ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

PENERAPAN DATA MINING PADA SISTEM REKOMENDASI DESTINASI WISATA BERBASIS TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY DAN VECTOR SPACE MODEL

I Kadek Nova Hary Saputra✉, Agus Adi Putrawan, Gde Brahupadhya Subiksa
Teknologi Rekayasa Perangkat Lunak, Jurusan Teknologi Informasi, Politeknik Negeri Bali, Badung, Indonesia
Email: 2215354064@pnb.ac.id

DOI: https://doi.org/10.46880/jmika.Vol10No2.pp524-534

ABSTRACT

Digital tourism transitions cause information overload, complicating tourist decisions with unstructured data. Collaborative filtering systems also encounter cold-start issues for new destinations that lack sufficient popularity metrics. To address this, this research develops a recommendation system for Bali tourist destinations using content-based filtering. The method integrates Term Frequency-Inverse Document Frequency (TF-IDF) to extract text weights and the Vector Space Model (VSM) via cosine similarity to measure user query and destination relevance. Built on a Model-View-Controller (MVC) architecture, the system dynamically links PHP Laravel and Python Flask via a REST API. Text preprocessing involves case folding, tokenizing, filtering, and stemming using the Sastrawi library. Evaluation results without Top N cutoff restrictions show an average recall rate of 90.20%, an average accuracy of 88.18%, and a mean average precision (MAP) of 72.82%. Although the absence of cutoff limits suppressed average precision to 34.43%, setting a 0.05 relevance threshold efficiently filtered out irrelevant noise. The system comprehensively recommends potential semantic destinations without relying on popularity, while maintaining high-ranking quality.

Keyword: Recommendation System, TF-IDF, VSM, Content-Based Filtering, Bali Tourism.

ABSTRAK

Industri pariwisata yang beralih ke ranah digital menyebabkan fenomena information overload, menyulitkan wisatawan mengambil keputusan akibat informasi destinasi yang melimpah dan tidak terstruktur. Sistem Collaborative Filtering juga rentan mengalami kendala cold-start saat merekomendasikan destinasi baru tanpa metrik popularitas memadai. Penelitian ini merancang sistem rekomendasi destinasi wisata di Bali menggunakan pendekatan Content-Based Filtering guna mengatasi masalah tersebut. Metode ini mengintegrasikan algoritma Term Frequency-Inverse Document Frequency (TF-IDF) untuk ekstraksi bobot teks deskripsi, serta Vector Space Model (VSM) dengan Cosine Similarity untuk mengukur derajat relevansi kueri pengguna dengan dokumen destinasi. Sistem dikembangkan pada arsitektur Model-View-Controller (MVC) yang mengintegrasikan kerangka kerja PHP Laravel dan Python Flask melalui REST API. Prapemrosesan teks dioptimasi menggunakan case folding, tokenizing, filtering, dan stemming dari pustaka Sastrawi. Evaluasi kinerja menggunakan kueri tanpa batasan pemotongan antarmuka (Top-N Cutoff) mencatatkan rata-rata performa yang andal: tingkat Recall 90,20%, Accuracy 88,18%, dan Mean Average Precision (MAP) 72,82%. Meskipun eksposur seluruh luaran menekan rata-rata nilai Precision ke angka 34,43%, ambang batas relevansi 0,05 terbukti efisien memfiltrasi noise. Sistem terbukti komprehensif dan mampu menjaga kualitas pemeringkatan dalam merekomendasikan destinasi potensial secara semantik tanpa bergantung pada popularitas.

Kata Kunci: Sistem Rekomendasi, TF-IDF, VSM, Content-Based Filtering, Pariwisata Bali.

PENDAHULUAN

Industri pariwisata modern telah mengalami transformasi fundamental dari model konvensional menuju ekosistem berbasis data yang sangat dinamis (Saputri et al., 2025). Pergeseran paradigma ini ditandai dengan munculnya perilaku Digital Independent Travel, di mana wisatawan kini memegang kendali penuh dalam perencanaan perjalanan mereka secara mandiri tanpa bergantung pada agen konvensional. Dalam kerangka Smart Tourism, wisatawan menuntut akses terhadap informasi yang tidak hanya cepat, tetapi juga terpersonalisasi sesuai dengan preferensi unik mereka (Saputri et al., 2025). Akan tetapi, masifnya digitalisasi ini memunculkan paradoks berupa *information overload* atau kelebihan beban informasi. Ledakan volume data di internet mengakibatkan informasi mengenai destinasi menjadi terlalu melimpah, tidak terstruktur, dan tumpang tindih, yang

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

justru memicu kebingungan kognitif dan menghambat efektivitas pengambilan keputusan wisatawan (Khusnah et al., 2025.).

Upaya mengatasi tantangan tersebut melalui sistem rekomendasi yang mendominasi saat ini, khususnya *Collaborative Filtering* yang bergantung pada rating atau riwayat interaksi pengguna lain, memiliki kelemahan fundamental ketika diterapkan pada ekosistem pariwisata yang dinamis (PUTRA & RACHMAN, 2024). Metode tersebut rentan terhadap masalah *cold-start*, yaitu kegagalan sistem dalam merekomendasikan destinasi wisata baru atau *hidden gems* yang belum memiliki metrik popularitas atau data ulasan yang memadai. Akibatnya, persebaran wisatawan hanya terkonsentrasi pada lokasi-lokasi yang sudah populer semata. Untuk mengatasi keterbatasan tersebut, penelitian ini merancang pendekatan alternatif berupa *Content-Based Filtering* yang mengevaluasi relevansi berdasarkan analisis karakteristik intrinsik pada teks deskriptif destinasi pariwisata (Gunarto et al., 2023).

