

Sistem Rekomendasi Desa Wisata di Sleman menggunakan Metode *Content-Based Filtering* berbasis *TF-IDF* dan *Cosine Similarity*

A Tourist Village Recommendation System in Sleman using Content-based Filtering, TF-IDF, and Cosine Similarity

¹Herda Rafif Arya Putra*, ²Arif Nur Rohman

^{1,2}Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta

^{1,2}Jl. Ring Road Utara Condong Catur Depok Sleman, Yogyakarta

*e-mail: herdarafif@students.amikom.ac.id

(received: 5 August 2026, revised: 27 August 2026, accepted: 28 August 2026)

Abstrak

Sleman dikenal sebagai salah satu kabupaten di Yogyakarta yang memiliki banyak desa wisata dengan potensi yang beragam, mulai dari wisata alam, budaya, hingga kuliner. Namun, banyak wisatawan masih kesulitan menemukan desa wisata yang sesuai dengan minat dan kebutuhan mereka karena informasi yang tersedia cenderung umum dan tidak personal. Penelitian ini membangun sistem rekomendasi desa wisata di Sleman menggunakan pendekatan Content-Based Filtering yang dikombinasikan dengan TF-IDF dan Cosine Similarity. Dataset terdiri dari 20 desa wisata aktif di Kabupaten Sleman yang dihimpun dari situs resmi Dinas Pariwisata Kabupaten Sleman tahun 2024, dengan atribut berupa nama desa wisata, kecamatan, jenis wisata, daya tarik utama, dan fasilitas. Atribut tersebut digabungkan menjadi satu teks deskriptif per desa wisata, diproses melalui case folding, tokenisasi, penghapusan stopword, dan stemming Sastrawi hingga menghasilkan 125 fitur unik, lalu diberi bobot menggunakan TF-IDF. Cosine Similarity digunakan untuk menghitung kemiripan antara query preferensi pengguna dan setiap desa wisata, dan lima desa wisata dengan skor tertinggi ditampilkan sebagai rekomendasi (Top-5). Evaluasi dilakukan pada tiga skenario query menggunakan Precision@5, Recall@5, dan F1-Score dengan ground truth yang ditetapkan secara manual berdasarkan kesesuaian atribut. Hasil pengujian menunjukkan rata-rata Precision@5, Recall@5, dan F1-Score sebesar 93,3%. Penelitian ini diharapkan dapat membantu wisatawan menemukan desa wisata yang tepat, sekaligus mendukung promosi potensi pariwisata desa di Kabupaten Sleman.

Kata kunci: content-based filtering, cosine similarity, sistem rekomendasi, sleman, TF-IDF, desa wisata

Abstract

Sleman Regency, located in the Special Region of Yogyakarta, is home to numerous tourist villages offering diverse attractions, ranging from natural and cultural tourism to culinary experiences. However, many tourists still face difficulties in finding tourist villages that match their interests and needs because the available information tends to be general rather than personalized. This study develops a tourist village recommendation system for Sleman using a Content-Based Filtering approach combined with Term Frequency-Inverse Document Frequency (TF-IDF) and cosine similarity. The dataset comprises 20 active tourist villages in Sleman Regency, compiled from the official website of the Sleman Regency Tourism Office in 2024. The dataset includes the name of each tourist village, subdistrict, tourism category, main attractions, and facilities. These attributes were combined into a single descriptive text for each village and processed through case folding, tokenization, stopword removal, and stemming using the Sastrawi library, resulting in 125 unique features. The resulting text representations were then weighted using TF-IDF. Cosine similarity was applied to measure the similarity between user preference queries and each tourist village, and the five villages with the highest similarity scores were presented as recommendations (Top-5). The system was evaluated using three query scenarios and the Precision@5, Recall@5, and F1-score metrics, with ground-truth relevance labels manually established based on attribute matching. The

<http://sistemasi.ftik.unisi.ac.id>

evaluation results yielded an average score of 93.3% across Precision@5, Recall@5, and F1-score. This study is expected to help tourists identify suitable tourist villages while supporting the promotion of village tourism potential in Sleman Regency.

Keywords: *Content-based filtering, cosine similarity, recommendation system, sleman, TF-IDF, tourist village*

1 Pendahuluan

Sektor pariwisata Indonesia dalam beberapa tahun belakangan menunjukkan geliat yang cukup menggembirakan. Di antara berbagai jenis wisata yang ada, desa wisata menjadi salah satu yang paling banyak menarik perhatian. Konsepnya sederhana namun menarik: kawasan pedesaan yang dikemas sebagai destinasi wisata dengan mengedepankan kearifan lokal, entah itu budaya, panorama alam, maupun cita rasa kuliner khas daerah setempat. Lebih dari sekadar tempat berkunjung, desa wisata juga terbukti mampu menggerakkan roda perekonomian warga desa secara nyata dan langsung [1].

Kabupaten Sleman yang berada di sisi utara Daerah Istimewa Yogyakarta tercatat sebagai salah satu kabupaten dengan koleksi desa wisata terbanyak di tanah air. Data dari Dinas Pariwisata setempat menunjukkan puluhan desa wisata aktif dengan beragam kategori dan daya tarik yang berbeda-beda. Banyaknya pilihan ini ibarat pisau bermata dua di satu sisi menjadi kekayaan tersendiri, namun di sisi lain justru membingungkan wisatawan yang ingin menemukan destinasi yang benar-benar cocok dengan selera mereka [1], [2].

Kenyataan di lapangan menunjukkan bahwa kebanyakan wisatawan masih bergantung pada rekomendasi dari orang-orang terdekat, unggahan di media sosial, atau hasil pencarian di internet yang cenderung bersifat general. Informasi semacam itu tidak dirancang untuk mempertimbangkan kebutuhan tiap individu secara spesifik, sehingga wisatawan terpaksa menyaring sendiri informasi dari berbagai sumber yang berserakan. Kondisi seperti ini jelas tidak efisien, dan tidak jarang membuat wisatawan melewati destinasi yang seharusnya sangat pas dengan keinginan mereka [3].

Di sinilah teknologi informasi hadir sebagai solusi. Perkembangan di bidang kecerdasan buatan, khususnya sistem rekomendasi, membuka jalan untuk menjawab tantangan tersebut. Sistem rekomendasi sudah lama terbukti bekerja baik di berbagai platform, mulai dari toko online, layanan hiburan digital, hingga aplikasi perjalanan wisata. Cara kerjanya adalah menganalisis data dan preferensi pengguna, lalu secara otomatis menyajikan pilihan yang paling relevan tanpa pengguna harus repot mencari sendiri [4], [15].

