Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.1, February 2026, pp. 935~945
e-ISSN: 2548-6861

Recommendation System Yogyakarta Tourism Using TF-IDF and Cosine Similarity Methods with Word Normalizer

Jauhar Fauzi Ulul Albab[1]*, Arif Nur Rohman[2]**,**
* Sistem Informasi, Universitas Amikom Yogyakarta
** Magister Teknik Informatika, Universitas Amikom Yogyakarta
jaufauzi@students.amikom.ac.id [1], arifrahman@amikom.ac.id [2]

Article Info

Article history:

Received 2025-11-12
Revised 2026-01-09
Accepted 2026-01-13

Keyword:

TF-IDF,
Cosine Similarity,
Word Normalizer.

ABSTRACT

The abundance of tourism information in Yogyakarta often overwhelms tourists due to non-standard text data. This research develops a tourism recommendation system using Content-Based Filtering by integrating TF-IDF and Cosine Similarity algorithms, enhanced with a Word Normalizer stage. The research method involves data preprocessing including case folding, filtering, stopword removal, and stemming combined with word normalization to standardize irregular spellings. Text feature representation is calculated using TF-IDF weighting, followed by measuring similarity between destinations through vector-based Cosine Similarity. The query testing of Pantai Parangtritis against Pantai Ngandong yielded the highest similarity score of 0.9397. System performance evaluation showed a Precision@5 of 0.84, Recall@5 of 0.10, and Mean Average Precision (MAP) of 0.81. In conclusion, strengthening the method with a Word Normalizer significantly improves the validity of top-ranked recommendations, enabling tourists to accurately find relevant attractions according to their preferences.

I. PENDAHULUAN

Daerah Istimewa Yogyakarta (DIY) merupakan salah satu destinasi wisata unggulan di Indonesia yang dikenal memiliki kekayaan budaya, sejarah, kuliner, dan keindahan alam yang memukau. Kota ini menjadi magnet utama bagi wisatawan domestik maupun mancanegara karena keunikannya sebagai kota pelajar yang sarat nilai tradisi sekaligus pusat kreativitas modern. Berdasarkan data Badan Pusat Statistik (BPS) tahun 2025, tercatat sebanyak 38.030.739 wisatawan nusantara berkunjung ke DIY sepanjang tahun 2024, meningkat sebesar 24,95% dibanding tahun sebelumnya. Angka ini memperlihatkan betapa besarnya daya tarik pariwisata Yogyakarta sebagai penggerak ekonomi daerah. Selain itu, menurut [1], selama periode long weekend tercatat 55.904 kendaraan memasuki kawasan pusat kota, menandakan ketergantungan wisatawan terhadap destinasi utama seperti Malioboro, Keraton, dan Tugu Yogyakarta. Fenomena ini menggambarkan bahwa sektor pariwisata tidak hanya berperan sebagai sumber pendapatan daerah, tetapi juga sebagai instrumen penting dalam memperluas lapangan kerja dan mendorong pembangunan berkelanjutan berbasis potensi lokal.

Berbagai upaya telah dilakukan untuk mendukung pengembangan dan penyebaran informasi destinasi wisata agar lebih efisien di era digital. Pemanfaatan teknologi informasi dan sistem berbasis data menjadi langkah strategis untuk memberikan kemudahan bagi wisatawan dalam menentukan destinasi sesuai minat dan preferensinya. Berbagai penelitian sebelumnya telah menunjukkan efektivitas pendekatan Content-Based Filtering dalam menghasilkan sistem rekomendasi wisata yang relevan. Misalnya, penelitian oleh [2]. mengembangkan sistem rekomendasi wisata berbasis *TF-IDF* dan *Cosine Similarity* yang berfokus pada kemiripan konten deskriptif antar destinasi. Temuan tersebut sejalan dengan hasil penelitian [3] yang membuktikan bahwa kombinasi kedua metode tersebut mampu mengukur tingkat kemiripan antar-item dengan baik, meskipun diterapkan pada domain olahraga. Selain itu, [4] dan [5] juga memperkuat bukti efektivitas *TF-IDF* dan *Cosine Similarity* dalam menghasilkan rekomendasi berbasis teks yang relevan di berbagai konteks aplikasi. Dengan demikian, dapat disimpulkan bahwa pendekatan berbasis konten masih menjadi salah satu metode yang paling banyak digunakan dalam pengembangan sistem rekomendasi modern.

Meskipun demikian, berbagai studi tersebut menunjukkan bahwa masih terdapat keterbatasan dalam aspek prapemrosesan data teks, khususnya dalam konteks bahasa alami pengguna Indonesia. Sebagian besar penelitian sebelumnya belum mengintegrasikan tahap *Word Normalizer*, yang berfungsi untuk menyesuaikan bentuk kata tidak baku, slang, singkatan, atau ejaan nonstandar yang kerap muncul dalam ulasan pengguna. Ketidakteraturan bentuk bahasa ini dapat menurunkan akurasi hasil pembobotan TF-IDF maupun perhitungan kesamaan menggunakan Cosine Similarity. Sejumlah penelitian seperti[6], [7], dan [8]memang telah memanfaatkan metode tersebut untuk pengembangan sistem rekomendasi wisata, namun belum secara khusus memperhatikan permasalahan variasi linguistik dalam data teks. Padahal, normalisasi bahasa memiliki peran penting untuk memastikan keseragaman bentuk kata sebelum tahap pembobotan. Selain itu, penelitian [9] menekankan pentingnya analisis opini publik menggunakan pendekatan *Natural Language Processing (NLP)* guna meningkatkan kualitas hasil analisis. Inovasi dalam penanganan kompleksitas dan nuansa bahasa ini, bahkan dalam domain yang sangat spesifik, telah menghasilkan pendekatan hibrida seperti penggabungan TF-IDF [10]. Sementara [11] berhasil menunjukkan efektivitas sistem rekomendasi berbasis konten video di YouTube yang menggunakan metode serupa. Fakta-fakta ini memperlihatkan bahwa penggabungan teknik *TF-IDF*, *Cosine Similarity*, dan *Word Normalizer* belum banyak diterapkan secara komprehensif dalam sistem rekomendasi berbasis teks untuk domain pariwisata di Indonesia.

Penelitian ini bertujuan untuk mengembangkan sistem rekomendasi wisata di Yogyakarta yang mengintegrasikan metode *TF-IDF* dan *Cosine Similarity* dengan tahapan Word Normalizer untuk meningkatkan ketepatan hasil rekomendasi berbasis konten deskriptif. Melalui pendekatan ini, sistem diharapkan mampu mengidentifikasi tingkat kemiripan antar destinasi wisata secara lebih tepat berdasarkan representasi teks yang telah dinormalisasi, sehingga rekomendasi yang dihasilkan menjadi lebih relevan. Penelitian ini secara khusus mencakup tiga tahapan utama: (1) pembangunan model pembobotan teks menggunakan *TF-IDF*, (2) pengukuran kemiripan antar destinasi wisata dengan *Cosine Similarity*, dan (3) penerapan *Word Normalizer* dalam proses prapemrosesan data untuk meningkatkan akurasi hasil analisis. Hasil akhir dari penelitian ini diharapkan dapat membantu wisatawan menemukan destinasi yang sesuai dengan minat dan preferensinya, sekaligus mendorong pemerataan promosi destinasi di seluruh wilayah Yogyakarta agar tidak hanya terpusat pada kawasan wisata populer.

Kebaruan (novelty) penelitian ini terletak pada pengembangan Word Normalizer berbasis lexical library hasil ekstraksi korpus lokal pariwisata Yogyakarta yang mampu melakukan koreksi typo, variasi ejaan, dan substitusi karakter non-alfabetik secara otomatis sebelum proses TF-IDF. Pendekatan ini berbeda dari teknik NLP modern berbasis embedding karena menekankan normalisasi linguistik berbasis domain-spesifik yang secara empiris meningkatkan konsistensi representasi teks.

II. METODE

Penelitian ini menggunakan metode penerapan metode sistem rekomendasi berbasis *konten (Content-Based Filtering)* yang dikombinasikan dengan teknik Term *Frequency–Inverse Document Frequency (TF-IDF), Cosine Similarity*, serta *Word Normalization*. Pendekatan ini dipilih karena mampu mengidentifikasi tingkat kemiripan antar deskripsi destinasi wisata secara matematis dan memberikan rekomendasi berbasis teks yang relevan terhadap preferensi pengguna.

