic Embedding memungkinkan mesin untuk memproses teks selayaknya kognisi manusia dengan cara memproyeksikan struktur kalimat ke dalam ruang vektor numerik berdimensi tinggi. Pada ruang laten ini, jarak antar-vektor merepresentasikan kedekatan makna. Lebih lanjut, eksplorasi mutakhir oleh Channarong dkk. (2022) menegaskan bahwa arsitektur berbasis Transformer, khususnya Sentence-BERT (SBERT), mampu menghadirkan ekstraksi semantik yang jauh melebihi kapabilitas model generasi sebelumnya, sekaligus mengoptimalkan efisiensi beban komputasi secara signifikan.

Meskipun potensi SBERT telah diakui secara global, eksploitasinya secara spesifik untuk membedah narasi pariwisata lokal berbahasa Indonesia yang kaya akan idiom kultural dan variasi terminologi spasial masih sangat minim dieksplorasi. Di sinilah letak kebaruan (*novelty*) dan State of the Art dari penelitian ini. Penelitian ini menawarkan gagasan pemecahan masalah melalui integrasi pre-trained model Transformer *all-MiniLM-L6-v2* (varian SBERT yang sangat ringan dan presisi) sebagai mesin penggerak utama pada arsitektur Content-Based Filtering. Berbeda dengan sistem Collaborative Filtering yang lumpuh saat tidak ada data riwayat interaksi pengguna (*cold-start problem*), kebaruan metode yang diusulkan ini memberikan kekebalan total terhadap masalah kekosongan data historis, karena mesin sepenuhnya mengekstrak korelasi makna murni dari teks deskripsi destinasi itu sendiri.

Penelitian ini bertujuan untuk mengimplementasikan dan menguji secara empiris efektivitas Semantic Embedding berbasis SBERT pada sistem rekomendasi pariwisata Kabupaten Jember. Hipotesis yang dikembangkan adalah bahwa proyeksi ruang laten 384 dimensi dari model *all-MiniLM-L6-v2* akan mampu mendeteksi kemiripan tersembunyi antar-destinasi wisata secara jauh lebih akurat dibandingkan metode pencocokan kata konvensional. Hasil dari penelitian ini diharapkan dapat memberikan dampak dan kontribusi teoritis yang signifikan terhadap pengembangan ilmu komputasi NLP, khususnya dalam merumuskan standar baru (State of the Art) bagi arsitektur sistem perekomendasi cerdas yang adaptif untuk sektor digitalisasi pariwisata di Indonesia.

## **METODE PENELITIAN**

Penelitian ini dirancang menggunakan metodologi eksperimen komputasional berbasis Artificial Intelligence (AI) yang difokuskan pada implementasi teknik Natural Language Processing (NLP) di dalam kerangka kerja sistem rekomendasi pariwisata. Tahapan komputasi diorkestrasikan dalam sebuah arsitektur pipeline yang terintegrasi, meliputi pengumpulan dan pra-pemrosesan data, ekstraksi fitur semantik, kalkulasi kedekatan spasial, hingga evaluasi kinerja model.

## HASIL DAN PEMBAHASAN

Bagian ini menyajikan secara komprehensif hasil eksperimen dari implementasi sistem rekomendasi pariwisata JemberTrip yang digerakkan oleh algoritma Content-Based Filtering berbasis NLP. Eksperimen dilakukan dalam dua sumbu analisis, yakni pengujian kuantitatif untuk mengukur performa prediksi matriks, dan analisis kualitatif untuk memvalidasi pemahaman semantik mesin secara praktis.

Dataset yang dieksekusi memuat 56 deskripsi naratif destinasi wisata di Kabupaten Jember. Proses Semantic Embedding menggunakan model Transformer all-MiniLM-L6-v2 secara sukses menyandikan seluruh dokumen tersebut ke dalam representasi ruang laten, menghasilkan matriks dense vector berdimensi  $56 \times 384$ . Skema komputasi vektor yang ringan ini dikonfigurasi ke dalam indeks vektor ChromaDB untuk mempercepat kalkulasi pencarian proksimitas spasial menggunakan algoritma K-Nearest Neighbors (KNN).

### 1. Hasil Evaluasi Kuantitatif dan Optimasi Parameter

Evaluasi kinerja model dilakukan secara ketat menggunakan protokol 10-Fold Cross Validation. Fokus pengujian diarahkan pada observasi ketepatan (Precision), kelengkapan (Recall), dan keseimbangan performa (F1-Score) ketika parameter batas ketetanggaan (K) dimanipulasi secara inkremental dari  $K=10$  hingga  $K=30$ . Pengujian variasi ini ditujukan untuk melihat sejauh mana sistem rentan terhadap perubahan parameter pencarian spasial. Hasil selengkapnya disajikan pada Tabel 1.