Berbagai upaya komputasi untuk menangani permasalahan pencarian informasi telah dieksplorasi oleh beberapa penelitian terdahulu. Studi yang dilakukan oleh Pratiwi et al. (2024) membuktikan bahwa pendekatan *Content-Based Filtering* dengan algoritma TF-IDF sangat relevan dalam mengekstraksi bobot teks deskriptif pada sistem rekomendasi pariwisata. Terkait infrastruktur sistemnya, Aziza et al. (2025) merancang sistem pariwisata berbasis website, namun menemukan bahwa fitur pencarian berbasis kategori statis konvensional memiliki keterbatasan absolut dalam merespons kueri spesifik pengguna berbasis bahasa alami (*free-text query*). Di sisi lain, keandalan Vector Space Model (VSM) dan metrik Cosine Similarity dalam merepresentasikan teks dan mengukur presisi kemiripannya telah tervalidasi dengan sangat baik oleh studi Sujjada et al. (2021) serta pengujian rekomendasi pada literatur oleh Ridhwanullah et al. (2024). Menindaklanjuti temuan-temuan tersebut, penelitian ini mengisi celah yang ada dengan mengintegrasikan algoritma pembobotan teks ke dalam arsitektur pencarian pariwisata untuk mengatasi kelemahan filter statis tersebut.

Sistem cerdas ini dikembangkan pada arsitektur perangkat lunak berbasis *website* yang difokuskan secara spesifik pada pemetaan potensi objek wisata di wilayah Bali. Mengingat kombinasi algoritma TF-IDF dan VSM telah umum diimplementasikan pada berbagai sistem temu kembali informasi, kebaruan dari penelitian ini terletak pada integrasi formulasi skema pembobotan multi-atribut (Nama, Deskripsi, dan Kategori) yang dikonstruksi secara khusus untuk membedah karakteristik ekosistem pariwisata Bali. Kontribusi unik penelitian ini adalah merancang kerangka kerja yang tidak sekadar mengukur kemiripan teks secara absolut, melainkan merekayasa arsitektur pembobotan tersebut untuk secara proaktif mengurai ketimpangan popularitas destinasi. Melalui pendekatan ini, sistem terbukti tangguh dalam mengangkat visibilitas destinasi potensial (*hidden gems*) di Bali murni melalui pemahaman semantik kueri bahasa alami wisatawan, sehingga mampu mengatasi kelemahan model konvensional yang bias terhadap destinasi *mainstream*.

KAJIAN LITERATUR

Penelitian Sebelumnya

Penelitian-penelitian terdahulu telah berupaya merespons tantangan penelusuran informasi, namun mayoritas masih menyisakan celah pada aspek fleksibilitas pencarian. Misalnya, studi oleh Aziza et al. (2025) yang merancang sistem pariwisata berbasis *website* mengungkap batasan fundamental dari metode konvensional: pencarian berbasis kategori statis gagal mengakomodasi intensi spesifik kueri bahasa alami (*free-text query*) dari pengguna. Celah keterbatasan infrastruktur ini menuntut pendekatan komputasi yang lebih dinamis berbasis teks, seperti yang diinisiasi oleh Pratiwi et al. (2024). Mereka membuktikan bahwa *Content-Based Filtering* dan algoritma TF-IDF efektif membedah dan mereduksi *noise* pada teks deskriptif pariwisata. Namun, ekstraksi bobot kata kunci semata belum cukup presisi jika tidak diintegrasikan dengan metode komputasi jarak vektor untuk memeringkatkan hasilnya secara hierarkis.

Untuk menyempurnakan akurasi kemiripan tersebut, kajian oleh Sujjada & Fergina (2021) dan Ridhwanullah et al. (2024) memvalidasi bahwa *Vector Space Model* (VSM) menggunakan metrik *Cosine Similarity* merupakan instrumen terkuat dalam memetakan relevansi multidimensi antara kueri dan dokumen. Melalui sintesis analisis literatur tersebut, terlihat jelas bahwa secara individual, TF-IDF kuat dalam ekstraksi *term* dan VSM andal dalam pemeringkatan. Oleh karena itu, penelitian ini tidak sekadar mereplikasi penggunaan algoritma tersebut secara homogen, melainkan menawarkan pendekatan analitis yang lebih tajam. Celah penelitian terdahulu diisi dengan mengintegrasikan kedua algoritma tersebut dan merekayasa ekstraksi komputasinya ke dalam skema multi-atribut (Nama, Deskripsi, Kategori) dengan proporsi bobot terkalibrasi. Pendekatan ini dirancang spesifik untuk membongkar dominasi popularitas dan menemukan probabilitas destinasi

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

hidden gems di Bali yang tidak dapat dijangkau oleh sistem pencarian konvensional.

Sistem Rekomendasi dan Data Mining

Sistem rekomendasi merupakan implementasi teknologi yang dirancang untuk memprediksi utilitas suatu entitas guna mengatasi ledakan informasi (Gunarto et al., 2023). Berbeda dengan *Collaborative Filtering* yang bergantung pada riwayat interaksi dan rating komunitas, penelitian ini mengadopsi *Content-Based Filtering* yang murni berfokus pada analisis atribut intrinsik (teks deskripsi) dari objek wisata (Pratiwi et al., 2024). Pendekatan ini merupakan manifestasi dari siklus *Data Mining* dan *Text Mining*, di mana prosedur komputasi sistematis digunakan untuk mengekstraksi pola tersembunyi dari data bahasa alami yang tidak terstruktur menjadi wawasan terstruktur yang relevan (Agustina et al., 2021). Perbedaan mendasar penelitian ini dibandingkan dengan studi-studi sistem rekomendasi pariwisata sebelumnya terletak pada rekayasa arsitektur penambangan datanya. Jika penelitian terdahulu umumnya mengekstraksi dan mengukur kemiripan secara tunggal pada satu korpus teks secara homogen, penelitian ini mendiferensiasikan proses *text mining* ke dalam skema multi-atribut (Nama, Deskripsi, dan Kategori) dengan proporsi bobot komputasi yang terkalibrasi. Melalui pembaruan pendekatan ini, sistem tidak hanya menambang data berdasarkan kecocokan leksikal sederhana, melainkan mampu mengorkestrasi relevansi semantik secara komprehensif untuk mendeteksi dan mengangkat destinasi wisata *hidden gems* yang selama ini gagal dijangkau oleh penambangan data pada algoritma konvensional.