Desain penelitian ini berfokus pada analisis data tekstual dengan tahapan metodologis yang terstruktur, meliputi:

1) Pengumpulan dan seleksi *dataset*, dilakukan untuk memperoleh sumber data yang representatif.
2) Prapemrosesan teks, meliputi pembersihan data, normalisasi kata, *tokenisasi*, dan penghapusan stopword.
3) Ekstraksi fitur menggunakan pembobotan *TF-IDF*, untuk mengubah teks menjadi representasi numerik yang dapat dianalisis secara komputasional.
4) Perhitungan tingkat kesamaan antar dokumen menggunakan *Cosine Similarity*, untuk mengukur kedekatan makna antar deskripsi wisata.
5) Implementasi dan pengujian sistem berbasis web, guna mengevaluasi efektivitas model dalam memberikan hasil rekomendasi destinasi yang relevan.

TF-IDF terbukti efektif sebagai metode ekstraksi fitur teks yang mampu mengubah data teks menjadi vektor numerik untuk analisis lebih lanjut, seperti yang diterapkan dalam penelitian analisis sentimen gambar AI generative [12]. Model eksperimen ini bertujuan menghasilkan sistem rekomendasi yang dapat mengidentifikasi hubungan semantik antar destinasi wisata berdasarkan deskripsi konten, sehingga dapat digunakan untuk mendukung pengambilan keputusan dalam promosi wisata berbasis data digital sejalan dengan penelitian sistem rekomendasi berbasis kesamaan semantik yang memanfaatkan Sentence Transformer dan algoritma Cosine Similarity pada data deskripsi konten [13].

A. Dataset dan Sumber Data

Dataset yang digunakan dalam penelitian ini bersumber dari platform Kaggle dengan judul "*Dataset* Wisata Jogja Sekitar" yang dikembangkan oleh Faris Rizqiawan (2023) dan diperbarui pada tahun 2025. *Dataset* ini berisi 475 data destinasi wisata yang mencakup wilayah Daerah Istimewa Yogyakarta dan sekitarnya, sehingga relevan dengan konteks penelitian dan memiliki validitas temporal yang tinggi.

Setiap data merepresentasikan satu destinasi wisata dan dilengkapi dengan deskripsi tekstual berbahasa Indonesia yang memuat informasi mengenai karakteristik, daya tarik, dan fasilitas destinasi. Panjang deskripsi setiap data bervariasi antara 36 hingga 334 kata. Dataset ini juga memiliki variasi

kategori wisata yang beragam, yang memungkinkan sistem yang dikembangkan memiliki kemampuan generalisasi yang baik dalam merekomendasikan atau mengklasifikasikan berbagai jenis destinasi wisata.

TABLE 1
RINGKASAN DATASET PARIWISATA

Aspek Dataset	Keterangan
Jumlah Destinasi	475 destinasi
Bahasa	Bahasa Indonesia
Bentuk Data	Teks deskripsi destinasi
Panjang Deskripsi	36 – 334 kata
Jumlah Kategori	7 kategori
Variasi Kategori	Pantai, Desa Wisata, Budaya dan Sejarah, Buatan, Wisata Air, Agrowisata, Alam

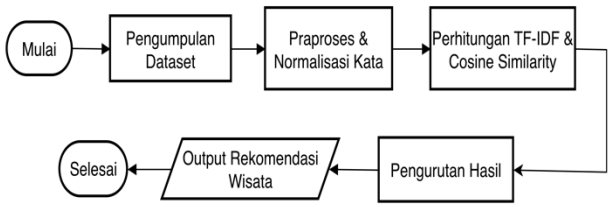
Keberadaan tujuh kategori wisata tersebut menunjukkan bahwa dataset ini tidak hanya merepresentasikan satu jenis pariwisata, tetapi mencakup berbagai tipe destinasi, mulai dari wisata alam, budaya, hingga wisata buatan. Hal ini memberikan cakupan data yang luas dan meningkatkan kemampuan generalisasi sistem yang dikembangkan dalam penelitian ini.

Data mencakup atribut utama seperti nama destinasi, kategori wisata, lokasi, dan deskripsi wisata. Kolom deskripsi digunakan sebagai sumber utama dalam proses analisis teks karena berisi uraian naratif yang mencerminkan keunikan setiap destinasi. Dataset ini kemudian dikonversi ke dalam format CSV untuk mempermudah proses integrasi ke dalam basis data MySQL, yang selanjutnya digunakan dalam sistem berbasis web PHP. Selanjutnya dilakukan tahap prapemrosesan teks yang mencakup pembersihan karakter non-alfabet, penghapusan kata tidak bermakna (*stopword*), serta normalisasi kata (*word normalization*) untuk menyeragamkan bentuk kata dasar.

Proses perhitungan *TF-IDF* dan *Cosine Similarity* dilakukan secara bertahap untuk mengukur bobot kata dan derajat kemiripan antar deskripsi wisata. Perhitungan dilakukan baik secara manual untuk verifikasi rumus maupun melalui implementasi sistem berbasis PHP, yang digunakan untuk menjalankan fungsi prapemrosesan, pembobotan, serta kalkulasi *Cosine Similarity*. Hasil perhitungan ini menjadi dasar dalam menentukan tingkat relevansi antar destinasi wisata yang kemudian ditampilkan dalam sistem rekomendasi berbasis web.

B. Tahapan Penelitian

Penelitian ini menggunakan pendekatan eksperimen berbasis sistem rekomendasi konten (*Content-Based Filtering*) dengan mengimplementasikan metode TF-IDF, *Cosine Similarity*, dan *Word Normalizer*. Tujuan utamanya adalah membangun sistem rekomendasi destinasi wisata di Daerah Istimewa Yogyakarta yang mampu memberikan rekomendasi relevan berdasarkan kesamaan deskripsi teks antar destinasi. Proses penelitian disusun dalam beberapa tahapan *sistematis* seperti yang ditunjukkan pada gambar 1.



Gambar 1. Flowchart Sistem

Gambar 1 Setiap tahap memiliki peran penting dalam menghasilkan sistem rekomendasi yang akurat dan dapat diandalkan.

C. Pengumpulan Dataset

Tahap pertama adalah pengumpulan dataset yang menjadi dasar seluruh proses analisis. Dataset diperoleh dari Kaggle dengan judul "*Dataset* Wisata Jogja Sekitar" (Faris Rizqiawan, 2023) serta dapat dikombinasikan dengan data terbuka dari Dinas Pariwisata Daerah Istimewa Yogyakarta untuk memperkaya informasi.

Data mencakup atribut penting seperti:

1) Nama tempat wisata,
2) Kategori wisata,
3) Deskripsi wisata.

Dataset dikonversi ke format CSV agar mudah diolah, kemudian diintegrasikan ke dalam basis data MySQL.

D. Praproses Data (Data Preprocessing)

Langkah ini bertujuan untuk memastikan bahwa data yang digunakan dalam sistem memiliki kualitas tinggi dan konsisten. Tahapan praproses mencakup:

1) Filtering kolom relevan, hanya mempertahankan kolom *id*, *nama*, dan *deskripsi*.
2) Penyederhanaan data, yaitu menghapus duplikasi, nilai kosong, serta teks yang tidak relevan.
3) Pembersihan karakter non-alfanumerik, termasuk penghapusan tanda baca dan spasi ganda.

Tahap ini menghasilkan dataset bersih yang siap untuk proses normalisasi dan ekstraksi fitur.

E. Word Normalizer

Pada penelitian ini, proses *word normalization* dilakukan secara hibrida dengan memanfaatkan dua lingkungan pemrosesan, yaitu Google Colab dan PHP. Google Colab digunakan pada tahap eksplorasi awal untuk melakukan ekstraksi pustaka kata serta pengujian proses *stemming* menggunakan *library* Sastrawi, sedangkan PHP digunakan untuk implementasi normalisasi yang terintegrasi di sisi *server*. Langkah awal dimulai dengan memanfaatkan Google Colab sebagai lingkungan komputasi berbasis Python untuk membaca seluruh data teks dari kolom deskripsi pada 475 data wisata dalam dataset. Seluruh kata dari kolom deskripsi tersebut diambil, kemudian dibersihkan menggunakan Sastrawi untuk menghilangkan imbuhan dan variasi bentuk kata. Setelah proses *stemming*, kata-kata yang identik dicatat hanya satu kali untuk membentuk kumpulan kata dasar unik

yang selanjutnya digunakan sebagai landasan perbaikan kata tidak baku atau kesalahan penulisan (*typo*) dalam sistem rekomendasi.