Salah satu pendekatan yang banyak digunakan dalam sistem rekomendasi adalah Content-Based Filtering. Metode ini bekerja dengan mencocokkan karakteristik dari suatu item dengan profil preferensi pengguna. Agar proses pencocokan pada data teks bisa berjalan secara terstruktur, digunakan TF-IDF (Term Frequency-Inverse Document Frequency) untuk memberikan bobot pada setiap kata berdasarkan seberapa sering kata itu muncul dan seberapa unik kata tersebut di antara keseluruhan data. Selanjutnya, Cosine Similarity dimanfaatkan untuk mengukur seberapa dekat kemiripan antara preferensi pengguna dengan profil setiap desa wisata yang ada [5], [7], [8].

Kombinasi ketiga metode ini dipilih bukan tanpa alasan. Content-Based Filtering dengan TF-IDF dan Cosine Similarity mampu menghasilkan rekomendasi yang personal hanya bermodalkan atribut item, tanpa harus mengandalkan data dari pengguna lain seperti yang diperlukan dalam Collaborative Filtering. Sejumlah penelitian terdahulu pun sudah memperlihatkan bahwa kombinasi ini bekerja efektif di berbagai bidang [9], [10], [11].

Sejumlah penelitian terdahulu telah menerapkan Content-Based Filtering untuk rekomendasi wisata di berbagai wilayah, antara lain di Aceh Tamiang [5], Aceh Tengah [8], Kabupaten Pekalongan [6], dan Gunungkidul [3], [13], serta pencarian destinasi wisata berbasis query dengan TF-IDF dan Cosine Similarity [7]. Namun, penelitian-penelitian tersebut menyasar wilayah, objek, atau kategori wisata yang berbeda, seperti objek wisata umum atau wisata kuliner, dan tidak berfokus pada desa wisata yang atributnya beragam (wisata alam, budaya, edukasi, agrowisata, hingga kuliner) dalam satu profil yang sama. Selain itu, dari literatur yang dikaji, sepengetahuan penulis belum ditemukan penelitian yang secara khusus membangun sistem rekomendasi untuk desa wisata di Kabupaten Sleman.

Berangkat dari celah tersebut, penelitian ini membangun sistem rekomendasi desa wisata di Kabupaten Sleman dengan Content-Based Filtering, TF-IDF, dan Cosine Similarity. Kontribusi

penelitian ini adalah (1) penerapan metode pada desa wisata Sleman yang direpresentasikan sebagai profil teks gabungan dari jenis wisata, daya tarik, dan fasilitas; (2) analisis bobot TF-IDF untuk menunjukkan kata yang menjadi ciri khas tiap desa wisata; dan (3) evaluasi Top-5 menggunakan Precision, Recall, dan F1-Score pada tiga skenario preferensi yang mewakili tipe wisatawan berbeda, disertai pembahasan keterbatasan hasil. Dengan demikian, penelitian ini diharapkan memudahkan wisatawan menemukan desa wisata yang sesuai dengan minatnya, sekaligus menjadi dasar awal bagi pengembangan sistem rekomendasi desa wisata yang lebih luas dan membantu mempromosikan potensi wisata desa Sleman [3], [6].

2 Tinjauan Literatur

Desa wisata merupakan suatu kawasan pedesaan yang memiliki keunikan dan daya tarik tersendiri, baik dari segi alam, budaya, maupun kehidupan sosial masyarakatnya, yang dapat dijadikan sebagai produk wisata bagi para pengunjung [1]. Konsep desa wisata tidak sekadar menjadikan desa sebagai objek kunjungan, tetapi juga menempatkan masyarakat lokal sebagai pelaku utama dalam pengelolaan dan pengembangan pariwisata itu sendiri.

Di Indonesia, pengembangan desa wisata telah menjadi salah satu prioritas dalam kebijakan pariwisata nasional. Kabupaten Sleman sendiri dikenal sebagai salah satu daerah dengan ekosistem desa wisata yang cukup matang, dengan puluhan desa wisata aktif yang masing-masing memiliki karakteristik dan keunggulan berbeda, mulai dari wisata edukasi pertanian, kerajinan batik, kesenian tradisional, hingga wisata alam lereng Gunung Merapi [1], [2].

Secara sederhana, sistem rekomendasi bisa dipahami sebagai perangkat lunak yang tugasnya membantu pengguna menemukan item atau informasi yang relevan dari tumpukan data yang sangat besar, disesuaikan dengan preferensi atau kebutuhan masing-masing pengguna [4], [15]. Saat ini, sistem semacam ini sudah menjadi komponen yang hampir selalu hadir di berbagai platform digital, mulai dari toko belanja online, layanan streaming musik dan film, hingga aplikasi pemesanan perjalanan wisata.

Jika dilihat dari cara kerjanya, sistem rekomendasi umumnya menggunakan tiga pendekatan yang berbeda. Pendekatan pertama adalah Collaborative Filtering, yang bekerja dengan mengidentifikasi kesamaan pola perilaku dan selera antara satu pengguna dengan pengguna-pengguna lainnya. Pendekatan kedua adalah Content-Based Filtering, yang tidak melihat pengguna lain sama sekali, melainkan langsung mencocokkan karakteristik atau atribut suatu item dengan apa yang disukai pengguna itu sendiri. Pendekatan ketiga adalah Hybrid, yang menggabungkan keduanya agar kelemahan masing-masing metode bisa saling ditutupi [4], [14].

Content-Based Filtering merupakan salah satu metode dalam sistem rekomendasi yang prinsip kerjanya bertumpu pada kemiripan konten. Artinya, sistem akan mencari item-item lain yang atributnya mirip dengan item yang sebelumnya pernah disukai atau dipilih oleh pengguna [5], [12]. Karena tidak bergantung pada data dari pengguna lain, metode ini relatif lebih mudah diterapkan dan tidak terdampak oleh masalah cold start pada sisi item, selama data atribut itemnya tersedia secara memadai.

