SISTEM REKOMENDASI PARIWISATA DI DAERAH ISTIMEWA YOGYAKARTA (DIY) BERBASIS HYBRID FILTERING DAN ANALISIS SENTIMEN

(Studi Kasus : Data destinasi pariwisata di DIY)

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana Program Studi Statistika



Disusun Oleh:

NOVI ELVA WIJAYANTI

22611168

PROGRAM STUDI STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2026

HALAMAN PERSETUJUAN PEMBIMBING

TUGAS AKHIR

Judul	:	**SISTEM REKOMENDASI PARIWISATA DI DAERAH ISTIMEWA YOGYAKARTA (DIY) BERBASIS HYBRID FILTERING DAN ANALISIS SENTIMEN** *(Studi Kasus: Data Destinasi Pariwisata Yogyakarta)*
Nama Mahasiswa	:	Novi Elva Wijayanti
NIM	:	22611168

TUGAS AKHIR INI TELAH DIPERIKSA DAN DISETUJUI UNTUK DIUJIKAN

Mengetahui	Yogyakarta, 30 Juni 2026
Ketua Prodi Statistika	Dosen Pembimbing
(Dr. Atina Ahdika, S.Si., M.Si.)	(Dr. Raden Bagus Fajriya Hakim, S.Si., M.Si.)

HALAMAN PENGESAHAN

TUGAS AKHIR

SISTEM REKOMENDASI PARIWISATA DI DAERAH ISTIMEWA YOGYAKARTA (DIY) BERBASIS HYBRID FILTERING DAN ANALISIS SENTIMEN

(Studi Kasus : Data destinasi pariwisata di DIY)

Nama Mahasiswa : Novi Elva Wijayanti

NIM : 22611168

TUGAS AKHIR INI TELAH DIUJIKAN

PADA TANGGAL : 23 Juli 2026

Nama Penguji	Tanda Tangan
1. Mujjati Dwi Kartikasari, S.Si., M.Sc.	.....
2. Ghiffari Ahnaf Danarwindu, M.Sc.	.....
3. Dr. Raden Bagus Fajriya Hakim, S.Si., M.Si.	.....

Mengetahui,

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Islam Indonesia

apt. Saepudin, S.Si., M.Si., Ph.D.

KATA PENGANTAR

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

Assalamu'alaikum Wr.Wb

Alhamdulillahirabbil 'alamin, segala puja dan puji syukur atas kehadirat Allah *Subhanahu Wa Ta'ala* yang telah melimpahkan rahmat dan inayah-Nya, sehingga dapat menyusun Laporan Tugas Akhir yang berjudul "SISTEM REKOMENDASI PARIWISATA DI DAERAH ISTIMEWA YOGYAKARTA (DIY) BERBASIS HYBRID FILTERING DAN ANALISIS SENTIMEN" dengan baik dan tepat waktu.

Laporan ini disusun sebagai salah satu syarat untuk melaksanakan penelitian Tugas Akhir sebagai syarat kelulusan program Sarjana (S1) pada Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia. Dalam proses penulisan laporan Tugas Akhir ini terdapat banyak pihak yang memberi bimbingan, dukungan, bantuan, serta masukan baik secara langsung maupun tidak langsung dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis menyampaikan ucapan terima kasih sebesar-besarnya kepada:

1. Allah SWT yang telah memberikan ridho dan rahmat-Nya sehingga Tugas Akhir ini dapat terselesaikan.
2. Kedua orang tua (Ayah Marjoko dan Ibu Siti Yuliati), Nenek (Kasmijah dan Menek), serta saudara dan keluarga besar atas segala dukungan dan kasih sayang yang senantiasa selalu mendoakan untuk kebahagiaan dan keberhasilan penulis.
3. Ibu Dr. Atina Ahdika, S.Si., M.Si., selaku Ketua Prodi Statistika yang telah memberikan dukungan akademik selama masa studi penulis.
4. Bapak Dr. Raden Bagus Fajriya Hakim, S.Si., M.Si., selaku dosen pembimbing yang telah memberikan arahan, bimbingan, serta masukan yang sangat berarti bagi penulis dalam menyelesaikan laporan ini.
5. Seluruh Dosen dan Staf Program Studi Statistika, Universitas Islam Indonesia.

6. Sahabat seperjuangan dari maba hingga selamanya Felita Nur Safira, Ichda Nisaail Ai'zza, Putriana Dwi Agustin, dan Zaina Bilqis Afifah atas semua proses dari awal hingga akhir selama kuliah.
7. Sahabat, saudara organisasi, dan rekan dekat penulis, yang selalu memberikan semangat, bantuan, diskusi, serta dukungan selama proses pengerjaan laporan.
8. Seluruh pihak lain yang telah membantu secara langsung maupun tidak langsung dalam penyusunan Tugas Akhir ini, yang tidak dapat disebutkan satu per satu. Secara khusus, penulis ingin berterima kasih kepada Diena Mukafasyadiah atas bantuan dan kontribusinya dalam pembuatan sistem berbasis *web* yang digunakan dalam penelitian ini.
9. Terima kasih juga untuk diri saya sendiri yang telah berusaha keras, berjuang, dan tidak menyerah, sehingga pada akhirnya mampu menyelesaikan tugas akhir ini dengan sebaik dan semaksimal mungkin.

Semoga amal kebaikan yang telah diberikan akan mendapat balasan dari Allah *Subhanahu Wa Ta'ala*. Dalam penulisan laporan Tugas Akhir ini, penulis sadar masih jauh dari kata sempurna dikarenakan masih terdapat banyak kesalahan. Sehingga, penulis sangat mengharapkan segala macam kritik dan saran yang membangun sebagai bahan evaluasi untuk penulis laporan selanjutnya, agar lebih baik lagi dan bermanfaat bagi pembaca dan penulis sendiri.

Wabillahitaufiq walhidayah,

Wassalamu'alaikum Warahmatullahi Wabarakatuh.

Yogyakarta, 30 Juni 2026



Novi Elva Wijayanti

DAFTAR ISI

HALAMAN JUDUL i
HALAMAN PERSETUJUAN PEMBIMBING TUGAS AKHIR ii
HALAMAN PENGESAHAN TUGAS AKHIR iii
KATA PENGANTAR iv
DAFTAR ISI vi
DAFTAR TABEL viii
DAFTAR GAMBAR ix
DAFTAR LAMPIRAN xi
PERNYATAAN xii
INTISARI xiii
ABSTRACT xiv
BAB I PENDAHULUAN 1
1.1. Latar Belakang Masalah 1
1.2. Rumusan Masalah 5
1.3. Batasan Masalah 5
1.4. Tujuan Penelitian 6
1.5. Manfaat Penelitian 6
BAB II TINJAUAN PUSTAKA 8
2.1 Penelitian Terdahulu 8
BAB III LANDASAN TEORI 24
3.1. Sistem Rekomendasi 24
3.2. Content-Based Filtering 24
3.2.1 Representasi Fitur Destinasi Wisata 25
3.2.2 Pengukuran Kemiripan (Similarity Measure) 27
3.3. Rule-Based Filtering 31
3.3.1 Pembentukan Aturan (*Rule Definition*) 32
3.3.2 Proses Penyaringan (*Filtering Process*) 33
3.3.3 Kelebihan dan Kekurangan Rule-Based Filtering 34
3.4. Hybrid Filtering 34
3.5. Analisis Sentimen Ulasan Pengunjung 36
3.5.1 Pra-Pemrosesan Teks (*Preprocessing Text*) 37
3.5.2 Ekstraksi Fitur Teks menggunakan TF-IDF 38
3.5.3 Ketidakseimbangan Data (*Imbalanced Dataset*) 39
3.5.4 Metode Klasifikasi Sentimen Support Vector Machine (SVM) 41
3.5.5 Metode Klasifikasi Sentimen Multinomial Naïve Bayes (MNB) 46
3.6. Keterkaitan Antar Metode 51
BAB IV METODOLOGI PENELITIAN 53
4.1. Populasi dan Sampel Penelitian 53
4.2. Tempat dan Waktu Penelitian 53
4.3. Definisi Operasional Variabel Penelitian 54

4.4. Alat dan cara organisir data 57
4.5. Metode Penelitian 58
BAB V HASIL DAN PEMBAHASAN 60
5.1 Pengumpulan Dataset 60
5.1.1 Pengumpulan Data Destinasi Wisata 60
5.1.2 Pengumpulan Data Ulasan Pengguna 62
5.1.3 Proses Scraping Data 63
5.2 *Exploratory Data Analysis* (EDA) 65
5.3 Persiapan *Dataset* 67
5.4 Analisis Sentimen 69
5.4.1 *Preprocessing* Data 69
5.4.2 Pelabelan Sentimen Menggunakan InSet Lexicon 70
5.4.3 Ekstraksi Fitur TF-IDF 74
5.4.4 Pemilihan Kernel SVM (Linier dan RBF) 77
5.4.5 Evaluasi Model Multinomial Naive Bayes (MNB) 81
5.4.6 Perbandingan Model dan Pemilihan Model Terbaik 83
5.4.7 Prediksi Seluruh Dataset dan Agregasi Skor Destinasi 85
5.5 Analisis Sistem Rekomendasi Hybrid Filtering 86
5.5.1 Rule Based Filtering 87
5.5.2 Content Based Filtering 87
5.5.3 Hybrid Filtering Score 89
5.6 Implementasi Sistem Rekomendasi 92
BAB VI PENUTUP 98
6.1 Kesimpulan 98
6.2 Saran 99
DAFTAR PUSTAKA 100
LAMPIRAN 103

DAFTAR TABEL

Tabel 2.1 Tabel Penelitian Sebelumnya 17
Tabel 3.1 Preferensi Pengguna 25
Tabel 3.2 Deskripsi Destinasi 26
Tabel 3.3 Kemunculan Destinasi 26
Tabel 3.4 Kata Destinasi 27
Tabel 3.5 Contoh TF-IDF Content-Based 27
Tabel 3.6 *Matriks Similarity* 30
Tabel 3.7 Kemiripan Destinasi 31
Tabel 3.8 Preferensi Rule-Based 33
Tabel 3.9 Hasil Destinasi Rule-Based 33
Tabel 3.10 Contoh Rule *Score* 36
Tabel 3.11 Contoh Similarity *Score* 36
Tabel 3.12 Contoh Hybrid *Score* 36
Tabel 3.13 Contoh TF-IDF Sentimen 39
Tabel 3.14 Contoh Sederhana SVM RBF 44
Tabel 3.15 Contoh Sederhana MNB 48
Tabel 3.16 Probabilitas Kata 49
Tabel 3.17 Confusion Matrix MNB 50
Tabel 4.1 Definisi Operasional variabel 54

DAFTAR GAMBAR

Gambar 3.1 Diagram Alir Rule-Based Filtering (Pre-*Filtering*)..... 34
Gambar 3.2 Diagram Alir Hybrid Filtering..... 35
Gambar 3.3 Diagram Alir Analisis Sentimen 37
Gambar 3.4 Diagram Alir Pra-Pemrosesan Teks 37
Gambar 4.1 Diagram Alir Penelitian.....59
Gambar 5.1 Database Inventarisasi DTW DIY 2025.....61
Gambar 5.2 Dataset Inventarisasi..... 62
Gambar 5.3 Data Link Google Maps 62
Gambar 5.4 Data Hasil Scraping..... 63
Gambar 5.5 *Website* Scraping 64
Gambar 5.6 Salah Satu Platform Digital Scraping..... 65
Gambar 5.7 Distribusi Rating Ulasan..... 66
Gambar 5.8 Distribusi Jumlah Kata per Ulasan 67
Gambar 5.9 Jumlah Destinasi 100 ulasan 67
Gambar 5.10 Statistik Awal *Dataset*..... 68
Gambar 5.11 Data *Cleaning*..... 68
Gambar 5.12 Filter Bahasa Indonesia 69
Gambar 5.13 *Preprocessing Text* 69
Gambar 5.14 Audit Lexicon..... 70
Gambar 5.15 Visualisasi Sentimen 71
Gambar 5.16 Word Cloud Positif..... 72
Gambar 5.17 Word Cloud Netral 73
Gambar 5.18 Word Cloud Negatif 74
Gambar 5.19 TF IDF Sentimen..... 75
Gambar 5.20 Visualisasi TF IDF 75
Gambar 5.21 *Training Testing* 76
Gambar 5.22 *Training Testing* RandomUnderSampler 77
Gambar 5.23 Analisis SVM Perbandingan Kernel 78
Gambar 5.24 Perbandingan Kernel 78
Gambar 5.25 Visualisasi Linier vs RBF..... 79

Gambar 5.26 Cross Validation 79
Gambar 5.27 Kombinasi Parameter SVM RBF 80
Gambar 5.28 Classification Report SVM RBF 80
Gambar 5.29 Confusion Matrix SVM RBF 81
Gambar 5.30 Kombinasi Parameter MNB 82
Gambar 5.31 Cross Validation MNB 82
Gambar 5.32 Classification Report MNB 83
Gambar 5.33 Confusion Matrix MNB 83
Gambar 5.34 *Comparison Summary* Analisis Sentimen 84
Gambar 5.35 Visualisasi Perbandingan SVM RBF vs MNB 84
Gambar 5.36 *Best* Model Analisis Sentimen 85
Gambar 5.37 Proporsi Analisis Sentimen 85
Gambar 5.38 Data Analisis Skor Sentimen 86
Gambar 5.39 TF IDF Sistem Rekomendasi 88
Gambar 5.40 Perbandingan Top-K dan Korelasi Spearman 89
Gambar 5.41 Visualisasi Keseluruhan Hybrid *Score* 90
Gambar 5.42 Ablation Study dan Top Destinasi 91
Gambar 5.43 *Final* Hybrid *Score* Rekomendasi 92
Gambar 5.44 Tampilan *Website* Sistem Rekomendasi Bagian Atas 93
Gambar 5.45 Pilihan *Website* Sistem Rekomendasi 94
Gambar 5.46 Hasil Sistem Rekomendasi 96

DAFTAR LAMPIRAN

Lampiran 1 103
Lampiran 2 104
Lampiran 3 105
Lampiran 4 106

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Tugas Akhir ini tidak terdapat karya karya yang sebelumnya pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan orang lain, kecuali yang diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 03 September 2026



Novi Elva Wijayanti

INTISARI

SISTEM REKOMENDASI PARIWISATA DI DAERAH ISTIMEWA YOGYAKARTA (DIY) BERBASIS HYBRID FILTERING DAN ANALISIS SENTIMEN

(Studi Kasus : Memanfaatkan data destinasi pariwisata di DIY)

Novi Elva Wijayanti
Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Islam Indonesia

Daerah Istimewa Yogyakarta (DIY) merupakan salah satu destinasi wisata unggulan di Indonesia dengan karakteristik objek wisata yang beragam. Namun, banyaknya pilihan destinasi dan informasi yang tersebar menyebabkan wisatawan mengalami kesulitan dalam menentukan objek wisata yang sesuai dengan preferensi dan kebutuhan mereka. Oleh karena itu, diperlukan suatu sistem rekomendasi yang mampu memberikan rekomendasi destinasi wisata secara personal dan informatif. Penelitian ini bertujuan untuk merancang dan mengembangkan sistem rekomendasi objek wisata dengan mengombinasikan pendekatan Content-Based Filtering dan Rule-Based Filtering. Data penelitian direncanakan berasal dari ulasan dan penilaian wisatawan yang diperoleh dari platform digital, yang menggambarkan pengalaman pengunjung secara langsung terhadap objek wisata di DIY. Metode Content-Based Filtering digunakan untuk mengukur kemiripan antar objek wisata berdasarkan deskripsi dan ulasan menggunakan representasi fitur Term Frequency–Inverse Document Frequency (TF-IDF) dan Cosine Similarity. Sementara, Rule-Based Filtering diterapkan untuk menyaring rekomendasi berdasarkan aturan preferensi pengguna, seperti kategori wisata dan fasilitas yang diinginkan. Selain itu, analisis sentimen terhadap ulasan wisatawan dilakukan menggunakan metode Support Vector Machine (SVM) vs Multinomial Naive Bayes (MNB) untuk mengidentifikasi kecenderungan opini pengguna. Penelitian ini diharapkan mampu menghasilkan sistem rekomendasi objek wisata yang tidak hanya mempertimbangkan kesesuaian konten, tetapi juga kualitas fasilitas yang dimiliki, sehingga dapat membantu wisatawan dalam pengambilan keputusan serta mendukung pengembangan pariwisata di Daerah Istimewa Yogyakarta.

Kata Kunci : Sistem Rekomendasi, Content-Based Filtering, Rule-Based Filtering, Support Vector Machine, Analisis Sentimen.

ABSTRACT

Tourism Recommendation System in the Special Region of Yogyakarta (DIY) Based on Hybrid Filtering and Sentiment Analysis
(Case Study : Utilizing Tourism Destination Data in the Special Region of Yogyakarta)

Novi Elva Wijayanti
Department of Statistics, Faculty of Matematics and Natural Sciences
Universitas Islam Indonesia

The Special Region of Yogyakarta (DIY) is one of Indonesia's leading tourist destinations, characterized by a wide variety of tourist attractions. However, the large number of available destinations and the dispersion of tourism information often make it difficult for tourists to identify attractions that align with their preferences and needs. Therefore, a recommendation system is required to provide personalized and informative tourism recommendations. This study aims to design and develop a tourist attraction recommendation system by integrating Content-Based Filtering and Rule-Based Filtering approaches. The research data are planned to be collected from tourists' reviews and ratings obtained from digital platforms, which reflect visitors' direct experiences with tourist attractions in DIY. The Content-Based Filtering method is employed to measure the similarity among tourist attractions based on textual descriptions and reviews using Term Frequency–Inverse Document Frequency (TF-IDF) and Cosine Similarity. Meanwhile, Rule-Based Filtering is applied to filter recommendations according to predefined user preference rules, such as tourism categories and desired facilities. In addition, sentiment analysis of tourists' reviews is conducted using the Support Vector Machine (SVM) vs Multinomial Naive Bayes (MNB) method to identify users' opinion tendencies. This study is expected to produce a tourist attraction recommendation system that not only considers content relevance but also the quality of available facilities, thereby supporting tourists' decision-making processes and contributing to tourism development in the Special Region of Yogyakarta.

Keywords: *Recommendation System, Content-Based Filtering, Rule-Based Filtering, Support Vector Machine, Analisis Sentimen.*

BAB I
PENDAHULUAN

1.1. Latar Belakang Masalah

Indonesia merupakan negara kepulauan yang memiliki kekayaan sumber daya alam melimpah, baik dalam sektor energi, pertambangan, maupun pariwisata. Keindahan alam, keberagaman budaya, dan warisan sejarah yang melimpah diberbagai wilayah Indonesia menjadikannya salah satu tujuan wisata yang diminati oleh wisatawan domestik maupun mancanegara.

Setiap tahun, jutaan wisatawan asing memadati Indonesia untuk menikmati beragam destinasi wisata yang tersebar di seluruh wilayah nusantara, hal tersebut menunjukkan tren peningkatan yang cukup signifikan. Berdasarkan data resmi sebelum COVID-19, jumlah kunjungan wisatawan mancanegara ke Indonesia menunjukkan tren peningkatan yaitu tahun 2018 sebanyak 15,81 juta kunjungan, meningkat sebesar 12,58% dibandingkan tahun 2017 yang mencapai 14,04 juta kunjungan (BPS, 2019). Meskipun sempat mengalami penurunan drastis akibat pandemi, sektor pariwisata berhasil pulih dengan meningkatnya jumlah kunjungan wisatawan mancanegara dari 1,56 juta kunjungan pada tahun 2021 menjadi 13,89 juta kunjungan pada tahun 2024 dan pada tahun 2025 kembali meningkat menjadi 15,39 juta kunjungan melebihi target (BPS, 2025).

Pemerintah Indonesia secara konsisten mengimplementasikan berbagai kebijakan serta program strategis, guna mengoptimalkan peran sektor pariwisata sebagai salah satu pilar penggerak ekonomi nasional. Pengembangan tersebut tidak hanya memperkenalkan destinasi yang populer di tingkat nasional, tetapi menjangkau hingga tingkat daerah agar potensi wisata lokal dapat ikut berkembang. Melalui kolaborasi antara pemerintah pusat dan daerah, promosi gigih dilakukan baik secara daring maupun kegiatan pameran international, yang bertujuan untuk memperkenalkan budaya ke pasar global.

Namun, efektivitas promosi tersebut kini sangat dipengaruhi oleh perkembangan teknologi informasi yang mengubah cara wisatawan dalam merencanakan perjalanan. Wisatawan cenderung meninggalkan metode

konvensional dan lebih mengandalkan platform digital, media sosial, dan ulasan pengunjung untuk menentukan destinasi yang akan dipilih. Perubahan perilaku tersebut menciptakan kebutuhan adanya sistem rekomendasi wisata yang mampu menyajikan informasi secara personal berbasis data.

Kondisi tersebut menjadi sangat relevan bagi Daerah Istimewa Yogyakarta (DIY), sebagai salah satu wilayah yang konsisten mengembangkan sektor pariwisata, yang dituntut beradaptasi dengan tren digitalisasi guna mempertahankan daya saing. Yogyakarta dikenal sebagai daerah yang memiliki potensi wisata sangat beragam, yang mencakup sektor alam, seni dan budaya, sejarah, hingga destinasi buatan yang dikembangkan secara inovatif. Keberagaman tersebut memperkuat posisi Yogyakarta sebagai salah satu destinasi utama di Indonesia. Selain itu, daerah ini juga memiliki ekosistem seni dan budaya yang aktif dengan terselenggaranya diberbagai event wisata, pertumbuhan ekomoni kreatif di tengah masyarakat, serta potensi bisnis dan investasi diberbagai usaha yang saling berkaitan dan menguntungkan (Wicaksono, 2020).

Disisi lain, Dinas Pariwisata Daerah Istimewa Yogyakarta (DIY) memiliki data pariwisata yang cukup lengkap, meliputi informasi mengenasi destinasi wisata, fasilitas pendukung, aksesibilitas, kapasitas, dan data kunjungan wisata. Akan tetapi, untuk data pendukung lainnya seperti rating, jam operasional, ulasan dan komentar menjadi sumber informasi penting dapat dilihat pada platform digital dan Google Maps. Sehingga, seringkali menimbulkan tantangan bagi wisatawan dalam menentukan pilihan yang sesuai dengan preferensi, kebutuhan, dan kondisi. Dikarenakan data-data tersebut masih tersebar dan belum dimanfaatkan secara optimal dalam bentuk sistem yang terintegrasi.

Berdasarkan kunjungan wisatawan ke Yogyakarta juga menunjukkan dinamika pertumbuhan yang signifikan. Data tersebut menunjukkan, bahwa sebelum terjadi pandemi yaitu tahun 2018 jumlah kunjungan wisatawan mancanegara sekitar 400 ribu, sedangkan wisatawan nusantara mencapai sekitar 5 juta (Tim Tugu Jogja, 2019). Data sebelum pandemi menunjukkan kondisi normal pada sektor pariwisata, sehingga digunakan untuk membandingkan proses pemulihan dari dampak pandemi pada tahun berikutnya. Data pada tahun 2024, terdapat kunjungan mancanegara sebanyak 236.463 jiwa dengan kunjungan

nusantara sebanyak 7,9 juta. Sementara pada tahun 2025, perbaruan data hingga bulan Oktober 2025 tercatat sebanyak 226.057 jiwa mancanegara dan 6,3 juta jiwa nusantara. Perubahan jumlah kunjungan tersebut menggambarkan besarnya volume pergerakan wisatawan, sehingga meningkatkan jumlah dan keragaman data pariwisata yang dihasilkan setiap tahunnya (BAPPERIDA DIY, 2025).

Sementara itu, terdapat fenomena terbaru di media sosial dan berbagai platform berita yang menunjukkan lonjakan minat yang cukup signifikan terhadap Daerah Istimewa Yogyakarta (DIY) menjelang libur natal 2025 hingga tahun baru 2026. Yogyakarta menjadi tujuan utama yang mengungguli destinasi populer kota lainnya, dengan peningkatan pencarian akomodasi sekitar 3,86 juta orang menggunakan kendaraan pribadi dan 3,16 juta orang menggunakan kendaraan umum (KOMPAS, 2025). Prediksi lonjakan pengunjung di DIY diperkirakan mencapai 9,38 juta orang dan kawasan Kota Yogyakarta mencapai 7 juta orang, sehingga volume lalu lintas rata-rata mencapai 20 persen daripada harian rata-rata selama libur Nataru 2025/2026 (detikNews, 2025).

Berdasarkan kondisi tersebut, diperlukan suatu sistem rekomendasi wisata yang mampu mengintegrasikan data destinasi, data fasilitas, data operasional, data ekonomi dan kunjungan wisata, serta analisis sentimen dari komentar pengunjung. Sistem rekomendasi tersebut diharapkan dapat membantu wisatawan dalam menentukan destinasi wisata yang sesuai preferensi yang diinginkan, serta dapat mendukung pengembangan pariwisata berbasis data di Daerah Istimewa Yogyakarta (DIY).

Penelitian sebelumnya yang telah diteliti oleh Husna, dkk. (2024) yaitu mengembangkan sistem rekomendasi pada sektor pariwisata menggunakan metode Content-Based Filtering di Kabupaten Gunung Kidul, DIY. Penelitian tersebut, mengembangkan sistem rekomendasi destinasi wisata sebagai dasar pemberian rekomendasi untuk pengguna. Pendekatan yang digunakan berfokus pada pengukuran kemiripan antar objek wisata atau preferensi pengguna berdasarkan data terstruktur. Walaupun dapat memberikan rekomendasi destinasi yang sesuai dengan karakteristik tertentu, tetapi penelitian tersebut masih memiliki keterbatasan dalam pemanfaatan sumber data, terutama terkait pemanfaatan data tidak terstruktur, seperti data komentar dan ulasan pengunjung yang tersedia di platform

digital. Sehingga, rekomendasi yang dihasilkan kurang akurat karena belum mempertimbangkan aspek persepsi dan pengalaman lansung dari pengunjung (Husna et al., 2024).

Berdasarkan penelitian yang pernah dilakukan sebelumnya, peneliti tertarik untuk melakukan penelitian mengenai sistem rekomendasi pariwisata di Daerah Istimewa Yogyakarta (DIY) yang dapat mengintegrasikan data destinasi wisata dari Dinas Pariwisata DIY, data fasilitas dan operasional, dan data tidak terstruktur yaitu komentar dan ulasan wisatawan dari platform digital. Oleh karena itu, peneliti mengusulkan sistem rekomendasi pariwisata di Daerah Istimewa Yogyakarta (DIY) berbasis Hybrid Filtering yang mengombinasikan Content-Based Filtering dan Rule-Based Filtering, serta analisis sentimen terhadap komentar wisatawan dari media digital Google Maps. Pendekatan tersebut diharapkan mampu menghasilkan rekomendasi destinasi wisata yang lebih relevan, kontekstual, dan sesuai dengan preferensi wisatawan, sekaligus mendukung pengembangan pariwisata berbasis data di wilayah Daerah Istimewa Yogyakarta (DIY).

Pada penelitian ini, analisis sentimen digunakan untuk mengubah komentar wisatawan menjadi informasi mengenai kecenderungan opini positif, negatif, dan netral yang selanjutnya diintegrasikan ke dalam sistem rekomendasi. Komentar wisatawan merupakan data teks tidak terstruktur yang terlebih dahulu direpresentasikan menjadi fitur numerik menggunakan Term Frequency-Inverse Document Frequency (TF-IDF), sehingga menghasilkan karakteristik data berdimensi tinggi dan *sparse*. Oleh karena itu, diperlukan metode klasifikasi yang sesuai dengan karakteristik data teks tersebut.

Metode yang dipilih dalam penelitian ini adalah *Support Vector Machine* (SVM) dan *Multinomial Naive Bayes* (MNB) karena keduanya memiliki pendekatan yang berbeda dalam klasifikasi teks. SVM bekerja dengan membentuk *decision boundary* optimal untuk memisahkan kelas pada data berdimensi tinggi, sedangkan MNB menggunakan pendekatan probabilistik berdasarkan distribusi kemunculan kata pada setiap kelas. Penelitian Martanto dan Istiono (2024) menunjukkan bahwa MNB mampu mengklasifikasikan sentimen ulasan menggunakan ekstraksi fitur TF-IDF. Sementara itu, penelitian Sinaga dkk. (2026) menunjukkan bahwa kombinasi TF-IDF dan SVM dapat digunakan secara efektif

untuk klasifikasi sentimen komentar YouTube, dengan kernel *linear* menghasilkan akurasi terbaik sebesar 74,95%. Hasil tersebut menunjukkan bahwa SVM memiliki kemampuan yang baik dalam menangani klasifikasi data teks berbasis TF-IDF, meskipun performanya dapat berbeda berdasarkan karakteristik data dan kernel yang digunakan. Berdasarkan perbedaan pendekatan serta hasil penelitian terdahulu tersebut, penelitian ini membandingkan SVM dan MNB untuk memperoleh model klasifikasi sentimen dengan performa terbaik. Model terpilih selanjutnya digunakan untuk menghasilkan informasi sentimen wisatawan yang diintegrasikan ke dalam sistem rekomendasi pariwisata di Daerah Istimewa Yogyakarta (DIY).

1.2. Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah dalam penelitian ini adalah:

1. Bagaimana memperoleh dan mengintegrasikan data destinasi wisata, fasilitas pendukung, dan informasi operasional wisata di Daerah Istimewa Yogyakarta (DIY) dari berbagai sumber berbasis *web* untuk membangun *database* yang mendukung sistem rekomendasi wisata berbasis Hybrid Filtering?
2. Bagaimana melakukan analisis sentimen terhadap komentar wisatawan yang diperoleh dari platform digital Google Maps menggunakan algoritma Support Vector Machine (SVM) dan Multinomial Naive Bayes (MNB) untuk memperoleh model klasifikasi dengan kinerja terbaik?
3. Bagaimana membangun sistem rekomendasi wisata berbasis Hybrid Filtering yang menggabungkan data terstruktur dan data tidak terstruktur untuk menghasilkan rekomendasi wisata yang relevan sesuai preferensi pengguna?

1.3. Batasan Masalah

Berdasarkan latar belakang di atas, rumusan masalah dalam penelitian ini adalah:

1. Penelitian hanya berfokus pada destinasi wisata yang berada di Daerah Istimewa Yogyakarta (DIY).

2. Data destinasi dan fasilitas wisata diperoleh dari Dinas Pariwisata DIY, sedangkan data variabel pendukung lainnya dan data komentar diperoleh melalui *scraping* platform digital.
3. Analisis sentimen hanya dilakukan pada komentar wisatawan dari platform digital Google Maps.
4. Model rekomendasi terbatasi pada pendekatan Content-Based Filtering, Rule-Based Filtering, dan analisis sentimen.
5. Sistem rekomendasi tidak mencakup rencana perjalanan wisata dalam rentang waktu hitungan hari (*itinerary planning*).
6. Penelitian ini tidak mengevaluasi perilaku pengguna secara langsung melalui *user study*.

1.4. Tujuan Penelitian

Penelitian ini bertujuan untuk:

1. Memperoleh dan mengintegrasikan data destinasi wisata, fasilitas pendukung, dan informasi operasional wisata di DIY dari berbagai sumber berbasis *web* untuk membangun *database* yang mendukung sistem rekomendasi berbasis Hybrid Filtering.
2. Melakukan analisis sentimen terhadap komentar wisatawan yang diperoleh dari platform digital Google Maps menggunakan algoritma Support Vector Machine (SVM) dan Multinomial Naive Bayes (MNB), serta menentukan model klasifikasi dengan performa terbaik sebagai dasar penilaian persepsi pengunjung terhadap destinasi wisata.
3. Membangun sistem rekomendasi wisata berbasis Hybrid Filtering yang menggabungkan data terstruktur dan tidak terstruktur untuk menghasilkan rekomendasi wisata yang relevan sesuai preferensi pengguna.

1.5. Manfaat Penelitian

Penelitian diharapkan dapat memberikan manfaat sebagai berikut:

1. Memberikan rekomendasi destinasi wisata yang relevan bagi wisatawan di Daerah Istimewa Yogyakarta (DIY), berdasarkan pengolahan data destinasi wisata, fasilitas pendukung, informasi operasional.

2. Memanfaatkan komentar wisatawan dari platform digital Google Maps melalui analisis sentimen untuk menggambarkan persepsi dan pengalaman pengunjung terhadap destinasi wisata.
3. Menghasilkan sistem rekomendasi wisata berbasis Hybrid Filtering yang menggabungkan pendekatan Content-Based Filtering, Rule-Based Filtering, dan analisis sentimen.
4. Membantu wisatawan dalam menentukan destinasi wisata yang sesuai dengan preferensi dan kebutuhan mereka berdasarkan informasi yang lebih komprehensif dan berbasis data.
5. Menyediakan informasi pendukung bagi pemerintah daerah dan pengelola destinasi wisata dalam memahami karakteristik destinasi serta persepsi wisatawan sebagai bahan evaluasi dan pengembangan pariwisata di Daerah Istimewa Yogyakarta (DIY).

BAB II
TINJAUAN PUSTAKA

Bab ini memaparkan tentang penelitian-penelitian terdahulu yang relevan dengan topik penelitian, khususnya yang berkaitan dengan sistem rekomendasi pariwisata dan metode yang digunakan. Tinjauan pustaka disusun untuk memberikan gambaran posisi penelitian yang akan dilakukan dibandingkan dengan penelitian sebelumnya, serta untuk menunjukkan kebaruan atau novelty dari penelitian ini.