Proses ini dilakukan dengan memanfaatkan *data frame* Pandas untuk memuat data dari berkas CSV, kemudian dilanjutkan dengan tahapan *tokenization* dan *stemming* menggunakan Sastrawi Stemmer, salah satu pustaka *Natural Language Processing* (NLP) berbahasa Indonesia yang umum digunakan dalam penelitian berbasis teks. Dari hasil eksekusi di Google Colab diperoleh sekitar 2.243 kata unik hasil *stemming*. Selanjutnya, data tersebut diekspor ke format Excel dan diimpor ke dalam program PHP dalam bentuk JSON untuk dijadikan sebagai lexical library—kumpulan kata dasar hasil *stemming* dan normalisasi yang berfungsi sebagai acuan sistem dalam mendeteksi serta memperbaiki kata tidak baku pada deskripsi wisata.

Setelah pustaka kata terbentuk di Google Colab, daftar kata tersebut disimpan dalam array PHP dan digunakan dalam proses normalisasi teks. Implementasi dilakukan melalui empat tahap transformasi utama sebagai berikut:

1) Penghapusan Kata Umum (*Stopword Removal*) : Menghapus kata-kata umum dalam Bahasa Indonesia seperti *yang*, *dan*, *di*, *ke*, *pada*, dan *dengan* menggunakan daftar *stopword* dari pustaka Sastrawi.
2) Normalisasi Teks : Menyamakan bentuk kata berbeda yang memiliki makna sama dengan mencocokkannya terhadap *lexical library* hasil ekstraksi di Google Colab.
3) Lowercasing : Mengubah seluruh huruf menjadi huruf kecil agar sistem tidak membedakan kapitalisasi, misalnya kata *"Pantai"* dan *"pantai"*.
4) Penyatuan Kata Identik : Menyelaraskan kata yang memiliki bentuk tidak baku atau salah ketik, seperti *"koootaa"*, *"kot4"*, atau *"k0t4"*, menjadi bentuk baku *"kota"* menggunakan fungsi *string matching* berbasis Levenshtein Distance.

Hasil dari keseluruhan proses ini berupa teks yang bersih, baku, dan seragam, sehingga siap digunakan pada tahap pembobotan TF-IDF. Penerapan *lexical library* berbasis hasil ekstraksi Colab memungkinkan sistem memahami kesamaan makna kata secara kontekstual berdasarkan struktur linguistik dataset, bukan sekadar melalui pencocokan literal antar karakter. Word Normalizer dalam penelitian ini menggunakan pendekatan hybrid rule-based normalization, yang terdiri dari: (1) stemming berbasis Sastrawi, (2) kamus kata dasar hasil ekstraksi korpus, (3) koreksi ejaan berbasis Levenshtein Distance, dan (4) substitusi karakter non-alfabetik berbasis aturan fonetik.

F. Ekstraksi Fitur Menggunakan TF-IDF

Tahap ini mengubah teks deskripsi menjadi bentuk numerik dengan metode *Term Frequency – Inverse Document Frequency (TF-IDF)*. Pendekatan pembobotan *TF-IDF* digunakan untuk menilai pentingnya setiap kata dalam deskripsi wisata terhadap keseluruhan korpus teks. Konsep ini sejalan dengan penelitian *Chatbot Helpdesk* yang menunjukkan bahwa *TF-IDF* efektif dalam mengukur frekuensi dan bobot kemunculan kata dalam dokumen untuk memahami makna pertanyaan pengguna serta membangun model representasi vektor dalam analisis teks [14].

1. Term Frequency (TF)

Term Frequency digunakan untuk mengukur seberapa sering suatu kata muncul dalam sebuah dokumen. Nilai *TF* dihitung dengan membagi jumlah kemunculan kata tersebut dengan total jumlah kata pada dokumen tersebut.
Rumusnya dituliskan sebagai:

$$tf_{i,j} = \frac{f_{i,j}}{d_j}$$

dengan:
$f_{i,j}$ = jumlah kemunculan term *i* pada dokumen *j*
d_j = total kata dalam dokumen *j*

2. Inverse Document Frequency (IDF)

IDF berfungsi untuk mengukur seberapa penting suatu kata dalam keseluruhan dokumen. Rumusnya adalah:

$$idf_i = \log(\frac{N+1}{df_i+1}) + 1$$

dengan:
N = jumlah total dokumen
df_i = jumlah dokumen yang memuat term *i*

3. TF-IDF

Bobot akhir dari setiap term dalam dokumen dihitung dengan mengalikan nilai *TF* dan *IDF*:

$$w_{i,j} = tf_{i,j}\ x\ idf_i$$

G. Perhitungan Cosine Similarity

Cosine Similarity digunakan untuk menghitung tingkat kesamaan antara dokumen *d_j* dan query *q*. Persamaannya dituliskan sebagai:

$$sim(q, d_j) = \frac{q \cdot d_j}{|q| \times |d_j|}$$
$$= \frac{\sum_{i=1}^{t} w_{i,q} \times w_{i,j}}{\sqrt{\sum_{i=1}^{t}(w_{i,q})^2} \times \sqrt{\sum_{i=1}^{t}(w_{i,j})^2}}$$

dengan:
q = nama wisata
d_j = deskripsi wisata
$w_{i,q}$ = bobot *TF-IDF* dari term ke-i dalam query nama wisata
$w_{i,j}$ = bobot *TF-IDF* dari term ke-I dalam deskripsi wisata
t = jumlah total term unik

H. Pengurutan dengan Top-N

Setelah diperoleh nilai *Cosine Similarity* antara *query* pengguna dan seluruh dokumen destinasi wisata dalam dataset, sistem akan melanjutkan tahap perangkingan. Proses ini bertujuan untuk menentukan tingkat relevansi setiap destinasi terhadap *preferensi* pengguna berdasarkan skor

kemiripan yang dihasilkan. Setiap destinasi wisata d_j akan memiliki nilai similarity score yang diperoleh dari hasil perhitungan sim(q, d_j). Nilai tersebut kemudian diurutkan secara menurun (descending order) sehingga destinasi dengan skor tertinggi menempati peringkat teratas. Mekanisme ini dikenal dengan pendekatan *Top-N Recommendation*, yaitu proses seleksi sejumlah N destinasi dengan tingkat kemiripan tertinggi untuk disajikan kepada pengguna.

I. Implementasi

Implementasi sistem rekomendasi wisata ini dilakukan dengan menggunakan bahasa pemrograman PHP yang berperan sebagai *server-side scripting language* dalam membangun aplikasi berbasis web. Proses implementasi dimulai dengan pembuatan modul prapemrosesan data yang berfungsi untuk membersihkan dan menormalisasi teks deskripsi wisata. Tahapan ini meliputi case folding, tokenizing, penghapusan stopwords, serta *word normalization* menggunakan aturan *linguistik* sederhana dan algoritma koreksi berbasis jarak antar kata. Hasil keluaran dari tahap ini berupa representasi teks yang telah siap untuk dilakukan pembobotan dan analisis kesamaan.

Selanjutnya, sistem menerapkan modul ekstraksi fitur dengan metode Term Frequency–Inverse Document Frequency (TF-IDF) untuk mengubah data teks menjadi bentuk vektor numerik. Representasi vektor ini kemudian diproses melalui modul perhitungan kesamaan (similarity computation) yang berfungsi untuk mengukur tingkat kemiripan antar deskripsi destinasi wisata berdasarkan profil pencarian pengguna (user query).

Hasil perhitungan kesamaan disimpan sementara dalam basis data MySQL yang terhubung langsung dengan antarmuka sistem berbasis web. Selanjutnya dilakukan proses perangkingan (ranking) untuk menampilkan sejumlah destinasi dengan nilai kemiripan tertinggi. Sistem menerapkan pendekatan Top-N Recommendation, di mana pengguna akan memperoleh daftar rekomendasi wisata yang paling relevan sesuai preferensi pencarian.