Secara teknis, alur kerjanya dimulai dari pembentukan representasi vektor untuk setiap item berdasarkan atribut-atributnya. Dari situ, sistem membangun profil preferensi pengguna berdasarkan riwayat item yang pernah dipilih atau diinteraksikan. Langkah terakhir adalah menghitung seberapa mirip profil pengguna tersebut dengan setiap item yang ada dalam sistem, lalu merekomendasikan item dengan skor kemiripan paling tinggi [5], [6], [8]. Dalam penelitian ini, setiap desa wisata diperlakukan sebagai sebuah dokumen teks yang mewakili informasi atributnya, mulai dari jenis wisata, kegiatan yang ditawarkan, fasilitas yang tersedia, hingga deskripsi umumnya.

Sejumlah penelitian terdahulu sudah memperlihatkan bahwa pendekatan ini cukup andal. Pratiwi, Asrianda, dan Rosnita [5] sukses membangun sistem rekomendasi objek wisata di Aceh Tamiang dengan Content-Based Filtering yang dikombinasikan bersama TF-IDF dan Cosine Similarity, dan hasilnya terbilang memuaskan. Hal serupa juga ditunjukkan oleh Salsabilla dan Utomo [6] yang menerapkan metode yang sama pada kasus wisata Kabupaten Pekalongan dan memperoleh tingkat relevansi rekomendasi yang baik.

TF-IDF adalah metode pembobotan kata yang sudah sangat lama dikenal dan banyak digunakan dalam dunia information retrieval maupun text mining. Cara kerjanya adalah mengukur seberapa penting suatu kata bagi sebuah dokumen, bukan hanya dilihat dari frekuensinya kemunculannya di dokumen itu saja, tetapi juga mempertimbangkan seberapa sering kata tersebut muncul di seluruh

<http://sistemasi.ftik.unisi.ac.id>

koleksi dokumen yang ada [7], [11]. Logikanya cukup masuk akal: kata yang sering muncul di satu dokumen tertentu tetapi jarang dijumpai di dokumen-dokumen lain kemungkinan besar adalah kata yang benar-benar mencirikan dokumen tersebut.

Metode ini terbentuk dari dua komponen yang bekerja bersama. Komponen pertama, menghitung seberapa sering sebuah kata muncul dalam satu dokumen semakin sering muncul, semakin besar nilainya. Komponen kedua, bertugas menilai keunikan sebuah kata di antara seluruh koleksi dokumen. Kata-kata umum seperti "dan", "yang", atau "di" yang hampir selalu hadir di mana-mana akan mendapat nilai IDF yang sangat kecil karena dianggap tidak terlalu bermakna. Sebaliknya, kata-kata yang hanya muncul di segelintir dokumen justru mendapat nilai IDF yang lebih besar, mencerminkan kekhasan katanya [9], [10].

Nilai TF-IDF akhir diperoleh dengan mengalikan kedua komponen tersebut, dan hasilnya berupa representasi vektor numerik dari setiap dokumen yang siap digunakan untuk proses perhitungan kemiripan [7]. Dalam penelitian ini, metode TF-IDF digunakan untuk merepresentasikan setiap desa wisata dalam bentuk vektor berdasarkan atribut teksnya, sehingga kemiripan antar desa wisata maupun antara preferensi pengguna dengan data desa wisata dapat dihitung secara matematis.

Cosine Similarity adalah metode yang digunakan untuk mengukur tingkat kemiripan antara dua buah vektor dengan cara menghitung nilai cosinus dari sudut yang terbentuk di antara keduanya [5], [7], [13]. Metode ini sangat populer digunakan dalam pengolahan teks dan sistem rekomendasi karena kemampuannya dalam membandingkan dua dokumen secara efektif, terlepas dari panjang atau ukuran dokumen tersebut.

Nilai Cosine Similarity berkisar antara 0 hingga 1. Nilai 1 menunjukkan bahwa kedua vektor memiliki arah yang identik, artinya dua dokumen tersebut sangat mirip satu sama lain. Sebaliknya, nilai 0 menunjukkan bahwa kedua vektor sama sekali tidak memiliki kesamaan. Dalam praktiknya, semakin tinggi nilai Cosine Similarity antara preferensi pengguna dengan suatu desa wisata, semakin besar kemungkinan desa wisata tersebut akan direkomendasikan kepada pengguna [8].

Salah satu keunggulan Cosine Similarity dibandingkan metode pengukuran jarak lainnya, seperti Euclidean Distance, adalah ketidakpekaan metode ini terhadap panjang dokumen. Dua dokumen yang memiliki konten serupa tetapi berbeda panjangnya tetap akan menghasilkan nilai kemiripan yang tinggi, asalkan distribusi kata-kata pentingnya relatif sama. Hal ini menjadikan Cosine Similarity sangat cocok dipadukan dengan TF-IDF dalam sistem rekomendasi berbasis teks seperti yang dibangun dalam penelitian ini [9], [15].

Beberapa penelitian sebelumnya yang relevan dengan penelitian ini perlu dikaji untuk melihat posisi dan kontribusi penelitian ini dalam peta keilmuan yang ada. Al Rasyid dan Ningsih [7] melakukan penelitian tentang penerapan algoritma TF-IDF dan Cosine Similarity untuk pencarian berbasis query pada dataset destinasi wisata. Penelitian tersebut menunjukkan bahwa kombinasi kedua metode ini mampu menghasilkan tingkat presisi yang cukup baik dalam menemukan destinasi wisata yang relevan berdasarkan kata kunci pencarian pengguna.

Oyadila, Abdullah, dan Razi [8] mengembangkan sistem rekomendasi destinasi wisata di Aceh Tengah menggunakan Content-Based Filtering dengan TF-IDF dan Cosine Similarity. Sistem yang dibangun memungkinkan pengguna memilih hingga lima atribut wisata seperti kuliner, piknik, atau hiking, dan menghasilkan rekomendasi berdasarkan skor kemiripan tertinggi. Penelitian ini membuktikan bahwa pendekatan tersebut dapat diimplementasikan secara efektif untuk rekomendasi wisata daerah.

Pratiwi, Asrianda, dan Rosnita [5] mengembangkan sistem rekomendasi objek wisata di Aceh Tamiang berbasis web dengan menganalisis ulasan dari Google Maps menggunakan Content-Based Filtering, TF-IDF, dan Cosine Similarity. Penelitian ini memperluas pendekatan konvensional dengan memanfaatkan data ulasan pengguna nyata sebagai sumber informasi atribut wisata.