2.1 Penelitian Terdahulu

Penelitian terdahulu sangat penting untuk mengidentifikasi perkembangan metode, pendekatan, serta keterbatasan yang masih terdapat pada penelitian sebelumnya. Kajian literatur ini bertujuan untuk memperoleh gambaran mengenai sistem rekomendasi yang menggunakan metode content-based filtering, rule-based filtering, maupun metode lainnya yang berkaitan dan sejalan. Dengan demikian, penelitian yang dilakukan dapat melengkapi atau memperbaiki kekurangan tersebut. Terdapat beberapa jurnal dan penelitian sebelumnya yang dijadikan acuan dan disajikan sebagai berikut.

Mondi dan Wijayanto. (2019) mengembangkan sistem rekomendasi wisata kuliner pada aplikasi MANGAN di Surakarta, Jawa Tengah, menggunakan metode content-based filtering dengan pendekatan cosine similarity. Fitur konten restoran yang digunakan meliputi nama menu, rasa, fasilitas, kategori, dan lokasi yang diperoleh dari basis data internal aplikasi. Hasil penelitian menunjukkan bahwa sistem mampu memberikan rekomendasi restoran dengan tingkat presisi rata-rata sebesar 89,15%. Namun, tingkat akurasi yang dihasilkan relatif rendah, yaitu sebesar 51,18%, yang disebabkan oleh keterbatasan jumlah data dan metode evaluasi yang digunakan. Selain itu, sistem belum memanfaatkan ulasan bebas dari pengguna dan hanya mengandalkan fitur konten terstruktur, sehingga kurang mampu merepresentasikan pengalaman subjektif pelanggan secara menyeluruh (Mondi & Wijayanto, 2019).

Penelitian yang dilakukan oleh Christofer dkk. (2020) mengembangkan sistem rekomendasi objek pariwisata berbasis Android di Kota Pontianak,

Kalimantan Barat, menggunakan metode content-based filtering dengan pendekatan *dot product*. Sistem rekomendasi memanfaatkan kategori wisata dan tag yang dipilih dari daftar favorit pengguna sebagai representasi preferensi implisit. Hasil penelitian menunjukkan bahwa sistem mampu memberikan rekomendasi yang sesuai dengan preferensi pengguna, dengan 89% responden menyatakan bahwa sistem mudah digunakan dan rekomendasi yang diberikan relevan. Namun demikian, sistem tidak memanfaatkan data teks ulasan maupun rating pengguna, serta hanya mengandalkan struktur tag yang ditentukan sebelumnya. Hal ini menyebabkan sistem kurang adaptif terhadap perubahan preferensi pengguna yang bersifat dinamis (Christofer et al., 2020).

Penelitian yang dilakukan oleh Cholil dkk. (2023) mengembangkan sistem rekomendasi tempat wisata di Kota Semarang, Jawa Tengah, menggunakan metode collaborative filtering berbasis pengguna (user-based collaborative filtering). Sistem rekomendasi dibangun berdasarkan data rating pengguna terhadap tempat wisata dengan skala 1 hingga 5. Hasil penelitian menunjukkan bahwa metode ini mampu merekomendasikan destinasi wisata berdasarkan kesamaan selera antar pengguna, serta memungkinkan pengguna memperoleh rekomendasi destinasi baru yang belum pernah dikunjungi sebelumnya. Namun, sistem mengalami permasalahan *cold-start* untuk pengguna maupun destinasi baru dan sangat bergantung pada ketersediaan data rating. Selain itu, jumlah data pengguna yang terbatas menyebabkan rekomendasi kurang stabil, serta sistem tidak mempertimbangkan atribut objek wisata seperti harga, fasilitas, dan aksesibilitas (Cholil et al., 2023).

Penelitian yang dilakukan oleh Faurina dkk. (2023) mengembangkan sistem rekomendasi wisata di Bali dengan mengombinasikan metode content-based filtering dan collaborative filtering. Metode content-based filtering digunakan untuk memberikan rekomendasi berdasarkan kesesuaian kategori, kota, dan harga destinasi wisata, sedangkan collaborative filtering berbasis neural network digunakan untuk mengolah data rating pengguna. Dataset yang digunakan terdiri atas 75 destinasi wisata dan sekitar 3.000 data rating dari 100 pengguna. Hasil penelitian menunjukkan bahwa metode collaborative filtering menghasilkan nilai RMSE sebesar 0,2427 dan mampu memberikan rekomendasi bertipe "*more like*

this", sementara metode content-based filtering efektif dalam merekomendasikan destinasi sesuai minat kategori pengguna. Namun, metode collaborative filtering memerlukan jumlah data rating yang besar, sedangkan performa content-based filtering menurun ketika data kategori terbatas. Selain itu, sistem belum memanfaatkan ulasan teks bebas dari pengguna dan hanya mengandalkan atribut terstruktur (Faurina et al., 2023).

Suhailah dkk. (2023) mengembangkan sistem rekomendasi pariwisata di Yogyakarta menggunakan metode Triangle Multiplying Jaccard (TMJ), yang merupakan kombinasi dari Jaccard similarity dan triangle similarity. Sistem rekomendasi dibangun berdasarkan data rating kunjungan dengan skala 0 hingga 5 dari 323 pengguna terhadap 25 destinasi wisata. Hasil penelitian menunjukkan bahwa sistem mampu memberikan rekomendasi destinasi wisata seperti Pantai Drini dan Keraton Yogyakarta dengan tingkat akurasi sebesar 70%. Namun, nilai *recall* yang dihasilkan sangat rendah, yaitu sebesar 1%, yang mengindikasikan bahwa sistem kurang mampu merekomendasikan destinasi yang jarang dikunjungi. Kondisi ini disebabkan oleh ketidakseimbangan dataset, sehingga rekomendasi cenderung bersifat umum (Suhailah, 2023).

Penelitian yang dilakukan oleh Aisah dkk. (2023) mengembangkan model analisis sentimen terhadap ulasan pengguna aplikasi Al-Qur'an digital pada Google Play Store menggunakan algoritma Support Vector Machine (SVM). Penelitian ini memanfaatkan ulasan pengguna sebagai sumber informasi untuk mengidentifikasi sentimen positif dan negatif, dengan SVM dipilih karena mampu melakukan klasifikasi berdasarkan konsep *margin* serta memiliki kemampuan dalam menangani data teks. Data yang diperoleh melalui Google Play Store diproses menggunakan tahapan *text preprocessing* sebelum dilakukan klasifikasi. Hasil penelitian menunjukkan bahwa algoritma SVM mampu mencapai akurasi sebesar 85,11%, sehingga menunjukkan bahwa SVM dapat digunakan untuk mengidentifikasi respons pengguna terhadap suatu aplikasi berdasarkan ulasan tekstual. Hasil analisis tersebut juga dapat dimanfaatkan untuk mengetahui aspek yang diapresiasi maupun yang masih perlu ditingkatkan oleh pengembang aplikasi (Aisah et al., 2023).

Penelitian yang dilakukan oleh Nurpila dkk. (2024) mengembangkan sistem rekomendasi tempat wisata berbasis website di Kabupaten Rembang, Jawa Tengah, menggunakan metode content-based filtering dengan pendekatan TF-IDF dan cosine similarity. Data yang digunakan berupa deskripsi tempat wisata yang diperoleh dari Dinas Pariwisata setempat, yang merupakan gabungan teks mengenai nama destinasi, fasilitas, lokasi, harga, dan jam operasional. Hasil penelitian menunjukkan bahwa sistem berhasil memberikan rekomendasi destinasi wisata yang relevan berdasarkan kesamaan deskripsi, dengan hasil uji *usability* yang menunjukkan tingkat kepuasan pengguna sebesar 89%. Meskipun demikian, sistem belum memproses data ulasan bebas maupun rating dinamis dari pengunjung, dan hanya mengandalkan teks deskriptif yang bersifat statis. Kondisi ini menyebabkan sistem rekomendasi kurang adaptif terhadap perubahan tren wisata maupun preferensi spesifik pengguna (Nurpila et al., 2024).

Penelitian yang dilakukan oleh Rizki dan Rianto. (2024) mengembangkan sistem rekomendasi *hybrid* menggunakan metode switching pada dataset produk mainan Amazon. Sistem ini mengombinasikan metode content-based filtering berbasis TF-IDF dan collaborative filtering berbasis K-Nearest Neighbors. Pendekatan *switching* digunakan untuk mengatasi permasalahan *cold-start*, di mana sistem menerapkan content-based filtering untuk pengguna baru dan collaborative filtering untuk pengguna lama. Hasil penelitian menunjukkan bahwa sistem mampu mencapai tingkat akurasi sebesar 83,62% untuk pengguna baru dan 74,9% untuk pengguna lama. Meskipun metode ini efektif dalam mengatasi *cold-start problem*, penelitian ini belum diterapkan pada domain pariwisata atau kuliner dan belum mempertimbangkan faktor kontekstual seperti lokasi geografis, cuaca, serta sentimen ulasan pengguna (Rizki & Rianto, 2024).

Penelitian yang dilakukan oleh Husna dkk. (2024) mengembangkan sistem rekomendasi pariwisata berbasis web di Kabupaten Gunungkidul, Daerah Istimewa Yogyakarta, menggunakan metode content-based filtering. Data yang digunakan berupa deskripsi tempat wisata yang diperoleh dari website resmi pemerintah, yang memuat informasi naratif mengenai daya tarik wisata, fasilitas, aksesibilitas, dan karakteristik destinasi lainnya. Proses representasi teks dilakukan menggunakan metode TF-IDF, sedangkan perhitungan tingkat kemiripan antar destinasi

dilakukan dengan cosine similarity. Hasil penelitian menunjukkan bahwa sistem mampu memberikan rekomendasi destinasi wisata yang memiliki kemiripan karakteristik, seperti Pantai Baron yang direkomendasikan bersama Pantai Ngandong, Nguyahan, Siung, dan Jogan. Namun, sistem hanya mengandalkan deskripsi resmi dan tidak memanfaatkan ulasan atau sentimen pengunjung, sehingga rekomendasi yang dihasilkan belum sepenuhnya mencerminkan pengalaman aktual wisatawan di lapangan (Husna et al., 2024).

Penelitian lain yang dilakukan oleh Faizah dan Sancoko. (2024) mengembangkan sistem rekomendasi pariwisata di Kabupaten Bantul menggunakan metode MOORA sebagai sistem pendukung keputusan berbasis Multi-Objective Optimization on the Basis of Ratio Analysis (MOORA). Penelitian tersebut memanfaatkan sejumlah kriteria terstruktur, seperti harga tiket, fasilitas, aksesbilitas, dan jenis wisata, untuk menentukan rekomendasi destinasi wisata yang optimal. Pendekatan tersebut dinilai efektif dalam membantu proses pengambilan keputusan secara objektif berdasarkan bobot kriteria yang telah ditentukan. Akan tetapi, sistem rekomendasi yang dikembangkan masih terbatas pada data terstruktur dan metode multi-kriteria, serta belum memanfaatkan data tidak terstruktur, seperti data komentar dan ulasan pengunjung yang tersedia di platform digital. Sehingga, rekomendasi yang dihasilkan kurang akurat karena belum mempertimbangkan aspek persepsi dan pengalaman langsung dari pengunjung (Faizah & Sancoko, 2024).

Penelitian yang dilakukan oleh Martanto dan Istiono (2024) menerapkan analisis sentimen pada ulasan pengguna aplikasi M-Paspor di Google Play Store menggunakan metode Multinomial Naive Bayes. Penelitian memanfaatkan 6.443 ulasan yang dikumpulkan melalui teknik web scraping, kemudian diproses menggunakan tahapan case folding, tokenization, normalization, stemming, serta pembobotan TF-IDF sebelum dilakukan klasifikasi ke dalam kelas positif, negatif, dan netral. Hasil penelitian menunjukkan bahwa pembagian data 80:20 memberikan performa terbaik dengan akurasi 66,45%, precision 72,24%, recall 64,12%, dan F1-score 64,79%. Meskipun metode Multinomial Naive Bayes mampu mengklasifikasikan data teks secara sederhana dan efisien, penelitian ini masih berfokus pada analisis sentimen aplikasi sehingga belum mengintegrasikan

hasil sentimen dengan sistem rekomendasi maupun atribut objek wisata sebagai dasar pengambilan keputusan (Martanto & Istiono, 2024).

Penelitian yang dilakukan oleh Atmadji dkk. (2025) mengembangkan platform evaluasi destinasi pariwisata berbasis teknologi digital untuk mendukung Destinasi Pariwisata Super Prioritas (DPSP) di Indonesia dengan memanfaatkan analisis sentimen menggunakan algoritma Naive Bayes Classifier. Data penelitian diperoleh melalui scraping 6.720 ulasan Google Maps dari lima destinasi prioritas, yaitu Danau Toba, Borobudur, Mandalika, Labuan Bajo, dan Likupang. Tahapan pengolahan data meliputi preprocessing teks, pembobotan TF-IDF, penyeimbangan data menggunakan SMOTE, serta klasifikasi sentimen ke dalam kelas positif, netral, dan negatif. Hasil penelitian menunjukkan model mencapai akurasi 75% dan berhasil diimplementasikan pada platform berbasis Next.js yang menyajikan dashboard evaluasi destinasi berdasarkan opini wisatawan. Namun, penelitian ini berorientasi pada evaluasi kualitas destinasi melalui visualisasi sentimen dan belum menghasilkan sistem rekomendasi wisata yang mempertimbangkan preferensi pengguna maupun pemeringkatan destinasi secara personal (Atmadji et al., 2025).

Penelitian yang dilakukan oleh Pratiwi dkk. (2025) mengkaji penerapan metode content-based filtering dalam sistem rekomendasi objek wisata di Kabupaten Aceh Tamiang, Provinsi Aceh. Sistem rekomendasi dikembangkan dengan memanfaatkan ulasan wisatawan yang diperoleh dari Google Maps sebagai sumber data utama. Representasi fitur dilakukan menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF), sedangkan tingkat kemiripan antar objek wisata dihitung menggunakan cosine similarity. Ulasan wisatawan dianalisis berdasarkan atribut-atribut tertentu, seperti pantai, piknik, sepi, dan sungai, sehingga sistem mampu menghasilkan rekomendasi yang bersifat sangat personal dan sesuai dengan preferensi spesifik pengguna. Sebagai contoh, destinasi Pantai Raja Muda direkomendasikan untuk kombinasi preferensi “pantai”, “sepi”, dan “piknik”. Meskipun demikian, penelitian ini hanya menggunakan satu sumber data, yaitu Google Maps, sehingga sangat bergantung pada kualitas dan kelengkapan ulasan yang tersedia. Selain itu, sistem belum mampu menangani permasalahan *cold-start* untuk destinasi wisata baru yang belum memiliki ulasan (Pratiwi et al., 2025).

Penelitian yang dilakukan oleh Maulana dkk. (2025) mengembangkan sistem rekomendasi destinasi wisata berbasis knowledge-based recommendation di Kabupaten Sukoharjo. Sistem rekomendasi tersebut menggunakan pendekatan constraint-based reasoning dengan mempertimbangkan beberapa kriteria, antara lain jenis wisata, fasilitas, harga tiket, lokasi kecamatan, dan jam operasional. Pendekatan ini memungkinkan sistem untuk menangani preferensi pengguna yang bersifat multidimensi secara eksplisit, seperti kombinasi jenis wisata, lokasi tertentu, dan batasan harga tiket. Metode ini dinilai sesuai untuk wilayah dengan keterbatasan data pengguna. Namun demikian, sistem rekomendasi yang dikembangkan belum mempertimbangkan pengalaman subjektif wisatawan, seperti ulasan atau sentimen pengunjung. Selain itu, rekomendasi yang dihasilkan masih bersifat rule-based sehingga kurang adaptif terhadap perubahan tren dan dinamika preferensi wisatawan (Maulana et al., 2025).

Penelitian yang dilakukan oleh Damanik dkk. (2025) mengembangkan sistem rekomendasi dan pemetaan destinasi wisata alam terbaik di Berastagi, Sumatera Utara, menggunakan metode Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) yang dikombinasikan dengan Web Geographic Information System (Web GIS). Sistem ini menggunakan sejumlah kriteria, antara lain akses kendaraan, biaya makan, biaya penginapan, keamanan, serta ketersediaan fasilitas toilet dan area parkir. Hasil penelitian menunjukkan bahwa sistem mampu menghasilkan rekomendasi destinasi wisata secara objektif dengan tingkat validasi sebesar 85% terhadap pilihan pengguna, serta menghasilkan lima destinasi wisata teratas. Meskipun demikian, kriteria dan bobot yang digunakan bersifat tetap dan ditentukan melalui metode AHP, sehingga sistem kurang adaptif terhadap preferensi pribadi pengguna yang dinamis. Selain itu, sistem belum menyediakan fitur rute optimal maupun navigasi secara *real-time* (Damanik et al., 2025).

Penelitian yang dilakukan oleh Putra dkk. (2025) mengembangkan sistem rekomendasi daftar wisata berbasis algoritma aturan bagi wisatawan di Kabupaten Pacitan, Jawa Timur. Sistem ini menggunakan pendekatan *rule-based algorithm* dengan mempertimbangkan berbagai faktor, seperti jarak, waktu tempuh, kondisi cuaca dari BMKG, rating, jenis kendaraan, kategori wisata, serta lokasi pengguna yang diperoleh melalui geolokasi menggunakan TomTom API. Hasil penelitian

menunjukkan bahwa sistem mampu menghasilkan itinerary wisata multi-hari yang adaptif terhadap kondisi cuaca dan jenis kendaraan, serta responsif terhadap perubahan lokasi pengguna secara *real-time*. Meskipun demikian, sistem ini masih bergantung pada layanan API pihak ketiga dan belum mempertimbangkan aspek anggaran atau biaya aktivitas secara eksplisit. Selain itu, sistem belum mengintegrasikan umpan balik pengguna (*user feedback loop*) untuk meningkatkan kualitas rekomendasi secara berkelanjutan (Putra et al., 2025).

Penelitian yang dilakukan oleh aljihadu dkk. (2025) mengembangkan sistem rekomendasi wisata kuliner di Kabupaten Gunungkidul, Daerah Istimewa Yogyakarta, menggunakan metode content-based filtering dengan pendekatan cosine similarity. Data yang digunakan berupa nama tempat dan deskripsi yang menggabungkan informasi rating, harga, jenis kuliner, lokasi, jam operasional, dan layanan. Hasil penelitian menunjukkan bahwa sistem mampu memberikan rekomendasi tempat kuliner yang relevan berdasarkan kesamaan nama dan deskripsi, seperti rekomendasi "Pasar Kuliner Logandeng" terhadap "Taman Kuliner Wonosari". Namun, sistem hanya menggunakan dua sumber teks utama tanpa pemisahan bobot antar kriteria dan belum mempertimbangkan ulasan tekstual bebas maupun analisis sentimen pengunjung (Aljihadu, 2025).

Penelitian yang dilakukan oleh Salsabilla dkk. (2025) dengan mengembangkan sistem rekomendasi wisata di Kabupaten Pekalongan menggunakan metode content-based filtering. Penelitian ini memanfaatkan data resmi dari Dinas Pariwisata Kabupaten Pekalongan yang mencakup delapan kategori wisata, yaitu alam, air, buatan, budaya, bukit, pantai, perkemahan, dan petualangan. Representasi fitur dilakukan menggunakan algoritma TF-IDF untuk memberikan bobot pada kategori wisata, sedangkan tingkat kemiripan antar destinasi dihitung menggunakan cosine similarity. Hasil penelitian menunjukkan bahwa sistem mampu memberikan rekomendasi destinasi wisata yang relevan berdasarkan kesesuaian kategori dengan preferensi pengguna, serta membantu wisatawan dalam menemukan destinasi yang sesuai minat. Akan tetapi, sistem ini masih terbatas pada penggunaan atribut kategori dan belum memanfaatkan data ulasan atau sentimen pengunjung, sehingga rekomendasi yang dihasilkan belum

sepenuhnya menggambarkan pengalaman wisatawan langsung (Salsabilla et al., 2025).

Penelitian yang dilakukan oleh Arfian dkk. (2025) menganalisis sentimen masyarakat terhadap penerima Kartu Indonesia Pintar-Kuliah (KIP-K) pada media sosial X menggunakan algoritma Support Vector Machine (SVM). Data diperoleh melalui proses *crawling* dengan kata kunci terkait KIP-K dan diberi label secara manual, kemudian melalui tahapan *preprocessing* yang meliputi *cleaning*, *stopword removal*, tokenisasi, stemming, dan normalisasi. Selanjutnya, teks direpresentasikan menggunakan TF-IDF dan model SVM dioptimalkan melalui *hyperparameter tuning* dengan Grid Search serta menggunakan kernel RBF. Hasil pengujian terhadap 941 data menunjukkan akurasi sebesar 59%, dengan kernel RBF menghasilkan presisi terbaik. Hasil tersebut menunjukkan bahwa SVM dapat digunakan untuk mengklasifikasikan sentimen pada data media sosial yang bersifat informal, meskipun variasi bahasa dan respons masyarakat masih menjadi tantangan dalam menghasilkan klasifikasi yang lebih akurat (Arfian et al., 2025).

Penelitian yang dilakukan oleh Sinaga dkk. (2026) menganalisis sentimen komentar YouTube terhadap kasus dugaan korupsi sektor minyak dan gas (migas) menggunakan metode Support Vector Machine (SVM). Data yang digunakan berupa komentar YouTube yang diproses melalui tahapan *preprocessing*, yaitu *cleaning*, *case folding*, *tokenizing*, *stopword removal*, dan *stemming*, kemudian direpresentasikan menggunakan Term Frequency-Inverse Document Frequency (TF-IDF). Klasifikasi dilakukan menggunakan empat jenis kernel SVM, yaitu *linear*, *sigmoid*, Radial Basis Function (RBF), dan *polynomial*. Hasil penelitian menunjukkan bahwa kernel *linear* memberikan performa terbaik dengan akurasi sebesar 74,95% dan *weighted F1-score* sebesar 0,75, sedangkan kernel *sigmoid*, RBF, dan *polynomial* masing-masing memperoleh akurasi sebesar 72,40%, 71,23%, dan 56,45%. Hasil tersebut menunjukkan bahwa kombinasi TF-IDF dan SVM efektif digunakan untuk mengklasifikasikan sentimen pada data teks media sosial, dengan karakteristik data penelitian cenderung dapat dipisahkan secara linear (Sinaga et al., 2026)

Tabel 2.1 Tabel Penelitian Sebelumnya

Tahun	Nama	Judul	Metode yang Digunakan	Hasil Penelitian
2019	Rhesa Havilah Mondi, Ardhi Wijayanto, Winarno	*Recommendation System With Content-Based Filtering Method For Culinary Tourism In Mangan Application*	Content-Based Filtering	Sistem rekomendasi mampu memberikan rekomendasi restoran dengan tingkat presisi rata-rata sebesar 89,15%. Akan tetapi, tingkat akurasi yang dihasilkan relatif rendah sebesar 51,18%. hal tersebut disebabkan oleh keterbatasan jumlah data dan metode evaluasi yang digunakan. Selain itu, sistem belum memanfaatkan data tidak terstruktur seperti komentar atau ulasan pengunjung, sehingga kurang mampu merepresentasikan pengalaman pelanggan secara menyeluruh.
2020	Kevin Christofer, Albertus Joko Santoso, Andi Wahju Rahardjo Emanuel	*Sistem Rekomendasi Objek Pariwisata di Pontianak Berbasis Android Menggunakan Metode Content-Based Filtering*	Content-Based Filtering	Sistem rekomendasi mampu memberikan rekomendasi yang sesuai dengan preferensi pengguna, denga 89% pengguna menyatakan bahwa sistem mudah digunakan dan relevan rekomendasinya. Akan tetapi, sistem tidak memanfaatkan data tidak terstruktur, seperti komentar atau ulasan pengunjung, sehingga kurang mampu merepresentasikan pengalaman pelanggan secara menyeluruh.
2023	Saifur Rohman Cholil, Novita Aria Rizki, Trinanda Fajri Hanifah	*Sistem Rekomendasi Tempat Wisata Di Kota Semarang Menggunakan Metode Collaborative Filtering*	Collaborative Filtering	Sistem rekomendasi mampu merekomendasikan destinasi berdasarkan kesamaan selera pengguna dan memperkenalkan destinasi baru. Akan tetapi, mengalami *cold-start problem* (bingung melayani pengguna baru), bergantung pada data rating terbatas, dan tidak mempertimbangkan atribut objek wisata.
2023	Ruvita Faurina,	*Implementasi Metode Content-Based Filtering dan*	Content-Based Filtering dan	Sistem rekomendasi mampu memberikan rekomendasi berbasis kategori dan rating

Tahun	Nama	Judul	Metode yang Digunakan	Hasil Penelitian
	Evlin Sitanggang	*Collaborative Filtering pada Sistem Rekomendasi Wisata di Bali*	Collaborative Filtering	besar, performa content-based filtering bergantung pada ketersediaan kategori, dan belum memproses ulasan teks bebas.
2023	Eva Suhailah, Hartatik	*Pembuatan Sistem Rekomendasi Pariwisata Yogyakarta Menggunakan Triangle Multiplaying Jaccard*	Triangle Multiplaying Jaccard	Sistem rekomendasi mampu memberikan rekomendasi destinasi, seperti pada ”Pantai Drini” dan ”Keraton Yogyakarta” dengan tingkat akurasi sebesar 70%. Akan tetapi, nilai *recall* yang dihasilkan sangat rendah, yaitu sebesar 1%, yang menunjukkan bahwa sistem kurang mampu merekomendasikan destinasi yang jarang dikunjungi. Hal tersebut dikarenakan ketidakseimbangan dari dataset, sehingga rekomendasi cenderung bersifat umum.
2023	Iis Siti Aisah, Bambang Irawan, Tati Suprapti	*Algoritma Support Vector Machine (SVM) untuk Analisis Sentimen Ulasan Aplikasi Al Qur'an Digital*	Support Vector Machine (SVM)	Algoritma SVM berhasil digunakan untuk menganalisis sentimen ulasan pengguna dan memperoleh akurasi sebesar 85,11%. Hasil tersebut menunjukkan bahwa SVM dapat digunakan untuk memahami respons pengguna terhadap aplikasi berdasarkan ulasan yang diberikan.
2024	Sinta Nurpila, Harries Arizonia Ismail, Satrio Agung Prakoso	*Rancang Bangun Rekomendasi Tempat Wisata Di Kabupaten Rembang Berbasis Website Menggunakan Metode Content Based Filtering*	Content-Based Filtering	Sistem rekomendasi mampu memberikan rekomendasi destinasi wisata berdasarkan kesamaan deskripsi, dengan 89% responden menyatakan bahwa rekomendasi yang diberikan sesuai dengan kebutuhan pengguna. Akan tetapi, sistem tidak memanfaatkan data tidak terstruktur, seperti komentar atau ulasan pengunjung, sehingga kurang mampu merepresentasikan pengalaman pelanggan secara menyeluruh.

Tahun	Nama	Judul	Metode yang Digunakan	Hasil Penelitian
2024	Muhammad Rizki, Rianto	*Sistem Rekomendasi Hybrid Menggunakan Metode Switching*	Switching	Sistem rekomendasi mampu memberikan rekomendasi destinasi wisata, dengan tingkat akurasi sebesar 83,62% untuk pengguna baru dan 74,9% untuk pengguna lama. Meskipun metode dalam penelitian tersebut efektif dalam mengatasi *cold-start problem* (memberikan rekomendasi akurat walaupun data kurang), tetapi belum diterapkan pada sektor pariwisata maupun kuliner dan belum mempertimbangkan factor kontekstual seperti lokasi geografis, cuaca, dan data tidak terstruktur, seperti komentar atau ulasan pengunjung, sehingga kurang mampu merepresentasikan pengalaman pelanggan secara menyeluruh.
2024	Siti Okta Sifa'ul Husna, Ika Nur Fajri, Yoga Pristyanto	*Sistem Rekomendasi Pariwisata Gunungkidul Berbasis Web Menggunakan Metode Content-Based Filtering*	Content-Based Filtering	Sistem rekomendasi mampu memberikan rekomendasi destinasi wisata berdasarkan karakteristik objek wisata dan preferensi pengguna dari data terstruktur. Akan tetapi, penelitian ini belum memanfaatkan data tidak terstruktur seperti komentar atau ulasan pengunjung, sehingga rekomendasi kurang akurat karena belum mempertimbangkan aspek persepsi dan pengalaman langsung dari pengunjung.
2024	Dinda Pratiwi, Asrianda, Lidya Rosnita	*Penerapan Metode Content-Based Filtering dalam Sistem Rekomendasi Objek Wisata di Aceh Tamiang*	Content-Based Filtering	Sistem berhasil memberikan rekomendasi objek wisata di Aceh Tamiang berdasarkan kemiripan fitur (kategori) dengan profil preferensi pengguna. Dengan cara kerjanya, apabila seorang pengguna menyukai satu objek wisata (misal: Pantai Ujung Tamiang), maka

Tahun	Nama	Judul	Metode yang Digunakan	Hasil Penelitian
				system akan merekomendasikan objek lainnya yang memiliki karakteristik serupa. Metode content-based filtering lebih direkomendasikan untuk sektor pariwisata, karena lebih sederhana diterapkan untuk data yang terbatas, tetapi masih memberikan hasil yang relevan dan sesuai preverensi pengguna. Untuk pengembangan lebih lanjut, disarankan menggunakan pendekatan Hybrid Filtering, memperluas data ulasan, dan menambahkan fitur analitik untuk meningkatkan akurasi dan relevansi rekomendasi.
2024	Anisa Nur Faizah, Sulistyo Dwi Sancoko	*Sistem Rekomendasi Wisata di Kabupaten Bantul Menggunakan Metode MOORA*	Multi-Objective Optimization on the Basis of Ratio Analysis (MOORA)	Sistem rekomendasi mampu memberikan rekomendasi destinasi wisata berbasis multi kriteria kuantitatif maupun kualitatif yang relevan dengan preferensi pengguna, juga telah berhasil diujikan secara fungsional menggunakan black-box testing. Akan tetapi, sistem belum mempertimbangkan komentar atau ulasan pengunjung, sehingga masih kurang terhadap persepsi dan pengalaman wisatawan yang bersifat dinamis.
2024	Muhammad Luthfi Fajri Martanto, Wirawan Istiono	*Sentiment Analysis of M-Paspor App Reviews Using Multinomial Naive Bayes*	Multinomial Naive Bayes, TF-IDF, Confusion Matrix	Analisis sentimen terhadap 6.443 ulasan Google Play menghasilkan akurasi terbaik 66,45% pada pembagian data 80:20. Penelitian berfokus pada klasifikasi sentimen dan belum menghasilkan sistem rekomendasi wisata.
2025	Ery Setiyawan Jullev Atmadji, Yuyun Wabula,	*The Use of Naive Bayes Classifier in Sentiment Analysis at Indonesia's Super Priority Tourism*	Naive Bayes Classifier, TF-IDF, SMOTE, Web Scraping	Platform evaluasi destinasi berbasis Google Maps mampu mengklasifikasikan 6.720 ulasan dengan akurasi 75%. Sistem menyajikan dashboard evaluasi destinasi,

Tahun	Nama	Judul	Metode yang Digunakan	Hasil Penelitian
	Hayuning Titi Karsanti, Kristopher	*Destinations Based on User Reviews*		tetapi belum memberikan rekomendasi wisata yang dipersonalisasi.
2025	Ridlo Maulana S, Sri Sumarlinda, Afu Ichsan Pradana	*Sistem Rekomendasi Berbasis Knowledge-Based Untuk Destinasi Wisata Di Kabupaten Sukoharjo*	Knowledge-Based	Sistem rekomendasi mampu menangani preferensi pengguna multidimensi (kriteria yang kompleks) secara jelasdan sesuai untuk wilayah dengan data pengguna terbatas. Namun, sistem belum mempertimbangkan ulasan dari pengunjung dan bersifat aturan (rule-based filtering), sehingga kurang adaptif terhadap perubahan tren wisata.
2025	Nadya Syahfitri Damanik, Safwandi, Fajriana	*Sistem Rekomendasi Dan Pemetaan Destinasi Wisata Alam Terbaik Di Berastagi Menggunakan Metode Technique For Order Preference By Similarity To Ideal Solution (Topsis) Berbasis Web*	Technique For Order Preference By Similirity To Ideal Solution (TOPSIS)	Sistem rekomendasi mampu menghasilkan objektif dengan validasi 85% dan lima destinasi teratas. Akan tetapi, bobot kriteria bersifat tetap, kurang adaptif terhadap preferensi dinamis, dan juga belum menyediakan rute optimal dan navigasi *real-time*.
2025	Berlian Juliartha Martin Putra, Anwar Fu'adi, Dwi Ariani Finda Yuniarti, Bagus Hikmahwan	*Perancangan Sistem Rekomendasi Daftar Wisata Berbasis Algoritma Aturan Bagi Wisatawan Kabupaten Pacitan*	Algoritma Aturan	Sistem rekomendasi mampu menghasilkan rencana perjalanan wisata dalam rentang waktu hitungan hari (*itinerary planning*) yang adaptif terhadap kondisi cuaca dan lokasi pengguna secara *real-time*. Namun, sistem bergantung pada API pihak ketiga, belum mempertimbangkan aspek biaya, dan belum menerima umpan balik pengguna.
2025	Lukman Hakim Aljihadu, Arif Nur Rohman	*Sistem Rekomendasi Wisata Kuliner Di Gunungkidul Menggunakan*	Content-Based Filtering	Sistem rekomendasi mampu memberikan rekomendasi kuliner berdasarkan kesamaan nama dan dekripsi. Akan tetapi, penelitian

Tahun	Nama	Judul	Metode yang Digunakan	Hasil Penelitian
		Metode Content Based Filtering		tersebut hanya menggunakan dua sumber teks utama, tidak memisahkan bobot kriteria, dan belum mempertimbangkan ulasan dan preferensi dengan jelas dari pengunjung.
2025	Cinta Salsabilla, Danang Wahyu Utomo	*Sistem Rekomendasi Wisata Kabupaten Pekalongan dengan Content Based Filtering*	Content-Based Filtering	Sistem rekomendasi mampu memberikan rekomendasi destinasi wisata yang relevan berdasarkan kesesuaian kategori dengan preferensi pengguna, serta membantu wisatawan dalam menemukan destinasi yang sesuai minat. Akan tetapi, sistem ini masih terbatas pada penggunaan atribut kategori dan belum memanfaatkan data ulasan atau sentimen pengunjung, sehingga rekomendasi yang dihasilkan belum sepenuhnya menggambarkan pengalaman pengunjung.
2025	Muhamad Hadi Arfian, Hendrian Sofu Maruhawa, Aulia Aisyah Putri, Dinah Ratulugina, Ellsza Ridzky Khoirunnisa, Verell Hermawan, Ahmad Julaibib Syahr	*Analisis Sentimen Pada Media Sosial Menggunakan Metode Support Vector Machine*	Support Vector Machine (SVM), TF-IDF, *n-gram*, dan *Grid Search* dengan kernel RBF	Pengujian terhadap 941 data menghasilkan akurasi sebesar 59%. Kernel RBF menghasilkan presisi terbaik. Hasil penelitian menunjukkan bahwa SVM dapat digunakan untuk menganalisis sentimen pada data media sosial, meskipun variasi bahasa informal dan perbedaan respons masyarakat menjadi tantangan dalam klasifikasi.
2026	Dion Bona Torang Sinaga, Agus Sulistyanto,	*Analisis Sentimen Komentar Youtube terhadap Kasus Dugaan Korupsi Migas*	Support Vector Machine (SVM) dengan	Kernel *linear* menghasilkan performa terbaik dengan akurasi 74,95% dan *weighted F1-score* 0,75. Sementara itu, kernel *sigmoid*, RBF, dan

Tahun	Nama	Judul	Metode yang Digunakan	Hasil Penelitian
	Hidayatul Ichwan	*Menggunakan Algoritma Support Vector Machine (SVM)*	empat jenis kernel, yaitu *linear*, *sigmoid*, RBF, dan *polynomial*	*polynomial* masing-masing memperoleh akurasi 72,40%, 71,23%, dan 56,45%. Hasil menunjukkan bahwa kombinasi TF-IDF dan SVM dapat digunakan untuk analisis sentimen komentar YouTube.