III. HASIL DAN PEMBAHASAN

A. Deskripsi Umum Penelitian

Penelitian ini menggunakan pendekatan *Content-Based Filtering (CBF)* yang dikombinasikan dengan metode Term *Frequency–Inverse Document Frequency (TF-IDF)* dan *Cosine Similarity* untuk membangun sistem rekomendasi destinasi wisata berbasis teks. Pendekatan ini dipilih karena mampu mengukur tingkat kemiripan antar deskripsi wisata secara matematis berdasarkan bobot kata yang muncul pada setiap deskripsi.

Melalui model ini, sistem dapat memberikan rekomendasi destinasi yang relevan dengan minat pengguna berdasarkan kesamaan makna dari deskripsi wisata yang dimasukkan. Sebagai contoh, jika pengguna mencari Pantai Parangtritis, sistem akan menampilkan pantai-pantai lain yang memiliki deskripsi dengan tingkat kemiripan tinggi, seperti Pantai Ngrenehan atau Pantai Cemara Sewu.

Dalam membuktikan keakuratan rumus dan algoritma yang digunakan, penelitian ini melakukan perhitungan manual *TF-IDF* dan *Cosine Similarity* pada sejumlah data contoh. Perhitungan manual ini dilakukan untuk memverifikasi bahwa implementasi sistem secara komputasional memberikan hasil yang sesuai dengan logika matematis dasar.

Tahapan awal uji coba dilakukan dengan menggunakan data *query* (Q) berupa satu destinasi wisata, yaitu Pantai Parangtritis, kemudian dibandingkan dengan beberapa destinasi pantai lain di Daerah Istimewa Yogyakarta. Data berikut digunakan untuk proses eksperimen manual awal guna menghitung tingkat kesamaan antar deskripsi wisata.

TABLE 2
DATA AWAL UNTUK UJI PERHITUNGAN MANUAL

Q	Pantai Parangtritis
W1	Pemecah Ombak Pantai Glagah
W2	Pantai Drini
W3	Pesona Pengklik Pantai Samas
W4	Pantai Nglambor
W5	Pantai Somandeng
W6	Pantai Ngandong
W7	Pantai Bukit Indah Nampu
W8	Pantai Cemara Sewu Bantul Yogyakarta
W9	Pantai Ngrenehan
W10	Pantai Ngrumput
W11	Pantai Watulawang
W12	Pantai Slili
W13	Pantai Pok Tunggal
W14	Pantai Siung
W15	Pantai Nampu
W16	Pantai Butuh
W17	Pantai Ngrawah
W18	Pantai Sepanjang
W19	Pantai Parangkusumo
W20	Pantai Wediombo

Data di atas akan digunakan untuk:

1) Melakukan pembersihan dan normalisasi teks agar setiap deskripsi memiliki format kata yang seragam.
2) Menghitung nilai *TF (Term Frequency)* dan *IDF (Inverse Document Frequency)* untuk masing-masing kata yang muncul dalam deskripsi.
3) Menghasilkan bobot *TF-IDF* dari setiap term terhadap dokumen.
4) Menghitung Cosine Similarity antara *Q = Pantai Parangtritis* dengan setiap *Wi* (W1–W20).
5) Mengurutkan hasil kemiripan (ranking) untuk menentukan *Top-N* rekomendasi destinasi wisata yang paling serupa.

B. Word Normalizer

Tahap *word normalization* merupakan bagian penting dalam proses prapemrosesan data teks yang bertujuan untuk memastikan bahwa setiap deskripsi destinasi wisata memiliki struktur linguistik yang seragam sebelum dilakukan proses

pembobotan menggunakan metode *TF-IDF*. Proses *word normalization* diawali dengan case folding, yaitu mengubah seluruh huruf dalam teks menjadi huruf kecil (lowercase) agar sistem tidak membedakan kata yang memiliki makna sama tetapi berbeda dalam penulisan, seperti “Pantai” dan “pantai”. Selanjutnya dilakukan pembersihan tanda baca dan spasi berlebih dengan menghapus karakter non-alfabetik seperti titik, koma, tanda seru, dan tanda tanya, sehingga dihasilkan teks yang lebih bersih dan terstruktur serta terhindar dari kesalahan penghitungan kata. Tahap terakhir adalah penghapusan stopwords, yaitu menghilangkan kata-kata umum yang sering muncul namun tidak memberikan makna penting terhadap konteks dokumen, sehingga hanya tersisa kata-kata bermakna yang relevan untuk proses analisis dan perhitungan bobot TF-IDF.

Daftar *stopwords* yang digunakan dalam penelitian ini meliputi kata-kata umum dalam bahasa Indonesia yang tidak memiliki nilai semantik penting terhadap konteks deskripsi wisata. Kata-kata tersebut antara lain:

"yang", "dan", "di", "ke", "dari", "pada", "dengan", "untuk", "adalah", "itu", "ini", "atau", "sebagai", ……..

Penghapusan kata-kata tersebut bertujuan untuk mempertahankan hanya istilah yang memiliki nilai semantik penting dalam menggambarkan karakteristik objek wisata. Setelah seluruh tahapan tersebut dijalankan, teks deskripsi wisata “Pantai Parangtritis” menghasilkan daftar kata yang telah siap untuk proses perhitungan frekuensi. Daftar kata hasil *tokenization* tersebut disajikan dalam table 2 berikut.

TABLE 3
TOKENISASI DESKRIPSI WISATA “PANTAI PARANGTRITIS” SETELAH NORMALISASI DAN PENGHAPUSAN STOPWORDS

Kode	Token Kata	Kode	Token Kata
K1	pantai	K12	kuda
K2	memiliki	K13	sisi
K3	parangtritis	K14	utara
K4	ikon	K15	timur
K5	kabupaten	K16	bukit
K6	bantul	K17	menikmati
K7	pasir	K18	pemandangan
K8	hitam	K19	ditemani
K9	bermain	K20	angin
K10	atv	K21	cerita
K11	menaiki	K22	misteri

C. Perhitungan Temp (Jumlah Kata per Dokumen)

Tahap ini dilakukan untuk menghitung jumlah kemunculan setiap kata *(term)* pada masing-masing dokumen setelah melalui proses *word normalization* dan penghapusan stopwords. Hasil perhitungan ini menjadi dasar bagi tahap selanjutnya, yaitu perhitungan *Term Frequency (TF).*

Setiap dokumen direpresentasikan sebagai sekumpulan kata hasil tokenisasi, di mana frekuensi kemunculan setiap kata dihitung dan disusun dalam bentuk matriks term-dokumen. Ringkasan hasil perhitungan jumlah kata per dokumen *(Temp)* ditampilkan pada Table berikut.

TABLE 4
PERHITUNGAN JUMLAH KATA (TEMP) PADA SETIAP DOKUMEN

Kode	K1	K2	K3	K4	...	K21	K22
Q	5	2	1	1	...	1	1
W1	4	0	0	0	...	0	0
W2	6	0	0	0	...	0	0
W3	10	2	1	0	...	1	0
W4	6	0	0	0	...	0	0
W5	2	1	0	0	...	0	0
W6	4	1	0	0	...	0	0
W7	1	0	0	0	...	0	0
W8	8	0	0	1	...	0	0
W9	5	1	0	0	...	0	0
W10	9	1	1	0	...	0	0
...	...	...	...	...	...	...	...
W19	9	0	0	0	...	1	0
W20	6	0	0	0	...	1	0

Table menunjukkan bahwa setiap dokumen memiliki variasi jumlah kemunculan kata yang berbeda. Nilai-nilai tersebut menggambarkan intensitas kemunculan suatu term dalam teks deskripsi wisata, yang selanjutnya digunakan untuk menghitung bobot relatif setiap kata dalam dokumen menggunakan metode *TF-IDF*.

Langkah ini memastikan bahwa sistem dapat menilai pentingnya setiap kata dalam konteks deskripsi destinasi wisata secara proporsional terhadap keseluruhan dokumen.