Teks Preprocessing

Sebelum data bahasa alami dapat diekstraksi oleh algoritma matematis, diperlukan rangkaian pra-pemrosesan teks (*text preprocessing*) untuk membersihkan data *noise* dan menstandardisasi variasi linguistik (Suhaeni et al., 2025). Tahapan krusial ini secara sekuensial meliputi *case folding* (penyeragaman karakter huruf kecil), *tokenizing* (pemecahan kalimat menjadi token tunggal), *filtering* atau *stopword removal* (penghapusan kosa kata umum yang tidak memiliki signifikansi analitis), serta *stemming* (reduksi kata berimbuhan menjadi kata dasar). Pada arsitektur sistem ini, proses morfologis bahasa diimplementasikan dengan memanfaatkan pustaka cerdas Sastrawi.

TF-IDF dan Vektor Space Model

Term Frequency-Inverse Document Frequency (TF-IDF) merupakan metode pembobotan statistik sentral dalam arsitektur penelusuran informasi (Al Rasyid et al., 2024). Metode ini mengkalkulasi frekuensi kemunculan leksikon pada dokumen spesifik (TF) dan mengukur tingkat kelangkaan kosa kata tersebut di seluruh ekosistem dokumen (IDF). Nilai bobot akhir suatu kata (W{i,j}) didapatkan dari perkalian kedua nilai tersebut seperti pada Persamaan 1.

$$W_{i,j} = TF_{i,j} \times IDF_i \qquad (1)$$

Lebih lanjut, *Vector Space Model* (VSM) dimanfaatkan sebagai kerangka kerja aljabar yang merepresentasikan kueri pencarian pengguna dan dokumen teks destinasi sebagai vektor di dalam ruang multidimensi (Panambunan et al., 2025). Derajat relevansi antara kueri dengan deskripsi destinasi dikomputasi secara akurat menggunakan *Cosine Similarity*, yakni dengan mengukur nilai kosinus sudut yang diapit oleh vektor kueri dan vektor dokumen. Secara matematis, rumus pengukuran jarak kedekatan ini dituliskan pada Persamaan 2.

$$CosSim(d_j, q) = \frac{\sum_{i=1}^{t}(w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \cdot \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \qquad (2)$$

Keterangan:

$CosSim(d_j, q)$ = Nilai kemiripan antara dokumen dj dan kueri q
w_{ij} = Bobot kata ke-i dalam dokumen ke-j
w_{iq} = Bobot kata ke-i dalam kueri q
t = Jumlah total kata unik

Kombinasi kedua metode matematis ini sanggup mengubah proses pencarian yang kualitatif (pemahaman bahasa) menjadi operasi kuantitatif yang terukur dan presisi, di mana nilai kemiripan akan berada pada rentang 0 hingga 1.

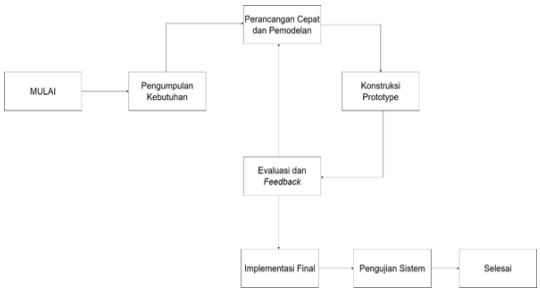
METODE PENELITIAN

Penelitian ini mengintegrasikan metode eksperimental pada ranah *data mining* dengan rekayasa perangkat lunak berlandaskan model *prototype*. Pendekatan ini dipilih karena pengembangan sistem berbasis penelusuran informasi menuntut validasi algoritma secara berulang untuk mencapai presisi optimal. Tahapan penelitian diawali dengan identifikasi masalah *information overload* dan studi literatur terkait algoritma, dilanjutkan dengan pengumpulan korpus

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

data destinasi pariwisata Bali melalui observasi digital. Data tersebut kemudian melalui tahap prapemrosesan teks guna mentransformasi narasi mentah menjadi format terstruktur. Fase pengembangan sistem menerapkan siklus *prototyping* yang mencakup perancangan arsitektur dan pemodelan komputasi, lalu diakhiri dengan pengujian untuk memverifikasi fungsionalitas perangkat lunak serta mengukur relevansi rekomendasi terhadap kueri pengguna. Alur pengembangan ini diilustrasikan pada Gambar 1.



Gambar 1. Alur Pengembangan System Prototype

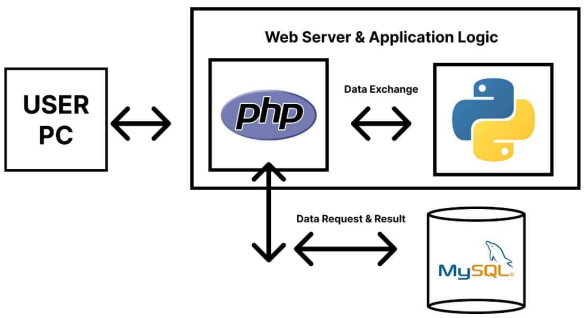
Objek dan Pengumpulan Data

Objek utama penelitian ini adalah data pariwisata yang mencakup destinasi wisata di wilayah Bali. Pemilihan wilayah ini didasarkan pada tingginya densitas potensi wisata yang sering kali mengalami ketimpangan popularitas, di mana banyak destinasi potensial (*hidden gems*) tertutup oleh minimnya informasi digital dan dominasi destinasi arus utama (*mainstream*). Secara spesifik, objek data diklasifikasikan menjadi dua kategori utama. Pertama, atribut destinasi yang merupakan data faktual (meliputi nama tempat, lokasi, kategori, kata kunci, dan gambar) untuk memberikan gambaran spesifik pada antarmuka aplikasi. Kedua, data teks deskripsi yang menjadi inti dari proses penambangan teks (*text mining*). Data ini berupa paragraf naratif yang mendeskripsikan suasana, aktivitas, dan karakteristik unik destinasi. Data deskripsi dikumpulkan secara mandiri melalui kurasi konten. Penyusunan narasi tersebut direkayasa agar kaya akan variasi kosakata dan sinonim guna mengoptimalkan kinerja algoritma TF-IDF dan VSM dalam menangkap konteks kueri pengguna yang beragam. Secara keseluruhan, korpus data yang dikumpulkan dan digunakan dalam ekosistem sistem ini berjumlah 488 entitas destinasi wisata. Skala kuantitatif dataset ini dikonstruksi secara representatif agar mampu menyimulasikan fenomena *information overload*, sehingga keandalan algoritma komputasi dalam memfiltrasi informasi yang relevan dan mengeliminasi data *noise* dapat divalidasi secara optimal.