Berdasarkan Tabel 2.1, dapat disimpulkan bahwa penelitian terdahulu masih memiliki keterbatasan dalam pemanfaatan data tidak terstruktur dan integrasi metode dalam sistem rekomendasi. Oleh karena itu, penelitian ini mengusulkan pengembangan sistem rekomendasi pariwisata menggunakan pendekatan Hybrid Filtering yang mengombinasikan Content-Based Filtering dan Rule-Based Filtering, serta analisis sentimen dari komentar dan ulasan pengunjung untuk menghasilkan sistem rekomendasi yang lebih akurat dan relevan.

BAB III
LANDASAN TEORI

3.1. Sistem Rekomendasi

Sistem rekomendasi merupakan sebuah sistem yang dirancang untuk memberikan kemudahan kepada pengguna dalam menemukan atau menentukan item yang sesuai dengan minat, kebutuhan, atau kriteria tertentu. Sistem tersebut bekerja dengan cara menyaring dan mengolah informasi dari pengguna maupun karakteristik item, sehingga mampu memberikan saran yang relevan dan bermanfaat (Faurina et al., 2023). Hingga saat ini, sistem rekomendasi telah banyak diterapkan pada berbagai platform digital, seperti marketplace, layanan pariwisata, hingga aplikasi berbasis lokasi untuk membantu pengguna dalam proses pengambilan keputusan (Aljihadu, 2025).

Menurut Melville dan Sindhwani (2010), sistem rekomendasi secara umum dibagi menjadi tiga pendekatan utama, diantaranya Collaborative Filtering, Content-Based Filtering, dan Hybrid Filtering. Selain ketiga pendekatan tersebut, terdapat juga pendekatan Rule Based Filtering yang menghasilkan rekomendasi berdasarkan aturan atau kriteria yang ditentukan sesuai kebutuhan pengguna (Faurina et al., 2023). Pada penelitian ini, sistem rekomendasi dibangun menggunakan pendekatan Hybrid Filtering yang mengombinasikan Rule-Based Filtering sebagai proses penyaringan awal dengan Content-Based Filtering sebagai proses perhitungan kemiripan destinasi sehingga diharapkan mampu menghasilkan rekomendasi yang lebih relevan.

Hal tersebut dikarenakan terdapat kemungkinan pemanfaatan atribut destinasi wisata secara langsung, seta memberikan rekomendasi yang transparan dan mudah diinterpretasikan. Dengan pendekatan ini, diharapkan sesuai dengan kondisi data destinasi wisata yang memiliki atribut eksplisit, seperti fasilitas dan persepsi dari pengunjung di, serta tidak bergantung pada data preferensi pengguna lain. Sehingga, sistem rekomendasi dalam penelitian ini berperan sebagai kerangka utama yang mengintegrasikan berbagai informasi terkait destinasi wisata, agar menghasilkan rekomendasi yang relevan, objektif, dan sesuai dengan kebutuhan pengguna.

3.2. Content-Based Filtering

Pendekatan Content-Based Filtering merupakan metode yang ada dalam sistem rekomendasi yang memberikan kepada pengguna berdasarkan kesamaan karakteristik antara item yang pernah diminati pengguna dengan item lain yang tersedia (Oyadila et al.,

2025). Pendekatan ini memanfaatkan atribut atau dekripsi konten dari setiap item untuk menentukan tingkat kesesuaian dengan preferensi pengguna, sehingga rekomendasi yang dihasilkan bersifat personal dan relevan (Salsabilla et al., 2025).

Secara konseptual, Content-Based Filtering bekerja dengan menganalisis fitur atau atribut yang melekat pada setiap item, seperti jenis wisata, fasilitas, harga, rating, maupun karakteristik lain yang juga relevan. Sistem kemudian membangun profil preferensi pengguna berdasarkan riwayat interaksi sebelumnya, baik secara eksplisit, maupun implisit (Oyadila et al., 2025). Dengan item yang memiliki tingkat kesesuaian tertinggi terhadap profil preferensi pengguna, selanjutnya yang akan direkomendasikan oleh sistem (Nurpila et al., 2024).

Landasan dasar Content-Based Filtering yaitu tingkat kemiripan (*similarity*), semakin tinggi kemiripan antara karakteristik item dengan preferensi pengguna, maka semakin besar kemungkinan item tersebut direkomendasikan (Faurina et al., 2023). Pendekatan tersebut bersifat *user-independence*, yang artinya sistem tidak bergantung pada data atau aktivitas pengguna lain, sehingga tetap dapat memberikan rekomendasi, meskipun item tersebut baru yang belum pernah dipilih oleh pengguna lain (Mondi & Wijayanto, 2019).

Tabel 3.1 Preferensi Pengguna

Destinasi	Karakteristik
Pantai Indah	Pantai, Indah, Bersih
Pantai Drini	Pantai, Indah, Ramai
Pantai Jungwok	Pantai, Kotor, Ramai

Misalkan Seorang pengguna memilih preferensi wisata alam dan pantai. Pada metode Content-Based Filtering, sistem akan membandingkan karakteristik preferensi pengguna dengan karakteristik setiap destinasi. Destinasi yang memiliki atribut paling mirip akan memperoleh nilai kemiripan yang lebih tinggi sehingga lebih berpeluang direkomendasikan. Berdasarkan Tabel 3.1, Pantai Indah dan Pantai Drini memiliki karakteristik yang lebih sesuai dengan preferensi pengguna dibandingkan Pantai Jungwok, sehingga kedua destinasi tersebut menjadi kandidat rekomendasi.

3.2.1 Representasi Fitur Destinasi Wisata

Dalam sistem rekomendasi berbasis konten, representasi fitur merupakan tahapan penting yang bertujuan untuk mengubah karakteristik objek ke dalam bentuk terstruktur dan numerik agar dapat diproses secara komputasional. Representasi fitur ini menentukan

bagaimana setiap destinasi wisata dideskripsikan berdasarkan atribut yang dimiliki, sehingga dapat dibandingkan dengan preferensi pengguna secara objektif (Ricci et al., 2011).

Pada destinasi wisata, fitur yang digunakan umumnya berbentuk atribut eksplisit pada objek wisata, khususnya pada fasilitas wisata. Atribut tersebut dipilih karena mampu mencerminkan kualitas dan kenyamanan destinasi wisata dari sudut pandang pengguna. Fasilitas meliputi ketersediaan area parker, toilet, tempat ibadah, tempat sampah, kantin/kafetaria, area istirahat pengunjung, serta fasilitas keamanan seperti CCTV.

Atribut-atribut fasilitas tersebut pada awalnya berskala nominal, yang memiliki 2 kategori, yaitu ada dan tidak ada. Agar dapat digunakan dalam proses perhitungan kemiripan pada sistem rekomendasi berbasis konten, atribut nominal dikonversi kedalam bentuk numerik melalui proses pengkodean (*encoding*). Pada penelitian ini, pengkodean dilakukan dengan skema biner, yaitu nilai 1 untuk fasilitas yang tersedia dan 0 untuk fasilitas yang tidak tersedia.

Tabel 3.2 Deskripsi Destinasi

Destinasi	Deskripsi
D1	Pantai indah bersih
D2	Pantai indah ramai
D3	Pantai kotor ramai

Berdasarkan Tabel 3.2, Sebelum dihitung menggunakan TF-IDF, setiap deskripsi masih berupa teks. TF-IDF kemudian mengubah setiap kata menjadi bobot numerik sehingga setiap destinasi dapat direpresentasikan dalam bentuk vektor. Dengan demikian, setiap destinasi memiliki representasi numerik yang dapat diproses oleh algoritma perhitungan kemiripan.

Jumlah dokumen diatas yaitu 3, mana N = 3

Tabel 3.3 Kemunculan Destinasi

Kata	Muncul pada	DF
pantai	D1, D2, D3	3
indah	D1, D2	2
bersih	D1	1
ramai	D2, D3	2
kotor	D3	1

$$IDF(t) = \ln\left(\frac{N}{DF(t)}\right) \quad (3.1)$$

Sehingga:

Tabel 3.4 Kata Destinasi

Kata	DF	IDF
pantai	3	0
indah	2	0,405
bersih	1	1,099
ramai	2	0,405
kotor	1	1,099

Karena setiap kata hanya muncul satu kali dalam setiap dokumen disederhanakan:

$$TF = 1$$

Maka TF-IDF:

D1 = "pantai indah bersih"

$$D1 = (0, 0.405, 1.099, 0, 0)$$

D2 = "pantai indah ramai"

$$D2 = (0, 0.405, 0, 0.405, 0)$$

D3 = "pantai kotor ramai"

$$D3 = (0, 0, 0, 0, 0.405, 1.099)$$

Kemudian terbentuk matriks:

Tabel 3.5 Contoh TF-IDF Content-Based

Destinasi	pantai	indah	bersih	ramai	kotor
D1	0	0,405	1,099	0	0
D2	0	0,405	0	0,405	0
D3	0	0	0	0,405	1,099

3.2.2 Pengukuran Kemiripan (Similarity Measure)

Pengukuran kemiripan merupakan tahap inti dalam metode Content-Based Filtering yang digunakan sebagai menentukan sejauh mana suatu destinasi wisata memiliki karakteristik yang mirip dengan preferensi pengguna dengan destinasi lain. Nilai kemiripan ini bertujuan untuk dasar dalam menentukan destinasi mana yang paling relevan untuk direkomendasikan kepada pengguna (Ricci et al., 2011).

Dalam sistem rekomendasi berbasis konten, pengukuran kemiripan dilakukan dengan membandingkan vektor fitur antar item. Setiap destinasi wisata direpresentasikan sebagai sebuah vektor yang berisi nilai numerik dari atribut-atribut yang telah ditentukan pada tahap representasi fitur. Semakin tinggi tingkat kemiripan antara dua vektor, maka semakin besar kemungkinan destinasi tersebut direkomendasikan (Han et al., 2023).

Terdapat beberapa metode yang dapat digunakan untuk mengukur kemiripan antar vektor fitur, seperti Euclidan Distance, Jaccard Similarity, dan Consine Similarity. Pada penelitian ini, metode yang digunakan adalah Consine Similarity karena mampu mengukur tingkat kesamaan pola antar vektor tanpa dipengaruhi oleh besarnya nilai absolut data (Wang et al., 2020).

3.2.2.1 Cosine Similarity

Cosine Similarity merupakan metode yang sering digunakan untuk menghitung kemiripan dua buah dokumen dengan cara menghitung nilai cosinus sudut yang terbentuk diantara dua buah vektor yang merepresentasikan kedua dokumen tersebut. Hasil dari pengukuran dokumen sesuai dengan kemiripan atau relevansi terhadap kata kunci. Nilai *Cosine Similarity* berkisar dari 0 hingga 1, dimana nilai 1 memiliki makna kemiripan antar dokumen tersebut sangat tinggi, begitupun sebaliknya. Berikut merupakan persamaan untuk menghitung nilai Consine Similary (Salsabila Hanifah, 2024).

$$Similarity(a,b) = \cos(\theta) = \frac{(a.b)}{\|a\|.\|b\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \quad (3.2)$$

dengan

$$\|a\| = \sqrt{\sum_{i=1}^{n} a_i^2} = \sqrt{a_1^2 + a_2^2 + a_3^2 + \cdots + a_n^2} \quad (3.3)$$

Keterangan:

a = Vektor a

b = Vektor b

$a.b$= *Dot product* antara vektor a dan b

$\|a\|$ = Panjang (norma) vektor a

$\|b\|$ = Panjang (norma) vektor b

θ = Sudut antara dua vektor

a_i = Komponen ke-i dari vektor a

i = Indeks komponen vektor, $i = 1, 2, \ldots, n$

n = Jumlah komponen (dimensi) pada vektor

Berikut contoh perhitungan Consine Similarity, misalkan hasil TF-IDF menghasilkan vektor berikut.

$D1 = (0, 0.405, 1.099, 0, 0)$

$D2 = (0, 0.405, 0, 0.405, 0)$

$D3 = (0, 0, 0, 0, 0.405, 1.099)$

Consine similarity D1 dengan D2, perkalian dot product:

$$D1.D2 = (0.405 \times 0.405) = 0.164$$

Panjang *vector*:

$$||D1|| = \sqrt{0.405^2 + 1.099^2} \approx 1.171$$

$$||D2|| = \sqrt{0.405^2 + 0.405^2} \approx 0.573$$

$$Consine(D1, D2) = \frac{0.164}{1.171 \times 0.573} \approx 0.245$$

Jadi kemiripan D1 dan D2 sekitar 0.245

Cosine similarity D1 dengan D3 tetapi tidak memiliki kata penting yang sama, maka:

$$D1 \times D3 = 0$$

$$Cosine(D1.D3) = 0$$

Yang berarti destinasi D1 dan D3 tidak memiliki kemiripan berdasarkan fitur yang digunakan, sehingga D2 lebih mirip dengan D1 dibandingkan D3.

3.2.2.2 Matriks *Similarity*

Matriks *similarity* merupakan matriks yang berisi nilai kemiripan antara setiap pasangan destinasi dalam sistem rekomendasi. Pada pendekatan Content-Based Filtering, setiap destinasi terlebih dahulu direpresentasikan sebagai vektor fitur menggunakan TF-IDF. Selanjutnya, tingkat kemiripan antarvektor destinasi dihitung menggunakan *Cosine Similarity* dengan menghitung kemiripan berdasarkan hasil perkalian titik (*dot product*) yang dinormalisasi oleh panjang masing-masing vektor. Pendekatan TF-IDF dan Cosine Similarity dapat digunakan untuk merepresentasikan karakteristik dokumen dalam ruang vektor dan mengukur tingkat kemiripannya (Widianto & Pebriyanto, 2024).

Hasil perhitungan kemiripan seluruh pasangan destinasi kemudian disusun ke dalam matriks berukuran $n \times n$, dengan $n$merupakan jumlah destinasi yang dibandingkan. Setiap baris dan kolom merepresentasikan destinasi yang sama, sedangkan nilai pada sel ke-(i, j) menunjukkan tingkat kemiripan antara destinasi ke-i dan destinasi ke-j. Karena perbandingan dilakukan terhadap himpunan destinasi yang sama, *matriks similarity* bersifat simetris, sehingga nilai kemiripan destinasi A terhadap destinasi B sama dengan destinasi B terhadap destinasi A. Nilai diagonal utama bernilai 1 karena setiap destinasi memiliki kemiripan sempurna dengan dirinya sendiri.

Pada representasi TF-IDF yang memiliki nilai fitur non-negatif, nilai *Cosine Similarity* berada pada rentang 0 hingga 1. Nilai yang semakin mendekati 1 menunjukkan tingkat kemiripan yang semakin tinggi, sedangkan nilai yang mendekati 0 menunjukkan tingkat kemiripan yang semakin rendah. Dalam sistem rekomendasi wisata, nilai kemiripan tersebut dapat digunakan sebagai dasar untuk menentukan destinasi yang memiliki karakteristik paling mirip dan selanjutnya digunakan dalam proses perangkingan rekomendasi (Scikit-learn Developers, 2026).

Contoh *matriks similarity* lanjutan dari *Cosine Similarity* sebagai berikut.

Tabel 3.6 *Matriks Similarity*

	D1	D2	D3
D1	1,000	0,245	0,000
D2	0,245	1,000	0,245
D3	0,000	0,245	1,000

atau dalam bentuk matriks:

$$S = \begin{bmatrix} 1 & 0.245 & 0 \\ 0.245 & 1 & 0.245 \\ 1 & 0.245 & 1 \end{bmatrix}$$

Matriks *cosine similarity* dibaca berdasarkan baris terhadap kolom, di mana setiap nilai menunjukkan tingkat kemiripan antara dua destinasi. Sebagai contoh, nilai S_{12}= 0,245 berarti Destinasi 1 memiliki kemiripan sebesar 0,245 dengan Destinasi 2, sedangkan S_{13} = 0 menunjukkan tidak adanya kemiripan pada representasi TF-IDF. Sementara itu, nilai pada diagonal matriks selalu 1 karena setiap destinasi memiliki kemiripan sempurna dengan dirinya sendiri.

Beberapa sifat penting dari matriks *similarity* adalah sebagai berikut.

- Nilai diagonal selalu 1, karena setiap destinasi dibandingkan dengan dirinya sendiri sehingga memiliki kemiripan sempurna.
- Matriks bersifat simetris, artinya nilai kemiripan D1 terhadap D2 sama dengan D2 terhadap D1. Sebagai contoh, nilai similarity D1 terhadap D2 sebesar 0,245, sehingga nilai D2 terhadap D1 juga sebesar 0,245.
- Nilai terbesar di luar diagonal menunjukkan destinasi yang memiliki tingkat kemiripan paling tinggi dan berpotensi menjadi rekomendasi berdasarkan pendekatan Content-Based Filtering.

Misalkan pengguna sedang melihat Destinasi D1. Sistem kemudian membaca baris D1 pada matriks *similarity*. Berdasarkan baris tersebut, diperoleh nilai kemiripan terhadap destinasi lainnya sebagai berikut.

Tabel 3.7 Kemiripan Destinasi

Destinasi Pembanding	Nilai Similarity	Keterangan
D2	0,245	Paling mirip
D3	0,000	Tidak memiliki kemiripan

D1 sendiri memiliki nilai *similarity* sebesar 1,000, tetapi tidak dimasukkan sebagai kandidat rekomendasi karena merupakan destinasi yang sedang dilihat oleh pengguna. Dengan demikian, apabila sistem menggunakan pendekatan Top-K Neighbor, maka D2 dipilih sebagai tetangga terdekat (*nearest neighbor*) dari D1 karena memiliki nilai *similarity* tertinggi, yaitu sebesar 0,245. Sementara itu, D3 memiliki nilai *similarity* sebesar 0,000, sehingga memiliki tingkat kemiripan paling rendah terhadap D1. Nilai similarity tersebut kemudian dapat digunakan sebagai Content *Score* dalam proses perangkingan rekomendasi.

Pada contoh ini, terdapat 3 destinasi, sehingga matriks similarity yang terbentuk berukuran 3 × 3. Artinya, setiap destinasi dibandingkan dengan seluruh destinasi lainnya menggunakan *Cosine Similarity*. Dalam penelitian sebenarnya yang menggunakan 315 destinasi, proses yang sama menghasilkan matriks *similarity* berukuran 315 × 315, sehingga setiap destinasi memiliki nilai kemiripan terhadap seluruh destinasi lainnya. Matriks tersebut kemudian digunakan untuk memperoleh Top-K Neighbor dan nilai Content *Score*, yang selanjutnya dapat digabungkan dengan Rule *Score* dalam perhitungan Hybrid *Score* untuk menghasilkan urutan rekomendasi akhir.

3.3. Rule-Based Filtering

Pendekatan Rule-Based Filtering merupakan salah satu metode dalam sistem rekomendasi yang bekerja berdasarkan sekumpulan aturan (*rules*) yang telah ditentukan sebelumnya. aturan tersebut digunakan untuk menyaring dan memilih item yang sesuai dengan kriteria tertentu dari pengguna atau sistem itu sendiri. Pendekatan ini sering kali digunakan ketika kebutuhan dari pengguna dapat diterjemahkan secara eksplisit dalam bentuk aturan logis yang jelas dan terukur (Ardhi Supratman et al., 2024).

Pendekatan ini juga sering dikategorikan sebagai bagian dari sistem pendukung keputusan (*Decisiom Support System*), dikarenakan *focus* utamanya merupakan membantu pengguna dalam mengambil keputusan dari kriteria yang telah ditentukan. Metode ini bersifat *deterministic*, yang berarti hasil rekomendasi sepenuhnya ditentukan oleh aturan

yang dilakukan, sehingga proses rekomendasi menjadi lebih transparan dan mudah dijelaskan kepada pengguna.

Akan tetapi, umumnya Rule-Based Filtering tidak bergantung pada pola perilaku pengguna lain ataupun perhitungan kemiripan antar item, tetapi mengandalkan kondisi logis berbentuk pernyataan *if-then*. Misalnya, apabila pengguna menginginkan destinasi wisata yang memiliki fasilitas parkir dan toilet, maka sistem hanya akan mempertimbangkan destinasi yang memenuhi kondisi tersebut. Sehingga, item yang tidak memenuhi aturan akan dieliminasi sejak awal proses rekomendasi (Ilmasik et al., 2025).

3.3.1 Pembentukan Aturan (*Rule Definition*)

Pada implementasinya, Rule-Based Filtering sangat sesuai digunakan untuk menyaring destinasi berdasarkan kebutuhan dan batasan tertentu. Umumnya, destinasi wisata memiliki atribut yang eksplisit dan diskrit, seperti ketersediaan fasilitas, jenis wisata, dan kondisi lingkungan. Atribut tersebut dapat diterjemahkan dengan mudah dalam aturan logis, sehingga sistem mampu menyesuaikan rekomendasi dengan kebutuhan pengguna secara langsung (Ilmasik et al., 2025). Setiap aturan dinyatakan dalam bentuk pernyataan logis yang terdiri dari kondisi (*antecedent*) dan hasil (*consequent*), aturan dapat diinterpretasikan sebagai berikut .

$$R_i : IF\ (x_1 = a_1) \wedge (x_2 = a_2) \wedge \ldots \wedge (x_n = a_n)\ THEN\ y = 1 \quad (3.4)$$

Atau

$$R_i : IF \bigwedge_{j=1}^{n} (x_j = a_j)\ THEN\ y = 1 \quad (3.5)$$

Keterangan:

R_i = Aturan ke-i

i = Indeks aturan, $i = 1, 2, \ldots, m$

m = Jumlah seluruh aturan

x_j = Atribut destinasi wisata ke-j

a_j = Nilai atribut yang diinginkan pengguna pada atribut ke-j

j = Indeks atribut, $j = 1, 2, \ldots, n$

n = Jumlah atribut yang digunakan dalam aturan

$y = 1$: destinasi memenuhi kriteria (direkomendasikan)

$y = 0$: destinasi tidak memenuhi kriteria (tidak direkomendasikan)

$\wedge$ = Operator logika AND

Apabila salah satu kondisi tidak memenuhi, maka nilai y = 0 artinya destinasi tidak direkomendasikan. Berikut ini adalah contoh tabel aturan yang disesuaikan dengan rumus diatas.

Tabel 3.1 Tabel Aturan Fasilitas

Kode Rule	Aturan
R_1	IF (Jenis Wisata = Alam) AND (Parkir = 1) AND (CCTV = 1) THEN $y = 1$ (Direkomendasikan)
R_2	IF (Jenis Wisata = Budaya) AND (Rating ≥ 4.0) THEN $y = 1$ (Direkomendasikan)
R_3	IF (Harga Tiket ≤ Rp25.000) AND (Toilet = 1) THEN $y = 1$ (Direkomendasikan)
R_4	IF (Jenis Wisata = Alam) AND (Parkir = 0) AND (CCTV = 1) THEN $y = 0$ (Tidak Direkomendasikan)

Contoh sederhana Rule Based Filtering, misalkan pengguna memberikan preferensi berikut.

Tabel 3.8 Preferensi Rule-Based

Preferensi	Nilai
Jenis wisata	Alam
Maksimum tiket	Rp25,000.00
Toilet	Ya

Kemudian terdapat tiga destinasi:

Tabel 3.9 Hasil Destinasi Rule-Based

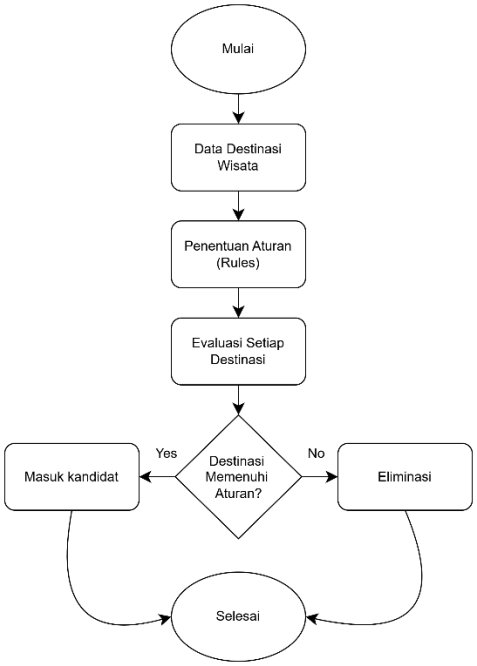
Destinasi	Hasil
Pantai Indah	Memenuhi
Air Terjun Sejuk	Memenuhi
Museum Jogja	Tidak memenuhi

Aturan yang digunakan dapat ditulis yaitu IF Jenis Wisata = Alam AND Harga Tiket ≤ Rp25.000 AND Toilet = Ya THEN Direkomendasikan. Destinasi yang tidak memenuhi aturan akan dieliminasi pada tahap penyaringan awal (pre-*filtering*).

3.3.2 Proses Penyaringan (*Filtering Process*)

Proses penyaringan pada Rule-Based Filtering dilakukan dengan mengevaluasi setiap destinasi wisata terhadap aturan yang telah ditentukan. Apabila destinasi tidak memenuhi satu atau lebih dari aturan, maka akan dilakukan eliminasi dari daftar kandidat rekomendasi. Pada proses ini, digunakan sebagai penyaringan awal (pre-*filtering*) sebelum

metode rekomendasi lainnya ditetapkan (Ricci et al., 2011). berikut merupakan alur dari diagram alir Rule-Based Filtering.



Gambar 3.1 Diagram Alir Rule-Based Filtering (Pre-*Filtering*)

3.3.3 Kelebihan dan Kekurangan Rule-Based Filtering

Kelebihan utama dari pendekatan ini adalah kemampuan dalam menghasilkan rekomendasi yang relevan sesuai dengan preferensi eksplisit pengguna tanpa perlu data historis atau interaksi dari pengguna lain. Dari hal tersebut, menjadikannya metode yang sangat efektif untuk mengatasi permasalahan *cold-start user*, yang berarti kondisi ketika *system* belum memiliki cukup data mengenai perilaku pengguna (Ilmasik et al., 2025).

Akan tetapi, pendekatan ini juga memiliki kekurangan, yaitu ketergantungan pada kelengkapan dan ketepatan aturan yang dirancang. Apabila aturan digunakan terlalu ketat, maka jumlah *item* yang direkomendasikan menjadi sangat terbatas, begitu juga sebaliknya (Ricci et al., 2011). Oleh karena itu, pada penelitian ini Rule-Based Filtering digunakan sebagai penyaringan awal (pre-*filtering*), untuk memastikan bahwa destinasi wisata yang direkomendasikan telah sesuai dengan kriteria minimum yang ditentukan.

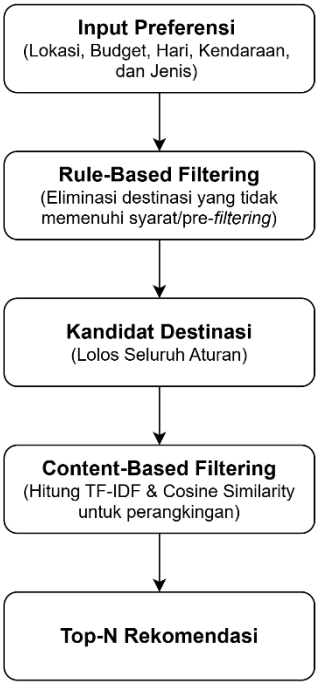
3.4. Hybrid Filtering

Hybrid Filtering merupakan pendekatan sistem rekomendasi yang mengombinasikan dua atau lebih teknik rekomendasi untuk memanfaatkan kelebihan dari masing-masing metode. Penggabungan beberapa teknik dilakukan karena setiap metode memiliki karakteristik dan keterbatasan yang berbeda sehingga dapat saling melengkapi dalam menghasilkan rekomendasi. Burke (2002) menjelaskan bahwa berbagai teknik

rekomendasi dapat dikombinasikan untuk membangun sistem hybrid dengan memanfaatkan keunggulan dari masing-masing pendekatan (Burke, 2002).

Pada penelitian ini, Hybrid Filtering dibangun dengan mengombinasikan Rule-Based Filtering dan Content-Based Filtering menggunakan prinsip Cascade Hybrid. Pada strategi cascade, proses rekomendasi dilakukan secara berurutan, sehingga hasil dari satu metode digunakan untuk mempersempit atau menyempurnakan hasil pada tahap berikutnya (Rebelo et al., 2022). Dalam penelitian ini, Rule-Based Filtering digunakan sebagai tahap penyaringan awal berdasarkan preferensi pengguna, seperti kategori, lokasi, kendaraan, dan anggaran. Destinasi yang memenuhi kriteria kemudian diproses menggunakan Content-Based Filtering berdasarkan representasi TF-IDF dan Cosine Similarity untuk memperoleh tingkat kemiripan antar destinasi. Pendekatan ini memungkinkan proses rekomendasi dilakukan pada kandidat yang telah sesuai dengan kebutuhan dasar pengguna.

Berikut adalah alur dari Hybrid Filtering.



Gambar 3.2 Diagram Alir Hybrid Filtering

Rumus yang digunakan yaitu

$$Hybrid\ Score = w_R R + w_C C \quad (3.6)$$

Dengan ketentuan

$$w_R + w_C = 1 \quad (3.6)$$

Keterangan:

w_R = Bobot Rule-Based Filtering

w_C = Bobot Content Based Filtering

R = Rule *Score*

C = Content *Score*

Contoh Hybrid Filtering, misalkan pengguna menginginkan wisata alam dengan anggaran Rp25.000.

Tahap Rule-Based Filtering:

Tabel 3.10 Contoh Rule *Score*

Destinasi	Rule *Score*
D1	1,000
D2	1,000
D3	0,000

Tahap Content-Based Filtering:

Tabel 3.11 Contoh Similarity *Score*

Destinasi	Similarity
D1	1,000
D2	0,245
D3	0,000

Tahap Hybid Filtering:

$$D1 = (0.4)(1) + (0.4)(1) = 0.8$$

$$D2 = (0.4)(1) + (0.4)(0.245) = 0.498$$

$$D3 = (0.4)(0) + (0.4)(0) = 0$$

Sehingga:

Tabel 3.12 Contoh Hybrid *Score*

Destinasi	Rule	Cosine	Hybrid
D1	1.000	1.000	0.800
D2	1.000	0.245	0.498
D3	0.000	0.000	0.000

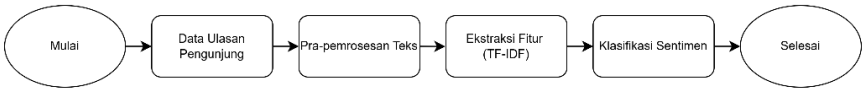
Karena D3 tidak memenuhi aturan pengguna, destinasi tersebut tidak menjadi kandidat rekomendasi. D1 dan D2 kemudian dihitung nilai kemiripannya menggunakan Cosine Similarity. D1 memiliki nilai similarity dan Hybrid *Score* tertinggi, yaitu masing-masing 1,000 dan 0,800, sehingga menjadi rekomendasi utama, sedangkan D2 berada pada urutan berikutnya dengan Hybrid *Score* 0,498.

3.5. Analisis Sentimen Ulasan Pengunjung

Analisis Sentimen atau *opinion mining* merupakan suatu metode yang digunakan untuk mengidentifikasi dan mengklasifikasikan opini berbentuk teks menjadi kategori positif, negatif, maupun netral. Analisis ini bertujuan untuk memahami persepsi dari

masyarakat terhadap suatu produk atau layanan yang dapat membantu dalam proses pengambilan keputusan. Studi kasus penggunaan Analisis Sentimen telah dilakukan pada berbagai platform, seperti Twitter atau X, TikTok, Instagram, Shopee, komentar pada Website pemerintah, dan lain sebagainya, sebagai evaluasi respon pengguna terhadap layanan dan produk tertentu (Lukmana, 2025).