D. Perhitungan TF (Term Frequency)

Tahap ini bertujuan untuk mengetahui seberapa sering suatu kata *(term)* muncul dalam setiap dokumen. Nilai *Term Frequency* (TF) menurut [15] menggambarkan tingkat kepentingan kata terhadap dokumen yang bersangkutan. Semakin sering suatu kata muncul, semakin besar pula pengaruhnya terhadap karakteristik isi dokumen tersebut.

TABLE 5
NILAI TERM FREQUENCY (TF) PADA SETIAP DOKUMEN

Kode	K1	K2	K3	K4	...	K21	K22
Q	0,132	0,053	0,026	0,026	...	0,026	0,026
W1	0,083	0	0	0	...	0	0
W2	0,181	0	0	0	...	0	0
W3	0,123	0,025	0,012	0	...	0,012	0
W4	0,162	0	0	0	...	0	0
W5	0,133	0,067	0	0	...	0	0
W6	0,102	0,026	0	0	...	0	0
W7	0,111	0	0	0	...	0	0
W8	0,114	0	0,014	0	...	0,014	0
W9	0,086	0,018	0	0	...	0	0
W10	0,098	0,011	0,011	0	...	0	0
...	...	...	...	...	...	...	...
W19	0,098	0	0	0	...	0	0
W20	0,113	0	0	0	...	0,019	0

Proses perhitungan dilakukan dengan membandingkan jumlah kemunculan setiap kata terhadap total jumlah kata dalam satu dokumen. Hasil perhitungan *TF* untuk *query* (Q) dan dokumen pembanding (W1–W20) ditunjukkan pada

Table 5 berikut. Tahap ini menjadi dasar pembentukan bobot kata pada proses TF-IDF sehingga memungkinkan sistem mengukur tingkat kesamaan antar dokumen secara kuantitatif dan objektif. Dengan demikian, kata-kata yang memiliki frekuensi tinggi pada suatu dokumen akan memberikan kontribusi lebih besar dalam menentukan tingkat relevansi dokumen tersebut terhadap kueri pengguna.

Table 3 menunjukkan bahwa kata K1 (pantai) memiliki nilai *TF* paling tinggi pada sebagian besar dokumen, menandakan bahwa kata ini menjadi kata dominan dalam deskripsi destinasi wisata yang dianalisis.

E. Perhitungan IDF (Inverse Document Frequency)

Tahap ini digunakan untuk menghitung seberapa penting suatu kata (*term*) terhadap keseluruhan koleksi dokumen yang dianalisis. Jika suatu kata muncul di banyak dokumen, maka nilai *Inverse Document Frequency* (IDF)-nya akan rendah karena dianggap kurang memiliki kekhususan. Sebaliknya, jika kata tersebut hanya muncul di beberapa dokumen, maka nilai *IDF*-nya tinggi. Pada proses ini, setiap kata dihitung berdasarkan dua parameter utama:

1) N = jumlah total dokumen dalam dataset (termasuk *query*), yaitu 21 dokumen (Q + W1–W20).
2) dfi = jumlah dokumen yang mengandung kata ke-i.

Perhitungan dilakukan terhadap setiap term yang telah diperoleh pada tahap sebelumnya. Nilai-nilai *IDF* untuk *query* dan seluruh dokumen dalam Table 3.4 berikut.

TABLE 6
NILAI INVERSE DOCUMENT FREQUENCY (IDF) PADA SETIAP DOKUMEN

Kode	K1	K2	K3	K4	...	K21	K22
Q	7.002	4.002	3.002	3.002	...	3.002	3.002
W1	6.002	2.002	2.002	2.002	...	2.002	2.002
W2	8.002	2.002	2.002	2.002	...	2.002	2.002
W3	12.002	4.002	3.002	2.002	...	3.002	2.002
W4	8.002	2.002	2.002	2.002	...	2.002	2.002
W5	4.002	3.002	2.002	2.002	...	2.002	2.002
W6	6.002	3.002	2.002	2.002	...	2.002	2.002
W7	3.002	2.002	2.002	2.002	...	2.002	2.002
W8	10.002	2.002	2.002	3.002	...	3.002	2.002
W9	7.002	3.002	2.002	2.002	...	2.002	2.002
W10	11.002	3.002	3.002	2.002	...	2.002	2.002
...	...	...	...	...	...	...	...
W19	11.002	2.002	2.002	2.002	...	2.002	2.002
W20	8.002	2.002	2.002	2.002	...	3.002	2.002

Dari Table 6 dapat dilihat bahwa nilai *IDF* tertinggi terdapat pada kata K1, yang menunjukkan bahwa kata ini muncul relatif jarang di sebagian dokumen namun sangat signifikan dalam mendeskripsikan konteks wisata pantai. Sebaliknya, term dengan nilai *IDF* rendah mengindikasikan kata yang sering muncul di berbagai dokumen sehingga kontribusinya terhadap pembeda antar-dokumen menjadi lebih kecil. Nilai-nilai *IDF* ini akan dikalikan dengan nilai *TF* pada tahap berikutnya untuk menghasilkan bobot akhir *TF-IDF*, yang menggambarkan tingkat kepentingan kata dalam masing-masing dokumen secara proporsional.

F. Perhitungan TF–IDF

Tahapan ini merupakan proses penggabungan antara dua komponen penting dalam analisis teks, yaitu *Term Frequency* (TF) dan *Inverse Document Frequency* (*IDF*). Nilai *TF-IDF* diperoleh dari hasil perkalian antara kedua faktor tersebut dan digunakan untuk menilai tingkat kepentingan suatu kata (*term*) dalam sebuah dokumen relatif terhadap seluruh koleksi dokumen.

Kata yang memiliki frekuensi tinggi dalam dokumen tertentu, namun jarang muncul di dokumen lain, akan menghasilkan nilai *TF-IDF* yang besar. Sebaliknya, kata yang sering muncul di hampir semua dokumen akan menghasilkan nilai *TF-IDF* yang kecil karena dianggap kurang memiliki nilai pembeda. Perhitungan ini diterapkan pada semua *term* dalam *query* dan 20 dokumen pembanding. Hasilnya ditunjukkan secara ringkas pada Table 6 berikut.

TABLE 7
HASIL PERHITUNGAN TF–IDF PADA SETIAP DOKUMEN

Kode	K1	K2	K3	K4	...	K21	K22
Q	0.921	0.211	0.079	0.079	...	0.079	0.079
W1	0.500	0.000	0.000	0.000	...	0.000	0.000
W2	1.455	0.000	0.000	0.000	...	0.000	0.000
W3	1.482	0.099	0.037	0.000	...	0.037	0.000
W4	1.298	0.000	0.000	0.000	...	0.000	0.000
W5	0.534	0.200	0.000	0.000	...	0.000	0.000
W6	0.616	0.077	0.000	0.000	...	0.000	0.000
W7	0.334	0.000	0.000	0.000	...	0.334	0.000
W8	1.143	0.000	0.000	0.043	...	0.043	0.000
W9	0.604	0.052	0.000	0.000	...	0.000	0.000
W10	1.076	0.033	0.033	0.000	...	0.000	0.000
...	...	...	...	...	...	...	...
W19	1.053	0.000	0.000	0.000	...	0.000	0.000
W20	0.906	0.000	0.000	0.000	...	0.057	0.000

Hasil pembobotan ini selanjutnya digunakan dalam tahap perhitungan *Cosine Similarity* untuk menentukan tingkat kemiripan antara dokumen *query* dengan dokumen-dokumen pembanding. Bobot vektor yang dihasilkan oleh TF-IDF ini kemudian digunakan sebagai input untuk perhitungan *Cosine Similarity*, yang berfungsi menentukan tingkat kesamaan antara *query* pengguna dan dokumen jawaban yang tersedia sebagaimana diterapkan dalam penelitian [16].

G. Perhitungan Cosine Similarity

Tahap akhir dari proses pembobotan teks ini adalah menghitung tingkat kesamaan (*similarity*) antara dokumen *query* dengan seluruh dokumen pembanding menggunakan metode *Cosine Similarity*. Dalam penelitian *Chatbot* Fiqih [17] , *TF-IDF* digunakan sebagai tahap pencocokan awal kata kunci sebelum dilakukan analisis lanjutan terhadap konteks pertanyaan. Metode ini berfungsi untuk mengukur seberapa besar kesamaan arah antara dua vektor TF–IDF, dengan nilai keluaran berkisar antara 0 hingga 1. Nilai mendekati 1 menunjukkan bahwa dua dokumen memiliki tingkat kemiripan yang tinggi, sedangkan nilai mendekati 0 menandakan kemiripan yang rendah.