Salsabilla dan Utomo [6] membangun sistem rekomendasi wisata Kabupaten Pekalongan dengan Content-Based Filtering dan membuktikan bahwa sistem tersebut mampu menghasilkan rekomendasi yang relevan dengan kategori wisata yang diminati pengguna. Sementara itu, Husna, Fajri, dan Pristyanto [3] mengembangkan sistem serupa untuk pariwisata Gunungkidul berbasis web dan membuktikan efektivitas Content-Based Filtering dalam konteks wisata daerah di Yogyakarta.

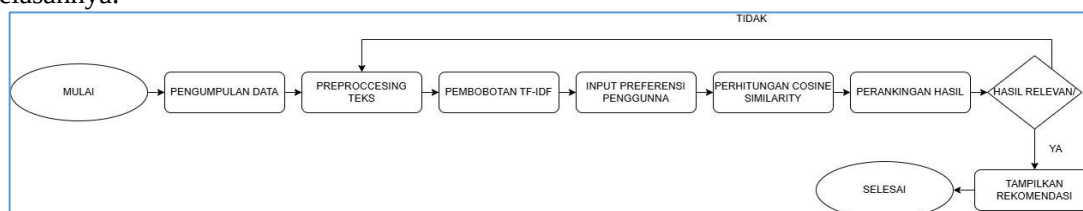
Dari kajian penelitian-penelitian tersebut, dapat disimpulkan bahwa pendekatan Content-Based Filtering berbasis TF-IDF dan Cosine Similarity telah terbukti efektif untuk sistem rekomendasi wisata [5], [6], [7], [8]. Namun, penelitian yang secara khusus berfokus pada desa wisata di Kabupaten Sleman

<http://sistemasi.ftik.unisi.ac.id>

masih belum banyak dilakukan. Penelitian ini hadir untuk mengisi celah tersebut dengan membangun sistem rekomendasi yang dirancang khusus sesuai dengan karakteristik dan keragaman desa wisata yang ada di Sleman.

3 Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode pengembangan sistem berbasis Content-Based Filtering. Secara garis besar, alur penelitian terbagi menjadi beberapa tahapan utama mulai dari pengumpulan data hingga pengujian sistem. Berikut adalah flowchart alur penelitian beserta penjelasannya.



Gambar 1 Alur penelitian

Langkah awal yang dilakukan dalam penelitian ini adalah menghimpun data seputar desa-desa wisata yang tersebar di Kabupaten Sleman. Dataset penelitian terdiri dari 20 desa wisata aktif di Kabupaten Sleman. Proses pengumpulan data dilakukan melalui beberapa jalur sekaligus, yakni dari laman resmi Dinas Pariwisata Kabupaten Sleman, portal desa wisata Yogyakarta, serta melalui observasi dan studi dokumentasi secara langsung. Setiap desa wisata dicatat dalam lima atribut, yaitu nama desa wisata, kecamatan, jenis wisata, daya tarik utama, dan fasilitas. Atribut jenis wisata, daya tarik utama, dan fasilitas digabungkan menjadi satu teks deskriptif sehingga setiap desa wisata menjadi satu dokumen, dengan total 20 dokumen.

Sebelum data teks bisa diolah lebih jauh, perlu ada tahap persiapan yang disebut preprocessing. Ini dilakukan karena data mentah dari berbagai sumber biasanya hadir dalam kondisi yang belum seragam dan masih bercampur dengan elemen-elemen yang tidak relevan, sehingga perlu dirapikan dan distandarisasi terlebih dahulu. Ada empat langkah preprocessing yang diterapkan dalam penelitian ini. Pertama, case folding seluruh teks diubah menjadi huruf kecil semua, supaya sistem tidak menganggap kata "Alam" dan "alam" sebagai dua hal berbeda. Kedua, tokenisasi teks dipotong-potong menjadi satuan kata individual yang disebut token. Ketiga, penghapusan stopwords kata-kata yang secara semantik tidak membawa makna berarti, seperti "dan", "yang", "di", dan "ke", disisihkan dari data. Keempat, stemming setiap kata berimbuhan dikembalikan ke bentuk dasarnya menggunakan algoritma Sastrawi, yang memang dikembangkan khusus untuk menangani bahasa Indonesia. Keluaran dari seluruh proses ini adalah kumpulan kata-kata bersih yang benar-benar mencerminkan isi dan karakter dari setiap desa wisata.

Setelah data teks selesai preprocessing, langkah berikutnya adalah mengonversi representasi teks tadi ke dalam bentuk angka menggunakan metode TF-IDF. Setiap kata yang masih tersisa setelah tahap preprocessing akan diberi bobot berdasarkan dua pertimbangan: pertama, seberapa kerap kata itu muncul dalam dokumen desa wisata tertentu yang dikenal sebagai Term Frequency; kedua, seberapa langka kata itu dijumpai di seluruh koleksi dokumen yang ada atau yang disebut Inverse Document Frequency. Nilai bobot akhir setiap kata diperoleh dari perkalian keduanya. Bobot-bobot ini kemudian disusun membentuk sebuah vektor numerik yang merepresentasikan masing-masing desa wisata dalam ruang berdimensi banyak. Dengan representasi berbentuk vektor inilah sistem baru bisa secara matematis membandingkan satu desa wisata dengan yang lainnya atau membandingkan preferensi pengguna dengan seluruh desa wisata yang ada dalam basis data.

Secara formal, perhitungan TF-IDF menggunakan rumus sebagai berikut:

$TF(t,d)$ = jumlah kemunculan kata t dalam dokumen d / total kata dalam dokumen d

$IDF(t) = \log(N / df(t))$

$TF-IDF(t,d) = TF(t,d) \times IDF(t)$

di mana N adalah total jumlah dokumen dan $df(t)$ adalah jumlah dokumen yang mengandung kata t .

Preferensi pengguna dimodelkan sebagai query teks pendek, tanpa akun maupun riwayat pengguna. Pengguna dapat mengetik kata kunci bebas (misalnya "wisata alam sungai" atau "kerajinan batik edukasi") atau memilih kategori yang disediakan sistem, yang diperlakukan sebagai kata kunci query. Query diproses dengan tahapan preprocessing yang sama seperti data desa wisata (case folding, tokenisasi, penghapusan stopword, dan stemming), lalu diubah menjadi vektor TF-IDF menggunakan kosakata dan nilai IDF dari 20 dokumen desa wisata. Karena vektor query berada pada ruang 125 dimensi yang sama dengan vektor desa wisata, kata pada query yang tidak terdapat dalam kosakata tidak memengaruhi skor kemiripan.