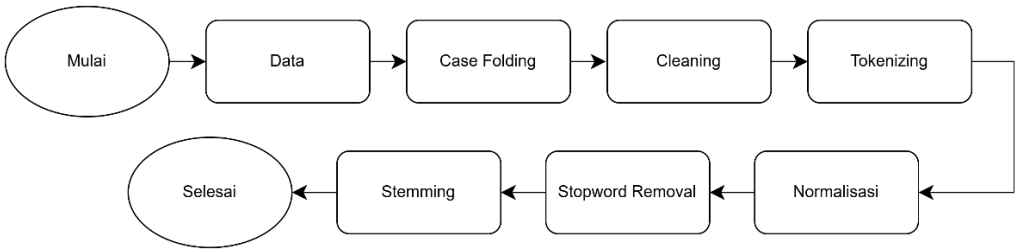
Pada penelitian ini, Analisis Sentimen dilakukan pada platform Google Maps. Dalam pariwisata, analisis sentimen memiliki peran penting dalam menggambarkan tingkat kepuasan dan pengalaman pengunjung terhadap suatu destinasi wisata, sehingga dapat dijadikan sebagai indikator kualitas destinasi dari sudut pandang pengguna. Analisis ini dilakukan beberapa tingkat analisis, diantaranya tingkat dokumen, tingkat kalimat, dan tingkat aspek (Mudya Yolanda & Tri Mulya, 2024). Jika dihubungkan dengan penelitian ini, tingkat dokumen gambarkan dengan setiap ulasan atau komentar dianggap sebagai opini yang memiliki satu label sentimen utama. Hal ini dipilih karena lebih sederhana dan sesuai dengan tujuan penelitian dalam mengidentifikasi persepsi umum pengunjung terhadap destinasi wisata. Berikut ini adalah alur dari diagram alir analisis sentimen.



Gambar 3.3 Diagram Alir Analisis Sentimen

3.5.1 Pra-Pemrosesan Teks (*Preprocessing Text*)

Pra-pemrosesan teks atau *preprocessing text* merupakan tahap awal dalam pengolahan data teks yang digunakan untuk menyederhanakan dan menstrukturkan data agar dapat diolah secara efisien. pada tahap ini, mengubah teks mentah baik terstruktur, semi-terstruktur, maupun tidak terstruktur menjadi format yang lebih bersih untuk proses analisis lanjutan. Dikarenakan pada platform digital, umumnya ulasan mengandung *noise*, seperti simbol, emoji, kata baku, dan variasi penulisan yang tidak konsisten. Berikut merupakan langkah-langkah pra-pemrosesan teks yang digunakan dalam penelitian ini (Oyadila et al., 2025).



Gambar 3.4 Diagram Alir Pra-Pemrosesan Teks

Dengan melakukan beberapa tahap, diantaranya (Oyadila et al., 2025):

1. *Case folding* merupakan proses perubahan huruf kapital menjadi huruf kecil semua, terlebih untuk teks yg memiliki struktur teks bervariasi dan tidak teratur.
2. Cleaning digunakan sebagai proses pembersihan teks dari karakter yang tidak diperlukan sehingga hanya terdapat informasi yang relevan, proses meliputi penghapusan URL, emoji, tanda baca, angka, karakter khusus, dan spasi berlebih.
3. *Tokenizing* yaitu proses untuk memisahkan kata pada kalimat atau suatu dokumen dalam bentuk token.
4. Normalisasi bertujuan untuk mengubah kata-kata tidak baku, singkatanm atau bahasa informal menjadi bentuk baku sesuai Kamus Besar Bahasa Indonesia (KBBI) atau kamus normalisasi yang digunakan.
5. *Stopword removal* adalah tahap menghilangkan kata-kata yang tidak memiliki makna dalam suatu kalimat, misalnya kata yang, di, adalah, dan lainnya.
6. *Stremming* digunakan untuk mendapatkan bentuk dasar dari kata turunan, dengan cara menghapus imbuhan pada kata tersebut, seperti sufiks, konfiks, dan infiks.

3.5.2 Ekstraksi Fitur Teks menggunakan TF-IDF

Pada tahap *vectorization* menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency*) merupakan teknik statistik dalam pengolahan bahasa alami yang digunakan untuk mengevaluasi seberapa penting suatu kata (*term*) dalam sebuah dokumen, terhadap koleksi dokumen yang lebih luas. Fungsi teknik ini adalah dapat memilih fitur dalam analisis teks dan pencarian informasi, serta sistem rekomendasi konten.

TF-IDF terdiri dari dua komponen utama, yaitu *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). *Term Frequency* (TF) untuk mengukur seberapa sering sebuah kata muncul dalam satu dokumen, dengan cara membagi jumlah kemunculan kata dengan total jumlah kata dalam dokumen tersebut. Berikut adalah rumus dari *Term Frequency*.

$$tf_{i,j} \frac{f_{i,j}}{\sum d_j} \quad (3.7)$$

Dengan keterangan $tf_{i,j}$ adalah jumlah kemunculan i dalam dokumen j, dan d_j adalah total kata dalam dokumen j.

Sedangkan *Inverse Document Frequency* (IDF), digunakan untuk mengukur seberapa penting kata tersebut dalam konteks koleksi dokumen yang lebih luas, dimana kata yang jarang muncul akan memiliki nilai IDF yang lebih tinggi karena dianggap lebih informatif. Berikut adalah rumus dari *Inverse Document Frequency.*

$$idf_i = \log\left(\frac{N}{df_i}\right) \quad (3.8)$$

Dengan keterangan N adalah total jumlah dokumen dalam koleksi, dan df_i adalah jumlah dokumen yang mengandung kata i.

Sehingga, bobot akhir dari TF-IDF pada sebuah kata dalam dokumen, diperoleh dengan mengalikan nilai TF dan IDF, seperti pada gambar berikut.

$$w = tf \times idf \quad (3.9)$$

Nilai dari w menunjukkan seberapa penting kata tersebut dalam sebuah dokumen tertentu dibandingkan pada dokumen lainnya (Oyadila et al., 2025).

Contoh TF-IDF untuk sentimen sebagai berikut.

Tabel 3.13 Contoh TF-IDF Sentimen

Komentar	Kata penting
D1	pantai indah bersih nyaman
D2	pantai kotor sampah bau

$$IDF(pantai) = \ln\left(\frac{2}{2}\right) = 0$$

$$IDF(indah) = \ln\left(\frac{2}{1}\right) = 0.693$$

$$TF(indah, D1) = \frac{1}{4} = 0.25$$

$$TF\ IDF\ (indah, D1) = (0.25)(0.693) = 0.173$$

Perhitungan yang sama dilakukan terhadap seluruh kata pada seluruh dokumen. Hasilnya membentuk matriks TF-IDF, di mana setiap baris merepresentasikan satu ulasan dan setiap kolom merepresentasikan satu fitur/kata. Dalam penelitian ini, setelah proses preprocessing diperoleh 14.327 ulasan dengan 9.774 fitur, artinya setiap dari 14.327 ulasan direpresentasikan menggunakan 9.774 fitur numerik. Matriks tersebut kemudian digunakan sebagai input fitur pada model klasifikasi, seperti SVM, untuk melakukan prediksi kelas sentimen positif, netral, dan negatif.

3.5.3 Ketidakseimbangan Data (*Imbalanced Dataset*)

Ketidakseimbangan data (*imbalanced dataset*) merupakan kondisi ketika jumlah data pada setiap kelas dalam suatu *dataset* tidak terdistribusi secara proporsional. Kondisi ini ditandai dengan adanya kelas yang memiliki jumlah data jauh lebih banyak (kelas mayoritas) dibandingkan kelas lainnya (kelas minoritas). Hal tersebut sering dijumpai pada

berbagai permasalahan klasifikasi, termasuk analisis sentimen, sehingga dapat memengaruhi proses pembelajaran model. Model yang dilatih menggunakan data tidak seimbang cenderung lebih mengenali pola pada kelas mayoritas dan memiliki kemampuan yang lebih rendah dalam mengklasifikasikan kelas minoritas (Osorio, Alvarez-Meza, Daza-Santacoloma, Orozco-Gutierrez, & Castellanos-Dominguez, 2021).

Salah satu pendekatan yang digunakan untuk mengatasi ketidakseimbangan yaitu data *balancing*. Teknik tersebut bertujuan untuk menghasilkan distribusi jumlah data yang lebih seimbang antar kelas sehingga model klasifikasi dapat mempelajari karakteristik masing-masing kelas secara lebih proporsional. Secara umum, data balancing dibedakan menjadi dua pendekatan, yaitu OverSampling dilakukan dengan menambah jumlah data pada kelas minoritas dan UnderSampling yang dilakukan dengan mengurangi jumlah data pada kelas mayoritas. Pemilihan teknik *balancing* perlu disesuaikan dengan karakteristik *dataset* dan tujuan penelitian agar tidak menurunkan kualitas informasi yang terkandung dalam data. OverSampling dapat meningkatkan risiko *overfitting* apabila dilakukan melalui duplikasi sampel, sedangkan UnderSampling memiliki risiko kehilangan sebagian informasi dari kelas mayoritas.

Pada penelitian ini, penanganan keseimbangan data menggunakan Random Under Sampling (RUS). RUS bekerja dengan memilih dan mengurangi sebagian data dari kelas mayoritas secara acak sehingga perbedaan jumlah data antar kelas menjadi lebih kecil. Data dari kelas minoritas tetap dipertahankan, sedangkan jumlah data pada kelas mayoritas dikurangi sesuai proporsi yang ditentukan. Metode ini tidak menambahkan data baru ke dalam dataset sehingga prosesnya relatif sederhana dan jumlah data yang digunakan dalam pelatihan menjadi lebih sedikit, yang dapat meningkatkan efisiensi waktu pelatihan model. Penerapan Random Under Sampling bertujuan mengurangi bias model terhadap kelas mayoritas sehingga model dapat mempelajari karakteristik kelas minoritas secara lebih baik. Namun, karena proses pengurangan data dilakukan secara acak, metode ini berpotensi menghilangkan sebagian informasi penting dari kelas mayoritas. Oleh sebab itu, penerapannya perlu dilakukan secara hati-hati agar keseimbangan data dapat tercapai tanpa menurunkan kinerja model (Dharmendra, Putra, & Atmojo, 2024).

Pemilihan Random Under Sampling pada penelitian ini didasarkan pada kondisi distribusi kelas sentimen yang tidak seimbang, dengan jumlah komentar pada kelas positif lebih besar dibandingkan kelas negatif dan netral. Kondisi tersebut berpotensi menyebabkan model lebih dipengaruhi oleh kelas mayoritas. RUS digunakan untuk

mengurangi dominasi kelas mayoritas dengan mengambil sampel secara acak hingga distribusi kelas menjadi lebih seimbang. Oleh karena itu, diperlukan tahap penyeimbangan data sebelum proses pelatihan model SVM dan MNB.

RUS dipilih karena merupakan teknik preprocessing yang sederhana, efisien, dan sesuai dengan ukuran dataset penelitian. Berbeda dengan metode Random Over Sampling maupun SMOTE yang menambahkan data baru melalui duplikasi atau pembentukan data sintetis, RUS mempertahankan seluruh data asli tanpa menghasilkan observasi buatan. Dengan demikian, model dilatih menggunakan data yang sepenuhnya berasal dari komentar wisatawan sehingga karakteristik bahasa dan opini tetap merepresentasikan kondisi sebenarnya (Almuayqil et al., 2022).

Selain itu, pengurangan jumlah sampel kelas mayoritas dapat mengurangi ukuran data latih sehingga proses pelatihan model menjadi lebih ringan secara komputasi. Penggunaan *random undersampling* juga telah diterapkan pada penelitian analisis sentimen untuk menangani dominasi kelas mayoritas. Meskipun demikian, RUS memiliki keterbatasan karena sebagian sampel kelas mayoritas dihilangkan secara acak sehingga terdapat kemungkinan informasi yang relevan ikut terhapus. Oleh karena itu, efektivitas penggunaan RUS dalam penelitian ini ditentukan berdasarkan hasil evaluasi model menggunakan *precision*, *recall*, dan *F1-score*, selain *accuracy* (Osorio, Alvarez-Meza, Daza-Santacoloma, Orozco-Gutierrez, & Castellanos-Dominguez, 2021).

3.5.4 Metode Klasifikasi Sentimen Support Vector Machine (SVM)

Support Vector Machine (SVM) merupakan teknik pembelajaran yang digunakan untuk menganalisis data dan mengenali pola untuk dilakukan klasifikasi, yang berfungsi membangun *hyperplane* dengan jangkauan Euclidean maksimum pada kasus latihan terdekat. Dengan suatu dataset diberi variabel x_i, sementara untuk kelas diberi variabel y_i. Metode SVM dibagi menjadi dua kelas, yaitu kelas pertama dipisah oleh *hyperplane* bernilai 1, sedangkan kelas kedua bernilai -1. Berikut merupakan rumus persamaan dari dua kelas yang dibahas sebelumnya pada persamaan 1 dan 2.

$$y_i(x_i \times w + b) \geq 1 \quad (3.10)$$

$$y_i(x_i \times w + b) \leq -1 \quad (3.11)$$

Keterangan:

x_i = data ke-i

w = nilai bobot *support vector* yang tegak lurus dengan *hyperplane*

b = nilai bias

y_i = kelas data ke-i

Bobot vektor (w) merupakan garis *vector* yang tegak lurus diantara titik pusat koordinat dengan garis *hyperplane*, bias adalah koordinat garis *relative* terhadap titik koordinat. Berikut merupakan persamaaan 3 untuk menghitung nilai b dan persamaan 4 untuk menghitung nilai w.

$$b = -\frac{1}{2}(w \times x^{+} + w.x) \quad (3.12)$$

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \quad (3.13)$$

Keterangan:

b = nilai bias

$w.x^{+}$ = nilai bobot untuk kelas data positif

$w.x^{-}$ = nilai bobot untuk kelas data negatif

w = bobot *vector*

a_i = nilai bobot data ke-i

Secara sederhana, algoritma SVM bertujuan untuk mengidentifikasi *hyperplane* paling optimal yang berfungsi untuk memisahkan dua kelas dalam ruang *input*. Gambar berikut merupakan pola-pola dalam dua kelas, yaitu positif (+1) ditunjukkan dalam bentuk lingkaran dan negatif (-1) dalam bentuk kotak. Dalam konteks klasifikasi, proses pembelajaran difokuskan pada upaya menemukan garis pembatas (*hyperplane*) terbaik, agar dapat membedakan kedua kelompok tersebut diantara berbagai pilihan garis pemisah (*discrimination boundaries*) yang tersedia seperti pada gambar (Sidupa & Dewi, 2025).

Terdapat empat kernel klasifikasi pada metode SVM, diantaranya (Scikit-learn Developers, 2026):

a. Kernel Linier

$$K(x_i, x_j) = x_i^T x_j \quad (3.14)$$

Keterangan:

$K(x_i, x_j)$ = nilai fungsi kernel antara data latih x_i dan data uji x_j.

x_i^T = transpose vektor fitur dari data latih ke-i.

x_j = vektor fitur dari data latih ke-j.

$x_i^T x_j$ = hasil perkalian titik (*dot product*).

b. Kernel Radial Basis Function

$$K(x_i, x_j) = \exp(-\gamma \left|\left|x_i - x_j\right|\right|^2) \quad (3.15)$$

Keterangan:

$K(x_i, x_j)$ = nilai fungsi kernel antara data latih x_i dan data uji x_j.

γ = parameter kernel.

x_i = vektor fitur dari data latih ke-i.

x_j = vektor fitur dari data latih ke-j.

$\left|\left|x_i - x_j\right|\right|^2$ = kuadrat jarak Euclidean antara $x_i$dan x_j.

exp = fungsi eksponensial dengan basis bilangan Euler (e).

c. Kernel Polynomial

$$K(x_i, x) = \left(\gamma x_i^T x_j + r\right)^d \quad (3.16)$$

Keterangan:

$K(x_i, x)$ = nilai fungsi kernel antara data latih x_i dan data uji x_j.

γ = parameter kernel.

x_i^T = transpose vektor fitur dari data latih ke-i.

x_j = vektor fitur dari data ke-j.

$x_i^T x_j$ = hasil perkalian titik antara kedua vektor.

d = derajat (*degree*) polinomial yang menentukan kompleksitas model.

d. Kernel Sigmoid

$$K(x_i, x) = \tanh(\gamma x_i^T x_j + r) \quad (3.17)$$

Keterangan:

$K(x_i, x)$ = nilai fungsi kernel antara data latih x_i dan data uji x_j.

γ = parameter kernel.

$tanh$ = fungsi hiperbolik tangen.

x_i^T = transpose vektor fitur dari data latih ke-i.

x_j = vektor fitur dari data ke-j.

r = konstanta ($coef0$).

Contoh sederhana pengaplikasian kernel RBF, didapatkan data sebagai berikut.

Ulasan	pantai	bagus	bersih	Label
U1	0,8	0,7	0,6	Positif
U2	0,7	0,6	0,5	Positif
U3	0,2	0,1	0,2	Negatif
U4	0,3	0,2	0,3	Negatif

U5	0,5	0,4	0,5	Netral

Misal SVM membandingkan U1 dengan U2

$$x_1 = (0.8, 0.7, 0.6)$$

$$x_1 = (0.7, 0.6, 0.5)$$

Dengan jarak kuadrat:

$$\left||x_1 - x_2|\right|^2 = (0.8 - 0.7)^2 + (0.7 - 0.6)^2 + (0.6 - 0.5)^2 = 0.03$$

Misal: $\gamma = 1$

$$K(x_1, x_2) = e^{-1(0.03)} \approx 0.970$$

Nilai yang mendekati 1 menunjukkan bahwa kedua vektor relatif mirip. Begitupun sebaliknya, apabila dua alasan mempunyai vektor TF-IDF yang sangat berbeda dengan nilai kernel mendekati 0.

Berikut ini contoh Confusion Matrix SVM RBF.

Tabel 3.14 Contoh Sederhana SVM RBF

Aktual \ Prediksi	Negatif	Netral	Positif	Total
Negatif	600	80	50	730
Netral	70	300	100	470
Positif	40	80	1.546	1.666
Total	710	460	1.696	2.866

Diagonal utama merupakan prediksi yang benar, diantaranya:

- Negatif-Negatif = 600
- Netral-Netral = 300
- Positif-Positif = 1546

Sehingga jumlah prediksi yang benar adalah 600 + 300 + 1546 = 2446.

Menghitung *Accuracy*:

$$Accuracy = \frac{jumlah\ prediksi\ benar}{jumlah\ seluruh\ data} = \frac{2446}{2866} = 0.8535 \approx 85{,}35\%$$

Menghitung *Precision*:

$$Precision = \frac{True\ Positive\ (TP)}{True\ Positif\ (TP) + False\ Positive\ (FP)}$$

- *Precision* kelas negatif

$$FP_{negatif} = 70 + 40 = 110$$

$$Precision_{negatif} = \frac{600}{600 + 110} = 0.8451 \approx 84{,}51\%$$

- *Precision* kelas netral

$$FP_{netral} = 80 + 80 = 160$$

$$Precision_{netral} = \frac{300}{300 + 160} = 0.6522 \approx 65{,}22\%$$

- *Precision* kelas positif

$$FP_{Positif} = 50 + 100 = 150$$

$$Precision_{Positif} = \frac{1546}{1546 + 150} = 0.9116 \approx 91{,}16\%$$

Menghitung *Precision Macro*:

$$Precision_{macro} = \frac{Precision_{negatif} + Precision_{netral} + Precision_{positif}}{3}$$

$$= \frac{0.8451 + 0.6522 + 0.9116}{3} = \frac{2.4089}{3} = 0.8030 \approx 80{,}30\%$$

Menghitung *Recall*:

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positif\ (TP) + \ False\ Negative\ (FN)}$$

- *Recall* kelas negatif

$$FN_{negatif} = 80 + 50 = 130$$

$$Recall_{negatif} = \frac{600}{600 + 130} = 0.8219 \approx 82{,}19\%$$

- *Recall* kelas netral

$$FN_{netral} = 70 + 100 = 170$$

$$Recall_{netral} = \frac{300}{300 + 170} = 0.6383 \approx 63{,}83\%$$

- *Recall* kelas positif

$$FN_{Positif} = 40 + 80 = 120$$

$$Recall_{Positif} = \frac{1546}{1546 + 120} = 0.9279 \approx 92{,}79\%$$

Menghitung *Recall Macro*:

$$Recall_{Macro} = \frac{0.8219 + 0.6383 + 0.9279}{3} = 0.7960 \approx 79{,}60\%$$

Menghitung F1-*Score*:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- Menghitung F1 negatif

$$F1_{negatif} = 2 \times \frac{0.8451 \times 0.8219}{0.8451 + 0.8219} = 0.8333 \approx 83{,}33\%$$

- Menghitung F1 netral

$$F1_{netral} = 2 \times \frac{0.6522 \times 0.6383}{0.6522 + 0.6383} = 0.6452 \approx 64{,}52\%$$

- Menghitung F1 positif

$$F1_{positif} = 2 \times \frac{0.9116 \times 0.9279}{0.9116 + 0.9279} = 0.9197 \approx 91{,}97\%$$

Menghitung F1-*Macro*:

$$F1_{Macro} = \frac{0.8333 + 0.6452 + 0.9197}{3} = 0.7994 \approx 79{,}94\%$$

3.5.5 Metode Klasifikasi Sentimen Multinomial Naïve Bayes (MNB)

Multinomial Naïve Bayes (MNB) merupakan teknik pengembangan dari algoritma Naive Bayes yang digunakan untuk menangani data teks. Metode ini menerapkan pendekatan *supervised learning* yang berbasis teori probabilitas Bayes, yang mana setiap fitur diasumsikan saling bebas (independen) satu sama lain. Apabila diterapkan pada klasifikasi dokumen, kata-kata yang terkandung dalam dokumen biasanya dijadikan sebagai fiturnya.

Menurut Hadaina dan Budiyanto (2022), Multinomial Naïve merupakan metode text mining yang digunakan pada proses klasifikasi teks yang memanfaatkan probabilitas suatu kelas dalam sebuah dokumen. Metode tersebut menghitung peluang kemunculan kata-kata tertentu pada setiap kelas sehingga dapat menentukan kategori yang paling sesuai pada suatu dokumen (Fadhlan Hadaina, 2022).

Dalam penelitian analisis sentimen, metode ini banyak digunakan karena memiliki tingkat kompleksitas yang rendah dan mampu memberikan hasil klasifikasi yang baik pada data teks. Dengan beberapa keunggulannya, yaitu mudah diterapkan pada data yang berukuran besar, mampu menangani klasifikasi multikelas, dan sangat sesuai untuk model Natural Language Processing (NLP) (Martanto & Istiono, 2024). Selain memiliki efisiensi komputasi yang mempercepat analisis sentimen dibandingkan dengan metode yang lain, juga dapat menangani dataset berukuran kecil hingga besar dengan kebutuhan komputasi yang rendah (Atmadji et al., 2025). Namun juga memiliki keterbatasan, diantaranya mengasumsikan antar kata bersifat independen, tidak mempertimbangkan hubungan semantik antar kata, sensitif terhadap kualitas preprocessing, dan performa dapat menurun apabila banyak kata ambigu atau kalimat yang kompleks (Martanto & Istiono, 2024).

Landasan metode ini adalah Teorema Bayes, yang digunakan untuk mengukur probabilitas suatu peristiwa dengan memanfaatkan informasi atau pengetahuan terdahulu. Sehingga saat diterapkan pada klasifikasi teks, rumus ini dapat memprediksi peluang suatu

dokumen yang kemudian masuk dalam kategori tertentu berdasarkan kemunculan kata-kata didalamnya. Rumus Teorema Bayes ditulis sebagai berikut.

$$P(C|D) = \frac{P(D|C) \times P(C)}{P(D)} \quad (3.18)$$

Keterangan:

$P(C|D)$ = Probabilitas dokumen D termasuk ke dalam kelas C.

$P(D|C)$ = Probabilitas dokumen D muncul pada kelas C.

$P(C)$ = Probabilitas awal (*prior probability*) dari kelas C.

$P(D)$ = Probabilitas kemunculan dokumen D.

Dikarenakan $P(D)$ bernilai konstan untuk seluruh data, maka proses klasifikasi cukup dilakukan dengan mencari nilai maksimum dari:

$$P(C|D) \propto P(D|C) \times P(C) \quad (3.19)$$

Dalam proses klasifikasi, nilai probabilitas posterior $P(C|D)$ akan dihitung untuk seluruh kelas yang tersedia, yang kemudian dimasukkan dalam kelas yang memiliki nilai probabilitas terbesar. Prinsip utama metode ini yaitu menghitung peluang suatu dokumen dalam kategori tertentu dengan melihat distribusi kata di dalamnya. Berbeda dengan Gaussian Naïve Bayes digunakan untuk data kontinu, Multinomial Naïve Bayes berfokus pada seberapa sering suatu kata muncul, sehingga sangat cocok untuk klasifikasi teks dan analisis sentimen. Asumsi Naïve pada setiap fitur (kata) bersifat independen terhadap fitur lain, misal suatu dokumen terdiri dari kata:

$$w_1, w_2, w_3, \dots, w_n \quad (3.20)$$

Maka probabilitas dokumen adalah berikut.

$$P(D|C) = P(w_1, w_2, w_3, \dots, w_n|C) \quad (3.21)$$

Dengan asumsi independensi.

$$P(D|C) = \prod_{i=1}^{n} P(w_i|C) \quad (3.22)$$

Langkah pertama dalam algoritma ini adalah menghitung probabilitas awal (*prior probability*) setiap kelas, yang dihitung menggunakan persamaan berikut (Hadaina & Budiyanto, 2022).

$$P(c) = \frac{N_c}{N_{doc}} \quad (3.23)$$

Keterangan:

c = Kategori atau kelas.

doc = Dokumen.

N_c = Jumlah dokumen pada kelas tertentu.

N_{doc} = Jumlah keseluruhan dokumen pelatihan.

Setelah dilakukan perhitungan probabilitas prior, langkah berikutnya adalah menghitung probabilitas kemunculan setiap kata pada masing-masing kelas (Likelihood Probability), berikut adalah persamaannya (Hadaina & Budiyanto, 2022).

$$P(w_i|c) = \frac{count(w_i, c) + 1}{\sum_w count(w, c) + |V|} \quad (3.24)$$

Keterangan:

w_i = Kata ke- i dalam seluruh dokumen yang berkategorikan c.

$count(w_i, c)$ = Jumlah kata tertentu yang muncul dalam suatu kategori atau kelas.

$\sum_w count(w, c)$ = Jumlah seluruh kata pada kelas.

$|V|$ = Jumlah seluruh kata unik pada kelas.

Berikut contoh sederhana dari metode Multinomial Naive Bayes.

Tabel 3.15 Contoh Sederhana MNB

Dokumen	Ulasan	Sentimen
D1	pantai bagus bersih	Positif
D2	pantai indah bagus	Positif
D3	pantai kotor buruk	Negatif
D4	tempat kotor jelek	Negatif
D5	pantai cukup bagus	Netral
D6	tempat cukup baik	Netral

Agar perhitungan menjadi lebih mudah, maka vocabulary kita batasi menjadi:

$$V = (pantai, bagus, bersih, indah, kotor, buruk, tempat, jelek, cukup, baik)$$

Jumlah dokumen yaitu N = 6, dengan masing-masing memiliki 2 dokumen, maka *prior probability*:

$$P(Negatif) = \frac{2}{6} = \frac{1}{3}$$

$$P(Netral) = \frac{2}{6} = \frac{1}{3}$$

$$P(Positif) = \frac{2}{6} = \frac{1}{3}$$

Pada tiga kelas tersebut memiliki probabilitas awal yang sama, kemudian melihat kata-kata dalam ulasan. Berikut menghitung jumlah kemunculan kata pada setiap kelas, contoh pada D1 dan D2 sebagai berikut.

- Positif

$D1$ = pantai bagus bersih

$D2$ = pantai indah bagus

Sehingga diperoleh pantai = 2, bagus = 2, bersih = 1, indah = 1, dan total seluruh kata = 6.

- Negatif

  $D3$ = pantai kotor buruk

  $D4$ = tempat kotor jelek

  Sehingga diperoleh pantai = 1, kotor = 2, buruk = 1, tempat = 1, jelek = 1, dan total seluruh kata = 6.

- Netral

  $D5$ = pantai cukup bagus

  $D6$ = tempat cukup baik

  Sehingga diperoleh pantai = 1, cukup = 2, bagus = 1, tempat = 1, baik = 1, dan total seluruh kata = 6.

Menghitung probabilitas kata:

$$\sum_{w} count(w, c) = 6$$

$$\alpha|V| = 1 \times 10 = 10$$

Maka, denominator 6 + 10 = 16.

Contoh pada kelas positif.

$$P(bagus|positif) = \frac{2+1}{16} = \frac{3}{16} = 0.1875$$

Menghitung probabilitas semua kata untuk data *training*.

Tabel 3.16 Probabilitas Kata

Kata	Positif	Negatif	Netral
pantai	3/16	2/16	2/16
bagus	3/16	1/16	2/16
bersih	2/16	1/16	1/16
indah	2/16	1/16	1/16
kotor	1/16	3/16	1/16
buruk	1/16	2/16	1/16
tempat	1/16	2/16	2/16
jelek	1/16	2/16	1/16
cukup	1/16	1/16	3/16
baik	1/16	1/16	2/16

Menghitung Skor Kelas dengan rumus Naive Bayes:

- Positif

$$P(Positif) = \frac{1}{3}$$

$$P(pantai|Positif) = \frac{3}{16}$$

$$P(bagus|Positif) = \frac{3}{16}$$

$$Score(Positif) = \frac{1}{3} \times \frac{3}{16} \times \frac{3}{16} = 0.01171875$$

- Negatif

$$P(Negatif) = \frac{1}{3}$$

$$P(pantai|Negatif) = \frac{2}{16}$$

$$P(bagus|Negatif) = \frac{1}{16}$$

$$Score(Negatif) = \frac{1}{3} \times \frac{2}{16} \times \frac{1}{16} = 0.00260417$$

- Netral

$$P(Netral) = \frac{1}{3}$$

$$P(pantai|Netral) = \frac{2}{16}$$

$$P(bagus|Netral) = \frac{2}{16}$$

$$Score(Netral) = \frac{1}{3} \times \frac{2}{16} \times \frac{2}{16} = 0.00520833$$

Berdasarkan ketiga skor tersebut, didapatkan yang terbesar yaitu kelas positif. Kemudian berikut contoh dari *Confusion Matrix*:

Tabel 3.17 Confusion Matrix MNB

Aktual \ Prediksi	Negatif	Netral	Positif
Negatif	2	0	1
Netral	0	2	1
Positif	0	1	3

Dengan total data yaitu N = 10 dan prediksi benar = 2 + 2 + 3 = 7.

Menghitung *Accuracy*:

$$Accuracy = \frac{7}{10} = 0.70 = 70\%$$

Menghitung *Precision, Recall*, dan F1 pada masing-masing kelas:

- Positif

$$TP = 3$$

$$FP = 1 + 1 = 2$$

$$FN = 1$$

$$Precision = \frac{3}{3+2} = 0.60$$

$$Recall = \frac{3}{3+1} = 0.75$$

$$F1 = 2 \times \frac{0.60 \times 0.75}{0.60 + 0.75} = 0.6667$$

- Negatif

$$TP = 2$$

$$FP = 0$$

$$FN = 1$$

$$Precision = \frac{2}{2+0} = 1$$

$$Recall = \frac{2}{2+1} = 0.6667$$

$$F1 = 2 \times \frac{1 \times 0.6667}{1 + 0.6667} = 0.80$$

- Netral

$$TP = 2$$

$$FP = 1$$

$$FN = 1$$

$$Precision = \frac{2}{2+1} = 0.6667$$

$$Recall = \frac{2}{2+1} = 0.6667$$

$$F1 = 2 \times \frac{0.6667 \times 0.6667}{0.6667 + 0.6667} = 0.6667$$

Menghitung F1-Macro:

$$F1_{Macro} = \frac{F1_{negatif} + F1_{netral} + F1_{positif}}{3}$$

$$= \frac{0.80 + 0.6667 + 0.6667}{3} = 0.7111 = 71{,}11\%$$

3.6. Keterkaitan Antar Metode

Dalam penelitian ini, pembentukan sistem rekomendasi destinasi wisata melibatkan beberapa metode utama yang saling berkaitan secara konseptual dan berurutan sesuai alur pemrosesan data, diantaranya analisis sentimen berbasis *Machine Learning* menggunakan antara metode Support Vector Machine (SVM) atau Multinomial Naïve Bayes (MVB), Rule-Based Filtering (RBF), dan Content-Based Filtering (CBF). Terintegrasinya metode-metode ini bertujuan untuk menghasilkan rekomendasi yang tidak hanya akurat secara numerik, tetapi juga relevan dan sesuai dengan preferensi pengguna.

Analisis Sentimen mengubah data tidak terstruktur, yang berupa ulasan pengunjung menjadi skor yang menggambarkan persepsi dan pengalaman pengguna terhadap destinasi.

Skor sentimen tersebut ditransformasikan menjadi variabel numerik yang dapat digabungkan dengan variabel/atribut destinasi lainnya, seperti fasilitas, kategori, dan operasional wisata. Seluruh atribut destinasi diproses melalui metode Rule-Based Filtering sebagai tahap awal penyaringan (pre-*filtering*).