Perhitungan dilakukan dengan mengalikan vektor TF–IDF antara *query (Q)* dan setiap dokumen *Wi*, kemudian membaginya dengan hasil kali panjang vektor kedua dokumen. Hasil perhitungan Cosine Similarity disajikan pada Table 7 berikut.

TABLE 8
NILAI COSINE SIMILARITY ANTARA QUERY (Q) DAN SETIAP DOKUMEN (WI)

Kode	sim(Q, Wi)	Kode	sim(Q, Wi)
W1	0,8971	W11	0,9157
W2	0,9162	W12	0,9243
W3	0,9365	W13	0,9184
W4	0,9162	W14	0,9177
W5	0,9018	W15	0,6176
W6	0,9397	W16	0,8825
W7	0,6176	W17	0,7010
W8	0,9177	W18	0,9320
W9	0,9285	W19	0,9151
W10	0,9243	W20	0,9162

Dari hasil perhitungan pada Table 7 dapat diamati bahwa sebagian besar dokumen memiliki nilai kemiripan di atas 0,90, yang menandakan hubungan semantik yang sangat dekat antara deskripsi *query* ("Pantai Parangtritis") dengan deskripsi pantai-pantai lainnya di dalam dataset.

Nilai tertinggi diperoleh pada dokumen W6 (0,9397) dan W3 (0,9365), yang menunjukkan bahwa kedua destinasi tersebut memiliki kemiripan konten paling besar terhadap *query*. Kedua dokumen tersebut kemungkinan menggambarkan karakteristik lokasi wisata pantai dengan fitur dan aktivitas serupa, seperti pasir hitam, wisata kuda, atau aktivitas ATV.

Sebaliknya, dokumen dengan nilai kemiripan rendah seperti W7 (0,6176) dan W15 (0,6176) cenderung memiliki deskripsi yang berbeda secara kontekstual atau menggunakan terminologi yang lebih umum sehingga relevansinya terhadap *query* menurun.

Hasil perhitungan *Cosine Similarity* ini menjadi dasar dalam tahap pengurutan (ranking) yang mana sejalan dengan penelitian [18] untuk menentukan daftar rekomendasi destinasi wisata dengan tingkat kemiripan tertinggi yang akan disajikan kepada pengguna dalam sistem rekomendasi berbasis web.

H. Pengurutan (Ranking)

Setelah seluruh nilai *Cosine Similarity* diperoleh, langkah selanjutnya adalah melakukan pengurutan (*ranking*) terhadap hasil tersebut. Tujuan dari tahap ini adalah untuk menampilkan daftar destinasi wisata yang paling relevan dengan *query* pengguna berdasarkan tingkat kesamaan deskripsi teks.

Proses perangkingan dilakukan dengan mengurutkan nilai *Cosine Similarity* dari yang tertinggi hingga terendah. Pendekatan ini dikenal sebagai *Top-N Recommendation*, di mana sistem hanya menampilkan sejumlah *N* destinasi dengan skor kemiripan tertinggi untuk memberikan hasil yang paling akurat dan efisien kepada pengguna. Table berikut menampilkan hasil Top-5 rekomendasi wisata berdasarkan skor *Cosine Similarity* tertinggi terhadap *query* "Pantai Parangtritis".

TABLE 9
HASIL PENGURUTAN (RANKING) BERDASARKAN NILAI COSINE SIMILARITY TERTINGGI

Peringkat	Kode Dokumen	Nilai Cosine Similarity
1	W6	0,9397
2	W3	0,9365
3	W18	0,9320
4	W9	0,9285
5	W10	0,9243

Berdasarkan hasil pada Table lima dokumen dengan nilai *Cosine Similarity* tertinggi menunjukkan tingkat kemiripan konten yang sangat kuat terhadap *query*. Dokumen W6, yang menempati peringkat pertama dengan skor 0,9397, diidentifikasi sebagai destinasi wisata yang paling relevan dengan "Pantai Parangtritis".

Pantai Ngandong (W6) memperoleh nilai kemiripan tertinggi karena deskripsinya mengandung sejumlah leksikon seperti "pantai", "bermain", "pemandangan", dan "pasir" yang juga ditemukan dalam deskripsi Pantai Parangtritis. Kemunculan bersama kata-kata tersebut mencerminkan adanya kesamaan semantik yang signifikan antara kedua entitas teks. Nilai kemiripan yang tinggi antara Pantai Parangtritis dan Pantai Ngandong (0,9397) menunjukkan bahwa kedua destinasi tersebut berbagi domain semantik serupa, yakni wisata pesisir selatan dengan karakteristik pemandangan alam, aktivitas rekreasi, dan elemen bentang alam berpasir yang identik. Temuan ini mengindikasikan bahwa model analisis tidak hanya mengidentifikasi kesamaan pada tingkat leksikal, tetapi juga pada tingkat konseptual atau tematik.

Sebaliknya, nilai kemiripan yang relatif rendah (misalnya W7 dan W15 sekitar 0,61) menandakan bahwa teks deskriptif tersebut bersifat generik dan memiliki keterbatasan dalam kekayaan linguistik yang dapat dibandingkan secara semantik. Hal ini mengisyaratkan bahwa sistem memiliki kepekaan terhadap heterogenitas dan densitas informasi dalam korpus teks yang dianalisis. Dengan demikian, hasil pengukuran kemiripan tidak semata-mata bersifat kuantitatif, melainkan juga memberikan indikasi kualitatif terhadap kedalaman dan kompleksitas deskripsi yang terkandung dalam dataset.

Hasil penelitian yang dilakukan oleh [19] memperlihatkan pola kemiripan yang sejalan dengan temuan pada Table 8. Dalam penelitian tersebut, dokumen dengan tema wisata pantai juga menempati posisi teratas dengan nilai Cosine Similarity mendekati 1, menunjukkan bahwa deskripsi yang mengandung leksikon serupa memiliki tingkat kedekatan semantik yang tinggi terhadap query. Pola ini sejajar dengan hasil penelitian ini, di mana dokumen W6 memperoleh skor 0,9397 karena memiliki representasi linguistik dan konteks makna yang serupa dengan query "Pantai Parangtritis". Kedua penelitian menunjukkan bahwa kombinasi metode TF-

IDF dan Cosine Similarity mampu mengenali keterkaitan semantik antar-dokumen berbasis deskripsi wisata secara konsisten, terutama ketika teks mengandung istilah yang berada dalam domain konseptual yang sama, yakni wisata pesisir dengan aktivitas rekreatif dan unsur alam berpasir.

Tahapan selanjutnya adalah proses implementasi sistem berbasis web, yang akan mengintegrasikan hasil perangkingan ini ke dalam antarmuka pengguna sehingga daftar rekomendasi dapat ditampilkan secara dinamis.

I. Evaluasi Kinerja Sistem Rekomendasi

1. Tahap evaluasi metrik

Evaluasi kinerja sistem rekomendasi dilakukan menggunakan metrik Precision@5, Recall@5, dan Mean Average Precision (MAP). Pengujian dilakukan sebanyak 10 kali dengan jenis destinasi wisata yang berbeda untuk menilai konsistensi dan relevansi rekomendasi yang dihasilkan sistem.

TABLE 10
HASIL EVALUASI METRIK

Metrik	Nilai Rata-rata
Precision@5	0.84
Recall@5	0.10
MAP	0.81

Berdasarkan tabel di atas, Precision@5 yang mencapai 0,84 menunjukkan bahwa sebagian besar rekomendasi Top-5 yang diberikan sistem termasuk kategori relevan. Dengan kata lain, pengguna cenderung menerima rekomendasi yang sesuai dengan preferensi destinasi wisata mereka. Nilai Recall@5 sebesar 0,10 menunjukkan bahwa sistem mampu menemukan sekitar 10% dari seluruh destinasi relevan dalam dataset. Rendahnya nilai recall ini wajar karena evaluasi hanya dilakukan pada Top-5 rekomendasi, sehingga beberapa destinasi relevan mungkin tidak muncul di urutan atas. MAP sebesar 0,81 menandakan kualitas peringkat rekomendasi secara keseluruhan cukup tinggi. Hal ini menunjukkan bahwa sebagian besar destinasi relevan muncul di posisi atas daftar rekomendasi, sehingga pengguna lebih cepat menemukan destinasi yang sesuai dengan preferensi mereka.