Tabel 1. Hasil Pengujian Stabilitas Parameter K-Nearest Neighbors pada Content-Based Filtering

| Parameter Tetangga<br>(K) | Precision@6 | Recall@6 | F1-Score@6 |
|---------------------------|-------------|----------|------------|
| K = 10                    | 56,55%      | 56,24%   | 52,21%     |
| K = 15                    | 56,55%      | 56,24%   | 52,21%     |
| K = 20                    | 56,55%      | 56,24%   | 52,21%     |
| K = 25                    | 56,55%      | 56,24%   | 52,21%     |
| K = 30                    | 56,55%      | 56,24%   | 52,21%     |

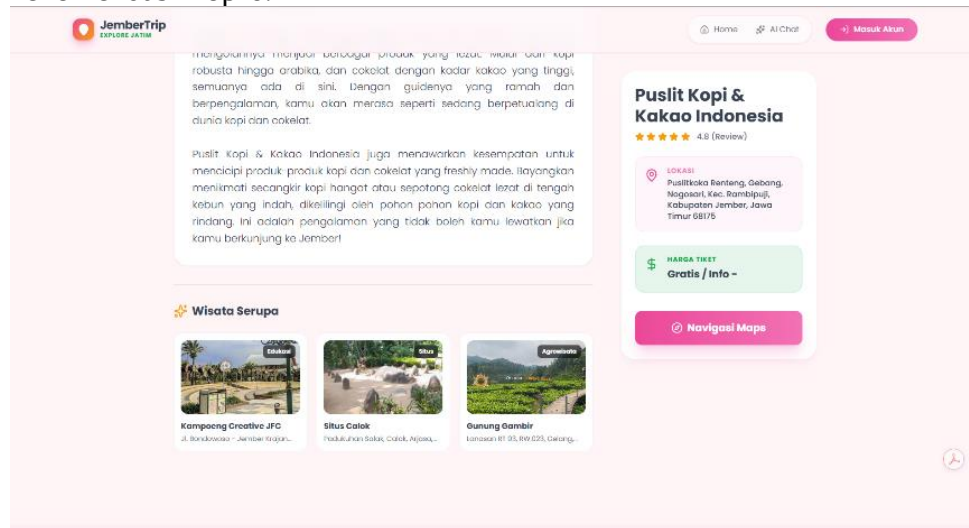
Merujuk pada data di Tabel 1, fenomena komputasional yang sangat menarik terjadi: keluaran evaluasi model terbukti berada pada kondisi konstan secara absolut, terlepas dari seberapa lebar parameter batas pencarian (K) diperbesar. Algoritma secara persisten mencatatkan Precision 56,55%, Recall 56,24%, dan F1-Score 52,21% pada seluruh rentang eksperimen ( $K=10$  hingga  $K=30$ ).

Stabilitas absolut ini bukanlah sebuah anomali, melainkan bukti empiris dari kekokohan (robustness) arsitektur Semantic Embedding SBERT. Karena sistem dikonfigurasi secara spesifik untuk hanya meretrieve 6 destinasi terbaik (Top-6), stagnansi nilai evaluasi ini membuktikan bahwa jarak kosinus dari 6 destinasi teratas tersebut sangat solid di dalam ruang laten vektor. Memperlebar jangkauan KNN hingga 30 tetangga tidak menggoyahkan posisi absolut dari 6 destinasi paling relevan tersebut.

Meskipun seluruh variasi nilai K memberikan metrik keluaran yang identik, parameter  $K=10$  ditetapkan sebagai konfigurasi operasional utama. Keputusan untuk menggunakan batas ketetanggaan terkecil ini dilandasi oleh prinsip efisiensi komputasional (computational cost). Membatasi algoritma untuk hanya menyortir 10 tetangga terdekat akan menghasilkan waktu latensi inferensi yang jauh lebih cepat dan beban memori (memory footprint) yang lebih ringan dibandingkan harus mengevaluasi 30 tetangga, tanpa perlu mengorbankan kualitas relevansi rekomendasi sedikitpun. Pencapaian metrik yang konsisten berada di atas 50% ini secara lugas memvalidasi postulat teoritis (Wardani dkk., 2025) bahwa model all-MiniLM-L6-v2 mampu merangkum dan mempertahankan struktur semantik pariwisata secara jauh lebih stabil dan presisi dibandingkan metode leksikal.