Pada tahap ini, sistem mengevaluasi setiap destinasi berdasarkan aturan logis yang merepresentasikan preferensi eksplisit pengguna, sehingga destinasi yang memenuhi kriteria minimum yang dipertahankan sebagai kandidat rekomendasi. Selanjutnya, kandidat tersebut diproses menggunakan Content-Based Filtering untuk menghitung tingkat kemiripan antara profil preferensi pengguna dan karakteristik destinasi wisata. Hasil perhitungan kemiripan digunakan sebagai dasar dalam proses perangkingan, sehingga sistem mampu menghasilkan rekomendasi destinasi wisata yang relevan dan sesuai dengan kebutuhan pengguna.

BAB IV
METODOLOGI PENELITIAN

4.1. Populasi dan Sampel Penelitian

Populasi pada penelitian ini yaitu seluruh destinasi wisata yang berada di Daerah Istimewa Yogyakarta (DIY) yang terdaftar dalam basis Data Inventarisasi DTW DIY 2025, dengan jumlah 315 destinasi wisata. Seluruh destinasi tersebut digunakan sebagai objek penelitian tanpa melakukan pemilihan sampel tertentu. Sampel pada penelitian ini adalah data ulasan atau komentar pengunjung, dikarenakan jumlah komentar pada platform digital bersifat besar dan dinamis. Platform digital yang digunakan, yaitu Google Maps, yang merepresentasikan persepsi dan pengalaman pengunjung terhadap masing-masing destinasi wisata.

Pada platform Google Maps, komentar diambil langsung dari halaman destinasi wisata terkait. Guna menjaga keseimbangan data antar destinasi dan menghindari biar popularitas, jumlah komentar dari kedua platform dibatasi dengan jumlah minimal 10 komentar dan maksimal 100 komentar setiap destinasi, tetapi tetap bergantung pada ketersediaan data.

Data destinasi wisata bersumber dari Dinas Pariwisata Daerah Istimewa Yogyakarta (DIY) dan *scraping* secara mandiri untuk melengkapi variabel yang belum tersedia. Selanjutnya untuk seluruh data komentar yang terkumpul akan diproses menggunakan analisis sentimen berbasis Support Vector Machine (SVM) dengan membandingkan kernel Linier vs RBF juga metode Multinomial Naive Bayes (MNB) untuk memperoleh label sentimen pada tingkat komentar. Hasil klasifikasi sentimen tersebut kemudian diagregasi pada tingkat destinasi, sehingga diperoleh skor sentimen yang merepresentasikan persepsi pengunjung terhadap masing-masing destinasi wisata. Skor sentimen tersebut diintegrasikan dengan variabel destinasi lain dalam sistem rekomendasi destinasi wisata berbasis Hybrid Filtering (perpaduan antara Content-Based Filtering dan Rule-Based Filtering) guna membangun sistem rekomendasi destinasi yang akurat.

4.2. Tempat dan Waktu Penelitian

Penelitian ini dilaksanakan secara non-eksperimental yaitu memanfaatkan data sekunder yang diperoleh dari Dinas Pariwisata DIY dan platform digital untuk pelengkap data yang kurang, serta komentar pengunjung destinasi. Proses pengolahan dan analisis data dilakukan pada lingkungan komputasi menggunakan laptop pribadi dengan bantuan

perangkat lunak (*software*) pendukung analisis data. Pengambilan data destinasi wisata di Dinas Pariwisata DIY dilakukan pada 20 Oktober 2025 hingga 19 November 2025, sedangkan data pelengkap lainnya dan komentar pada platform digital dilakukan pada bulan november 2025 hingga awal Mei 2026, dengan disesuaikan ketersediaan data pada masing-masing platform.

4.3. Definisi Operasional Variabel Penelitian

Penelitian ini memiliki dua kelompok variabel, yaitu variabel *input* dan variabel *output*. Variabel *input* merupakan karakteristik destinasi wisata yang digunakan sebagai dasar dalam sistem rekomendasi, yang terdiri dari informasi dasar destinasi, seperti nama dan jenis destinasi, aspek lokasi, operasional, serta fasilitas. Sementara itu, variabel *output* adalah hasil analisis yang dihasilkan dari permodelan sistem rekomendasi, berupa skor sentimen komentar pengunjung sebagai evaluasi dari pengunjung secara nyata terhadap destinasi wisata. Definisi operasional variabel dijelaskan sebagai berikut.

Tabel 4.1 Definisi Operasional variabel

Nama Variabel	Definisi Operasional	Skala Data	Sumber Data	Satuan
Nama Destinasi	Nama resmi destinasi wisata, sebagai informasi dasar untuk menggambarkan karakteristik umum dari destinasi.	Nominal	Database Inventarisasi DTW DIY	-
Kategori/Jenis Wisata	Kategori utama destinasi wisata (alam, budaya, taman rekreasi, sejarah), sebagai informasi dasar untuk menggambarkan karakteristik umum dari destinasi.	Nominal	Database Inventarisasi DTW DIY	-
Longitude	Koordinat geografis bujur destinasi, untuk menghitung jarak atau sebaran lokasi secara *real-time.*	Rasio	Scraping Google Maps	Derajat

Nama Variabel	Definisi Operasional	Skala Data	Sumber Data	Satuan
Latitude	Koordinat geografis lintang destinasi, untuk menghitung jarak atau sebaran lokasi secara *real-time.*	Rasio	Scraping Google Maps	Derajat
Harga Tiket Weekday	Biaya tiket masuk destinasi wisata hari kerja.	Rasio	Scraping Google Maps	Rupiah
Harga Tiket Weekend dan Libur Nasional	Biaya tiket masuk destinasi wisata akhir pekan.	Rasio	Scraping Google Maps	Rupiah
Jam Operasional Weekday	Lama waktu operasional hari kerja.	Rasio	Database Inventarisasi DTW DIY dan Scraping Google Maps	Jam
Jam Operasional Weekend dan Libur Nasional	Lama waktu operasional akhir pekan.	Nominal	Database Inventarisasi DTW DIY dan Scraping Google Maps	Jam
Parkir Motor	Biaya parkir motor pada area destinasi wisata.	Nominal	Scraping Google Maps, Google *Search*, Sosial Media	Rupiah
Parkir Mobil	Biaya parkir mobil pada area destinasi wisata.	Nominal	Scraping Google Maps, Google *Search*, Sosial Media	Rupiah
Mode Transportasi Motor	Biaya perjalanan menggunakan motor yang dihitung per kilometer (km).	Nominal	Google *Search*	Rupiah

Nama Variabel	Definisi Operasional	Skala Data	Sumber Data	Satuan
Mode Transportasi Mobil	Biaya perjalanan menggunakan mobil yang dihitung per kilometer (km).	Nominal	Google *Search*	Rupiah
Toilet	Ketersediaan fasilitas toilet, sebagai batasan minimum dalam proses rule-based filtering untuk menjamin standar kelayakan destinasi.	Nominal	Database Inventarisasi DTW DIY	0/1
Area Parkir	Ketersediaan area parkir, sebagai batasan minimum dalam proses rule-based filtering untuk menjamin standar kelayakan destinasi.	Nominal	Database Inventarisasi DTW DIY	0/1
Ruang Ibadah	Ketersediaan fasilitas tempat ibadah di destinasi wisata, sebagai batasan minimum dalam proses rule-based filtering untuk menjamin kenyamanan pengunjung dalam menjalankan ibadah.	Nominal	Database Inventarisasi DTW DIY	0/1
Tempat Sampah	Ketersediaan fasilitas tempat sampah di area destinasi, sebagai indikator dukungan kebersihan dan sebagai batasan minimum dalam rule-based filtering.	Nominal	Database Inventarisasi DTW DIY	0/1
Area Makan	Ketersediaan fasilitas kafetaria atau tempat	Nominal	Database Inventarisasi	0/1

Nama Variabel	Definisi Operasional	Skala Data	Sumber Data	Satuan
Area Makan	makan di destinasi wisata, sebagai batasan minimum dalam rule-based filtering untuk mendukung kebutuhan konsumsi pengunjung.		DTW DIY	
Area Istirahat	Ketersediaan area khusus untuk istirahat pengunjung, sebagai batasan minimum dalam rule-based filtering untuk meningkatkan kenyamanan kunjungan wisata.	Nominal	Database Inventarisasi DTW DIY	0/1
CCTV	Ketersediaan kamera pengawas (CCTV) di area destinasi wisata, sebagai batasan minimum dalam rule-based filtering dan indikator keamanan.	Nominal	Database Inventarisasi DTW DIY	0/1
Sentimen Pengunjung	Nilai sentimen dari komentar pengunjung (positif, netral, atau negatif)	Interval	Scraping TikTok dan Google Maps	-1 s.d. +1

4.4. Alat dan cara organisir data

Penelitian ini dilaksanakan melalui beberapa tahapan yang dimulai dari pengumpulan data, pengolahan data, hingga penyusunan data untuk analisis dan pemodelan sistem rekomendasi. Data penelitian tersebut terdiri dari data destinasi wisata dan data komentar dari pengunjung yang diperoleh dari platform digital. Data destinasi wisata diperoleh dari data inventarisasi Daya Tarik Wisata (DTW) Daerah Istimewa Yogyakarta (2025) dan *scraping* secara manual pada beberapa platform digital. Setelah itu, data diseleksi dan dibersihkan (data *cleaning*) untuk memastikan kelengkapan dan

konsistensi format, pada variabel numerik (harga dan koordinat) dan variabel kategorikal (jenis wisata dan fasilitas).

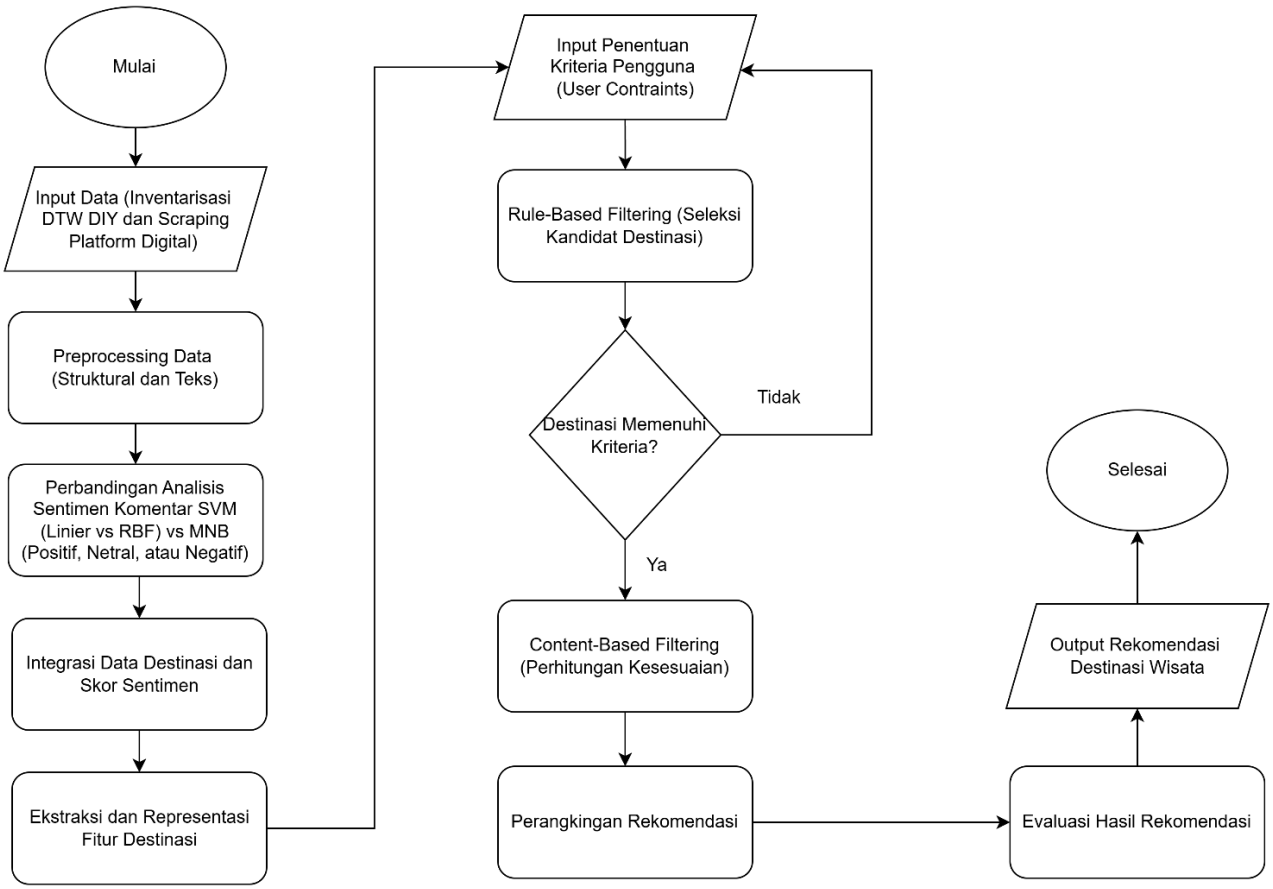
Sementara, data komentar yang diperoleh dari pengunjung dikumpulkan melalui proses *website scraping* pada platform Google Maps, disesuaikan dengan kata kunci nama destinasi wisata. Proses pengambilan dilakukan pada rentang waktu penelitian yang telah ditentukan, data komentar yang didapatkan diseleksi untuk menghilangkan duplikasi, komentar yang tidak relevan, maupun tidak mengandung opini terhadap destinasi. Setelah itu, data komentar dilakukan *preprocessing* teks yang meliputi *case folding*, *cleaning*, *tokenizing*, dan lainnya. Kemudian dianalisis menggunakan metode Analisis Sentimen yang membandingkan metode Support Vector Machine (SVM) kernel Linier dan RBF dengan Multinomial Naive Bayes (MNB) untuk memperoleh skor sentimen masing-masing destinasi.

Hasil klasifikasi sentimen pada tingkat komentar selanjutnya diagregasi pada tingkat destinasi wisata, dengan skor sentimen tersebut digunakan sebagai salah satu variabel numerik dalam sistem rekomendasi. Selanjutnya, data destinasi dan hasil analisis sentimen digabungkan berdasarkan kesesuaian nama destinasi. Variabel fasilitas wisata dilakukan *scoring* 1 untuk tersedia dan 0 untuk tidak tersedia, sebagai batasan (*constraint*) dalam proses Rule-Based Filtering untuk menjamin standar kelayakan destinasi wisata. Setelah itu, seluruh data yang telah diproses disusun dalam bentuk dataset terstruktur yang disimpan dalam format spreadsheet atau software *Excel* untuk dilakukan analisis lebih lanjut, yang akan diolah menggunakan bahasa pemrograman Python di software *Visual Studio Code* untuk membangun sistem rekomendasi berbasis Content-Based Filtering dan Rule-Based Filtering (Hybrid). Model yang telah dibangun diimplementasikan dalam bentuk prototipe website sistem rekomendasi, menggunakan framework *web* berbasis Next.js sebagai *backend* serta antarmuka pengguna berbasis *website* untuk menampilkan hasil rekomendasi destinasi wisata.

4.5. Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif berbasis data sekunder dan data platform digital, dengan metode sistem rekomendasi yang mengombinasikan Rule-Based Filtering dan Content-Based Filtering, serta dengan dukungan dari analisis sentimen komentar pengunjung. Penelitian ini menggunakan pendekatan sensus, yaitu seluruh data destinasi wisata yang tersedia pada data inventarisasi (DTW) Daerah Istimewa Yogyakarta (DIY). Dan tidak melakukan survey atau eksperimen, sehingga tidak melakukan

perhitungan besar sampel secara statistik. Hasil dari pemodelan sistem rekomendasi diimplementasikan dalam bentuk prototipe sistem sebagai media simulasi rekomendasi destinasi wisata. Alur metodologi penelitian secara rinci disajikan dalam diagram alir pada Gambar 4.1.



Gambar 4.1 Diagram Alir Penelitian

BAB V
HASIL DAN PEMBAHASAN

5.1 Pengumpulan Dataset

Tahap pengumpulan dataset merupakan langkah awal dalam melakukan penelitian sistem rekomendasi pariwisata berbasis hybrid filtering dan analisis sentimen. Proses ini mencakup pengumpulan informasi destinasi serta ulasan pengguna sebagai sumber informasi utama dalam membangun sistem rekomendasi. Data yang dikumpulkan meliputi data struktural destinasi wisata dan data teks berupa opini pengguna yang bersumber dari berbagai platform digital.

Data struktural destinasi wisata digunakan untuk merepresentasikan karakteristik masing-masing destinasi wisata, sedangkan ulasan pengguna untuk memperoleh informasi mengenai opini publik terhadap destinasi wisata melalui analisis sentimen. Pengumpulan data tersebut dilakukan secara bertahap agar data memiliki struktur yang sesuai untuk membangun sistem rekomendasi berbasis hybrid filtering dan analisis sentimen.

5.1.1 Pengumpulan Data Destinasi Wisata

Penelitian ini berfokus pada destinasi wisata yang berada di wilayah Daerah Istimewa Yogyakarta (DIY), yang meliputi Kota Yogyakarta, Kabupaten Sleman, Kabupaten Bantul, Kabupaten Gunungkidul, dan Kabupaten Kulon Progo. Pemilihan wilayah berdasarkan tingginya aktivitas sektor pariwisata di DIY dan keberagaman jenis destinasi wisata yang tersedia, seperti wisata alam, taman rekreasi, sejarah, dan budaya. Keberagaman tersebut menjadi faktor penting pengembangan sistem rekomendasi yang memungkinkan proses pemilihan preferensi pengguna terhadap karakteristik destinasi wisata yang berbeda-beda.

Dataset pada penelitian ini menggunakan data pariwisata yang diambil dari Dinas Pariwisata provinsi Daerah Istimewa Yogyakarta (DIY), yang dilakukan secara langsung di kantor dinas. Serta, melakukan scraping pada platform digital seperti Google Maps, Google *Search*, Instagram, Tiktok, dan WhatsApp untuk menentukan beberapa variabel penting dalam sistem.



	Destinasi	12. Jenis Wisata	Sub Kategori	13. Daya Tarik Ut	14. Daya Tarik Pe	15. Event Wisata	16. Kapasitas Ob	17. Jumlah Wisat	18. Loke
2	Embung Jlamprong	Taman Rekreasi	Taman Air/Tirta	Embung di dataran	Kuliner	-	+-500 org	500 org/tahun	Tidak Ad
3	Hutan Mangrove Kadilangu	Alam	Bahari	Hutan Mangrove ya	Jembatan Merah	-	-	109.500/ tahun (da	Ada
4	Air Terjun Curug Sidoharjo	Alam	Tirta	Air Terjun	Dekat dengan embu	-	-	Tidak Menentu	Tidak Ad
5	Grojogan Watu Purba	Alam	Tirta	Air terjun buatan ya	Pemandangan alam	-	500	1.200-2.000 orang	Ada
6	Air Terjun Grojogan Sewu	Alam	Tirta	Air terjun dengan a	Tidak ada	tidak ada	20 orang	tidak teridentifikasi	Tidak Ad
7	Air Terjun Luweng Sampang	Alam	Tirta	Air terjun dengan formasi bebatuan uni		-	-	-	Tidak Ad
8	Eks. Pengepul Nila Bulurejo	Sejarah		Bangunan eks Peng	Pesanggrahan Paku	-	Kurang lebih 100 orang		Gratis
9	Air Terjun Banyunibo	Alam	Tirta	Air terjun tersembu	Pemandangan alam	-	-	-	Tidak Ad
10	Gunung Batur	Alam	Gua	Formasi pegunungai	Mercusuar	-	100	kurang lebih 120	Tidak Ad
11	Bukit Mintorogo	Alam	Pegunungan	Bukit Mintorogo de	Cagar budaya rerur	Ritual Jumat kliwon	500	+- 150 orang ketika	Tidak Ad
12	Gua Tapan	Alam	Gua	Gua alam yang mas	-	-	-	TIDAK MENENTU	Tidak Ad
13	Air Terjun Kedung Pedut	Alam	Tirta	Air terjun bertingka	kolam pemandian	padusan menjelang	±500	±1000/bulan	Ada
14	Bendungan Kamijoro	Taman Rekreasi	taman air/tirta	Bendungan yang m	Kuliner, Spot Manci	-	Kurang lebih 1.000	-	Tidak Ad
15	Bukit Wisata Pule Payung	Alam	Pegunungan	Wisata alam denga	Kampung wisata ec	-	2.000 orang	13.300 orang/tahur	Ada
16	Pantai Mbongosan	Alam	Bahari	Pantai yang tersem	Spot memancing	-	-	-	Tidak Ad
17	Geoforest Watu Payung Turunan	Alam	Pegunungan	Formasi batuan uni	-	-	500 orang	-	Ada
18	Gua senen	Alam	Gua	Gua yang memiliki	-	-	50-80	-	Ada
19	He Ha Ocean View	Taman Rekreasi	Taman Rekreasi Ke	Restoran dan kafe c	Jeep Adventure, Ou	Ocean Culture/ Tar	2.000	365.000 orang/tah	Ada
20	Air Terjun Jurug Gede	Alam	Tirta	Air terjun bertingka	Pemandangan gard	Kenduri Bumi	30	180/BULAN	Ada
21	Pantai Kayu Arum	Alam	Bahari	Pantai	Camping ground	-	+-500 org	+-500-1.000 org/t	Ada
22	Gua Grubug	Alam	Gua	Gua	-	-	-	-	Tidak Ad
23	Pantai Bugel	Alam	Bahari	Pantai dengan ham	-	-	-	-	Tidak Ad
24	Pantai Midodaren	Alam	Bahari	Pantai	Camping ground	-	+-1.000 org	+-50 org/bulan	Ada
25	Air Terjun Kembang Soka	Alam	Tirta	Air terjun berundak	kolam pemandian	-	±300	300/bulan	Ada
26	Bunga Chrysant	Taman Rekreasi	Agrowisata	Hamparan perkebu	kulineran	-	200	kurang lebih 360	Tidak Ad

Gambar 5.1 Database Inventarisasi DTW DIY 2025

Pada Gambar 5.1 di atas, merupakan database inventarisasi Daerah Tujuan Wisata (DTW) Daerah Istimewa Yogyakarta (DIY) tahun 2025 yang didapatkan dari Dinas Pariwisata DIY. Data tersebut sebagai acuan untuk mendapatkan variabel penting yang akan digunakan dan scraping selanjutnya, terutama pada variabel nama-nama destinasi yang menjadi bagian penting. Selain itu, variabel yang digunakan lainnya yaitu, jenis wisata, daya tarik utama, area toilet, ruang ibadah, tempat sampah, tempat istirahat pengunjung, area parkir, tempat makan/kafetaria, dan cctv.

Data destinasi wisata yang dikumpulkan melalui scraping meliputi rating maps, latitude, longitude, harga tiket weekday, harga tiket weekend, harga parkir motor, harga parkir mobil, mode transportasi motor, mode transportasi mobil, jam operasional weekday, dan jam operasional weekend destinasi wisata. Variabel-variabel tersebut dipilih karena dianggap dapat merepresentasikan faktor-faktor yang dipertimbangkan oleh pengguna dalam memilih destinasi wisata.

Variabel jenis wisata untuk menggambarkan kategori utama destinasi wisata, seperti wisata alam, budaya, sejarah, dan taman rekreasi. Variabel tersebut menjadi salah satu fitur utama dalam Content Based Filtering karena memiliki pengaruh besar terhadap preferensi pengguna. Untuk variabel harga tiket weekday dan harga tiket weekend merepresentasikan tingkat keterjangkauan destinasi wisata, pemisahan harga dikarenakan beberapa destinasi menerapkan harga yang berbeda berdasarkan hari kunjungan. Sedangkan variabel rating sebagai indikator kualitas umum destinasi dari penilaian pengguna. Sehingga sistem rekomendasi memberikan rekomendasi yang realistis atau sesuai dengan kondisi aktual.



	Destinasi	Jenis Wisata	Rating Maps	Latitude	Longitude	Daya Tarik Utama	Harga Tiket Weekd	Harga Tiket Weeke	Parkir Motor	Parkir Mobil	Jam Buka Weekday	Jam Buka Weekend	Toilet	Ter
2	Embung Nglanggeran	Taman Rekreasi	4.6	-7.845785995	110.5467331	Embung/danau bu	Rp15.000	Rp15.000	Rp2.000	Rp5.000	08.00-17.00	00.00-23.59	Ada	Ad
3	Grand Puri Waterpark	Taman Rekreasi	4.4	-7.8700729	110.3524795	Taman rekreasi air	Rp30.000	Rp40.000	Rp3.000	Rp5.000	09.00-17.00	08.00-17.00	Tidak Ada	Tid
4	Embung Ilamprong	Taman Rekreasi	4.4	-8.009973699	110.6768828	Embung di dataran	Rp0	Rp0	Rp0	Rp0	00.00-23.59	00.00-23.59	Ada	Ad
5	Taman Jamu Naturindo	Taman Rekreasi	4.6	-7.826493864	110.1520153	Koleksi Tanaman O	Rp0	Rp0	Rp3.000	Rp5.000	07.30-15.30	07.30-12.30	Ada	Ad
6	Taman Tebing Breksi	Taman Rekreasi	4.5	-7.78110343	110.5045843	Tebing bekas tamb	Rp10.000	Rp15.000	Rp2.000	Rp5.000	06.00-20.00	06.00-20.00	Ada	Ad
7	Ekowisata Nologaten	Taman Rekreasi	4.7	-7.77453276	110.398615	Suasanan area pen	Rp0	Rp0	Rp3.000	Rp5.000	09.00-22.00	09.00-22.00	Ada	Ad
8	Jogja Mini Zoo Exotarium	Taman Rekreasi	4.5	-7.727732472	110.3584692	Kebun Binatang dir	Rp25.000	Rp25.000	Rp3.000	Rp5.000	08.00-17.00	08.00-17.00	Ada	Ad
9	Embung Batara Sriten	Taman Rekreasi	4.5	-7.838024558	110.537942	Embung (danau bu	Rp5.000	Rp5.000	Rp3.000	Rp5.000	05.00-18.00	05.00-18.00	Ada	Ad
10	Lembah Sorony	Taman Rekreasi	4.5	-7.945050528	110.3657413	kolam renang, taur	Rp6.000	Rp6.000	Rp3.000	Rp5.000	08.00-18.00	06.00-17.00	Ada	Ad
11	Bejiharjo Edupark	Taman Rekreasi	4.4	-7.935385202	110.6.362378	Wisata edukasi yan	Rp150.000	Rp150.000	Rp3.000	Rp5.000	00.00-23.59	00.00-23.59	Ada	Ad
12	The Mangtlung View	Taman Rekreasi	4.4	-7.84476338	110.4825423	View Jogja	Rp0	Rp0	Rp3.000	Rp5.000	15.00-22.00	12.00-22.00	Ada	Ad
13	Bendungan Kamijoro	Taman Rekreasi	4.3	-7.878962136	110.265559	Bendungan yang m	Rp5.000	Rp5.000	Rp2.000	Rp5.000	07.00-18.00	07.00-18.00	Ada	Tid
14	Wisata Bojong Asri	Taman Rekreasi	4.8	-8.008575974	110.295505	Kuliner, outbound &	Rp15.000	Rp15.000	Rp3.000	Rp5.000	00.00-23.59	00.00-23.59	Ada	Ad
15	Gamplong Studio Alam Sleman	Taman Rekreasi	4.5	-7.805229621	110.2367553	Studio pembuatan	Rp0	Rp0	Rp2.000	Rp5.000	09.00-17.00	09.00-17.00	Ada	Ad
16	Ledok Sambi Ecopark	Taman Rekreasi	4.7	-7.646882269	110.428427	Area rekreasi alam	Rp0	Rp0	Rp3.000	Rp5.000	09.00-16.30	09.00-16.30	Ada	Ad
17	De Mangol View	Taman Rekreasi	4.5	-7.83732818	110.4897886	Spot kuliner dan fot	Rp20.000	Rp20.000	Rp3.000	Rp5.000	10.00-21.00	10.00-21.00	Tidak Ada	Tid
18	Waduk Sermo	Taman Rekreasi	4.6	-7.82397576	110.1262868	Waduk besar yang	Rp30.000	Rp30.000	Rp2.000	Rp5.000	00.00-23.59	00.00-23.59	Tidak Ada	Tid
19	Suraloka Interactive Zoo	Taman Rekreasi	4.5	-7.605468375	110.4294007	Kebun binatang inte	Rp45.000	Rp45.000	Rp3.000	Rp5.000	09.00-17.00	09.00-17.00	Ada	Ad
20	Obelix Village	Taman Rekreasi	4.5	-7.686866695	110.3728014	Kombinasi wisata a	Rp25.000	Rp30.000	Rp3.000	Rp5.000	10.00-21.00	08.00-21.00	Ada	Ad
21	Galaxy Waterpark Jogja	Taman Rekreasi	4.4	-7.815290221	110.413616	Wisata air (waterpa	Rp40.000	Rp45.000	Rp3.000	Rp5.000	09.00-16.00	08.00-16.00	Ada	Ad
22	He Ha Sky View	Taman Rekreasi	4.5	-7.849328351	110.478214	Restoran dan kafe	Rp30.000	Rp35.000	Rp3.000	Rp5.000	10.00-21.00	08.00-21.00	Ada	Ad
23	Jogja Agro Park	Taman Rekreasi	4.5	-7.78594228	110.2213618	Edukasi pertanian	Rp0	Rp0	Rp3.000	Rp5.000	06.00-18.00	06.00-18.00	Ada	Ad
24	Embung Tonogoro	Taman Rekreasi	4.5	-7.658081737	110.2342457	Embung di atas buk	Rp3.000	Rp3.000	Rp3.000	Rp5.000	08.00-18.00	08.00-18.00	Ada	Ad
25	Stone Valley	Taman Rekreasi	4.5	-8.150659226	110.6265728	Tidak Ada	Rp20.000	Rp20.000	Rp3.000	Rp5.000	10.00-18.00	10.00-18.00	Tidak Ada	Tid
26	Agrowisata Bhumi Merapi	Taman Rekreasi	4.5	-7.63948906	110.4262867	Spot Foto berupa re	Rp30.000	Rp35.000	Rp3.000	Rp5.000	09.00-17.00	09.00-17.00	Ada	Ad
27	The World Landmark Merapi Park	Taman Rekreasi	4.3	-7.620362954	110.4215207	Taman bunga, Repl	Rp35.000	Rp35.000	Rp2.000	Rp5.000	09.00-16.00	09.00-16.00	Ada	Ad
28	Kampung Emas	Taman Rekreasi	4.5	-7.83483355	110.5430643	Kuliner	Rp35.000	Rp35.000	Rp3.000	Rp5.000	10.00-17.00	10.00-17.00	Ada	Ad
29	Taman Pintar	Taman Rekreasi	4.5	-7.79953726	110.367775	Zona Interaktif Bert	Rp35.000	Rp35.000	Rp3.000	Rp5.000	08.30-16.00	08.30-16.00	Ada	Ad
30	Embung Kleco	Taman Rekreasi	4.1	-7.77718640	110.16.352277	Embung (danau bu	Rp3.000	Rp3.000	Rp3.000	Rp5.000	05.30-19.00	05.30-19.00	Ada	Ad
31	Gembira Loka Zoo	Taman Rekreasi	4.6	-7.803699398	110.398076	Koleksi Satwa yang	Rp70.000	Rp85.000	Rp3.000	Rp5.000	08.00-16.00	08.00-16.00	Ada	Ad
32	Kalilurang Park Botanical Garden	Taman Rekreasi	4.7	-7.597688341	110.4263087	Alam, Koleksi flora	Rp25.000	Rp25.000	Rp3.000	Rp5.000	09.00-17.00	07.00-17.00	Ada	Ad
33	Satoloka	Taman Rekreasi	4.3	-7.62379286	110.4250642	Kebun binatang mir	Rp30.000	Rp30.000	Rp3.000	Rp5.000	09.00-17.00	09.00-17.00	Tidak Ada	Tid
34	Drini Park	Taman Rekreasi	4.4	-8.13513116	110.5770648	Taman rekreasi kelu	Rp25.000	Rp35.000	Rp3.000	Rp5.000	08.00-18.00	08.00-18.00	Ada	Ad
35	Telaga Kemuning	Taman Rekreasi	4.4	-7.895943761	110.5201955	Telaga alami denga	Rp2.000	Rp2.000	Rp3.000	Rp5.000	07.00-20.00	07.00-20.00	Ada	Ad
36	Omah Salak	Taman Rekreasi	4.2	-7.644594964	110.37105	Wisata Edukasi Pert	Rp10.000	Rp10.000	Rp3.000	Rp5.000	08.00-16.00	08.00-16.00	Ada	Ad
37	HEHA Forest	Taman Rekreasi	4.5	-7.62522254	110.4243221	Wisata kuliner dan	Rp20.000	Rp25.000	Rp3.000	Rp5.000	11.00-20.00	10.00-20.00	Ada	Ad

Gambar 5.2 Dataset Inventarisasi

Gambar 5.2 di atas, merupakan agregasi tampilkan dataset yang telah dikumpulkan. Dataset berjumlah 315 entri, yang berisi data dari berbagai destinasi pariwisata yang ada di wilayah Yogyakarta dengan memperlihatkan struktur data. Setiap destinasi memiliki karakteristik yang berbeda sehingga dapat digunakan dalam proses pencocokan preferensi pengguna (similarity) pada sistem rekomendasi.

5.1.2 Pengumpulan Data Ulasan Pengguna

Selain menggunakan data destinasi wisata, penelitian ini juga menggunakan data ulasan pengguna sebagai sumber data analisis sentimen. Data tersebut digunakan untuk memetakan tren opini publik pada tiap destinasi wisata. Pendekatan ini memungkinkan sistem untuk mengintegrasikan pengalaman serta persepsi riil pengguna, tetapi tetap mempertimbangkan karakteristik bawaan dari destinasi tersebut.