Penentuan komposisi bobot pada perhitungan kemiripan akhir yakni atribut Nama (40%), Deskripsi (35%), dan Kategori (25%) ditetapkan berdasarkan tingkat spesifisitas informasi terhadap intensi pencarian wisatawan. Atribut Nama diberikan bobot tertinggi (40%) untuk mengakomodasi pencarian eksplisit (*exact match*); apabila pengguna memasukkan kueri yang identik dengan nama objek wisata, sistem secara arsitektural harus memprioritaskan dokumen tersebut sebagai luaran utama. Selanjutnya, atribut Deskripsi memegang proporsi yang signifikan (35%) karena atribut ini merupakan inti dari penambangan teks (TF-IDF). Bobot ini memastikan sistem mampu menangkap kedekatan semantik, variasi kosakata, serta karakteristik tersembunyi (*hidden gems*) yang dicari pengguna meski tidak terwakili oleh penamaan destinasi. Sementara itu, atribut Kategori diberikan bobot paling kecil (25%) lantaran sifatnya yang terlalu umum. Kecocokan pada kategori berfungsi krusial sebagai penyaring awal agar konteks rekomendasi tetap relevan (misalnya memisahkan entitas wisata gunung dan pantai), namun tidak mendominasi penentuan peringkat akhir.

Arsitektur Sistem

Sistem dikembangkan berbasis *website* menggunakan pola desain *Model-View-Controller* (MVC) dengan mengintegrasikan dua lingkungan komputasi secara dinamis melalui antarmuka *Representational State Transfer* (REST) API. Kerangka kerja PHP Laravel difungsikan sebagai pengelola basis data MySQL dan penyaji presentasi visual (*front-end*). Sementara itu, modul Python Flask dialokasikan secara eksklusif sebagai unit pemrosesan cerdas (*backend*) untuk mengeksekusi beban komputasi matriks algoritma TF-IDF dan VSM secara waktu nyata (*real-time*). Mekanisme pertukaran data ini diilustrasikan pada Gambar 2.



Gambar 2. Diagram Arsitektur Sistem

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

HASIL DAN PEMBAHASAN

Preprocessing Data

Dalam implementasi penelitian ini, *dataset* yang digunakan secara spesifik terdiri dari 488 entitas destinasi wisata di wilayah Bali. Atribut faktual dari destinasi tersebut dikumpulkan melalui teknik *web scraping* dari platform Google Maps. Distribusi kategori wisata dari dataset ini sangat beragam, mencakup Wisata Bukit, Cafe / Resto, Wisata Air Terjun, Wisata Pantai, Taman Wisata, Wisata Pantai / Alam, Beach Club, Taman Wisata / Rekreasi, Wisata Gunung / Pendakian, Bar / Lounge, Wisata Air / Alam, Bumi Perkemahan, Glamping, Internet Cafe / Coworking, Klub Malam, Wisata Danau, Wisata Air Panas, Wisata Hiburan / Rekreasi, Camping Ground, hingga Monkey Forest. Guna mengoptimalkan kinerja algoritma dalam menangkap konteks kueri wisatawan, data teks deskripsi untuk setiap destinasi dikurasi dan divalidasi secara mandiri. Proses validasi ini dilakukan dengan merancang ulang narasi deskripsi agar sekomprehensif mungkin dalam memuat variasi kosakata, sinonim, karakteristik unik, serta suasana destinasi sehingga mampu merepresentasikan kondisi aktual destinasi tersebut dan siap untuk diproses oleh algoritma penambangan teks.

Sebelum algoritma komputasi dapat mengekstraksi bobot informasi dari korpus data yang telah dikurasi tersebut, data teks bahasa alami harus melalui tahapan pra-pemrosesan (*text preprocessing*) guna mereduksi data *noise* dan menstandardisasi variasi morfologis kata. Proses ini merupakan syarat mutlak agar sistem dapat menerjemahkan kueri pengguna menjadi representasi yang dapat diukur secara kuantitatif. Berikut adalah 3 data query dalam tahapan preprocessing:

Tabel 1. *Preprocessing Data Query 1*

Tahap Preprocessing	Deskripsi Proses / Detail	Output Kalimat / Kata
Kueri Awal	Penerimaan kueri input asli dari pengguna.	BERIKAN REKOMENDASI SPOT MENYELAM DAN DIVINGG DI PANTAI YANG MEMILKI SPOT ESTETIK UNTUK MELIHAT TERUMBU KARANG!!!
Case Folding	Penyeragaman seluruh karakter menjadi huruf kecil (*lowercase*).	berikan rekomendasi spot menyelam dan divingg di pantai yang memilki spot estetik untuk melihat terumbu karang!!!
Typo Correction	Perbaikan kesalahan pengetikan leksikal (memilki → memiliki, divingg → diving).	berikan rekomendasi spot menyelam dan diving di pantai yang memiliki spot estetik untuk melihat terumbu karang!!!
Cleaning	Eliminasi karakter non-alfabetik, tanda baca (!!!), dan spasi berlebih.	berikan rekomendasi spot menyelam dan diving di pantai yang memiliki spot estetik untuk melihat terumbu karang
Stopwords Removal	Penghapusan kosa kata fungsional yang tidak memiliki signifikansi analitis (berikan, dan, di, yang, memiliki, untuk, melihat).	rekomendasi spot menyelam diving pantai spot estetik terumbu karang
Stemming Sastrawi	Reduksi kata berimbuhan menjadi bentuk kata dasar (menyelam → selam) memanfaatkan pustaka Sastrawi.	rekomendasi spot selam diving pantai spot estetik terumbu karang

Tabel 2. *Preprocessing Data Query 2*

Tahap Preprocessing	Deskripsi Proses / Detail	Output Kalimat / Kata
Kueri Awal	Penerimaan kueri input asli dari pengguna.	beach club
Case Folding	Penyeragaman seluruh karakter menjadi huruf kecil (*lowercase*). Dalam kasus ini, teks sudah murni dalam format *lowercase*.	beach club