Tahap ini menjadi inti dari keseluruhan proses rekomendasi. Ketika preferensi pengguna dan data desa wisata sudah sama-sama berbentuk vektor TF-IDF, kemiripan di antara keduanya dihitung menggunakan Cosine Similarity. Metode ini mengukur seberapa besar sudut yang terbentuk antara dua vektor semakin sempit sudutnya, semakin besar nilai cosine yang dihasilkan, dan itu berarti kedua vektor semakin berdekatan secara semantik. Rentang nilai Cosine Similarity berada di antara 0 dan 1, di mana angka 1 menandakan kesesuaian penuh, sedangkan angka 0 berarti tidak ada irisan konten sama sekali.

Rumus Cosine Similarity yang digunakan adalah sebagai berikut:

$$\text{Cosine Similarity}(A, B) = (A \cdot B) / (\|A\| \times \|B\|)$$

di mana A adalah vektor preferensi pengguna, B adalah vektor desa wisata, $A \cdot B$ adalah hasil dot product kedua vektor, dan $\|A\|$ serta $\|B\|$ adalah panjang (magnitude) masing-masing vektor.

Setelah nilai Cosine Similarity diperoleh untuk seluruh desa wisata yang ada dalam database, sistem kemudian mengurutkan desa-desa wisata tersebut dari nilai kemiripan tertinggi hingga terendah. Desa wisata dengan nilai Cosine Similarity tertinggi berarti paling mirip dengan preferensi yang dimasukkan pengguna, dan itulah yang akan ditampilkan sebagai rekomendasi utama. Sistem akan menampilkan sejumlah desa wisata teratas sebagai hasil rekomendasi, yang biasa disebut Top-N Recommendation. Nilai N bisa disesuaikan sesuai kebutuhan, misalnya menampilkan 5 atau 10 desa wisata terbaik. Untuk mengevaluasi kualitas rekomendasi yang dihasilkan, penelitian ini menggunakan beberapa metrik evaluasi standar dalam sistem rekomendasi, yaitu Precision, Recall, dan F1-Score. Apabila hasil evaluasi belum memenuhi ambang batas yang ditetapkan, maka proses akan dikembalikan ke tahap preprocessing untuk dilakukan penyempurnaan lebih lanjut, misalnya dengan memperluas kamus stopword atau menyesuaikan parameter stemming.

Evaluasi dilakukan dengan tiga skenario query yang mewakili tiga tipe minat wisatawan, yaitu wisata alam dan petualangan, wisata budaya-kerajinan-edukasi, dan wisata kuliner tradisional. Untuk setiap query, ground truth berupa himpunan lima desa wisata relevan ditetapkan secara manual oleh peneliti berdasarkan kesesuaian atribut jenis wisata, daya tarik utama, dan fasilitas setiap desa terhadap query. Sistem menampilkan lima desa wisata teratas (Top-5), kemudian dihitung: $\text{Precision@5} = \text{jumlah rekomendasi relevan pada Top-5} / 5$; $\text{Recall@5} = \text{jumlah rekomendasi relevan pada Top-5} / \text{jumlah desa relevan pada ground truth}$; dan $\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$. Karena jumlah desa relevan pada ground truth sama dengan jumlah rekomendasi yang ditampilkan (lima), nilai Precision@5 dan Recall@5 selalu sama.

Tahap akhir dari alur penelitian ini adalah menampilkan hasil rekomendasi kepada pengguna dalam tampilan yang mudah dipahami. Setiap desa wisata yang direkomendasikan akan ditampilkan lengkap dengan informasi penting seperti nama desa, lokasi, jenis wisata, daya tarik, fasilitas, serta nilai kemiripan (similarity score) sebagai indikator seberapa cocok desa wisata tersebut dengan preferensi pengguna. Tampilan ini diharapkan dapat membantu wisatawan dalam mengambil keputusan dengan lebih mudah dan percaya diri.

4. Hasil dan Pembahasan

Penelitian ini menggunakan dataset yang terdiri dari 20 desa wisata aktif di Kabupaten Sleman. Data dikumpulkan dari website resmi Dinas Pariwisata Kabupaten Sleman tahun 2024, mencakup atribut nama desa wisata, kecamatan, daya tarik wisata, jenis wisata, dan fasilitas. Tabel 1 menyajikan sebagian data mentah sebelum dilakukan proses pra-pemrosesan.

Tabel 1 Dataset desa wisata sleman sebelum pra-pemrosesan

No	Nama Desa Wisata	Kecamatan	Jenis Wisata	Daya Tarik Utama
1	Desa Wisata Pentingsari	Cangkringan	Wisata Alam, Budaya, Alam Agrowisata	Merapi, lereng outbound, homestay
2	Desa Wisata Kelor	Banguntapan	Wisata Budaya, Agrowisata	Edukasi, Batik tulis, gerabah, kesenian Jawa
3	Desa Wisata Sambi	Ngaglik	Wisata Budaya, Agrowisata	Kuliner, Seni pertunjukan, kuliner tradisional
4	Desa Wisata Trumpon	Tempel	Agrowisata, Edukasi	Wisata Kebun salak, pengolahan salak pondoh,
5	Desa Wisata Brayut	Sleman	Wisata Edukasi, Kuliner	Budaya, Gamelan, tari tradisional, kuliner Jawa
6	Desa Wisata Pulesari	Turi	Wisata Alam, Agrowisata	Outbound, Sungai, camping, flying fox, outbound
7	Desa Wisata Tanjung	Mlati	Wisata Budaya, Agrowisata	Edukasi, Kerajinan bambu, gerabah, pertanian
8	Desa Wisata Krebet	Pajangan	Wisata Budaya, Belanja	Edukasi, Batik kayu, workshop, galeri seni

Berdasarkan Tabel 1 terlihat bahwa setiap desa wisata memiliki profil yang khas dan berbeda satu sama lain. Desa Wisata Pulesari menonjol pada aktivitas alam sungai dan outbound, sementara Desa Wisata Krebet memiliki identitas kuat pada kerajinan batik kayu. Perbedaan atribut inilah yang menjadi dasar representasi teks dalam sistem rekomendasi.