GOOGLE MAPS

No	Destinasi	Link
1	Embung Ilamprong	https://maps.app.goo.gl/9qyZ9Dxj6P4WHvHN9
2	Hutan Mangrove Kadilangu	https://maps.app.goo.gl/1mHQmA4wxGtPUdjj8
3	Air Terjun Curug Sidoharjo	https://maps.app.goo.gl/BEApte5PDLRanhSo6 https://maps.app.goo.gl/CzSvFZ7y4dcdkSbu7 embung canggal maps
4	Grojogan Watu Purba	https://maps.app.goo.gl/uKqDs2C59mh1ifiu5
5	Air Terjun Grojogan Sewu	https://maps.app.goo.gl/LcKZcm7bijS3D19m8
6	Air Terjun Luweng Sampang	https://maps.app.goo.gl/NzTokPjN5a1XaWHZ8
7	Eks. Pengepul Nila Bulurejo	https://maps.app.goo.gl/DVMJV5GHb9Wzmql69
8	Air Terjun Banyunibo	https://maps.app.goo.gl/nskSaPZewXzHXDeF8
9	Gunung Batur (5)	https://maps.app.goo.gl/RVGF61xQJRtnZKpU8
10	Bukit Mintorogo	https://maps.app.goo.gl/quPCtGERQeBkTwaf8
11	Gua Tapan/Seluman	https://maps.app.goo.gl/LUE2z5gfSNdAM9tr8
12	Air Terjun Kedung Pedut	https://maps.app.goo.gl/EFGp9V4zfeU8n2Jt5

Gambar 5.3 Data Link Google Maps

Pada Gambar 5.3, menunjukkan daftar link dari platform digital yaitu Google Maps, yang akan dilakukan sebagai langkah pertama untuk *scraping* komentar dan penilaian pengguna. Data ulasan dikumpulkan dan menggunakan teknik *scraping* untuk memperoleh data dalam jumlah besar secara lebih efisien dan terstruktur, kemudian diperoleh *output* pada Gambar 5.4.

	title	url	stars	name	reviewUrl	text
2	Air Terjun	https://w	2	Kinanthi M	https://w	Saya tidak sampai di air terjunnya, karena akses jalan yang cukup menanjak dan sempit
3	Air Terjun	https://w	5	Antoni Ab	https://w	Syahdu dan alami
4	Air Terjun	https://w	5	Widuri Wi	https://www.google.com/maps/reviews/data=!4m8!14m7!1m6!2m5!1sChdDSUhNMG9nS0VJQ0FnSUMyenNLeG1BRRAB!2m1!1s0x0:0x
5	Air Terjun	https://w	3	Dipa Wija	https://w	Air terjun nya belum mengalir, masih kering
6	Air Terjun	https://w	5	Dian Fitria	https://w	salah satu objek wisata seru nih, wajib banget sih ke sini 👍😍
7	Air Terjun	https://w	4	Dayat Day	https://w	Akses untuk gowes ada tapi tidak terawat
8	Air Terjun	https://w	5	ari nekno	https://w	Alami, nyaman, sejuk,,
9	Air Terjun	https://w	5	Saryoto FF	https://w	Air terjun alami di puncak Dusun Batur. Batu besar ini masih 1 kawasan
10	Air Terjun	https://w	1	Iwan_kocl	https://www.google.com/maps/reviews/data=!4m8!14m7!1m6!2m5!1sChdDSUhNMG9nS0VJQ0FnSURTcDdQS19RRRAB!2m1!1s0x0:0x7
11	Air Terjun	https://w	5	tyas zahra	https://w	Sip
12	Air Terjun	https://w	5	Hasan Ash	https://www.google.com/maps/reviews/data=!4m8!14m7!1m6!2m5!1sChdDSUhNMG9nS0VJQ0FnSUNDNjhUOHI3RRAB!2m1!1s0x0:0x7
13	Air Terjun	https://w	5	Zhidane A	https://w	Sakjane apik..tapi kurang danane Yo gx apik
14	Air Terjun	https://w	5	Sigit Ardia	https://w	Tempatnya bagus nan asri, namun masih agak terpencil dan agak susah aksesnya Debit airnya tergantung musim, saat m
15	Air Terjun	https://w	1	Luqqman	https://www.google.com/maps/reviews/data=!4m8!14m7!1m6!2m5!1sChdDSUhNMG9nS0VJQ0FnSUNBa19uWDFRRRAB!2m1!1s0x0:0x
16	Air Terjun	https://w	5	Soeryanto	https://www.google.com/maps/reviews/data=!4m8!14m7!1m6!2m5!1sChdDSUhNMG9nS0VJQ0FnSUNjN00zZTlRRRAB!2m1!1s0x0:0x7e
17	Air Terjun	https://w	5	Ben Ina	https://www.google.com/maps/reviews/data=!4m8!14m7!1m6!2m5!1sChdDSUhNMG9nS0VJQ0FnSURzak1TYXpnRRAB!2m1!1s0x0:0x7
18	Air Terjun	https://w	4	Purwaning	https://www.google.com/maps/reviews/data=!4m8!14m7!1m6!2m5!1sChZDSUhNMG9nS0VJQ0FnSUNzaDhETlJnEAE!2m1!1s0x0:0x7ef7
19	Air Terjun	https://w	5	Nor fatikh	https://w	Yang suka hiking dan sekaligus suka berenang cocok banget dateng kesini karena selain menyuguhkan pesona alam yang ind
20	Air Terjun	https://w	4	Danen Jim	https://w	Pertama kali kesini sekitar 3-4 tahun yg lalu..masih sepi bgt
21	Air Terjun	https://w	4	Evi Widias	https://w	Lumayan bagus, tp jalan nya jauh..
22	Air Terjun	https://w	3	Buku Hari	https://w	Bagus buat yang suka air
23	Air Terjun	https://w	5	Waone Ra	https://w	Airterjun ini bagus dan sekaligus unik dengan adanya sebuah batu yg mungkin di pahat menyerupai bentuk wajah manusia ,
24	Air Terjun	https://w	1	Alvina	https://www.google.com/maps/reviews/data=!4m8!14m7!1m6!2m5!1sChdDSUhNMG9nS0VJQ0FnSURJNXJHU3ZBRRAB!2m1!1s0x0:0x7
25	Air Terjun	https://w	5	fajar ajay	https://www.google.com/maps/reviews/data=!4m8!14m7!1m6!2m5!1sChZDSUhNMG9nS0VJQ0FnSURJcG9QQ0VBEAE!2m1!1s0x0:0x7e

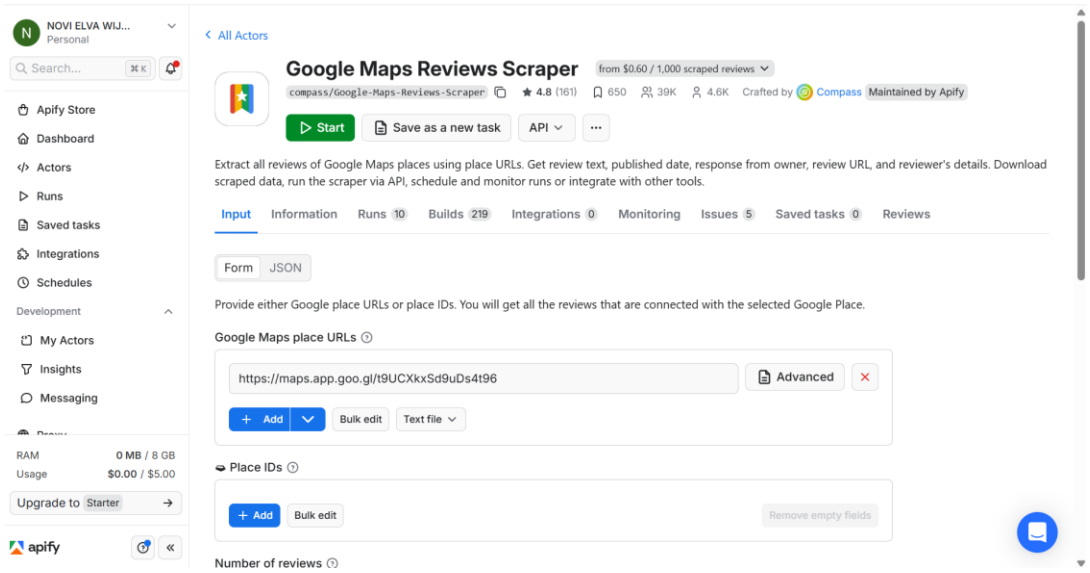
Gambar 5.4 Data Hasil Scraping

Data ulasan yang dikumpulkan berupa komentar pengguna mengenai pengalaman mereka ketika mengunjungi destinasi wisata tertentu. Komentar tersebut memiliki variasi opini, baik positif, netral, maupun negatif. Oleh karena itu, diperlukan proses analisis sentimen untuk mengidentifikasi kecenderungan opini pengguna terhadap masing-masing destinasi wisata.

5.1.3 Proses Scraping Data

Proses scraping data dilakukan untuk mengumpulkan data ulasan pengguna serta informasi destinasi wisata secara otomatis dari platform digital. Metode ini dipilih karena kemampuannya dalam menghimpun data berskala besar dengan waktu yang singkat dan efisien daripada proses manual, sehingga lebih efisien untuk menghimpun data dalam jumlah besar dibandingkan metode manual.

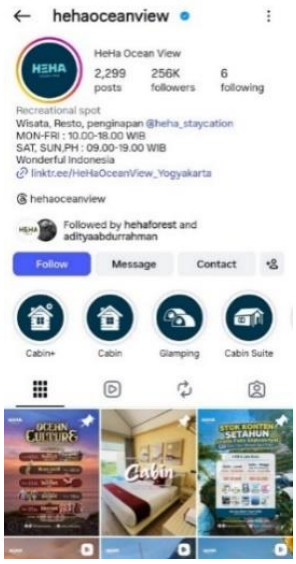
Proses ini menggunakan bahasa pemrograman Python dengan bantuan beberapa pustaka (*library*) pendukung, dengan menggunakan link berikut https://apify.com/compass/google-maps-reviews-scraper. Langkah awal dimulai dengan mengakses halaman destinasi wisata pada Google Maps berdasarkan nama destinasi yang telah ditentukan sebelumnya, kemudian sistem melakukan proses pengambilan data umum destinasi wisata, seperti nama destinasi, link destinasi, rating, nama pengulas komentar, link komentar, dan komentar pengguna.



Gambar 5.5 *Website* Scraping

Pada proses pengambilan ulasan pengguna, sistem melakukan scrolling halaman secara otomatis untuk memastikan komentar dapat diambil secara lengkap, data yang berhasil diperoleh kemudian akan disimpan ke dalam dataset sementara sebelum digabungkan dengan seluruh destinasi dan dilakukan proses pembersihan data.

Tahap pembersihan data dilakukan untuk memastikan kualitas dataset dalam penelitian. Proses pembersihan meliputi penghapusan data kosong, penyesuaian format data, penghapusan semua variabel kecuali data destinasi dan komentar, serta pengecekan missing value agar data lebih konsisten dan siap di proses pada tahap selanjutnya. Selanjutnya data dikonversi dalam bentuk dataframe, yang bertujuan untuk mempermudah proses pengolahan data, preprocessing, analisis sentimen, dan integrasi dataset secara keseluruhan.

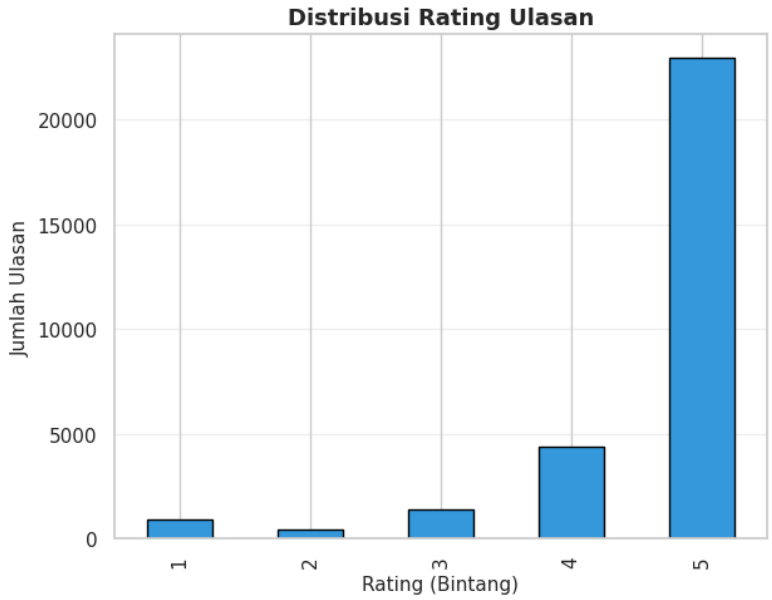


Gambar 5.6 Salah Satu Platform Digital Scraping

Selain scraping untuk ulasan pengguna, peneliti juga melakukan scraping untuk variabel yang akan digunakan tetapi tidak terdapat di database inventarisasi DTW DIY 2025. Proses scraping tersebut dilakukan secara manual satu persatu pada platform digital seperti Google Maps, Google *Search*, Instagram, TikTok, dan WhatsApp. Dengan variabel yang dilakukan, diantaranya rating umum destinasi, longitude, latitude, harga tiket weekday, harga tiket weekend/hari libur nasional, harga parkir motor, harga parkir mobil, jam buka/operasional weekday dan weekend.

5.2 *Exploratory Data Analysis* (EDA)

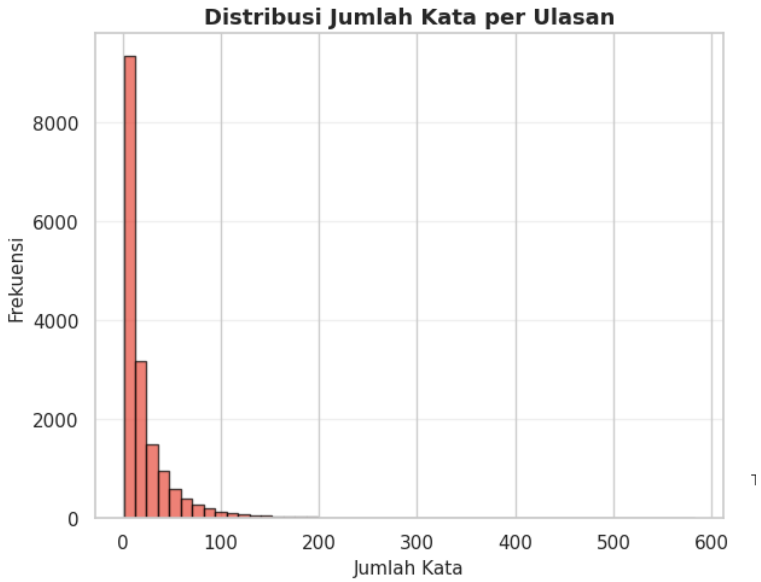
Exploratory Data Analysis (EDA) merupakan langkah awal dalam analisis data yang digunakan untuk memahami karakteristik *dataset* sebelum melakukan *preprocessing* pada analisis sentimen dan pemodelan. Langkah ini dapat memberikan gambaran terhadap distribusi data, pola distribusi yang muncul, ketidakseimbangan, karakteristik teks, ketersediaan data setiap destinasi, dan identifikasi permasalahan pada data yang dapat mempengaruhi hasil analisis sentimen. Hasil EDA kemudian menjadi dasar dalam menentukan perlakuan terhadap data dan memastikan metode yang digunakan sesuai dengan karakteristik dataset. Hal tersebut dilakukan terhadap beberapa aspek, diantaranya distribusi rating ulasan, distribusi jumlah kata pada setiap ulasan, dan jumlah ulasan pada masing-masing destinasi wisata.



Gambar 5.7 Distribusi Rating Ulasan

Pada Gambar 5.7, diketahui bahwa distribusi rating ulasan pengguna terhadap destinasi wisata didominasi oleh rating dengan bintang 5 dengan jumlah yang berbeda sangat signifikan daripada rating 1, 2, 3, dan 4. Kondisi tersebut menunjukkan bahwa sebagian besar pengguna memberikan penilaian positif terhadap destinasi wisata yang mereka kunjungi. Tingginya frekuensi rating 5 mencerminkan kepuasan mayoritas pengunjung terhadap berbagai aspek destinasi wisata, seperti fasilitas, kebersihan, pelayanan, maupun daya tarik wisata yang tersedia. Namun, distribusi rating tersebut tidak seimbang dengan kecenderungan ulasan positif, sehingga berpotensi ketidakseimbangan kelas (class imbalance) pada analisis sentimen apabila menjadi acuan awal dalam memahami karakteristis data. Oleh karena itu, dilakukan pelabelan sentimen tidak berdasarkan rating, tetapi pendekatan InSet Lexicon sehingga setiap komentar diklasifikasikan berdasarkan kandungan kata positif dan negatif pada ulasan.

Selanjutnya, Gambar 5.8 menunjukkan sebagian besar komentar terdiri atas jumlah kata yang *relative* sedikit. Grafik tersebut menggambarkan bahwa pengguna lebih sering memberi komentar yang ringkas dan menyampaikan ini dari pengalaman mereka selama mengunjungi destinasi, tetapi terdapat beberapa komentar dengan jumlah kata yang cukup banyak yang biasanya berisi penjelasan detail mengenai kondisi objek wisata, fasilitas yang tersedia, pelayanan selama berkunjung, maupun merekomendasikan pada pengunjung lain.



Gambar 5.8 Distribusi Jumlah Kata per Ulasan

Berdasarkan Gambar 5.9, terdapat 282 destinasi wisata yang masing-masing memiliki 100 ulasan pada dataset penelitian. Angka tersebut bukan jumlah ulasan keseluruhan di Google Maps, tetapi jumlah ulasan yang digunakan setelah proses *scraping*. Penelitian ini membatasi pengambilan data sebanyak 10 hingga 100 ulasan pada setiap destinasi sehingga setiap objek wisata memiliki jumlah data yang relatif seragam. Pembatasan digunakan untuk menjaga keseimbangan bobot data antar destinasi, mengurangi bias akibat dominasi destinasi yang memiliki ribuan ulasan, serta meningkatkan efisiensi proses komputasi pada tahap *preprocessing* dan analisis sentimen menggunakan pendekatan InSet Lexicon. Sehingga, setiap destinasi memiliki kontribusi yang lebih proporsional dalam pembentukan model analisis sentimen.

Jumlah destinasi ≥100 ulasan: 282

	title	jumlah_ulasan
322	Wunung Giriselokondo	100
0	Museum Benteng Vredeburg	100
1	Museum Dirgantara	100
2	Affandi Museum	100
3	Agro Tourism Bhumi Merapi	100
...	...	...
19	Batu Alien	100
18	Balong Waterpark	100
14	AloeLand,Rasane Vera,Aloe Station's "Pabrik Ol...	100
12	Air Terjun Sri Gethuk	100
11	Air Terjun Sidoharjo	100

282 rows × 2 columns

Gambar 5.9 Jumlah Destinasi 100 ulasan

5.3 Persiapan *Dataset*

Sebelum melakukan *preprocessing*, dilakukan tahap persiapan data untuk memastikan data yang digunakan memiliki kualitas yang baik. Tahap tersebut meliputi

pemeriksaan kondisi awal *dataset*, pembersihan data (data *cleaning*), dan penyaringan komentar berbahasa Indonesia. Hal tersebut agar data yang tidak lengkap dan tidak relevan dapat dihapus sehingga hanya data layak yang digunakan pada proses analisis sentimen.

Berdasarkan Gambar 5.10, memperlihatkan dataset hasil *scraping* Google Maps yang berisi komentar pengguna terhadap berbagai destinasi wisata di Daerah Istimewa Yogyakarta. Sebelum dilakukan pengolahan lebih lanjut, terlebih dahulu peneliti melakukan pemeriksaan terhadap kondisi awal dataset yang meliputi jumlah data, jumlah variabel, keberadaan nilai kosong (*missing value*), data duplikat, serta jumlah destinasi wisata yang berhasil dikumpulkan. Diketahui terdapat 30.071 baris data dengan 6 variabel, tetapi terdapat 13.023 *missing value*.

```
Shape: (30071, 6)
Columns: ['title', 'url', 'stars', 'name', 'reviewUrl', 'text']

📊 STATISTIK AWAL DATASET:
Jumlah baris        : 30,071
Jumlah kolom        : 6
Missing values      : 13023
Duplikat            : 0
```

Gambar 5.10 Statistik Awal *Dataset*

Setelah itu, dilakukan cleaning data untuk membersihkan dataset yang tidak layak dilakukan analisis, sehingga jumlah data berkurang secara signifikan karena sebagian besar data yang dihapus merupakan baris yang tidak memiliki isi komentar. Selain itu, sebanyak 1.139 data duplikat dihapus untuk menghindari adanya komentar yang sama muncul lebih dari satu kali. Dari proses cleaning tersebut, didapatkan 15.909 komentar.

```
Data cleaning selesai
Duplikat dihapus: 1139
Total data: 15,909
```

Gambar 5.11 Data *Cleaning*

Dikarenakan penelitian menggunakan pendekatan InSet Lexicon Bahasa Indoensia, sehingga seluruh komentar yang digunakan harus menggunakan bahasa Indonesia. Oleh karena itu, dilakukan proses identifikasi bahasa (*language detection*) untuk menyaring komentar yaang menggunakan bahasa selain bahasa Indonesia. Didapatkan sebanyak 1.508 atau 9,5% dari seluruh data dihapus karena teridentifikasi menggunakan bahasa selain bahasa Indonesia dan diperoleh 14.401 komentar yang berbahasa Indonesia yang akan digunakan sebagai data utama dalam proses *text preprocessing* dan analisis sentimen pada Gambar 5.12.

```
Filter Bahasa Indonesia selesai
Data sebelum: 15,909
Data dihapus: 1,508 (9.5%)
Data sekarang: 14,401
```

Gambar 5.12 Filter Bahasa Indonesia

5.4 Analisis Sentimen

5.4.1 *Preprocessing* Data

Tahap *preprocessing* data diterapkan untuk membersihkan dan mempersiapkan data teks sebelum dilakukan proses analisis sentiment menggunakan InSet Lexicon dan ekstraksi fitur menggunakan Term Frequency-Inverse Document Frequency (TF-IDF). Data hasil scraping umumnya masih memuat elemen yang tidak relevan, seperti simbol, huruf kapital, URL, *emoticon*, dan spasi berlebih. Oleh karena itu, dibutuhkan preprocessing agar data menjadi lebih terstruktur dan siap diproses secara efisien oleh model klasifikasi.

Pada penelitian ini, proses preprocessing data dilakukan menggunakan satu fungsi utama (*pipeline preprocessing*) yang menggabungkan beberapa tahapan pengolahan teks secara berurutan, diantaranya *case folding, cleaning, tokenizing*, normalisasi kata, *stopword removal,* dan *stemming*. Penggabungan seluruh tahapan ke dalam satu fungsi dilakukan untuk menjaga konsistensi perlakuan terhadap seluruh ulasan yang dianalisis dalam penelitian ini.

```
✓ Preprocessing selesai
  Dokumen kosong dihapus: 87
  Total dokumen: 14,327

📝 CONTOH TEXT SEBELUM & SESUDAH PREPROCESSING:

Sampel 1:
  Original: yikes!...
  Cleaned:  yikes...

Sampel 2:
  Original: Spot asik buat mancing tpi sayang kemarin ombak lagi besar...
  Cleaned:  spot asik buat mancing tpi sayang kemarin ombak besar...

Sampel 3:
  Original: Tempatnya enak buat nyantai apalagi kalau gak ada orang lain.
udaranya...
  Cleaned:  tempat enak buat nyantai kalau tidak orang udara banyak orang bakar sa...

Sampel 4:
  Original: Mantabb.. Deket banget dari rumah n masih sepii pas dateng kesana. Nex...
  Cleaned:  mantabb deket banget rumah masih sepi pas dateng kesana next time kala...

Sampel 5:
  Original: Area pantai berpasir putih dan ombak yang kalem sangat rekomen untuk b...
  Cleaned:  area pantai pasir putih ombak kalem sangat rekomen main sama anak kelu...
```

Gambar 5.13 *Preprocessing Text*

Berdasarkan Gambar 5.13, menunjukkan bahwa *preprocessing text* berhasil menghapus dokumen kosong sebanyak 87 data dan menyederhanakan struktur kalimat

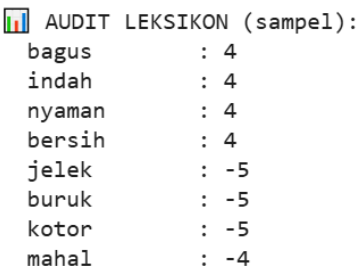
tanpa menghilangkan informasi utama yang terkandung dalam ulasan pengguna. Seperti pada sampel 3, setelah diproses kata-kata tersebut berubah menjadi huruf kecil, kata tidak baku “gak” menjadi “tidak”, dan kata yang dianggap tidak memberikan informasi penting dihilangkan melalui proses stopword removal.

Pada sampel lain juga terlihat bahwa kata yang bersifat umum, tanda baca, maupun karakter yang tidak relevan juga dihapus sehingga teks menjadi lebih ringkas tapi tetap mempertahankan informasi utama. Serta, penggunaan kata tidak baku telah dinormalisasikan, sehingga menghasilkan representasi teks yang lebih seragam dan konsisten. Kemudian, hasil preprocessing tersebut digunakan sebagai *input* utama dalam proses pelabelan sentimen menggunakan metode InSet Lexicon untuk menentukan bobot atau skor sentimen dari masing-masing ulasan.

5.4.2 Pelabelan Sentimen Menggunakan InSet Lexicon

Setelah dilakukan tahap *preprocessing*, tahap berikutnya yaitu melakukan pelabelan sentimen. Pada penelitian ini pelabelan dilakukan menggunakan pendekatan Lexicon Based dengan memanfaatkan InSet Lexicon, sebuah kamus sentimen bahasa Indonesia yang disusun oleh Koto dan Rahman (2017). Penelitian ini menggunakan pelabelan manual yang mengandalkan penilaian manusia yang ditulis *custom* dan dipadukan dengan pendekatan yang menentukan polaritas sentimen berdasarkan skor setiap kata yang terdapat dalam kamus Lexicon.

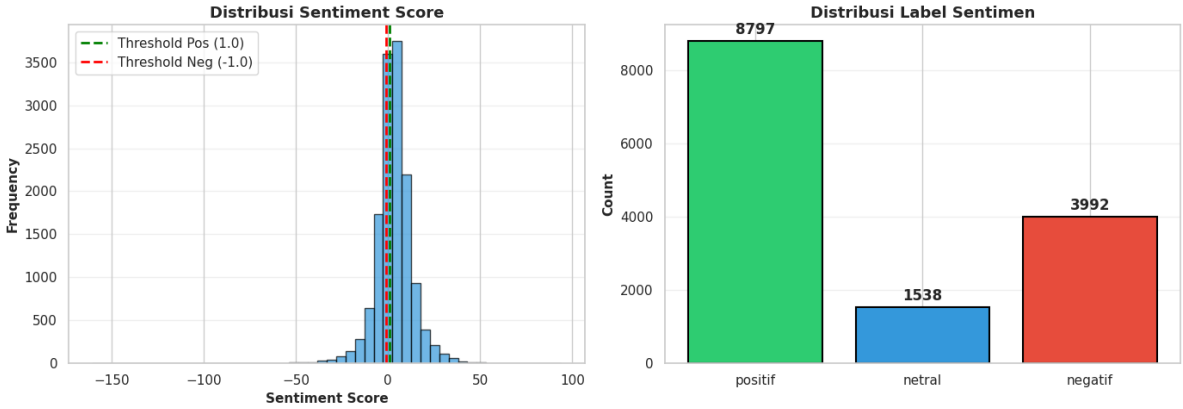
Setiap kata pelabelan manual dan pada InSet Lexicon memiliki bobot sentimen positif maupun negatif. Kata dengan label positif diberikan skor positif, dan sebaliknya. Sebagai contoh pada Gambar 5.14, dilakukan audit terhadap sebagian isi kamus untuk memastikan bahwa proses pembacaan berhasil dilakukan sebelum digunakan pada proses pelabelan, seperti kata bagus, indah, dan nyaman memiliki skor positif, sedangkan kata jelek, buruk, dan kotor memiliki skor *negative*.



Gambar 5.14 Audit Lexicon

Perhitungan skor sentimen dilakukan dengan menjumlahkan seluruh skor kata yang terdapat pada setiap ulasan, semakin besar jumlah skor positif maka kecenderungan

setimen semakin positif, begitupun sebaliknya. Dan penelitian ini dilakukan secara otomatis terhadap seluruh ulasan yaitu 14.401 yang telah melalui tahap *preprocessing*, sehingga didapatkan skor sentimen untuk setiap ulasan. Kemudian, dilakukan penentuan kelas sentimen berdasarkan nilai skor sentimen dengan label positif (skor > 0), netral (skor $= 0$), dan negatif (skor < 0). Pendekatan tersebut dipilih karena sederhana, transparan, dan sesuai dengan karakteristik metode Lexicon Based yang tidak memerlukan data latih berlabel.



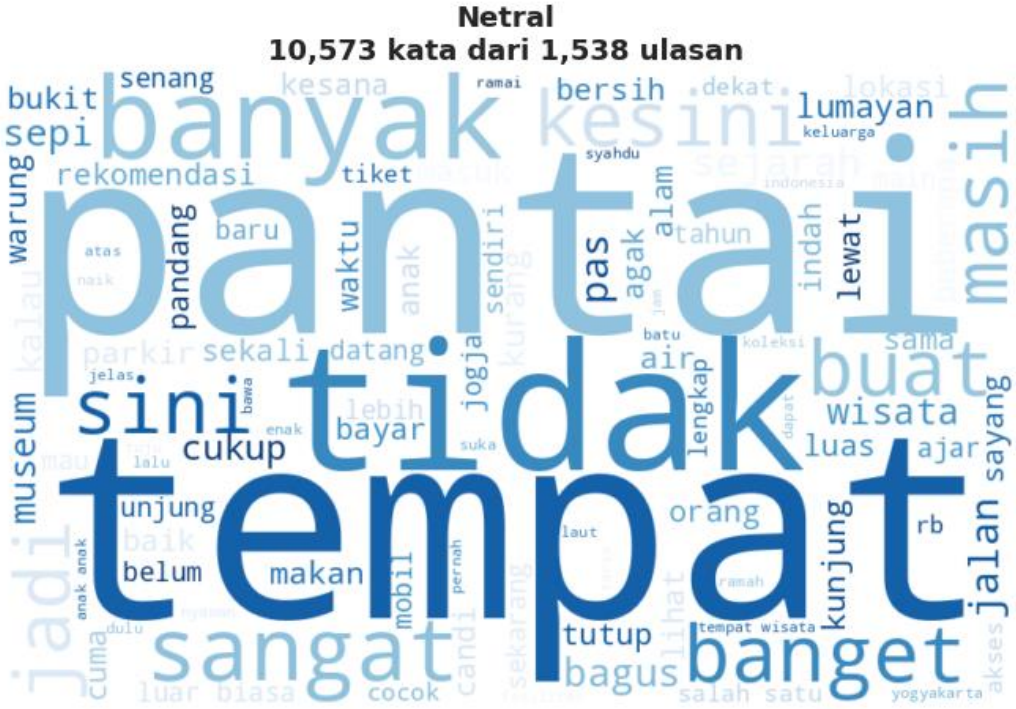
Gambar 5.15 Visualisasi Sentimen

Berdasarkan Gambar 5.15, diketahui bahwa sebagian besar ulasan termasuk kategori positif sebanyak 8.797 ulasan, kemudian sebanyak 3.992 ulasan negatif, dan sebanyak 1.538 ulasan netral. Hal tersebut menunjukkan bahwa distribusi kelas *relative* didominasi oleh dua kelas utama yaitu positif diikuti *negative*, sedangkan netral memiliki proporsi lebih sedikit. Sehingga menggambarkan adanya ketidakseimbangan distribusi kelas (*class imbalance*) yang perlu diperhatikan pada tahap pembangunan model klasifikasi.



Gambar 5.16 Word Cloud Positif

Berdasarkan Gambar 5.16, menunjukkan bahwa terdapat 37.193 kata dari 8.797 ulasan kategori positif dengan semakin besar ukuran suatu kata pada visualisasi WordCloud, semakin sering kata tersebut muncul dalam kelompok sentimen. Kata bagus, baik, cocok, dan indah merupakan kata yang paling sering muncul. Hal tersebut karena sebagian besar wisatawan memberikan apresiasi terhadap kebersihan, kenyamanan, dan keindahan destinasi wisata di Daerah Istimewa Yogyakarta.



Gambar 5.17 Word Cloud Netral

Berdasarkan Gambar 5.17, menunjukkan bahwa terdapat 10.573 kata dari 1.538 ulasan kategori netral. Kata *tempat*, *parkir*, *enak*, *tidak*, *sangat*, *banyak*, *baik*, *kesini*, dan *buat* merupakan kata yang paling sering muncul. Kemunculan kata-kata tersebut menunjukkan bahwa ulasan dengan kategori netral cenderung berisi informasi mengenai kondisi tempat, fasilitas, akses, serta pengalaman kunjungan yang disampaikan secara deskriptif tanpa menunjukkan kecenderungan sentimen yang kuat.