2. Analisis Kinerja pada Berbagai Jenis Kueri

Konsistensi performa sistem pengujian tidak hanya dilakukan pada satu kueri, tetapi juga pada beberapa kueri representatif dari kategori wisata yang berbeda, yaitu wisata pantai, budaya, alam, dan buatan. Hasil pengujian menunjukkan bahwa nilai Precision@5 pada seluruh kueri berada pada rentang 0,50–1,00, yang mengindikasikan bahwa sistem secara konsisten mampu menghasilkan rekomendasi relevan pada berbagai jenis destinasi.

3. Validasi Kualitatif

Validasi kualitatif dilakukan terhadap 10 responden mahasiswa menggunakan skala Likert 1–5.

TABLE 11
VALIDASI KUALITATIF

Aspek Penilaian	Rata-rata Skor (1–5)
Relevansi rekomendasi	4.65
Kualitas informasi	4.05
Kemudahan penggunaan	4.10
Kepuasan pengguna	4.55
Skor keseluruhan	4.70

Hasil pengujian menunjukkan bahwa sistem memperoleh skor rata-rata keseluruhan sebesar 4,70, yang mengindikasikan bahwa sistem rekomendasi dinilai sangat relevan, mudah digunakan, dan memberikan informasi yang membantu pengguna dalam menemukan alternatif destinasi wisata. Aspek relevansi rekomendasi memperoleh skor tertinggi sebesar 4,65, yang menunjukkan bahwa hasil rekomendasi sesuai dengan kata kunci pencarian pengguna.

4. Analisis Distribusi Rekomendasi

Analisis distribusi menunjukkan bahwa 37% dari destinasi yang muncul pada Top-10 rekomendasi berasal dari kategori destinasi kurang populer, yang mengindikasikan bahwa sistem tidak hanya mempromosikan destinasi utama tetapi juga memberikan eksposur pada destinasi sekunder.

J. Implementasi Sistem

Implementasi sistem dilakukan dalam bentuk aplikasi web berbasis PHP yang berfungsi untuk menampilkan rekomendasi destinasi wisata di Daerah Istimewa Yogyakarta secara dinamis berdasarkan hasil perhitungan *TF–IDF* dan *Cosine Similarity*. Sistem ini dirancang dengan struktur antarmuka yang sederhana, interaktif, dan mudah dipahami pengguna, serta memanfaatkan basis data MySQL sebagai penyimpanan utama dataset wisata. Tujuan utama dari implementasi ini adalah agar hasil analisis yang telah dilakukan pada tahap sebelumnya dapat diterapkan secara langsung dalam lingkungan aplikasi web yang dapat diakses oleh pengguna umum.

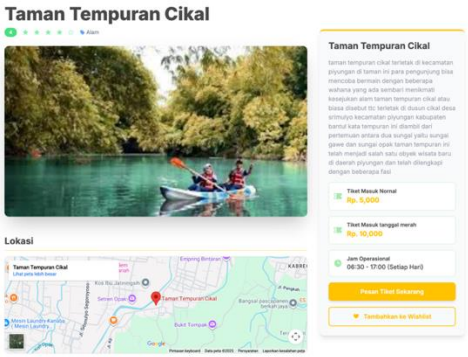
1. Tahap Word Normalization

Word normalization menjadi tahap krusial dalam penelitian ini karena bahasa Indonesia memiliki fleksibilitas morfologis yang tinggi serta sering mengalami kesalahan penulisan akibat variasi gaya pengguna atau ketidaktepatan pengetikan. Secara umum, proses *word normalization* dalam penelitian ini meliputi prapemrosesan dasar dan implementasi normalizer. Tahap prapemrosesan dimulai dengan konversi seluruh karakter menjadi huruf kecil untuk menghindari pembedaan kata berdasarkan kapitalisasi, sehingga kata seperti “Pantai” dan “pantai” diperlakukan sama. Selanjutnya, teks dibersihkan dari tanda baca, karakter non-alfabetik, dan spasi berlebih menggunakan regular expression agar data lebih konsisten dan siap untuk pemrosesan lebih lanjut. Selain itu, kata-kata umum atau stopwords yang tidak berkontribusi terhadap makna, seperti “yang”, “dan”, “di”, “ke”, “dengan”, “itu”, dan “adalah”, dihapus berdasarkan pustaka resmi Bahasa Indonesia sehingga analisis lebih fokus pada kata-kata

yang bermakna dan relevan dengan karakteristik destinasi wisata.

Keunggulan utama word normalizer yang dikembangkan dalam penelitian ini adalah kemampuannya untuk mendeteksi dan memperbaiki kesalahan penulisan secara otomatis, terutama pengulangan huruf vokal atau kesalahan input lainnya, dengan memanfaatkan pendekatan berbasis pola fonetik dan string similarity. Sistem ini juga mampu menangani fenomena penggunaan angka sebagai pengganti huruf yang umum ditemukan dalam teks media sosial atau ulasan daring, seperti mengubah "kot4" menjadi "kota" atau "k0t4" menjadi "kota". Lebih lanjut, word normalizer dapat mengenali kombinasi kompleks antara huruf dan angka serta pengulangan huruf berlebihan, kemudian menormalkan kata tersebut menjadi bentuk standar yang benar, misalnya mengubah "Kot4 yogyakartaaaaa" menjadi "kota yogyakarta" dan "P4nTai P4r4tritis" menjadi "pantai parangtritis". Implementasi word normalizer ini menghasilkan korpus teks yang bersih, seragam, dan memiliki akurasi semantik tinggi dibandingkan pendekatan normalisasi konvensional, sehingga siap digunakan pada tahap pembobotan TF-IDF untuk mendukung sistem rekomendasi destinasi wisata.

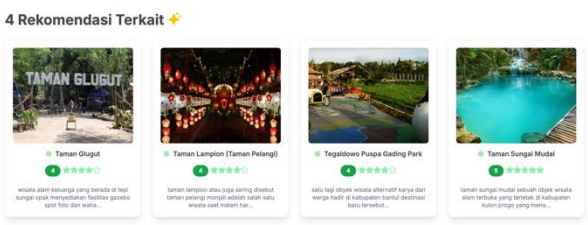
2. Halaman Detail Wisata (Destination Detail Page)



Gambar 2 Halaman Detail Wisata (Destination Detail Page)

Halaman ini menampilkan informasi lengkap mengenai destinasi wisata yang dipilih pengguna. Tampilan halaman ini dirancang agar informatif dan menarik, sehingga pengguna dapat memperoleh gambaran menyeluruh tentang destinasi yang akan dikunjungi.

3. Rekomendasi Wisata Terkait (Related Recommendation Section)



Gambar 3 Rekomendasi Wisata Terkait (Related Recommendation Section)

Di bagian bawah halaman detail, sistem secara otomatis menampilkan daftar rekomendasi wisata terkait berdasarkan hasil perhitungan *Cosine Similarity* antara deskripsi destinasi yang sedang dibuka dengan seluruh destinasi lain dalam dataset.Destinasi yang memiliki nilai kesamaan tertinggi *(Top-N)* ditampilkan dalam bentuk *horizontal scroll card* berisi gambar dan nama wisata.

IV. KESIMPULAN

A. Kesimpulan

Penelitian ini berhasil mengembangkan sistem rekomendasi wisata Yogyakarta berbasis Content-Based Filtering dengan mengintegrasikan metode Term Frequency–Inverse Document Frequency (TF-IDF), Cosine Similarity, dan Word Normalizer. Berdasarkan hasil analisis dan implementasi, sistem mampu memberikan rekomendasi destinasi wisata yang relevan terhadap minat pengguna melalui pengolahan deskripsi teks secara matematis dan linguistik.