Kolom Daya Tarik Wisata, Jenis Wisata, dan Fasilitas, digabung menjadi satu teks deskriptif untuk masing-masing desa wisata. Teks gabungan ini selanjutnya melewati beberapa tahap pemrosesan: case folding, tokenisasi, penghapusan stopword berbahasa Indonesia, dan stemming dengan algoritma Sastrawi. Hasil preprocessing menghasilkan 125 kata fitur unik yang menjadi kosakata dasar model TFIDF, di antaranya: alam, batik, budaya, camping, gamelan, kerajinan, kuliner, outbound, salak, tenun, dan tracking.

Setiap desa wisata direpresentasikan dalam bentuk vektor TF-IDF berukuran 125 dimensi. Nilai TF-IDF yang tinggi menandakan kata tersebut sangat menonjol dan menjadi ciri khas desa wisata bersangkutan dibandingkan yang lain. Tabel 2 menyajikan tiga kata dengan bobot TF-IDF tertinggi untuk masing-masing desa wisata.

Tabel 2 Hasil pembobotan TF-IDF

No	Nama Desa Wisata	Kata ke-1	Kata ke-2	Kata ke-3
1	Desa Wisata Pentingsari	alam (0,340)	homestay (0,320)	aula (0,308)
2	Desa Wisata Kelor	batik (0,467)	kerajinan (0,402)	kesenian (0,321)
3	Desa Wisata Sambi	kuliner (0,378)	lapangan (0,363)	dapur (0,363)
4	Desa Wisata Trumpon	salak (0,568)	olah (0,478)	kebun (0,347)
5	Desa Wisata Brayut	gamelan (0,547)	budaya (0,344)	kuliner (0,284)
6	Desa Wisata Pulesari	camping (0,516)	outbound (0,395)	alam (0,324)

7	Desa Wisata Tanjung	kerajinan (0,436)	produk (0,348)	gerabah (0,306)
8	Desa Wisata Krebet	kayu (0,504)	seni (0,443)	batik (0,366)
9	Desa Wisata Nganggri	outbound (0,402)	alam (0,330)	gunung (0,299)
10	Desa Wisata Srowolan	pasar (0,444)	tradisional (0,417)	kuliner (0,394)
11	Desa Wisata Kaliurang	taman (0,462)	museum (0,406)	alam (0,254)
12	Desa Wisata Gamplong	tenun (0,615)	studio (0,410)	film (0,410)
13	Desa Wisata Turgu	tracking (0,424)	camping (0,424)	alam (0,266)
14	Desa Wisata Wonokerto	salak (0,575)	olah (0,483)	kebun (0,351)
15	Desa Wisata Lopati	bambu (0,700)	galeri (0,333)	workshop (0,333)
16	Desa Wisata Kembangarum	bunga (0,499)	keluarga (0,499)	kebun (0,363)
17	Desa Wisata Garongan	seafood (0,429)	pantai (0,429)	bahari (0,429)
18	Desa Wisata Jagalan	perak (0,480)	sejarah (0,349)	batik (0,349)
19	Desa Wisata Bokoharjo	candi (0,485)	sejarah (0,353)	souvenir (0,243)
20	Desa Wisata Sleman Kota	alun (0,705)	malam (0,353)	pasar (0,310)

Berdasarkan Tabel 2 terlihat bahwa setiap desa wisata memiliki kata pembeda yang unik. Desa Wisata Sleman Kota didominasi kata alun dengan bobot 0,705, Desa Wisata Lopati oleh kata bambu (0,700), dan Desa Wisata Gamplong oleh kata tenun (0,615). Sebaliknya, kata umum seperti alam dan budaya mendapat bobot lebih rendah karena muncul di banyak dokumen sekaligus, sehingga nilai IDF-nya kecil. Ini membuktikan TF-IDF berhasil memilah kata yang benar-benar informatif dan spesifik per desa wisata.

Hasil Perhitungan Cosine Similarity

Setelah seluruh desa wisata direpresentasikan sebagai vektor TF-IDF, dibangun matriks kemiripan antar dokumen menggunakan Cosine Similarity. Tabel 3 menyajikan penggalan matriks untuk lima desa wisata pertama sebagai ilustrasi.

Tabel 3 Penggalan matriks cosine similarity (5×5)

Desa Wisata	Pentingsari	Kelor	Sambi	Trumpon	Brayut
Pentingsari	1,0000	0,2686	0,3433	0,1109	0,2776
Kelor	0,2686	1,0000	0,2551	0,0854	0,1863
Sambi	0,3433	0,2551	1,0000	0,0966	0,4473
Trumpon	0,1109	0,0854	0,0966	1,0000	0,0684
Brayut	0,2776	0,1863	0,4473	0,0684	1,0000

Berdasarkan Tabel 3, pasangan Desa Wisata Sambu dan Brayut memiliki kemiripan tertinggi sebesar 0,4473 karena keduanya sama-sama menawarkan budaya Jawa dan kuliner tradisional. Sebaliknya, Desa Wisata Trumpon memiliki kemiripan sangat rendah dengan desa lain karena kontennya sangat spesifik pada kebun salak. Nilai rata-rata Cosine Similarity keseluruhan dataset adalah 0,1332, menunjukkan tingkat diversitas konten yang tinggi dan kondisi yang ideal bagi sistem rekomendasi untuk bekerja secara akurat.

Pengujian sistem dilakukan dengan tiga skenario query yang mewakili tipe preferensi wisatawan berbeda. Setiap query diproses melalui preprocessing dan TF-IDF yang sama, lalu dihitung nilai Cosine Similarity-nya terhadap seluruh 20 desa wisata dalam dataset.

Skenario 1 Query: "wisata alam sungai outbound camping"

Tabel 4 Hasil rekomendasi skenario 1

Rank	Nama Desa Wisata	Cosine Similarity
1	Desa Wisata Pulesari	0,7409
2	Desa Wisata Turgo	0,3422
3	Desa Wisata Nganggring	0,2980
4	Desa Wisata Wonokerto	0,2199
5	Desa Wisata Pentingsari	0,2153

Desa Wisata Pulesari muncul di posisi pertama dengan skor 0,7409 karena memang menawarkan outbound, aktivitas sungai, camping, dan flying fox. Desa Wisata Turgo di posisi kedua juga relevan dengan jalur tracking dan area camping di hutan pinus Merapi. Seluruh lima rekomendasi pada skenario ini dinilai relevan.