Gambar 5.18 Word Cloud Negatif

Pada Gambar 5.18, terlihat bahwa terdapat 21.780 kata dari 3.992 ulasan kategori negatif dengan kata besar, jalan, jauh, sampah, pagi, saran, hujan, tinggal, terjun, jaga, dan kotor merupakan kata yang paling sering muncul. Hal tersebut karena sebagian besar wisatawan memberikan apresiasi terhadap akses jalan, lokasi, kebersihan fasilitas umum, kondisi lingkungan, maupun pengalaman pengunjung. Terdapat kata bagus, senang, maupun fasilitas terkadang tidak selalu bernilai negatif, karena terkadang karena berada di dalam ulasan yang secara keseluruhan memperoleh label sentimen *negative*.

5.4.3 Ekstraksi Fitur TF-IDF

Setelah melalui tahap *preprocessing text* dan pelabelan sentimen, tahap selanjutnya yaitu melakukan ekstraksi fitur menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF). Tahap ini digunakan untuk mengubah data teks menjadi representasi numerik sehingga dapat diproses oleh algoritma pembelajaran mesin (*machine learning*), khususnya metode Support Vector Machine (SVM) dan Multinomial Naïve Bayes (MNB).

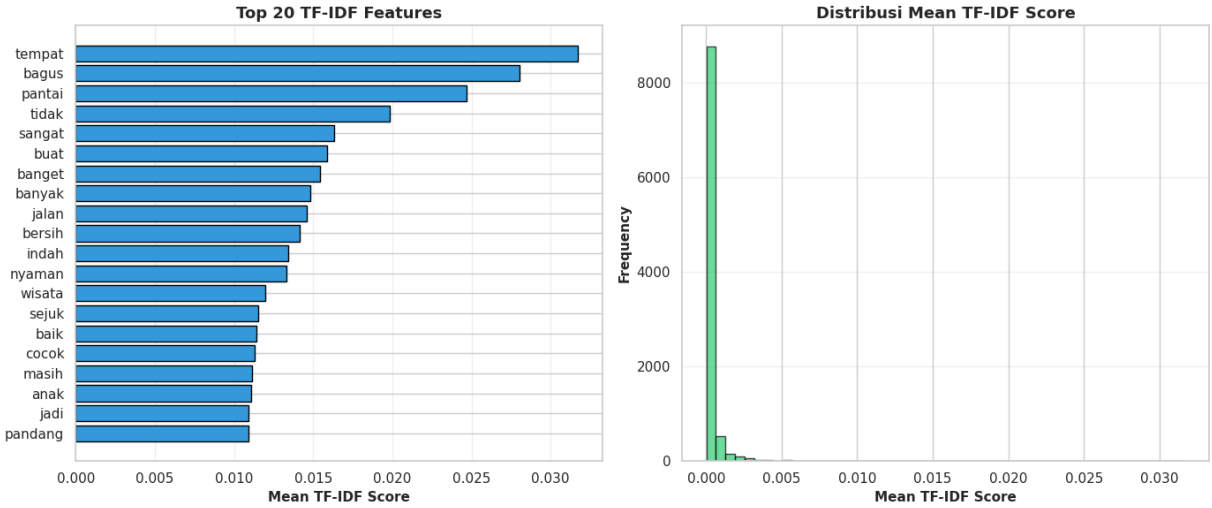
Proses ekstraksi fitur menggunakan metode TF-IDF menghasilkan representasi numerik dari seluruh dokumen ulasan yang telah melalui tahap *preprocessing*. Berdasarkan Gambar 5.19, terdapat sebanyak 14.327 dokumen berhasil dikonversi menjadi matriks TF-IDF berdimensi 14.327×9.774. Setiap baris pada matriks merepresentasikan satu dokumen ulasan, sedangkan setiap kolom menunjukkan fitur berupa kata unik yang diperoleh setelah proses *preprocessing* selesai dilakukan. Terdapat 9.774 fitur sebagai representasi numerik dari

seluruh dokumen ulasan, nilai sparsity sebesar 99,79% berarti sebagian besar elemen dalam matriks TF-IDF bernilai nol yang menjadi karakteristik umum pada data teks dikarenakan setiap dokumen hanya mengandung sebagian kecil dari keseluruhan kosakata yang terdapat pada dataset.

```
✓ TF-IDF selesai
  Dimensi matriks    : (14327, 9774)
  Jumlah dokumen     : 14,327
  Jumlah fitur       : 9,774
  Sparsity           : 99.79%
```

Gambar 5.19 TF IDF Sentimen

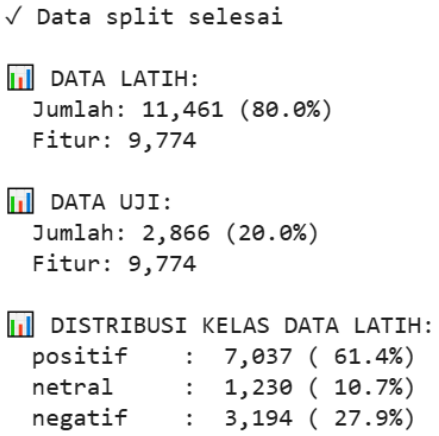
Selain itu, dilakukan identifikasi terhadap dua puluh kata dengan rata-rata bobot TF-IDF tertinggi seperti pada Gambar 5.20. Gambar menunjukkan kata tempat memiliki bobot TF-IDF tertinggi sebesar 0.031698, diikuti kata bagus sebesar 0.028056, kata pantai sebesar 0.024715, dan beberapa kata lainnya. Hal tersebut menunjukkan sebagian besar ulasan membahas kondisi tempat, keindahan alam, kebersihan, kenyamanan, serta akses menuju lokasi wisata. Akan tetapi, munculnya kata tidak sebagai salah satu fitur dengan bobot tinggi sehingga kata negasi memiliki kontribusi penting dalam membedakan sentimen positif dan negatif, serta dapat membuktikan bahwa mempertahankan kata negasi pada *preprocessing* memberikan informasi yang relevan bagi proses klasifikasi.



Gambar 5.20 Visualisasi TF IDF

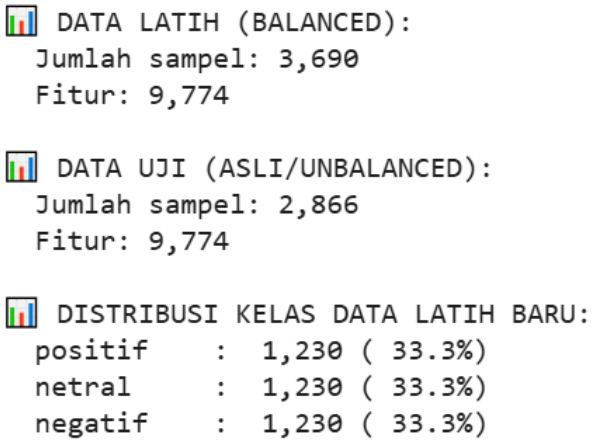
Selanjutnya dilakukan pembagian data menjadi data latih (*training*) dan data uji (*testing*) menggunakan proporsi 80:20, yang bertujuan agar model dapat dilatih menggunakan sebagian besar data dan sisanya untuk mengevaluasi kemampuan model terhadap data yang belum pernah dilihat sebelumnya. Berdasarkan Gambar 5.21, data uji dan data latih menggunakan seluruh dokumen yaitu 9.774 fitur TF-IDF, sehingga tidak ada perubahan jumlah fitur pada

kedua dataset. Pada distribusi label sentimen terdapat ketidakseimbangan kelas (*class imbalance*) dengan kelas positif mendominasi sebanyak 7.037 data (61,4%), diikuti kelas negative sebanyak 3.194 data (27,9%), dan kelas netral sebanyak 1.230 data (10,7%). Dari ketidakseimbangan tersebut, menyebabkan model lebih cenderung mempelajari kelas mayoritas sehingga performa klasifikasi terhadap minoritas kurang optimal.



Gambar 5.21 *Training Testing*

Dari ketidakseimbangan tersebut, menyebabkan model lebih cenderung mempelajari kelas mayoritas sehingga performa klasifikasi terhadap minoritas kurang optimal. Untuk mengatasi hal tersebut, dilakukan proses penyeimbangan data menggunakan metode RandomUnderSampler untuk mengurangi sampel pada kelas mayoritas hingga sama dengan kelas minoritas sehingga seluruh kelas memiliki proporsi yang seimbang. Gambar 5.22, menunjukkan bahwa kelas minoritas yaitu netral terdiri dari 1.230 data sehingga jumlah kelas positif dan *negative* disesuaikan menjadi 1.230 data. Pada data latih yang sebelumnya 11.461 berubah menjadi 3.690 data, tetapi data uji tetap menggunakan 2.866 data karena menggambarkan kondisi distribusi data yang sebenarnya pada proses evaluasi model. Setelah dilakukan proses RandomUnderSampler, data latih menjadi seimbang dan diharapkan mengurangi bias model terhadap kelas mayoritas, juga meningkatkan kemampuan model dalam mengenali seluruh kelas sentimen secara lebih proporsional.



Gambar 5.22 *Training Testing* RandomUnderSampler

5.4.4 Pemilihan Kernel SVM (Linier dan RBF)

Sebelum menggunakan model klasifikasi sentimen Support Vector Machine (SVM), peneliti membandingkan antara dua fungsi kernel yaitu kernel Linier dan kernel Radial Basis Function (RBF) dikarenakan untuk menentukan kernel yang mampu memberikan performa klasifikasi terbaik terhadap data ulasan wisata di DIY. Kernel Linier dipilih karena merupakan kernel paling sederhana dan umum digunakan pada data teks ekstraksi TF-IDF, yang bekerja mencari sebuah hyperplane linier yang mampu memisahkan masing-masing kelas. Sedangkan, kernel RBF karena memiliki kemampuan memetakan data ke ruang dimensi lebih tinggu sehingga mampu menangkap hubungan non-linier antar fitur. Keberagaman ulasan pengguna dan pola bahasa yang kompleks, memungkinkan hubungan antar kelas sentimen tidak dapat dipisahkan secara linier.

```
================================================
Testing: LINEAR (C=10)
================================================

CV F1-macro: 0.6812 ± 0.0117
  Per fold: [0.6936 0.6606 0.6764 0.6882 0.687 ]

Test Set Metrics:
  Accuracy  : 0.7125
  F1-macro  : 0.6476
  Precision : 0.6413
  Recall    : 0.7111

================================================
Testing: RBF (C=100, γ=scale)
================================================

CV F1-macro: 0.6948 ± 0.0170
  Per fold: [0.7028 0.7066 0.6692 0.6807 0.7145]

Test Set Metrics:
  Accuracy  : 0.7268
  F1-macro  : 0.6652
  Precision : 0.6502
  Recall    : 0.7241
```

Gambar 5.23 Analisis SVM Perbandingan Kernel

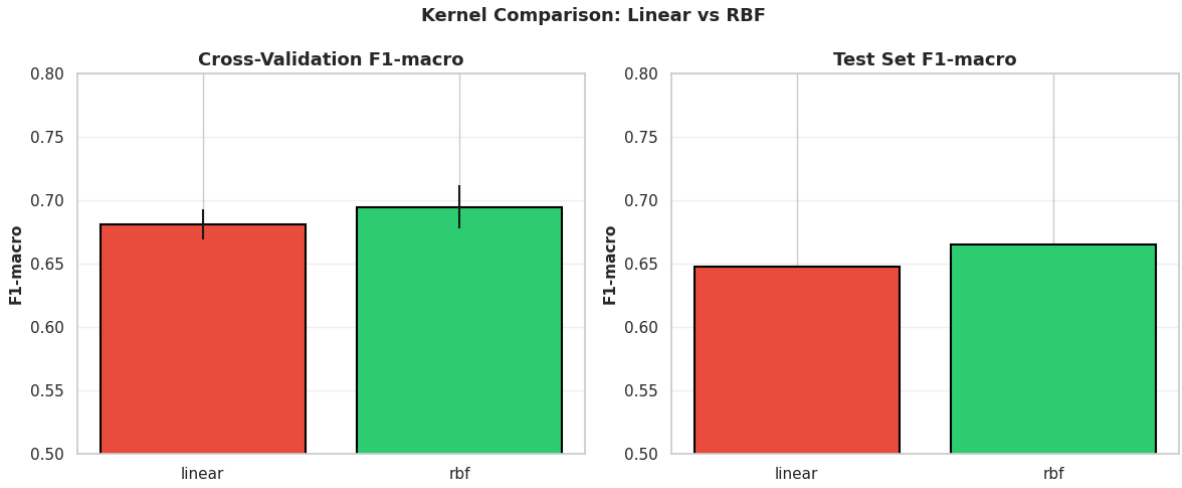
Berdasarkan Gambar 5.23, dilakukan evaluasi pada kedua kernel menggunakan data uji yang tidak digunakan selama pelatihan model, dengan hasil kernel RBF yang menghasilkan performa lebih baik dibandingkan kernel linier. Terlihat seluruh metrik evaluasi pada kernel RBF lebih tinggi dibandingkan kernel linier, yang berarti mampu mengenali pola sentimen dengan lebih baik. Meskipun nilai simpangan baku (standard deviation) pada kernel Linier sebesar 0,0117 dan kernel RBF sebesar 0,0170 sehingga lebih besar kernel Linier, tetapi perbedaan tersebut relatif kecil dan dianggap stabil pada setiap fold.

```
==================================
PERBANDINGAN LINEAR vs RBF
==================================
kernel  cv_f1_mean  test_f1_macro
linear    0.681178       0.647617
   rbf    0.694763       0.665174

✓ KERNEL TERPILIH: RBF
  F1-macro: 0.6652
```

Gambar 5.24 Perbandingan Kernel

Hasil evaluasi menggunakan 5-Fold Cross Validation pada Gambar 5.24, menunjukkan bahwa kernel RBF memperoleh F1-macro sebesar 0.694763, sedangkan kernel Linier sebesar 0.681178. Perbedaan tersebut dikarenakan kernel RBF memiliki kemampuan lebih baik dalam mengklasifikasi ketiga kelas sentimen secara seimbang.



Gambar 5.25 Visualisasi Linier vs RBF

Visualisasi pada Gambar 5.25, terlihat bahwa kernel RBF menghasilkan nilai Cross Validation dan F1-macro lebih tinggi dibandingkan dengan kernel Linier. Selisih menunjukkan kernel lebih mampu memodelkan hubungan antar fitur TF-IDF yang kompleks, walaupun data telah direpresentasikan dalam bentuk vektor numerik tetapi distribusi fitur hasil TF-IDF memungkinkan terbentuknya batas keputusan yang tidak sepenuhnya linier. Juga visualisasi Cross Validation menunjukkan kedua kernel memiliki variasi performa yang relatif kecil antar fold, sehingga peningkatan performa tidak hanya pada satu fold tertentu namun cukup konsisten selama proses validasi silang.

```
🔄 Running 5-fold stratified cross-validation...

[Parallel(n_jobs=-1)]: Using backend LokyBackend with 2 concurrent workers.

================================================================================
CROSS-VALIDATION RESULTS
================================================================================

📊 F1-MACRO:
   Per fold: [0.7028 0.708  0.6679 0.6807 0.7158]
   Mean ± Std: 0.6950 ± 0.0179

📊 ACCURACY:
   Per fold: [0.7019 0.7073 0.6667 0.6789 0.7154]
   Mean ± Std: 0.6940 ± 0.0183

📊 PRECISION-MACRO:
   Per fold: [0.7057 0.7098 0.6752 0.686  0.7196]
   Mean ± Std: 0.6993 ± 0.0162

📊 RECALL-MACRO:
   Per fold: [0.7019 0.7073 0.6667 0.6789 0.7154]
   Mean ± Std: 0.6940 ± 0.0183
[Parallel(n_jobs=-1)]: Done   5 out of   5 | elapsed:   59.1s finished
```

Gambar 5.26 Cross Validation

Gambar 5.26 menunjukkan bahwa penelitian kernel terbaik didasarkan pada nilai *F1-macro* karena melibatkan tiga kelas sentimen dengan distribusi data yang tidak seimbang dan metrik tersebut memberikan bobot yang sama pada setiap kelas, sehingga lebih representatif

dibandingkan hanya menggunakan akurasi. Kemudian akan dioptimalkan menggunakan proses Grid Search dengan 5-Fold Cross Validation untuk memperoleh kombinasi parameter terbaik.

```
🔄 Grid Search untuk SVM RBF...

  Parameter combinations: 25
  Cross-validation: 5-fold

Fitting 5 folds for each of 25 candidates, totalling 125 fits

✓ Grid Search selesai (1281.23 detik)

  Best parameters: {'C': 5, 'gamma': 'scale'}
  Best CV F1-macro: 0.7067
```

Gambar 5.27 Kombinasi Parameter SVM RBF

Selanjutnya didapatkan hasil pada Gambar 5.27, yang mengombinasikan nilai parameter C dan gamma menggunakan 5-Fold Cross Validation. Dengan 25 kombinasi parameter diuji lalu menghasilkan total 125 proses pelatihan model karena 25 kombinasi × 5 fold. Kombinasi parameter tersebut menghasilkan nilai Cross Validation *F1-macro* sebesar 0,7067 yang mana lebih tinggi dibandingkan parameter awal yaitu sebesar 0,6948. Peningkatan tersebut menunjukkan bahwa proses optimasi parameter berhasil meningkatkan kemampuan model dalam mengklasifikasikan ketiga kelas sentimen.

```
✓ Prediction time: 4891.55 ms

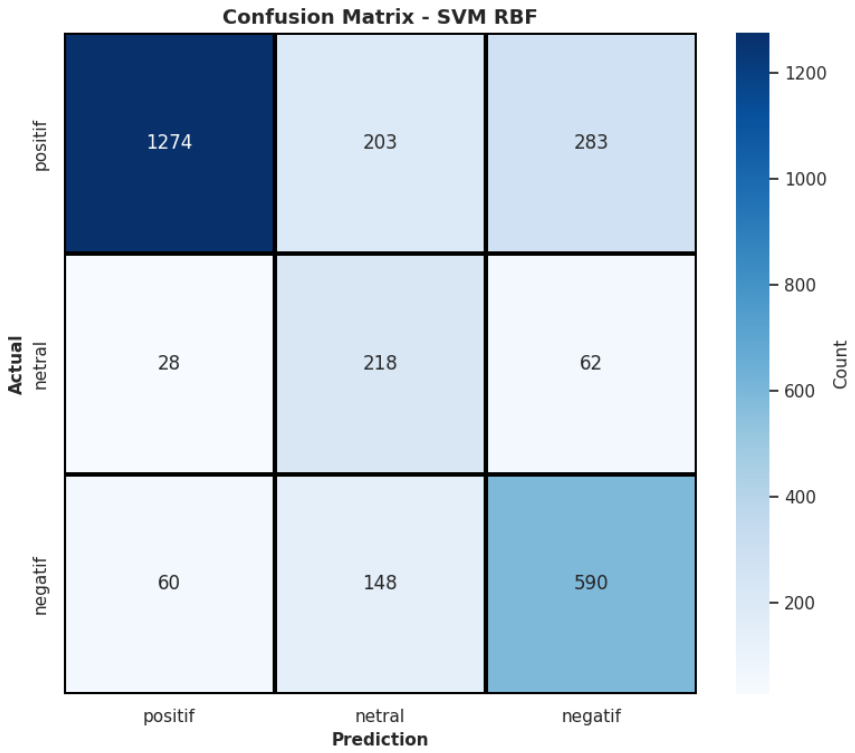
============================================================
CLASSIFICATION REPORT
============================================================
              precision    recall  f1-score   support

     negatif     0.6310    0.7393    0.6809       798
      netral     0.3831    0.7078    0.4971       308
     positif     0.9354    0.7239    0.8161      1760

    accuracy                         0.7264      2866
   macro avg     0.6498    0.7237    0.6647      2866
weighted avg     0.7913    0.7264    0.7442      2866
```

Gambar 5.28 Classification Report SVM RBF

Setelah didapatkan hasil performa klasifikasi yang cukup baik, Gambar 5.28 menunjukkan bahwa SVM RBF mampu menghasilkan nilai akurasi sebesar 0,7264 dan nilai *F1-macro* sebesar 0,6647 berarti mampu mengklasifikasikan ketiga kelas sentimen secara seimbang. Dengan nilai CV *F1-macro* sebesar 0,6950 dan simpangan baku 0,0179 sebelumnya dari Cross Validation memperlihatkan bahwa performa model *relative* konsisten pada setiap fold validasi silang.



Gambar 5.29 Confusion Matrix SVM RBF

Pada Gambar 5.29, model SVM dengan kernel RBF mampu mengklasifikasikan sebagian besar ulasan pada ketiga kelas sentimen dengan baik, yang ditunjukkan oleh dominannya nilai pada diagonal utama *confusion matrix*. Model berhasil mengklasifikasikan 1.274 ulasan positif, 218 ulasan netral, dan 590 ulasan negatif secara benar, meskipun masih terdapat beberapa kesalahan prediksi akibat kemiripan karakteristik antar ulasan. Namun secara keseluruhan, model memperoleh *accuracy* sebesar 72,64%, *precision* makro sebesar 64,98%, *recall* makro sebesar 72,37%, dan *F1-score* makro sebesar 66,47%. Hasil tersebut menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam mengklasifikasikan ketiga kelas sentimen sehingga dipilih sebagai model terbaik.

5.4.5 Evaluasi Model Multinomial Naive Bayes (MNB)

Selain menggunakan metode SVM, penelitian ini membandingkan dengan metode Multinomial Naïve Bayes (MNB) untuk melakukan klasifikasi sentimen. Memilih metode ini karena salah satu metode yag banyak digunakan pada klasifikasi dokumen berbasis teks dan memiliki proses komputasi yang sangat cepat, yang juga bekerja berdasarkan probabilitas kemunculan kata pada masing-masing kelas sehingga sesuai dengan data hasil fitur TF-IDF. Berdasarkan Gambar 5.30, grid search metode ini menggunakan parameter alpha dengan validasi silang sebanyak lipatan yaitu 5 fold, parameter tersebut untuk mengatasi kemungkinan munculnya probabilitas nol pada kata-kata yang tidak ditemukanpada suatu

kelas selama proses pelatihan. Nilai alpha sebesar 0,5 memberikan performa terbaik berdasarkan rata-rata nilai F1-macro selama validasi silang dan didapatkan nilai terbaiknya yaitu 0,6024.

```
  Parameter combinations: 5
Fitting 5 folds for each of 5 candidates, totalling 25 fits

✓ Grid Search selesai (0.38 detik)

  Best parameters: {'alpha': 0.5}
  Best CV F1-macro: 0.6024
```

Gambar 5.30 Kombinasi Parameter MNB

Selanjutnya metode MNB dievaluasi menggunakan Stratified 5-Fold Cross Validation untuk mengukur konsistensi performa model pada data pelatihan, didapatkan hasil pada Gambar 5.31. Nilai simpangan baku (*standard deviation*) sebesar 0,0272 pada F1-macro yang berarti performa model relatif stabil pada setiap fold, walaupun variasinya lebih sedikit dibandingkan dengan metode SVM RBF. Sehingga menunjukkan kemampuan model dalam melakukan generalisasi terhadap data baru masih cukup konsisten, tapi performanya belum optimal. Nilai F1-macro pada 0,5929 berarti kemampuan model mengklasifikasikan ketiga kelas sentimen masih tergolong sedang dan nilai akurasi tersebut berarti model mampu menebak data dengan benar sebesar 60,33%.

```
🔄 Running 5-fold cross-validation (MNB)...

[Parallel(n_jobs=-1)]: Using backend LokyBackend with 2 concurrent workers.

📊 F1-MACRO:
   Mean ± Std: 0.5929 ± 0.0272

📊 ACCURACY:
   Mean ± Std: 0.6033 ± 0.0261
```

Gambar 5.31 Cross Validation MNB

Kemudian dilakukan evaluasi model pada Gambar 5.32, menunjukkan bahwa nilai *accuracy* sebesar 68,91% berarti kemampuan cukup baik dalam mengklasifikasikan sentimen pada data uji. Dengan kelas positif nilai *precision* sebesar 89,50% dan F1-*macro* sebesar 78,42% yang berarti prediksi sentimen cenderung akurat, sementara kelas negatif memiliki nilai precision sebesar 55,23% dan nilai recall sebesar 78,07% yang berarti mampu mengenali sebagian besar data negatif meskipun masih cukup banyak prediksi negatif yang kurang tepat. Secara keseluruhan, nilai macro average F1-macro sebesar 59,97% menunjukkan performa model belum merata pada seluruh kelas, sedangkan nilai weighted average F1-macro berarti kinerja model lebih dipengaruhi oleh keberhasilan klasifikasi pada kelas mayoritas yaitu sentimen positif.

```
==========================================================================
CLASSIFICATION REPORT (MNB)
==========================================================================
              precision    recall  f1-score   support

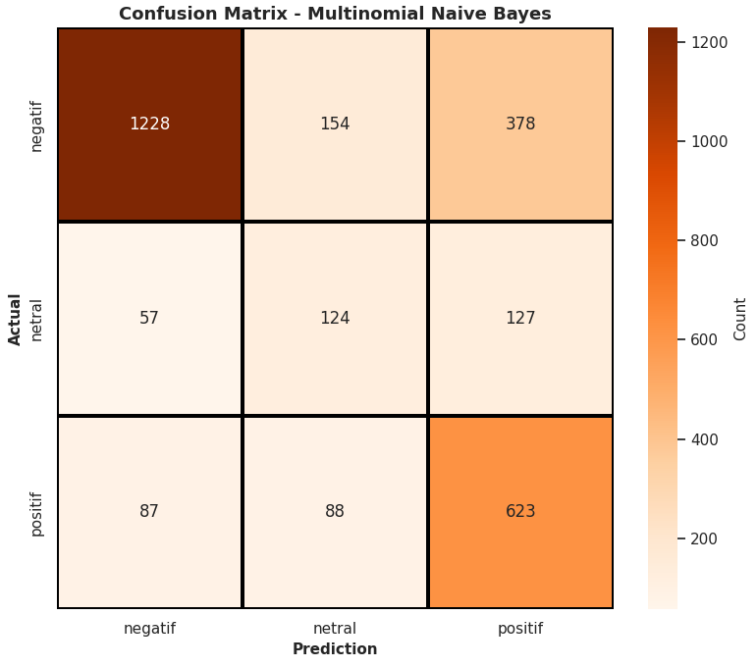
     negatif     0.5523    0.7807    0.6469       798
      netral     0.3388    0.4026    0.3680       308
     positif     0.8950    0.6977    0.7842      1760

    accuracy                         0.6891      2866
   macro avg     0.5954    0.6270    0.5997      2866
weighted avg     0.7398    0.6891    0.7012      2866

[Parallel(n_jobs=-1)]: Done   5 out of   5 | elapsed:    0.4s finished
```

Gambar 5.32 Classification Report MNB

Berdasarkan Gambar 5.33, menunjukkan bahwa metode MNB dapat mengklasifikasikan sebagian besar ulasan pada kelas positif dan negatif cukup baik, tetapi kesulitan pada kelas netral karena banyaknya data netral yang salah diklasifikasikan pada kelas positif dan negatif. Model berhasil mengklasifikasikan 1.228 ulasan negatif, 623 ulasan positif dengan benar, dan 124 ulasan netral dengan tepat. Secara keseluruhan, confusion matrix lebih efektif mengenali kelas mayoritas, tetapi kesulitan membedakan karakteristik sentimen netral sehingga sering terjadi salah klasifikasi pada kelas tersebut.



Gambar 5.33 Confusion Matrix MNB

5.4.6 Perbandingan Model dan Pemilihan Model Terbaik

Tahap terakhir pada proses klasifikasi sentimen yaitu membandingkan performa metode Support Vector Machine (SVM) kernel Radial Basis Function (RBF) dengan metode Multinomial Naïve Bayes (MNB) untuk menentukan model dengan kemampuan terbaik

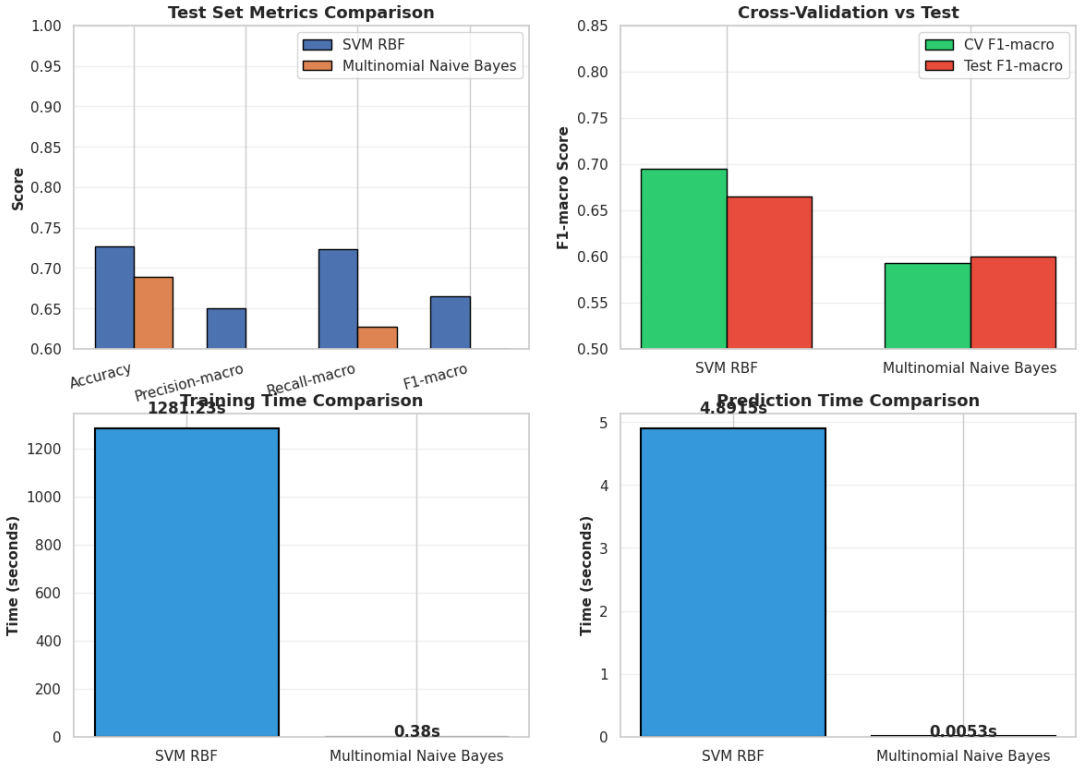
dalam mengklasifikasikan ulasan wisata sebelum hasil prediksi digunakan sebagai komponen pada sistem rekomendasi pariwisata. Berdasarkan Gambar 5.34, terlihat bahwa model SVM kernel RBF menghasilkan performa yang lebih baik dibandingkan dengan model MNB pada hampir seluruh metrik evaluasi.

```
================================================================================
COMPARISON SUMMARY
================================================================================
                    Model  Accuracy  Precision-macro  Recall-macro  F1-macro
                  SVM RBF  0.726448         0.649844      0.723668  0.664731
Multinomial Naive Bayes  0.689114         0.595382      0.627009  0.599684

CV F1-macro Mean  CV F1-macro Std              Best Params  Training Time (s)  Prediction Time (s)
        0.695031        0.017911 {'C': 5, 'gamma': 'scale'}        1281.230652             4.891549
        0.592863        0.027220           {'alpha': 0.5}           0.378673             0.005271
```

Gambar 5.34 *Comparison Summary* Analisis Sentimen

Model SVM kernel RBF memperoleh nilai akurasi lebih tinggi dibandingkan MNB, berarti model SVM RBF mampu mengklasifikasikan ulasan dengan tingkat ketepatan yang lebih tinggi dibandingkan MNB. Selain itu, nilai *Precision-macro*, nilai *Recall-macro*, dan *F1-macro* menunjukkan keunggulan SVM RBF dengan nilai F1-macro menjadi *indicator* utama dalam mengevaluasi performa model.