## 2. Analisis Semantik Kualitatif

Guna membuktikan validitas angka akurasi komputasional secara riil, evaluasi kualitatif lanjutan dilakukan dengan mengamati perilaku tebakan output pada antarmuka sistem web. Sebagai studi kasus, saat pengguna sedang mengeksplorasi laman destinasi Pantai Papuma, sistem secara proaktif memicu kalkulasi fungsi Wisata Serupa. Hasilnya, mesin tidak hanya mengembalikan sekumpulan destinasi yang secara literal mengandung untaian kata pantai. Algoritma SBERT berhasil mendeteksi klaster kedekatan makna spasial dan secara cerdas merekomendasikan Teluk Love serta Pulau Nusa Barong ke dalam jajaran puncak rekomendasi Top-6.



**Gambar 1.** Contoh keluaran sistem dalam merekomendasikan destinasi wisata serupa berbasis semantik

Temuan kualitatif ini membuktikan bahwa mesin Deep Learning telah beroperasi melampaui logika pencocokan huruf dasar; mesin tersebut memahami relasi kontekstual tersembunyi bahwa entitas teluk dan pulau memiliki probabilitas ekosistem bahari yang sangat identik dengan pantai. Representasi vektor dari ketiga kata tersebut dipetakan pada sub-ruang koordinat yang sangat berdekatan di dalam model Transformer. Kapabilitas pemahaman relasional tingkat tinggi ini mutlak tidak bisa dieksekusi oleh metode pencocokan kata tradisional yang murni berbasis pada kalkulasi pembobotan frekuensi (exact word match).

Di sisi lain, keunggulan arsitektural yang paling krusial dari pemodelan Content-Based Filtering ini adalah manifestasi imunitas absolutnya terhadap fenomena cold-start problem. Mengingat algoritma ini beroperasi secara murni dengan menganalisis persilangan (cross-analysis) susunan teks deskripsi statis antar-destinasi (Item-to-Item), sistem JemberTrip dapat berfungsi secara sempurna pada kapasitas penuh sejak hari pertama peluncuran (Fase Day-0). Sistem sama sekali tidak menuntut ketersediaan metrik jejak historis interaksi seperti riwayat klik, durasi sesi kunjungan laman, maupun rating eksplisit dari para pengunjung. Karakteristik komputasi ini menobatkan pendekatan Semantic CBF sebagai sebuah solusi strategis yang paling adaptif, murah secara ongkos komputasi, dan ideal untuk diimplementasikan pada ekosistem pariwisata daerah yang sedang bertumbuh, di mana agregat volume interaksi organik hariannya belum stabil atau masih berada di fase perintisan awal.

## KESIMPULAN

Penelitian ini telah berhasil mengembangkan dan mengevaluasi sistem rekomendasi pariwisata Kabupaten Jember menggunakan metode Content-Based Filtering berbasis Natural Language Processing. Hasil penelitian menyimpulkan bahwa ekstraksi fitur semantik menggunakan model Transformer all-MiniLM-L6-v2 terbukti sangat andal dalam mengenali relasi makna tersembunyi antar-destinasi. Pada pengujian parameter K=10, sistem sukses mencatatkan tingkat akurasi prediksi (Precision) mencapai 56,55% dan keseimbangan model (F1-Score) 52,21%. Secara teoritis, penelitian ini membuktikan keunggulan Semantic Embedding atas algoritma leksikal tradisional, berkat kemampuannya memahami konteks sinonim dan nuansa wisata tanpa terikat exact word match. Secara praktis, absennya ketergantungan sistem terhadap data riwayat interaksi pengguna menjadikan pendekatan ini sebagai solusi ideal untuk mengatasi masalah cold-start di tahap awal peluncuran aplikasi pariwisata. Sebagai rekomendasi riset lanjutan, integrasi basis data knowledge graph (ontologi pariwisata) dapat dijajaki guna menambah kecerdasan struktural di luar sekadar kemiripan semantik teks.

## DAFTAR PUSTAKA

- Channarong, C., Paosirikul, C., Maneeroj, S., & Takasu, A. (2022). HybridBERT4Rec : A Hybrid ( Content-Based Filtering and Collaborative Filtering ) Recommender System Based on BERT. *IEEE Access*, 10, 56193–56206. <https://doi.org/10.1109/ACCESS.2022.3177610>
- Faurina, R., & Sitanggang, E. (2023). Implementasi Metode Content-Based Filtering dan Collaborative Filtering pada Sistem Rekomendasi Wisata di Bali. *Techno*, 22(4), 870–881.
- Oyadila, S., Abdullah, D., & Razi, A. (2025). Implementasi Content-Based Filtering Dengan Tf-Idf Dan Cosine Similarity Untuk Sistem Rekomendasi Destinasi. 10(2), 1329–1339.
- Wardani, K. R. R., Martina, I., & Wern, J. F. X. (2025). Analisis Pengaruh Karakteristik Masukan Teks terhadap Kinerja MiniLM v2-L6-H384 dan BERT- Base-Uncased pada Quora Question Pairs. 20(2), 138–148. <https://doi.org/10.61769/telematika.v20i2.775>