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Typo Correction	Perbaikan kesalahan pengetikan leksikal. Karena kosakata yang dimasukkan tidak memiliki kesalahan ejaan, tidak ada *string* yang diganti.	beach club
Cleaning	Eliminasi karakter non-alfabetik, tanda baca, dan spasi berlebih.	beach club
Stopwords Removal	Penghapusan kosa kata fungsional. Karena kueri ini merupakan istilah bahasa Inggris utuh dan tidak mengandung Custom Stopword pariwisata, frasa dibiarkan tetap utuh.	beach club
Stemming Sastrawi	Reduksi kata berimbuhan menjadi bentuk kata dasar. Pustaka Sastrawi secara arsitektur tidak mengenali afiksasi morfologis bahasa asing, sehingga struktur kata ini berhasil lolos 100%.	beach club

Tabel 3. *Preprocessing Data Query 3*

Tahap Preprocessing	Deskripsi Proses / Detail	Output Kalimat / Kata
Kueri Awal	Penerimaan kueri input asli dari pengguna.	pegunungan yang cocok di daki oleh pendaki pemula dengan view yang indah
Case Folding	Penyeragaman seluruh karakter menjadi huruf kecil (*lowercase*).	pegunungan yang cocok di daki oleh pendaki pemula dengan view yang indah
Typo Correction	Perbaikan kesalahan pengetikan leksikal. Seluruh ejaan kalimat yang diinput pengguna sudah berstatus baku, sehingga tidak ada yang perlu diperbaiki.	pegunungan yang cocok di daki oleh pendaki pemula dengan view yang indah
Cleaning	Eliminasi karakter non-alfabetik dan tanda baca. Kalimat ini hanya mengandung spasi dan alfabet dasar.	pegunungan yang cocok di daki oleh pendaki pemula dengan view yang indah
Stopwords Removal	Penghapusan kosa kata fungsional yang tidak memiliki signifikansi analitis (yang, di, oleh, dengan, indah). Menghasilkan penyusutan kalimat yang drastis.	pegunungan cocok daki pendaki pemula view
Stemming Sastrawi	Reduksi kata berimbuhan menjadi bentuk kata dasar (pegunungan -> gunung, pendaki -> daki, pemula -> mula) memanfaatkan pustaka Sastrawi.	gunung cocok daki daki mula view

Skenario Uji Model dan Hasil Rekomendasi

Pengujian performa dan tingkat relevansi model dieksekusi menggunakan representasi kueri yang telah melalui tahapan prapemrosesan (*preprocessing*) guna merepresentasikan intensi pencarian wisatawan secara natural. Dalam komputasi pemeringkatan, sistem menerapkan metode *Cosine Similarity* untuk mengukur jarak kedekatan vektor antara kueri preferensi pengguna dengan korpus dokumen destinasi wisata. Nilai skor akhir dikalkulasi secara objektif berdasarkan distribusi pembobotan multi-atribut.

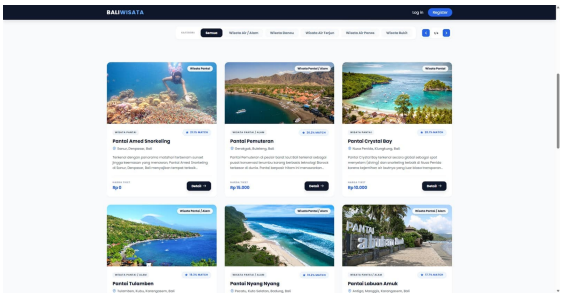
Penentuan komposisi bobot pada perhitungan kemiripan akhir yakni Nama (40%), Deskripsi (35%), dan Kategori (25%) ditetapkan melalui serangkaian eksperimen heuristik dan evaluasi empiris secara berulang selama fase pengujian *prototype*. Atribut Nama diberikan bobot tertinggi (40%) untuk mengakomodasi pencarian eksplisit (*exact match*); apabila pengguna memasukkan kueri yang identik dengan nama objek wisata, sistem secara arsitektural harus memprioritaskan dokumen tersebut sebagai luaran utama. Selanjutnya, atribut Deskripsi memegang proporsi yang signifikan (35%) karena atribut ini merupakan inti dari penambangan teks (TF-IDF). Bobot ini memastikan sistem mampu menangkap kedekatan semantik, variasi kosakata, serta karakteristik tersembunyi (*hidden gems*) yang dicari pengguna meski tidak terwakili oleh penamaan destinasi. Sementara itu, atribut Kategori diberikan bobot paling kecil (25%) lantaran sifatnya yang terlalu umum. Kecocokan pada kategori berfungsi krusial sebagai penyaring awal agar konteks rekomendasi tetap relevan (misalnya memisahkan entitas wisata gunung dan pantai), namun tidak mendominasi penentuan peringkat akhir.

Implementasi dari skema pemeringkatan algoritmik tersebut selanjutnya disajikan secara interaktif kepada pengguna melalui antarmuka sistem

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

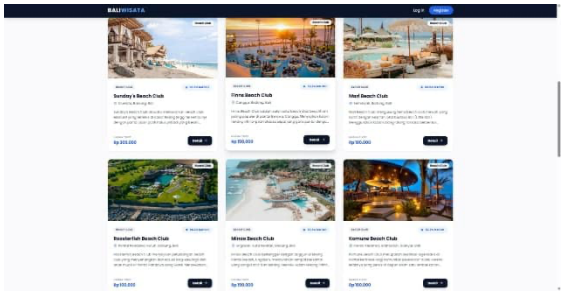
ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

berbasis *website*. Ketika wisatawan memasukkan kueri naratif terkait pencarian spot menyelam, sistem mengeksekusi kalkulasi beban matriks di sisi *back-end* dan menampilkan deretan rekomendasi terbaik secara *real-time*. Representasi visual dari hasil penelusuran sistem tersebut dapat dilihat pada gambar dibawah.

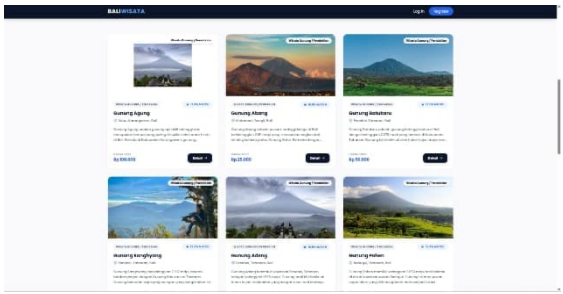


Gambar 3. Tampilan Antarmuka Hasil Rekomendasi Destinasi Wisata Query 1

Di balik representasi visual pada Gambar 3, 4, dan 5, sistem melakukan kalkulasi matematis yang ketat untuk mengurutkan destinasi. Tabel 4, 5, dan 6 memaparkan secara rinci metrik komputasi luaran algoritmik yang menghasilkan lima rekomendasi teratas (*Top-5 Recommendations*) pada pengujian kueri tersebut.