Rumusan masalah pertama terkait bagaimana cara menentukan relevansi antar destinasi wisata berbasis deskripsi teks telah terjawab melalui penerapan metode TF-IDF. Pembobotan ini secara efektif mengukur pentingnya setiap kata terhadap keseluruhan dokumen, sehingga menghasilkan representasi numerik yang menggambarkan karakteristik unik tiap destinasi.

Rumusan masalah kedua, yaitu bagaimana mengukur tingkat kesamaan antar destinasi wisata, diselesaikan dengan penerapan metode Cosine Similarity. Hasil pengujian menunjukkan nilai kesamaan yang tinggi pada beberapa destinasi pantai seperti W6 (0,9397) dan W3 (0,9365), yang menunjukkan bahwa sistem mampu mengenali hubungan semantik antar deskripsi dengan akurasi tinggi.

Rumusan masalah ketiga mengenai peningkatan akurasi hasil rekomendasi berhasil dijawab melalui penerapan Word Normalizer. Proses normalisasi kata (*case folding, tokenisasi*, dan penghapusan *stopwords*) terbukti meningkatkan konsistensi bentuk bahasa, sehingga memperbaiki hasil pembobotan dan perhitungan kesamaan antar dokumen.

Secara keseluruhan, sistem rekomendasi yang dibangun mampu menampilkan hasil perangkingan destinasi wisata secara dinamis berdasarkan nilai kemiripan tertinggi dan terintegrasi dalam aplikasi web berbasis PHP. Pendekatan ini terbukti efektif dalam membantu wisatawan menemukan destinasi sesuai preferensi serta mendukung pemerataan promosi wisata di seluruh wilayah Yogyakarta. Dengan demikian, penggabungan metode *TF-IDF*, *Cosine Similarity*, dan *Word Normalizer* dapat dijadikan dasar yang kuat dalam pengembangan sistem rekomendasi berbasis teks di bidang pariwisata.

B. Keterbatasan Penelitian

Meskipun sistem rekomendasi yang dikembangkan menunjukkan hasil kemiripan yang tinggi, penelitian ini masih memiliki beberapa keterbatasan. Sistem hanya

menggunakan pendekatan content-based filtering berbasis deskripsi teks sehingga belum mempertimbangkan preferensi historis pengguna, rating, maupun interaksi pengguna lain. Selain itu, pendekatan TF-IDF masih bersifat berbasis frekuensi kata dan belum mampu menangkap konteks semantik mendalam seperti hubungan makna antar kata secara implisit. Dataset juga masih terbatas pada satu sumber data sehingga berpotensi mempengaruhi generalisasi sistem pada data yang lebih heterogen.

C. Arah Pengembangan

Pengembangan penelitian selanjutnya dapat dilakukan dengan mengintegrasikan metode collaborative filtering berbasis interaksi pengguna, embedding semantik seperti Word2Vec atau FastText, serta model deep learning berbasis transformer seperti BERT untuk menangkap makna kontekstual. Selain itu, sistem dapat dikembangkan menjadi sistem rekomendasi hibrida yang menggabungkan content-based dan collaborative filtering untuk meningkatkan personalisasi rekomendasi.

DAFTAR PUSTAKA

[1] Serly Putri Jumbadi, "55 Ribu Kendaraan Masuk Jogja Saat Liburan, Kawasan Gumaton Padat," *Detik.com*, Yogyakarta, Jun. 28, 2025.

[2] M. Tamam Huda and A. Permana Wibowo, "Recommendation System For Mobile Application Tour Guide And Travel Services Using Demographic Filtering And Content-Based Filtering Methods Based On Android", [Online]. Available: https://jws.rivierapublishing.id/index.php/jws

[3] K. Samosir and F. Ginting, "A Comparative Analysis of Content-Based Filtering and TF-IDF Approaches for Enhancing Sports Recommendation Systems," vol. 6, no. 2, pp. 90–97, 2024, [Online]. Available: http://innovatics.unsil.ac.id

[4] A. O. Rahmawati, R. Susanto, and H. Hasanah, "Sistem Rekomendasi Bursa Kerja Khusus (Bkk) Menggunakan Metode Content Based Filtering Pada Smk Tunas Bangsa," *Jurnal Informatika Teknologi Dan Sains*.

[5] W. Ferbiansyah, A. Muhammad Irwan, B. Santoso, and S. Kacung, "Implementasi Metode Content Based Filtering Menggunakan Synopsis Similarity Untuk Pemilihan Anime," *Jurnal Informatika Teknologi Dan Sains*, VOL. 7, NO. 2, P. 955, 2025.

[6] S. Oyadila, D. Abdullah, and A. Razi, "Implementasi Content-Based Filtering Dengan TF-IDF Dan Cosine Similarity Untuk Sistem Rekomendasi Destinasi Wisata Di Aceh Tengah," *RABIT : Jurnal Teknologi Dan Sistem Informasi Univrab*, Vol. 10, NO. 2, PP. 1329–1339, Jul. 2025, doi: 10.36341/rabit.v10i2.6532.

[7] L. H. Aljihadu, "Sistem Rekomendasi Wisata Kuliner Di Gunungkidul Menggunakan Metode Content Based Filtering," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 1, Jan. 2025, doi: 10.23960/jitet.v13i1.5955.

[8] A. Dwi Aryanto, A. Primadewi, N. Agung, and A. D. Aryanto, "Rekomendasi Wisata Kabupaten Magelang menggunakan Metode Content-Based Filtering dan Location-Based Service," *JURNAL FASILKOM*, vol. 15, pp. 172–78, 2025.

[9] F. Farasalsabila, E. Utami, and M. Hanafi, "Analysis Of Public Opinion On Indonesian Television Shows Using Support Vector Machine," *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, vol. 10, no. 2, pp. 239–246, Mar. 2024, doi: 10.33330/jurteksi.v10i2.2935.

[10] T. Rafah Masuzzahra, K. Umam, H. Mustofa, and M. R. Handayani, "HANA: An AI Chatbot for Islamic Jurisprudence on Menstruation using SBERT and TF-IDF," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 3, p. 1013, 2025, [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[11] Yuliana, Mira, and A. Hari Kristianto, "Machine Learning Content-Based Filtering Women Empowering Recommendations On Youtube," *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, vol. XI, no. 4, pp. 2407–1811, 2025, doi: 10.33330/jurteksi.v11i4.4154.

[12] R. Saputra, Y. Pristyanto, and I. N. Fajri, "Generative AI Image Sentiment Analysis on Social Media X using TF-IDF and FastText," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[13] A. Pannadhika Putra, D. Purnami Singgih Putri, and Aak. Cahyawan Wiranatha, "Scientific Paper Recommendation System: Application of Sentence Transformers and Cosine Similarity Using arXiv Data," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 4, 2025, [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[14] G. H. Setiawan, I. Made, and B. Adnyana, "Improving Helpdesk Chatbot Performance with Term Frequency-Inverse Document Frequency (TF-IDF) and Cosine Similarity Models," 2023. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[15] D. Septiani and I. Isabela, "SINTESIA: Jurnal Sistem dan Teknologi Informasi Indonesia Analisis Term Frequency Inverse Document Frequency (TF-IDF) Dalam Temu Kembali Informasi Pada Dokumen Teks".

[16] G. H. Setiawan, I. Made, and B. Adnyana, "Improving Helpdesk Chatbot Performance with Term Frequency-Inverse Document Frequency (TF-IDF) and Cosine Similarity Models," *Journal of Applied Informatics and Computing (JAIC)*, vol. 7, no. 2, p. 252, 2023, [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[17] T. Rafah Masuzzahra, K. Umam, H. Mustofa, and M. R. Handayani, "HANA: An AI Chatbot for Islamic Jurisprudence on Menstruation using SBERT and TF-IDF," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[18] T. Wahyu Intan Permadani, D. Adi Prasetya, E. Daniati, and P. Korespondens, "Sistem Rekomendasi Film Berdasarkan Genre Menggunakan Metode Content-Based Filtering dengan Algoritma Cosine Similarity," 2025.

[19] R. Al Rasyid, D. Handayani, and U. Ningsih, "Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian Pada Dataset Destinasi Wisata," *Jurnal Teknologi Informasi dan Komunikasi)*, vol. 8, no. 1, p. 2024, 2024, doi: 10.35870/jti.