Skenario 2 Query: "batik kerajinan budaya edukasi workshop"

Tabel 5 Hasil rekomendasi skenario 2

Rank	Nama Desa Wisata	Cosine Similarity
1	Desa Wisata Kelor	0,6442
2	Desa Wisata Tanjung	0,5574
3	Desa Wisata Krebet	0,5401
4	Desa Wisata Jagalan	0,4818
5	Desa Wisata Lopati	,3167

Desa Wisata Kelor tampil sebagai rekomendasi utama dengan skor 0,6442 karena menonjolkan batik tulis, gerabah, dan workshop kesenian Jawa. Tanjung, Krebet, dan Jagalan di posisi berikutnya juga relevan sebagai sentra kerajinan bambu-gerabah, batik kayu, dan perak Kotagede. Seluruh lima rekomendasi dinyatakan relevan.

Skenario 3 Query: "kuliner tradisional pasar budaya Jawa"

Tabel 6 Hasil rekomendasi skenario 3

Rank	Nama Desa Wisata	Cosine Similarity
------	------------------	-------------------

1	Desa Wisata Srowolan	0,7322
2	Desa Wisata Sambi	0,4093
3	Desa Wisata Sleman Kota	0,3660
4	Desa Wisata Brayut	0,3434
5	Desa Wisata Pentingsari	0,2255

Desa Wisata Srowolan meraih posisi pertama dengan skor 0,7322 karena identitasnya sangat kuat pada pasar tradisional, kuliner asli Jawa, dan permainan rakyat. Sambi, Sleman Kota, dan Brayut di posisi berikutnya juga relevan. Satu catatan adalah Pentingsari di peringkat lima yang fokus utamanya lebih ke wisata alam, sehingga relevansinya terhadap query ini sedikit lebih lemah.

Evaluasi dilakukan menggunakan Precision@5, Recall@5, dan F1-Score dengan ground truth yang ditetapkan secara manual berdasarkan kesesuaian atribut desa wisata terhadap setiap query. Hasil evaluasi disajikan pada Tabel 7.

Tabel 7 Hasil evaluasi sistem rekomendasi

Query	Precision@5	Recall@5	F1-Score
Wisata alam sungai outbound camping	1,00 (100%)	1,00 (100%)	1,00
Batik kerajinan budaya edukasi workshop	1,00 (100%)	1,00 (100%)	1,00
Kuliner tradisional pasar budaya Jawa	0,80 (80%)	0,80 (80%)	0,80
Rata-rata (Mean)	0,933 (93,3%)	0,933 (93,3%)	0,933

Berdasarkan tabel 7, nilai rata-rata Precision@5, Recall@5, dan F1-Score sebesar 93,3% menunjukkan bahwa pendekatan ini mampu menempatkan sebagian besar desa wisata relevan pada lima peringkat teratas untuk ketiga skenario yang diuji. Sebagai pembandingan, dengan 20 desa wisata dan lima desa relevan per query, pemilihan lima desa secara acak hanya menghasilkan Precision@5 sekitar 0,25 (5/20), sehingga capaian 0,933 berada jauh di atas peluang acak. Nilai 0,80 pada Skenario 3 terjadi karena Desa Wisata Pentingsari berada di peringkat kelima meskipun fokus utamanya wisata alam. Hal ini kemungkinan disebabkan oleh kata "budaya" pada query yang juga muncul pada jenis wisata Pentingsari (alam, budaya, agrowisata), yang menunjukkan bahwa TF-IDF mencocokkan kata, bukan makna atau konteks.

Hasil tersebut perlu ditafsirkan dengan hati-hati. Pertama, metode ini bertumpu pada kecocokan kata (lexical matching), sehingga sinonim, kata di luar kosakata 125 fitur, maupun kata yang salah ketik tidak dapat dikenali. Skenario uji menggunakan kata kunci yang selaras dengan kosakata dataset (misalnya camping, outbound, dan batik), sehingga performa yang tinggi belum tentu berlaku untuk query berbahasa bebas. Kedua, kualitas rekomendasi bergantung pada kelengkapan dan kualitas atribut. Pada Tabel 2, kata seperti aula, lapangan, dan dapur muncul sebagai kata berbobot tertinggi pada beberapa desa, padahal kata tersebut berasal dari atribut fasilitas dan kurang mencerminkan minat wisata. Ketiga, seperti Content-Based Filtering pada umumnya, sistem cenderung hanya merekomendasikan desa dengan atribut yang mirip dengan query sehingga keragaman rekomendasi terbatas, dan belum mempertimbangkan faktor lokasi, biaya, maupun ulasan pengunjung.

Jumlah dataset yang kecil juga memengaruhi hasil. Dengan hanya 20 dokumen, nilai IDF dihitung dari koleksi yang sempit sehingga kata yang hanya muncul pada sedikit desa (misalnya alun, bambu, dan tenun) memperoleh bobot sangat tinggi dan dapat mendominasi skor kemiripan. Ketika jumlah desa wisata bertambah, estimasi IDF akan lebih stabil, tetapi persaingan antardesa dalam Top-5 juga lebih ketat sehingga nilai Precision@5 dapat berbeda dari hasil penelitian ini. Selain itu,

pengujian hanya menggunakan tiga query dengan ground truth yang ditetapkan manual oleh peneliti, sehingga skor 93,3% belum dapat digeneralisasi dan sebaiknya dipandang sebagai bukti awal kelayakan (proof of concept).

Dibandingkan pendekatan lain, Content-Based Filtering dipilih karena tidak memerlukan data interaksi pengguna. Collaborative Filtering membutuhkan riwayat atau rating dari banyak pengguna yang belum tersedia untuk desa wisata Sleman sehingga rentan terhadap masalah cold start, sedangkan pendekatan Hybrid dapat menutupi kelemahan kedua metode tetapi memerlukan data perilaku pengguna tambahan [4], [14]. Metode berbasis representasi semantik (misalnya word embedding atau model bahasa) berpotensi mengatasi keterbatasan pencocokan kata, namun membutuhkan sumber daya dan data yang lebih besar. Hasil penelitian ini sejalan dengan temuan Oyadila, Abdullah, dan Razi [8] bahwa kombinasi TF-IDF dan Cosine Similarity dapat bekerja baik untuk rekomendasi wisata, tetapi perbandingan angka secara langsung tidak dilakukan karena dataset, atribut, dan skenario pengujian berbeda. Perbandingan kuantitatif dengan baseline lain pada dataset yang sama menjadi pekerjaan lanjutan.