Gambar 4. Tampilan Antarmuka Hasil Rekomendasi Destinasi Wisata Query 2



Gambar 5. Tampilan Antarmuka Hasil Rekomendasi Destinasi Wisata Query 3

Tabel 4. Hasil Rekomendasi Sistem Query 1

Rank	ID	Nama Destinasi Wisata	Kategori Wisata	Sim_{Nama}	Sim_{Kat}	Sim_{Desc}	Skor Akhir	Relevansi
1	476	Pantai Amed Snorkeling	Wisata Pantai	0,2391	0,2525	0,1504	0,2114	R
2	425	Pantai Pemuteran	Wisata Pantai / Alam	0,1516	0,2203	0,2462	0,2019	R
3	455	Pantai Crystal Bay	Wisata Pantai	0,1239	0,2525	0,2532	0,2013	R
4	431	Pantai Tulamben	Wisata Pantai / Alam	0,1504	0,2203	0,2216	0,1928	R
5	414	Pantai Nyang Nyang	Wisata Pantai / Alam	0,1142	0,2203	0,2609	0,1921	R

Tabel 5. Hasil Rekomendasi Sistem Query 2

Rank	ID	Nama Destinasi Wisata	Kategori Wisata	Sim_{Nama}	Sim_{Kat}	Sim_{Desc}	Skor Akhir	Relevansi
1	203	Sunday's Beach Club	Beach Club	0,5801	1,0000	0,2451	**0,5678**	R
2	186	Finns Beach Club	Beach Club	0,5682	1,0000	0,2413	**0,5617**	R

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

3	204	Mari Beach Club	Beach Club	0,5682	1,0000	0,2294	**0,5576**	R
4	232	Roosterfish Beach Club	Beach Club	0,5682	1,0000	0,2119	**0,5515**	R
5	231	Minoo Beach Club	Beach Club	0,5682	1,0000	0,1350	**0,5245**	R

Tabel 6. Hasil Rekomendasi Sistem Query 3

Rank	**ID**	**Nama Destinasi Wisata**	**Kategori Wisata**	Sim_{Nama}	Sim_{Kat}	Sim_{Desc}	**Skor Akhir**	**Relevansi**
1	97	Gunung Agung	Wisata Gunung / Pendakian	0,0773	0,4170	0,1113	**0,1741**	R
2	98	Gunung Abang	Wisata Gunung / Pendakian	0,0694	0,4170	0,1048	**0,1687**	R
3	99	Gunung Batukaru	Wisata Gunung / Pendakian	0,0626	0,4170	0,0776	**0,1564**	R
4	121	Gunung Sanghyang	Wisata Gunung / Pendakian	0,0579	0,4170	0,0818	**0,1560**	R
5	117	Gunung Adeng	Wisata Gunung / Pendakian	0,0579	0,4170	0,0797	**0,1553**	R

Pengujian Confusion Matrix

Evaluasi kinerja kuantitatif diimplementasikan menggunakan tiga skenario kueri pengujian dengan tingkat kompleksitas yang bervariasi guna menguji ketahanan model. Ketiga kueri ini direkayasa untuk merepresentasikan berbagai intensi pencarian wisatawan, mulai dari kueri spesifik yang singkat, kueri dengan kategori luas, hingga kueri naratif berskala kompleks. Khusus pada kueri naratif, pengujian menyertakan kesalahan pengetikan leksikal (*typo*), kapitalisasi tidak beraturan, penggunaan tanda baca berlebih, serta penggabungan kata fungsional (*stopword*) dengan kata kunci semantik utama. Skenario pengujian ini dieksekusi tanpa menerapkan pembatasan visualisasi rekomendasi peringkat teratas (*Top-N Cutoff*) pada antarmuka. Pendekatan ini ditujukan untuk mengukur efektivitas absolut dari sistem penyaringan ambang batas (*threshold*) kemiripan pada skor > 0,05. Hasil komputasi parameter *Confusion Matrix* dari ketiga pengujian kueri tersebut disajikan pada Tabel 7.

Tabel 7. *Pengujian Confusion Matrix*

No	**Skenario Kueri Uji**	**TP**	**FP**	**FN**	**TN**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
1	BERIKAN REKOMENDASI SPOT MENYELAM DAN DIVINGG DI PANTAI YANG MEMILKI SPOT ESTETIK UNTUK MELIHAT TERUMBU KARANG!!!	12	100	5	371	78,48%	10,71%	70,59%	18,60%
2	beach club	26	16	0	446	96,72%	61,90%	100,00%	76,47%
3	pegunungan yang cocok di daki oleh pendaki pemula dengan view yang indah	23	52	0	413	89,34%	30,67%	100,00%	46,94%

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Berdasarkan klasifikasi parameter pada Tabel 7, arsitektur sistem memperlihatkan dinamika kemampuan identifikasi dan filtrasi yang adaptif terhadap berbagai variasi kueri:

1. True Positive (TP) & False Negative (FN): Pada Skenario 2 dan Skenario 3, sistem berhasil mengidentifikasi seluruh destinasi yang relevan dengan sempurna, dibuktikan dengan parameter FN yang bernilai 0 (*Recall* 100,00%). Hal ini mengartikan tidak ada satupun destinasi relevan yang tereliminasi. Namun, pada Skenario 1 (Kueri Panjang dan *Noise*), sistem mencatatkan 5 FN, yang mengindikasikan adanya segelintir destinasi relevan yang tidak lolos ambang batas karena kompleksitas kueri yang terlalu panjang terhadap teks deskripsi destinasi.
2. True Negative (TN): Model terbukti konsisten dan sukses memfiltrasi entitas destinasi yang secara absolut tidak relevan melintasi berbagai skenario kueri (dengan rentang TN tinggi, yakni 371 hingga 446 dokumen) dengan menetapkan skor di bawah ambang batas ($\leq 0{,}05$). Hal ini membuktikan keandalan sistem dalam mengenali dan membuang konteks *noise* utama dari pangkalan data secara efektif.
3. False Positive (FP): Tingkat FP sangat bergantung pada keluasan semantik kueri. Pada Skenario 1, terdapat 100 destinasi (FP) yang diklasifikasikan relevan semata-mata karena memiliki kesamaan sebagian kata kunci pada kata umum seperti "pantai" dan "estetik". Skenario 3 juga mencatatkan 52 FP karena sifat pencarian yang umum (pegunungan). Sebaliknya, pada kueri yang lebih singkat dan terarah seperti Skenario 2, sistem berhasil menekan angka FP secara signifikan menjadi hanya 16 dokumen.

Fenomena komputasi matriks di atas melahirkan rasio *trade-off* yang bervariasi dalam ranah sistem penemuan informasi (*Information Retrieval*). Pada kueri yang spesifik (Skenario 2 dan 3), sistem mencatatkan *Recall* absolut 100,00% yang secara empiris memvalidasi keandalan integrasi modul prapemrosesan NLP. Khusus pada Skenario 1, meskipun *Recall* tertekan ke angka 70,59% akibat kueri panjang, algoritma *Typo Correction* terbukti adaptif merekonstruksi masukan ("DIVINGG" dikoreksi menjadi "diving", "MEMILKI" menjadi "memiliki"). Proses ini, yang dipadukan dengan tahapan *stemming*, memastikan pencocokan term berjalan maksimal sehingga sistem tetap mampu mengamankan 12 destinasi relevan (TP) dari kueri mentah tersebut. Di sisi lain, Skenario 2 menunjukkan performa evaluasi paling proporsional dengan *Precision* yang solid (61,90%) dan *F1-Score* tertinggi (76,47%).

Secara keseluruhan, ketiadaan batasan pemotongan antarmuka (*Top-N Cutoff*) mengharuskan sistem untuk merilis seluruh dokumen yang lolos *threshold*. Meskipun kebijakan ini menekan metrik *Precision* secara drastis pada kueri yang menggunakan kosakata umum (seperti pada Skenario 1 dengan *Precision* 10,71%), arsitektur ini secara sengaja memprioritaskan ketersediaan seluruh alternatif pilihan. Pendekatan ini diambil agar semua destinasi yang berpotensi relevan, terutama *hidden gems*, dapat tersaji secara komprehensif kepada wisatawan, tanpa adanya informasi krusial yang tertinggal oleh pembatasan antarmuka.

Evaluasi Kinerja dengan Mean Average Precision (MAP)

Guna memastikan bahwa rendahnya nilai presisi matriks tidak merusak pengalaman pengguna (*User Experience*), metrik *Mean Average Precision* (MAP) diaplikasikan untuk mengukur kualitas urutan pemeringkatan (*ranking quality*) dokumen di antarmuka. Rekapitulasi hasil komputasi *Average Precision* (AP) disajikan pada Tabel 8.

Tabel 8. Rekapitulasi *Average Precision (AP)*

No	Skenario Kueri Uji	Average Precision (AP)
1	BERIKAN REKOMENDASI SPOT MENYELAM DAN DIVINGG DI PANTAI YANG MEMILKI SPOT ESTETIK UNTUK MELIHAT TERUMBU KARANG!!!	34,54%
2	beach club	100,00%
3	pegunungan yang cocok di daki oleh pendaki pemula dengan view yang indah	83,91%
-	Mean Average Precision (MAP)	72,82%

Mengacu pada Tabel 8, arsitektur sistem berhasil mendokumentasikan skor *Mean Average Precision* (MAP) sebesar 72,82%. Dalam konteks pengujian kueri masif tanpa batasan jumlah luaran, metrik ini memberikan justifikasi akademis yang kuat mengenai integritas algoritma pemeringkatan *Vector Space Model* (VSM). Nilai MAP tersebut mengindikasikan bahwa sistem secara konsisten sukses menempatkan destinasi-destinasi yang paling relevan (TP) dari berbagai variasi kueri untuk menduduki klasemen hierarki teratas pada daftar hasil penelusuran. Keandalan pemeringkatan ini dibuktikan secara empiris pada Skenario 2 yang mencatatkan nilai

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Average Precision (AP) sempurna sebesar 100,00%, serta Skenario 3 dengan skor AP sebesar 83,91%. Sebaliknya, meskipun pada pengujian kueri panjang yang bising (Skenario 1) nilai AP tertekan menjadi 34,54% akibat tingginya rasio *False Positive*, dokumen-dokumen yang merupakan data *noise* (FP) tersebut berhasil diisolasi oleh sistem pada urutan terbawah antarmuka karena memiliki akumulasi pembobotan kosinus yang jauh lebih kecil dan mendekati batas *threshold* 0,05. Distribusi pemeringkatan yang presisi ini menjamin pengguna tidak akan terganggu oleh eksposur data yang tidak relevan, karena informasi dengan nilai relevansi terkuat akan selalu disajikan pada fokus pandangan utama.

KESIMPULAN

Berdasarkan hasil pengujian dan pembahasan yang telah dilakukan, penerapan algoritma *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Vector Space Model* (VSM) dengan pendekatan *Content-Based Filtering* berhasil diimplementasikan secara komprehensif untuk membangun sistem rekomendasi destinasi wisata di Bali. Sistem ini terbukti efektif dalam mengatasi masalah *information overload*, kesenjangan semantik (*semantic gap*), serta kendala *cold-start* dengan mengekstraksi karakteristik intrinsik dari kueri bahasa alami pengguna terhadap teks deskripsi destinasi. Dengan demikian, rekomendasi untuk mengangkat visibilitas destinasi potensial dan *hidden gems* tidak lagi bergantung pada metrik popularitas atau *rating*. Berdasarkan evaluasi kinerja menggunakan berbagai variasi skenario kueri tanpa batasan pemotongan antarmuka (*Top-N Cutoff*), model menunjukkan kemampuan identifikasi yang adaptif, mencatatkan tingkat *Recall* absolut hingga 100,00% pada kueri yang spesifik. Hal ini membuktikan keberhasilan sistem dalam menyajikan seluruh destinasi yang relevan tanpa ada informasi krusial yang tertinggal. Penggunaan ambang batas relevansi (*threshold*) pada skor 0,05 juga terbukti konsisten dan andal dalam memfiltrasi ratusan dokumen *noise* dari pangkalan data di berbagai kondisi pencarian.