5. Kesimpulan

Penelitian ini membangun sistem rekomendasi desa wisata di Kabupaten Sleman menggunakan Content-Based Filtering yang dikombinasikan dengan TF-IDF dan Cosine Similarity. Dataset terdiri dari 20 desa wisata aktif dengan 125 kata fitur unik setelah preprocessing teks. Pembobotan TF-IDF mampu menonjolkan kata khas tiap desa wisata, seperti bambu pada Desa Wisata Lopati (0,700), alun pada Desa Wisata Sleman Kota (0,705), dan tenun pada Desa Wisata Gamplong (0,615). Pada tiga skenario query, sistem memperoleh rata-rata Precision@5, Recall@5, dan F1-Score sebesar 93,3%, yang menunjukkan bahwa pendekatan ini menjanjikan sebagai dasar sistem rekomendasi desa wisata pada skala data yang diuji. Namun, hasil tersebut masih bersifat awal karena hanya menggunakan 20 desa wisata, tiga query, dan ground truth yang ditetapkan manual oleh peneliti, serta metode yang bertumpu pada kecocokan kata sehingga belum menangkap makna dan sinonim. Penelitian selanjutnya dapat memperluas dataset dan jumlah query uji, melibatkan beberapa penilai atau pengujian langsung pada wisatawan, membandingkan dengan baseline atau metode lain (misalnya representasi semantik), mengintegrasikan pendekatan hybrid dengan Collaborative Filtering, serta menambahkan fitur lokasi seperti jarak geografis menggunakan formula Haversine.

Referensi

- [1] A. Riyanti dan A. C. Lesmana, "Pengembangan Daya Tarik Wisata di Kaliurang, Yogyakarta," J. Indones. Tour. Hosp. Recreat., Vol. 5, No. 1, 2022, DOI: 10.17509/jithor.v5i1.45008.
- [2] A. S. N. S. Ningrum, "Content Based dan Collaborative Filtering pada Rekomendasi Tujuan Pariwisata di Daerah Yogyakarta," Telematika, Vol. 16, No. 1, 2019, DOI: 10.31315/telematika.v16i1.3023.
- [3] S. O. Husna, I. N. Fajri, dan Y. Pristyanto, "Sistem Rekomendasi Pariwisata Gunungkidul berbasis Web menggunakan Metode Content-Based Filtering," JURSISTEKNI, Vol. 6, No. 3, 2024, DOI: 10.52005/jursistekni.v6i3.363.
- [4] R. Faurina dan E. Sitanggang, "Implementasi Metode Content-Based Filtering dan Collaborative Filtering pada Sistem Rekomendasi Wisata di Bali," Techno.Com, Vol. 22, No. 4, 2023, DOI: 10.33633/tc.v22i4.8556.
- [5] D. Pratiwi, Asrianda, dan L. Rosnita, "Penerapan Metode Content-based Filtering dalam Sistem Rekomendasi Objek Wisata di Aceh Tamiang," J. Ilmu Komput. dan Inform. (JIKI), Vol. 4, No. 2, 2024, DOI: 10.54082/jiki.169.
- [6] C. Salsabilla dan D. W. Utomo, "Sistem Rekomendasi Wisata Kabupaten Pekalongan dengan Content Based Filtering," SISTEMASI, Vol. 14, No. 1, 2025, DOI: 10.32520/stmsi.v14i1.4839.
- [7] R. Al Rasyid dan D. H. U. Ningsih, "Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian pada Dataset Destinasi Wisata," J. Teknol. Inf. dan Komun. (JTIK), Vol. 8, No. 1, 2024. [Online]. Tersedia: <https://www.researchgate.net/publication/379646614>

- [8] S. Oyadila, D. Abdullah, dan A. Razi, "Implementasi *Content-Based Filtering* dengan *TF-IDF* dan *Cosine Similarity* untuk Sistem Rekomendasi Destinasi Wisata di Aceh Tengah," *Rabit: J. Teknol. dan Sist. Inf. Univrab*, Vol. 10, No. 2, 2025.
- [9] Y. R. Ramadhan, Y. Heryadi, L. Lukas, dan W. Budiharto, "Implementasi *Content-Based Filtering* menggunakan *TF-IDF* dan *Cosine Similarity* untuk Sistem Rekomendasi Resep Masakan," *J. Teknol. Inf.*, Vol. 18, No. 1, 2024.
- [10] R. H. D. Reswara, Evanita, dan A. Susanto, "Implementasi *Content-Based Filtering* pada Sistem Rekomendasi Buku Perpustakaan," *JATI (J. Mhs. Tek. Inform.)*, Vol. 9, No. 2, 2025, DOI: 10.36040/jati.v9i2.13312.
- [11] A. A. Huda, R. Fajarudin, dan A. Hadinegoro, "Sistem Rekomendasi *Content-Based Filtering* menggunakan *TF-IDF Vector Similarity* untuk Rekomendasi Artikel Berita," *Build. Informatics, Technol. Sci. (BITS)*, Vol. 4, No. 3, 2022, DOI: 10.47065/bits.v4i3.2511.
- [12] S. M. Iqbal dan I. Pratama, "Penerapan Metode *Content based Filtering* pada Sistem Rekomendasi Pemilihan Produk Skincare," *Smart Comp: Jurnalnya Orang Pintar Komput.*, Vol. 13, No. 3, 2024, DOI: 10.30591/smartcomp.v13i3.7156.
- [13] L. H. Aljihadu dan A. N. Rohman, "Sistem Rekomendasi Wisata Kuliner di Gunungkidul menggunakan Metode *Content Based Filtering*," *J. Inform. dan Tek. Elektro Terapan*, Vol. 13, No. 1, 2025, DOI: 10.23960/jitet.v13i1.5955.
- [14] R. Al Fajar dan R. R. Andhika, "Perancangan Aplikasi Rekomendasi Pariwisata Belinyu berbasis Android menggunakan Metode *Content-Based Filtering* dengan *Firestore Cloud Computing*," *Reputasi: J. Rekayasa Perangkat Lunak*, Vol. 6, No. 2, 2025, DOI: 10.31294/reputasi.v6i2.11171.
- [15] S. I. Adam dan W. G. Mokodaser, "Implementasi Sistem Rekomendasi Produk *E-Commerce* menggunakan *Content-Based Filtering* berbasis *Cosine Similarity*," *Simtek: J. Sist. Inf. dan Tek. Komput.*, Vol. 10, No. 2, 2025, DOI: 10.51876/simtek.v10i2.1665.