Meskipun eksposur seluruh luaran memicu rasio *trade-off* yang menekan nilai *Precision* pada kueri berskala semantik luas (hingga menyentuh angka 10,71%), arsitektur sistem secara keseluruhan tetap mampu menjaga kualitas pemeringkatan (*ranking quality*) yang sangat baik dengan skor *Mean Average Precision* (MAP) sebesar 72,82%. Pencapaian ini mengonfirmasi bahwa algoritma secara konsisten berhasil mengisolasi data *noise* dan menempatkan destinasi-destinasi yang paling relevan pada klasemen hierarki teratas di antarmuka pencarian pengguna.

Sebagai saran pengembangan untuk penelitian selanjutnya, performa dan personalisasi sistem dapat ditingkatkan dengan mengintegrasikan metode *Collaborative Filtering* bersama dengan *Content-Based Filtering* yang telah ada, guna membentuk *Hybrid Recommender System*. Selain itu, penambahan fitur berbasis lokasi geografis (*Location-Based Services*) direkomendasikan agar sistem tidak hanya mempertimbangkan relevansi semantik, tetapi juga dapat menyarankan destinasi secara optimal berdasarkan kedekatan jarak dari lokasi wisatawan.

DAFTAR PUSTAKA

Agustina, D. A., Subanti, S., & Zukhronah, E. (2021). Implementasi Text Mining Pada Analisis Sentimen Pengguna Twitter Terhadap Marketplace di Indonesia Menggunakan Algoritma Support Vector Machine. *Indonesian Journal of Applied Statistics, 3*(2), 109-122. https://doi.org/10.13057/IJAS.V3I2.44337

Al Rasyid, R., Handayani, D., & Ningsih, U. (2024). Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian Pada Dataset Destinasi Wisata. *Jurnal Teknologi Informasi Dan Komunikasi)*, *8*(1), 2024. https://doi.org/10.35870/jti k.v8i1.1416

Aziza, A. N., Jamil, M., & Aris, V. (2025). Rancang Bangun Sistem Informasi Pariwisata dan UMKM Lokal Kabupaten Wajo Berbasis Website. *RIGGS: Journal of Artificial Intelligence and Digital Business, 4*(2), 1117–1124. https://doi.org/10.31004/riggs.v4i2.628

Gunarto, S. A., Honggara, E. S., & Purwanto, D. D. (2023). Website Sistem Rekomendasi dengan Content Based Filtering pada Produk Perawatan Kulit. *Jurnal Sistem Dan Teknologi Informasi (JustIN), 11*(3), 399. https://doi.org/10.26418/JUSTIN.V11I3.59049

Hidayat, U., Fatah, Z., & Fajriyanto. (2024). Sistem Informasi Geografis (SIG) Pemetaan Destinasi Wisata di Kabupaten Buleleng Bali. *Prosiding Seminar Nasional Sains Dan Teknologi "SainTek," 1*(2), 324–331.

Khusnah, R. S., & Roosdhani, M. R. (2025). The Influence of Information Overload on Decision Difficulty Mediated by Social Media Exhaustion on Generation Z in Jepara. *Jurnal Manajemen Stratejik dan Simulasi Bisnis, 6*(1), 1-14.

Maulana, T. I., & Abdillah, A. R. (2022). Pemanfaatan Sistem Temu Kembali Informasi dalam Pencarian Dokumen Menggunakan Vector Space Model. *Sistem dan Teknologi Informasi Indonesia (SINTESIA), 1*(2), 89-95.

Panambunan, S., Hasibuan, A., & Santa, K. (2025). Implementasi Vector Space Model pada Aplikasi Pengarsipan Berbasis Web di POLDA

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

Sulut. *Jurnal Minfo Polgan*, *14*(2), 2131–2137. https://doi.org/10.33395/JMP.V14I2.15203

Pratiwi, D., Asrianda, A., & Rosnita, L. (2024). Penerapan Metode Content-Based Filtering dalam Sistem Rekomendasi Objek Wisata di Aceh Tamiang. *Jurnal Ilmu Komputer Dan Informatika, 4*(2), 85–96. https://doi.org/10.54082/JIKI.169

Putra, K. R., & Rachman, M. A. (2024). Perbandingan Metode Content-based, Collaborative dan Hybrid Filtering pada Sistem Rekomendasi Lagu. *MIND (Multimedia Artificial Intelligent Networking Database) Journal, 9*(2), 179–193. https://doi.org/10.26760/MINDJOURNAL.V9I2.179-193

Ridhwanullah, D., Kumarahadi, Y. K., & Raharja, B. D. (2024). Content-Based Filtering pada Sistem Rekomendasi Buku Informatika. *Jurnal Ilmiah SINUS, 22*(2), 57–66. https://doi.org/10.30646/SINUS.V22I2.840

Saputri, D. R., Travis, D., Hilmi, N., Afriadi, A., Chidmahdjati, A. B., & Apriyenson, H. (2025). Pengembangan Smart Tourism Melalui Digitalisasi Promosi Wisata (Studi Kasus Pantai Tanjung Siambang). *Jurnal Pengabdian Masyarakat Bangsa, 2*(12), 5802–5808. https://doi.org/10.59837/JPMBA.V2I12.2070

Suhaeni, C., Kamila, S. A., Fahira, F., Yusran, M., & Alfa Dito, G. (2025). Exploring a Large Language Model on the ChatGPT Platform for Indonesian Text Preprocessing Tasks. *Indonesian Journal of Statistics & Its Applications*, *9*(1), 100. https://doi.org/10.29244/IJSA.V9I1P100-116

Sujjada, A., & Fergina, A. (2021). Implementasi Metode Vector Space Model untuk Deteksi Emosi Menggunakan Data Teks Twitter. Jurnal RESTIKOM: Riset Teknik Informatika Dan Komputer, 3(3), 116–129. https://doi.org/10.52005/RESTIKOM.V3I3.89

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.