Gambar 5.35 Visualisasi Perbandingan SVM RBF vs MNB

Berdasarkan Gambar 5.35, dari visualisasi Test Set Metrics Comparison terlihat bahwa model SVM kernel RBF memiliki nilai yang lebih tinggi dibandingkan MNB pada seluruh

metrik evaluasi, sehingga dapat disimpulkan model SVM RBF mampu membanbun batas keputusan (*decision boundary*) yang lebih baik terhadap data ulasan wisata dibandingkan MNB. Pada visualisasi Cross Validation vs Test menunjukkan bahwa nilai CV F1-macro dan Test F1-macro SVM RBF lebih berdekatan, sehingga model mampu meratakan dengan cukup baik karena performa data pelatihan dan pengujian tidak mengalami penurunan yang terlalu besar. Sementara, visualisasi *Training Time Comparison* dan *Prediction Time Comparison* model SVM membutuhkan waktu yang jauh lebih lama dibandingkan dengan MNB. Sehingga penelitian ini memilih model SVM kernel RBF sebagai klasifikasi sentimen, dapat dilihat pada Gambar 5.36.

```
================================
✓ BEST MODEL: SVM RBF
  F1-macro: 0.6647
================================

✓ Final model terkunci: SVM RBF
  F1-macro: 0.6647
```

Gambar 5.36 *Best* Model Analisis Sentimen

5.4.7 Prediksi Seluruh Dataset dan Agregasi Skor Destinasi

Setelah dilakukan pemilihan model terbaik dan klasifikasi sentimen, peneliti melakukan proses prediksi untuk menghasilkan label sentimen bagi setiap ulasan sehingga bermanfaat pada tahap agregasi skor rekomendasi destinasi wisata. Pada tahap ini melakukan seluruh data ulasan yaitu 14.327 sehingga setiap ulasan memperoleh prediksi berupa sentimen positif, netral, dan negatif. Hal tersebut berbeda saat tahap evaluasi model yang hanya menggunakan data uji. Seperti Gambar 5.37, hasil tersebut digunakan sebagai dasar dalam menghitung proporsi sentiment pada masing-masing destinasi wisata, yang kemudian dikombinasikan dengan variabel lain seperti nilai rata-rata rating, jumlah ulasan, dan tingkat kepercayaan (*confidence score*) untuk membentuk skor rekomendasi akhir setiap destinasi pada analisis sentimen.

```
📊 DISTRIBUTION OF PREDICTIONS:
  positif    :   7,029 ( 49.1%)
  netral     :   2,737 ( 19.1%)
  negatif    :   4,561 ( 31.8%)
```

Gambar 5.37 Proporsi Analisis Sentimen

Setelah seluruh ulasan diprediksi, kemudian peneliti melakukan agregasi berdasarkan nama destinasi wisata dengan nilai rata-rata skor sentimen. Pada tahap agregasi, seluruh ulasan yang berasal dari destinasi yang sama dikelompokkan menjadi satu kesatuan lalu

dihitung dengan beberapa indikator diantaranya rata-rata rating (*rata2_bintang*), standar deviasi rating (*std_bintang*), jumlah ulasan (*jumlah_ulasan*), proporsi sentimen positif (*rasio_positif*), proporsi sentimen netral (*rasio_netral*), proporsi sentimen negatif (*rasio_negatif*), nilai rata-rata tingkat kepercayaan model (*confidence_avg*), serta skor rekomendasi akhir (*skor_rekomendasi*). Hasil agregasi menghasilkan 315 destinasi wisata, dengan setiap destinasi direpresentasikan oleh satu baris data yang memuat seluruh indikator tersebut.

	title	rata2_bintang	std_bintang	jumlah_ulasan	rasio_positif	rasio_netral	rasio_negatif	confidence_avg	skor_rekomendasi
2	Kawasan Gua Cerme	5		1	1	0	0	0.999992169	1
3	Ekowisata Tritis Gunungkidul	5	0	7	0.857142857	0.14285714	0	0.876550918	0.928571
4	Gunung Pinus Pengger	4.771428571	0.49024089	35	0.828571429	0.14285714	0.02857143	0.838910049	0.891429
5	Hutan Pinus Mangunan	4.925925926	0.26435053	54	0.777777778	0.14814815	0.07407407	0.837106684	0.881481
6	Museum Anaia Alisha	4.893333333	0.58294067	75	0.76	0.10666667	0.13333333	0.859017716	0.869333
7	Taman Jamu Naturindo	4.916666667	0.47104941	84	0.75	0.0952381	0.1547619	0.849612028	0.866667
8	Watu Goyang	4.857142857	0.52132236	42	0.738095238	0.07142857	0.19047619	0.778720849	0.854762
9	Wisata Bojong Asri	4.875	0.38435306	56	0.732142857	0.14285714	0.125	0.801046366	0.853571
10	Makam Ki Ageng Giring Sukodono	5	0	17	0.705882353	0	0.29411765	0.852279271	0.852941
11	Puncak Becici	4.822222222	0.44153085	45	0.733333333	0.13333333	0.13333333	0.802388777	0.848889
12	De MANGOL VIEW	4.509433962	0.74993952	53	0.79245283	0.13207547	0.0754717	0.868849804	0.84717
13	Mengger Kemuning	4.803571429	0.40089186	56	0.732142857	0.08928571	0.17857143	0.819581965	0.846429
14	Museum Anak Bajang	5	0	6	0.666666667	0	0.33333333	0.928893711	0.833333
15	Museum Gerabah Kasongan	4.938461538	0.24218556	65	0.676923077	0.13846154	0.18461538	0.785135787	0.832308
16	Grand Puri Waterpark	4.861538462	0.60922397	65	0.692307692	0.07692308	0.23076923	0.866766508	0.832308
17	Rumah Sandi Negara	4.984848485	0.12309149	66	0.666666667	0.15151515	0.18181818	0.808556761	0.831818
18	Panorama Setro	4.775862069	0.65005002	58	0.706896552	0.06896552	0.22413793	0.823791021	0.831034
19	HeHa Ocean View	4.555555556	1.1244891	72	0.75	0.06944444	0.18055556	0.838957201	0.830556
20	Pantai Watu Kodok	4.65	1.00128123	40	0.725	0.15	0.125	0.798899519	0.8275
21	Laguna Pantai Depok	4.636363636	0.6990253	33	0.727272727	0.15151515	0.12121212	0.747185115	0.827273

Gambar 5.38 Data Analisis Skor Sentimen

Nilai skor sentimen berada pada rentang 0 hingga 1, yang menunjukkan proporsi hasil prediksi sentimen terhadap seluruh ulasan pada setiap destinasi wisata. Nilai yang semakin mendekati 1 pada kolom rasio_positif menunjukkan bahwa hampir seluruh ulasan destinasi tersebut diprediksi sebagai sentimen positif, begitu juga untuk rasio_netral dan rasio_negatif. Misalnya pada destinasi Kawasan Gua Cerme pada Gambar 5.38, nilai skor rekomendasi sentimen 1 karena memiliki rating rata-rata maksimum sebesar 5, seluruh ulasan yang tersedia diprediksi positif (*rasio* positif = 1), dan nilai *confidence* model sangat tinggi yaitu 0,999. Sementara, destinasi Gunung Pinus Pengger dan Hutan Pinus Mangunan tetap memperoleh skor rekomendasi yang tinggi walaupun sebagian kecil ulasan bersentimen netral maupun negatif. Hasil agregasi tersebut digunakan sebagai dasar untuk menentukan urutas rekomendasi destinasi wisata terbaik di Daerah Istimewa Yogyakarta.

5.5 Analisis Sistem Rekomendasi Hybrid Filtering

Pada tahap ini dilakukan pembangunan sistem rekomendasi menggunakan metode Hybrid Filtering, kombinasi Rule Based Filtering dan Content Based Filtering. Sehingga system rekomendasi tidak hanya mempertimbangkan kualitas destinasi berdasarkan atribut

numerik, tetapi juga tingkat kemiripan karakteristik antar destinasi wisata. Dengan menggunakan data yang menggabungkan data inventarisasi destinasi wisata dan hasil analisis sentimen sebelumnya, sehingga setiap destinasi memiliki informasi yang lengkap untuk proses perhitungan rekomendasi.

Variabel yang diambil dari data inventarisasi berisi nama destinasi, jenis wisata, kategori wisata, daya tarik utama, fasilitas, harga tiket weekend dan weekday, serta rating di Google Maps tiap destinasi. Sementara, untuk data hasil analisis sentiment hanya memuat nilai skor sentimen dari agregasi ulasan pengguna. Penggabungan dataset tersebut digunakan sebagai dasar dalam proses perhitungan sistem rekomendasi.

5.5.1 Rule Based Filtering

Tahap ini melakukan perhitungan Rule Based Score sebagai representasi kualitas setiap destinasi berdasarkan aturan yang telah ditentukan sebelumnya. Metode Rule Based Filtering digunakan untuk memberikan penilaian terhadap destinasi dengan beberapa atribut numerik yang dianggap berpengaruh terhadap kualitas destinasi wisata, diantaranya rating Google Maps, skor sentimen, dan fasilitas.

Sebelum melakukan perhitungan skor, atribut tersebut dinormalisasi menggunakan Min-Max Normalization agar berada pada rentang nilai yang sama yaitu 0 hingga 1. Kemudian diberikan bobot tertentu sesuai tingkat kepentingannya dan dijumlahkan, sehingga menghasilkan nilai Rule Based Filtering yang akan menjadi salah satu komponen dalam membentuk Hybrid Score.

Berikutnya dilakukan pembentukan representasi konten setiap destinasi wisata, untuk mengubah karakteristik setiap destinasi menjadi representasi teks sehingga dapat dihitung kemiripannya. Representasi konten tersebut dilakukan dengan menggabungkan beberapa variabel, diantaranya jenis wisata, kategori wisata, daya tarik, dan lainnya untuk menjadi sebuah dokumen teks.

5.5.2 Content Based Filtering

Setelah terbentuk, dilakukan ekstraksi fitur menggunakan TF-IDF untuk mengubah representasi teks menjadi bentuk numerik, sehingga dapat dihitung tingkat kemiripannya menggunakan Consine Similarity. Berdasarkan Gambar 5.39, terdapat 315 destinasi wisata dengan 14 fitur utama hasil ekstraksi TF-IDF. Ukuran matriks tersebut *relative* kecil, sehingga representasi konten yang digunakan cukup ringkas sehingga proses komputasi menjadi lebih efisien. Selanjutnya dilakukan perhitungan Consine Similarity untuk mengukur tingkat kemiripan antar destinasi wisata, dengan ukuran matriksnya 315×315

yang berarti setiap destinasi dibandingkan dengan seluruh destinasi lainnya. Diperoleh nilai rata-rata sebesar 0,6791 yang berarti berada pada kategori kemiripan sedang hingga tinggi, sehingga secara umum destinasi wisata memiliki karakteristik yang cukup mirip berdasarkan fitur yang digunakan.

```
🔄 Membangun Matriks Kemiripan Konten (TF-IDF Combined)...
TF-IDF shape: (315, 14)
✓ Dimensi Matriks Cosine Similarity: (315, 315)
  Rata-rata cosine similarity antar destinasi: 0.6791
count    49455.000000
mean         0.679099
std          0.201468
min          0.198120
25%          0.482012
50%          0.681510
75%          0.751837
max          1.000000
dtype: float64
  ⚠️ Similarity medium (0.6-0.8).
  Persentil 10: 0.4606
  Persentil 25: 0.4820
  Persentil 50: 0.6815
  Persentil 75: 0.7518
  Persentil 90: 1.0000
  Persentil 95: 1.0000
  Persentil 99: 1.0000
```

Gambar 5.39 TF IDF Sistem Rekomendasi

Selanjutnya dilakukan perhitungan Consine Similarity untuk mengukur tingkat kemiripan antar destinasi wisata, dengan ukuran matriksnya 315 × 315 yang berarti setiap destinasi dibandingkan dengan seluruh destinasi lainnya. Diperoleh nilai rata-rata sebesar 0,6791 yang berarti berada pada kategori kemiripan sedang hingga tinggi, sehingga secara umum destinasi wisata memiliki karakteristik yang cukup mirip berdasarkan fitur yang digunakan. Pada gambar diatas, sebagian besar pasangan destinasi memiliki nilai Cosine Similarity pada rentang 0,60–0,80, menunjukkan bahwa mayoritas destinasi wisata di DIY memiliki tingkat kemiripan yang cukup tinggi berdasarkan karakteristik kontennya. Sementara itu, hanya sebagian kecil pasangan destinasi yang memiliki nilai similarity di bawah 0,40, yang berarti adanya destinasi dengan karakteristik yang berbeda dibandingkan destinasi lainnya. Adanya sejumlah nilai similarity sebesar 1,00 menunjukkan terdapat beberapa destinasi yang memiliki representasi konten yang identik sehingga menghasilkan tingkat kemiripan maksimum.

```
📊 Perbandingan Top-K:
    K  Mean_Content_Score  Std_Content_Score  Mean_Hybrid_Score  \
0   3            0.993114           0.076118           0.664440
1   5            0.993258           0.075436           0.664498
2  10            0.993343           0.075056           0.664532

   Std_Hybrid_Score
0          0.086651
1          0.086545
2          0.086485

📊 Korelasi Spearman antar pilihan K (Hybrid Ranking):
  K=3 vs K=5: rho=1.0000 (p=0)
  K=3 vs K=10: rho=0.9999 (p=0)
  K=5 vs K=10: rho=1.0000 (p=0)
```

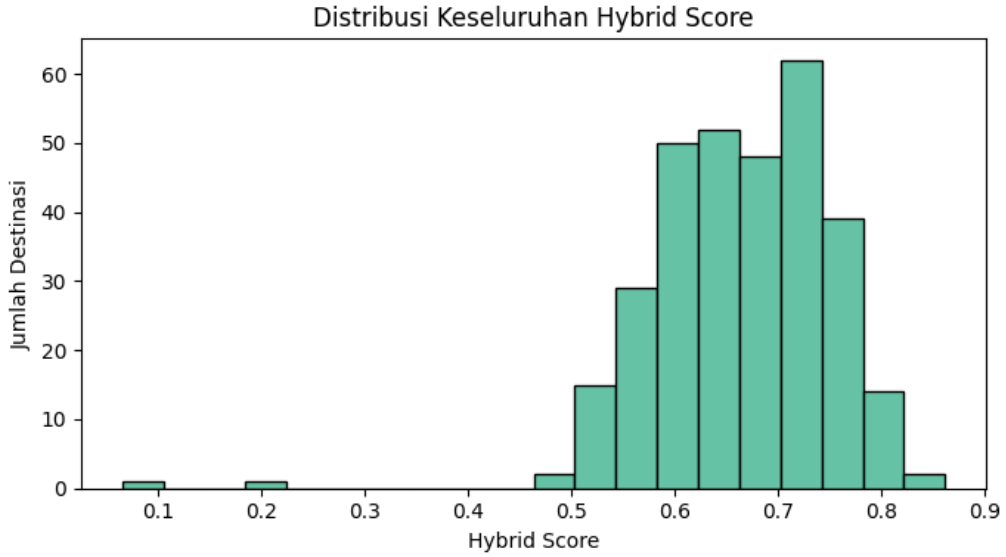
Gambar 5.40 Perbandingan Top-K dan Korelasi Spearman

Tahap berikutnya yaitu melakukan pengujian beberapa nilai Top-K Neighbor yaitu K = 3, K = 5, dan K = 10 yang bertujuan untuk mengetahui jumlah tetangga terdekat yang paling stabil dalam membentuk Content Score. Gambar 5.40 memperlihatkan bahwa rata-rata Content Score dan Hybrid Score pada ketiga nilai K memiliki perbedaan yang sangat kecil, sehingga perubahan jumlah tetangga terdekat tidak memberikan perubahan yang signifikan terhadap hasil rekomendasi. Kemudian dilakukan pengujian Korelasi Spearman dan diperoleh nilai korelasi yang mendekati satu dan 1, yang berarti perubahan pada nilai K hampir tidak mengubah urutan rekomendasi destinasi wisata, sehingga sistem rekomendasi dibangun dengan tingkat stabilitas ranking yang sangat tinggi. Lalu, Content Score dihitung berdasarkan rata-rata nilai similarity dari K Neighbor terdekat setiap destinasi dan didapatkan nilai 0,993 pada seluruh variasi nilai K, yang berarti sebagian besar destinasi memiliki tingkat kemiripan konten yang sangat tinggi dengan destinasi lain.

5.5.3 Hybrid Filtering Score

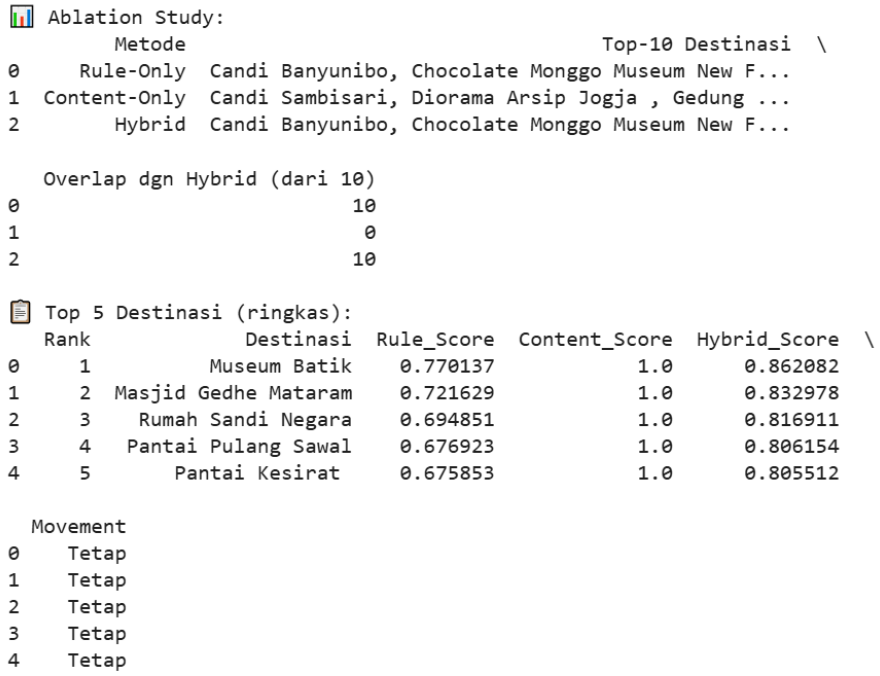
Selanjutnya menggabungkan Rule Based Score dengan Content Score menjadi Hybrid Score. Hybrid Score dihitung menggunakan kombinasi linier kedua metode sehingga mampu mempertimbangkan kualitas destinasi sekaligus tingkat kemiripan antar destinasi. Pada Gambar 5.41, terlihat bahwa sebagian besar destinasi memiliki Hybrid Score pada rentang sekitar 0,60 hingga 0,75. Hal ini menunjukkan bahwa mayoritas destinasi memperoleh nilai rekomendasi yang relatif tinggi sebagai hasil kombinasi antara Rule-Based Filtering dan Content-Based Filtering. Selain itu, hanya terdapat beberapa destinasi dengan nilai Hybrid Score yang sangat rendah, yang mengindikasikan bahwa

destinasi tersebut memiliki kualitas maupun tingkat kemiripan yang lebih rendah dibandingkan destinasi lainnya.



Gambar 5.41 Visualisasi Keseluruhan Hybrid *Score*

Setelah Hybrid *Score* diperoleh, seluruh destinasi diurutkan berdasarkan nilai Hybrid *Score* dari yang tertinggi hingga terendah. Gambar 5.42 menunjukkan lima destinasi wisata dengan nilai Hybrid *Score* tertinggi, seperti Museum Batik menempati peringkat pertama dengan nilai sebesar 0,8621, diikuti Masjid Gedhe Mataram sebesar 0,8330, Rumah Sandi Negara sebesar 0,8169, Pantai Pulang Sawal sebesar 0,8062, dan Pantai Kesirat sebesar 0,8055. Nilai Hybrid Score yang tinggi menunjukkan bahwa destinasi tersebut memiliki kualitas atribut yang baik sekaligus memiliki karakteristik konten yang relevan dalam sistem rekomendasi. Pada gambar tersebut, Ablation Study digunakan untuk membandingkan performa Rule-Based Filtering, Content-Based Filtering, dan Hybrid Filtering



```
📊 Ablation Study:
        Metode                                      Top-10 Destinasi  \
0     Rule-Only  Candi Banyunibo, Chocolate Monggo Museum New F...
1  Content-Only  Candi Sambisari, Diorama Arsip Jogja , Gedung ...
2        Hybrid  Candi Banyunibo, Chocolate Monggo Museum New F...

   Overlap dgn Hybrid (dari 10)
0                            10
1                             0
2                            10

📋 Top 5 Destinasi (ringkas):
   Rank             Destinasi  Rule_Score  Content_Score  Hybrid_Score  \
0     1          Museum Batik    0.770137            1.0      0.862082
1     2  Masjid Gedhe Mataram    0.721629            1.0      0.832978
2     3    Rumah Sandi Negara    0.694851            1.0      0.816911
3     4   Pantai Pulang Sawal    0.676923            1.0      0.806154
4     5       Pantai Kesirat     0.675853            1.0      0.805512

  Movement
0    Tetap
1    Tetap
2    Tetap
3    Tetap
4    Tetap
```

Gambar 5.42 Ablation Study dan Top Destinasi

Secara Keseluruhan, menunjukkan bahwa rekomendasi Hybrid memiliki Top-10 destinasi yang identik dengan Rule-Based Filtering, sedangkan Content-Based Filtering menghasilkan daftar rekomendasi yang berbeda. Hal tersebut karena pada penelitian ini komponen Rule-Based memberikan kontribusi yang lebih dominan terhadap pembentukan rekomendasi akhir, sementara Content-Based Filtering berfungsi sebagai pendukung dalam mempertimbangkan kemiripan karakteristik antar destinasi.

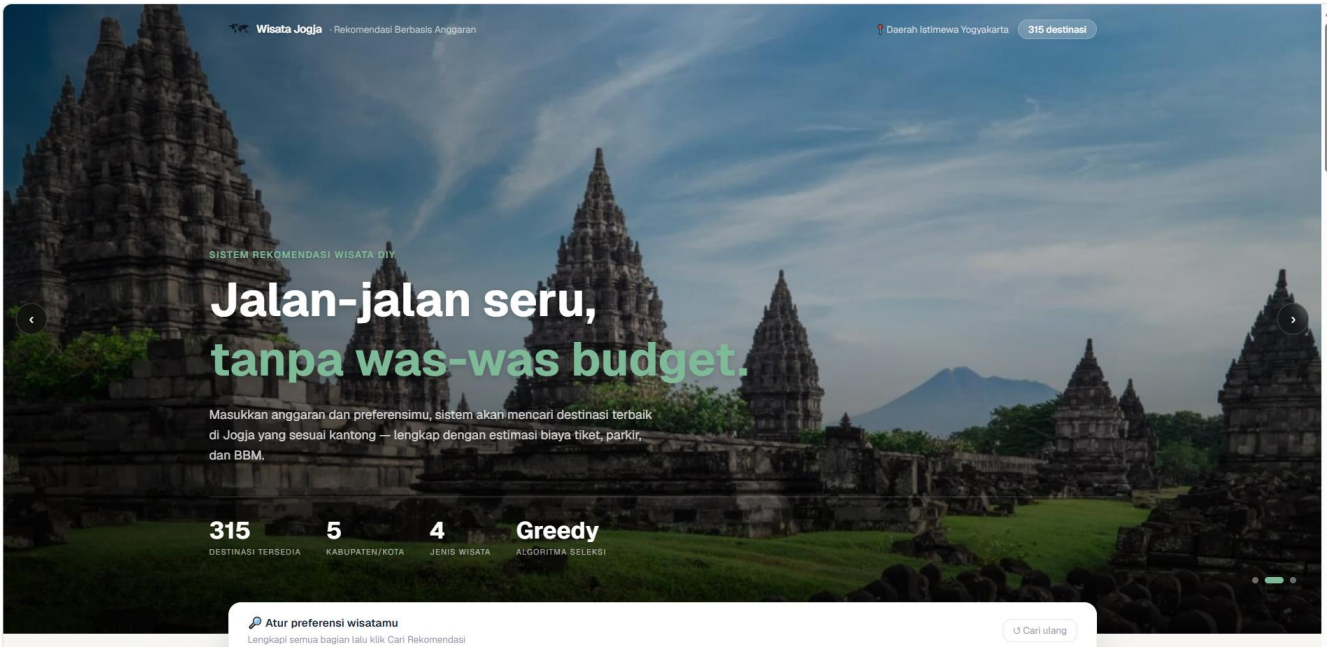
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Destinasi	Rule_Scor	Content_Scor	Hybrid_Sc	Rank_Rule	Rank_Hyb	Delta_Ran	Movement	Rule_Cont	Content_C	Rank	Recomme	Recommendation_Label		
2	Museum Batik	0.770137	1	0.862082	1	1	0	Tetap	0.462082	0.4	1	86.21	Sangat Direkomendasikan		
3	Masjid Gedhe Mataram	0.721629	1	0.832978	2	2	0	Tetap	0.432978	0.4	2	83.3	Sangat Direkomendasikan		
4	Rumah Sandi Negara	0.694851	1	0.816911	3	3	0	Tetap	0.416911	0.4	3	81.69	Sangat Direkomendasikan		
5	Pantai Pulang Sawal	0.676923	1	0.806154	4	4	0	Tetap	0.406154	0.4	4	80.62	Sangat Direkomendasikan		
6	Pantai Kesirat	0.675853	1	0.805512	5	5	0	Tetap	0.405512	0.4	5	80.55	Sangat Direkomendasikan		
7	Candi Banyunibo	0.674227	1	0.804536	6	6	0	Tetap	0.404536	0.4	6	80.45	Sangat Direkomendasikan		
8	Pantai Nguyahan	0.671536	1	0.802922	7	7	0	Tetap	0.402922	0.4	7	80.29	Sangat Direkomendasikan		
9	Museum Sonobudoyo	0.670168	1	0.802101	8	8	0	Tetap	0.402101	0.4	8	80.21	Sangat Direkomendasikan		
10	Pantai Ngobaran	0.664728	1	0.798837	9	9	0	Tetap	0.398837	0.4	9	79.88	Direkomendasikan		
11	Chocolate Monggo Muse	0.656726	1	0.794036	10	10	0	Tetap	0.394036	0.4	10	79.4	Direkomendasikan		
12	Pantai Kayu Arum	0.650792	1	0.790475	11	11	0	Tetap	0.390475	0.4	11	79.05	Direkomendasikan		
13	Museum Taman Tino Sid	0.649086	1	0.789452	12	12	0	Tetap	0.389452	0.4	12	78.95	Direkomendasikan		
14	Makam Ki Ageng Giring S	0.648039	1	0.788824	13	13	0	Tetap	0.388824	0.4	13	78.88	Direkomendasikan		
15	Pantai Jungwok	0.641186	1	0.784712	14	14	0	Tetap	0.384712	0.4	14	78.47	Direkomendasikan		
16	Air Terjun Kembang Soka	0.640746	1	0.784448	15	15	0	Tetap	0.384448	0.4	15	78.44	Direkomendasikan		
17	Taman Nasional Gunung	0.639914	1	0.783949	16	16	0	Tetap	0.383949	0.4	16	78.39	Direkomendasikan		
18	Museum Purbakala Tam	0.633264	1	0.779958	17	17	0	Tetap	0.379958	0.4	17	78	Direkomendasikan		
19	Taman Tebing Breksi	0.629204	1	0.777522	18	18	0	Tetap	0.377522	0.4	18	77.75	Direkomendasikan		
20	Museum Pendidikan Ind	0.627068	1	0.776241	19	19	0	Tetap	0.376241	0.4	19	77.62	Direkomendasikan		
21	Candi Sambisari	0.623305	1	0.773983	20	20	0	Tetap	0.373983	0.4	20	77.4	Direkomendasikan		
22	Museum Gerabah Kason	0.621468	1	0.772881	21	21	0	Tetap	0.372881	0.4	21	77.29	Direkomendasikan		
23	Gunung Ireng	0.621408	1	0.772845	22	22	0	Tetap	0.372845	0.4	22	77.28	Direkomendasikan		
24	Panorama Setro	0.620464	1	0.772278	23	23	0	Tetap	0.372278	0.4	23	77.23	Direkomendasikan		

Gambar 5.43 *Final* Hybrid *Score* Rekomendasi

5.6 Implementasi Sistem Rekomendasi

Pada tahap implementasi, sistem rekomendasi dikembangkan dalam platform aplikasi berbasis website yang dapat memfasilitasi pengguna dengan menginput preferensi wisata secara personal. Desain antarmuka (*interface*) dikembangkan dengan mengutamakan aspek kemudahan pengguna (*user-friendly*) agar pengguna dapat mengatur kriteria destinasi sesuai keinginan sebelum sistem rekomendasi memproses. Berdasarkan Gambar 5.44, halaman utama menyajikan informasi umum terkait sistem rekomendasi yang telah dibuat.

Halaman tersebut menampilkan informasi jumlah destinasi wisata yang tersedia dalam basis data, yaitu sebanyak 315 destinasi, cakupan wilayah penelitian yang meliputi lima kabupaten/kota di Daerah Istimewa Yogyakarta (DIY), dan empat jenis wisata yang menjadi objek rekomendasi. Selain itu, halaman utama juga menampilkan informasi bahwa sistem menerapkan algoritma Greedy sebagai metode seleksi akhir dalam menentukan kombinasi destinasi yang paling sesuai dengan anggaran pengguna setelah proses rekomendasi Hybrid Filtering dilakukan.



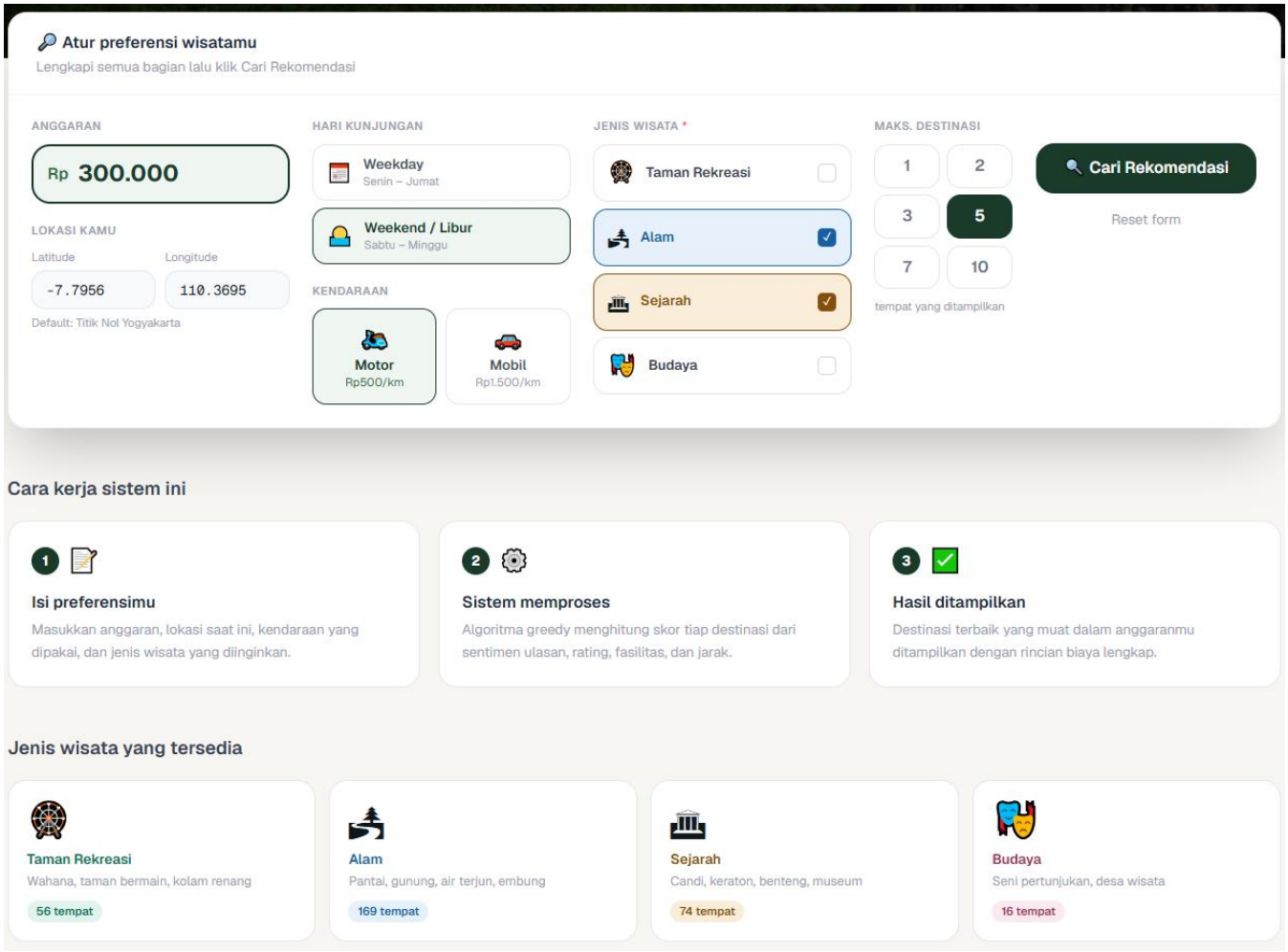
Gambar 5.44 Tampilan *Website* Sistem Rekomendasi Bagian Atas

Pada bagian tengah halaman ditampilkan slogan "Jalan-jalan seru, tanpa was-was budget" beserta deskripsi singkat mengenai tujuan sistem, yaitu membantu pengguna memperoleh rekomendasi destinasi wisata yang sesuai dengan anggaran dan preferensi yang dimiliki. Deskripsi tersebut juga menjelaskan bahwa hasil rekomendasi disertai estimasi biaya perjalanan, meliputi biaya tiket masuk, parkir, dan bahan bakar (BBM).

Selanjutnya, Gambar 5.45 pengguna dapat men-*scroll* halaman ke bawah yang akan menampilkan formulir pengaturan preferensi wisata yang menjadi hal utama dalam proses rekomendasi. Pada bagian ini, pengguna dapat menentukan beberapa parameter yang akan digunakan sistem dalam menghitung rekomendasi destinasi wisata. Parameter pertama yang diisikan adalah anggaran perjalanan, yaitu jumlah dana yang disediakan pengguna untuk melakukan perjalanan wisata. Nilai anggaran tersebut digunakan sebagai salah satu parameter dalam proses Rule-Based Filtering untuk menghitung estimasi biaya perjalanan setiap destinasi, hasil perhitungan tersebut digabungkan dengan skor Content-Based Filtering sehingga menghasilkan nilai Hybrid *Score*. Berdasarkan nilai tersebut, algoritma Greedy melakukan proses seleksi akhir untuk memilih destinasi yang memiliki skor rekomendasi tertinggi dengan tetap mempertimbangkan batas anggaran yang dimiliki pengguna.

Parameter kedua yaitu hari kunjungan, yaitu *weekday* atau *weekend*/libur. Pilihan ini untuk menyesuaikan biaya tiket masuk karena beberapa destinasi memiliki tarif yang berbeda antara hari kerja dan akhir pekan atau hari libur. sistem juga menyediakan pilihan jenis kendaraan, yaitu sepeda motor atau mobil. Informasi ini digunakan untuk menghitung

estimasi biaya bahan bakar berdasarkan tarif konsumsi kendaraan yang telah ditentukan pada sistem. Pada bagian jenis wisata, pengguna dapat memilih satu atau lebih kategori destinasi yang diinginkan, yaitu Taman Rekreasi, Alam, Sejarah, dan Budaya. Preferensi tersebut digunakan sebagai salah satu kriteria dalam proses *Content-Based Filtering* sehingga rekomendasi yang dihasilkan sesuai dengan minat pengguna.



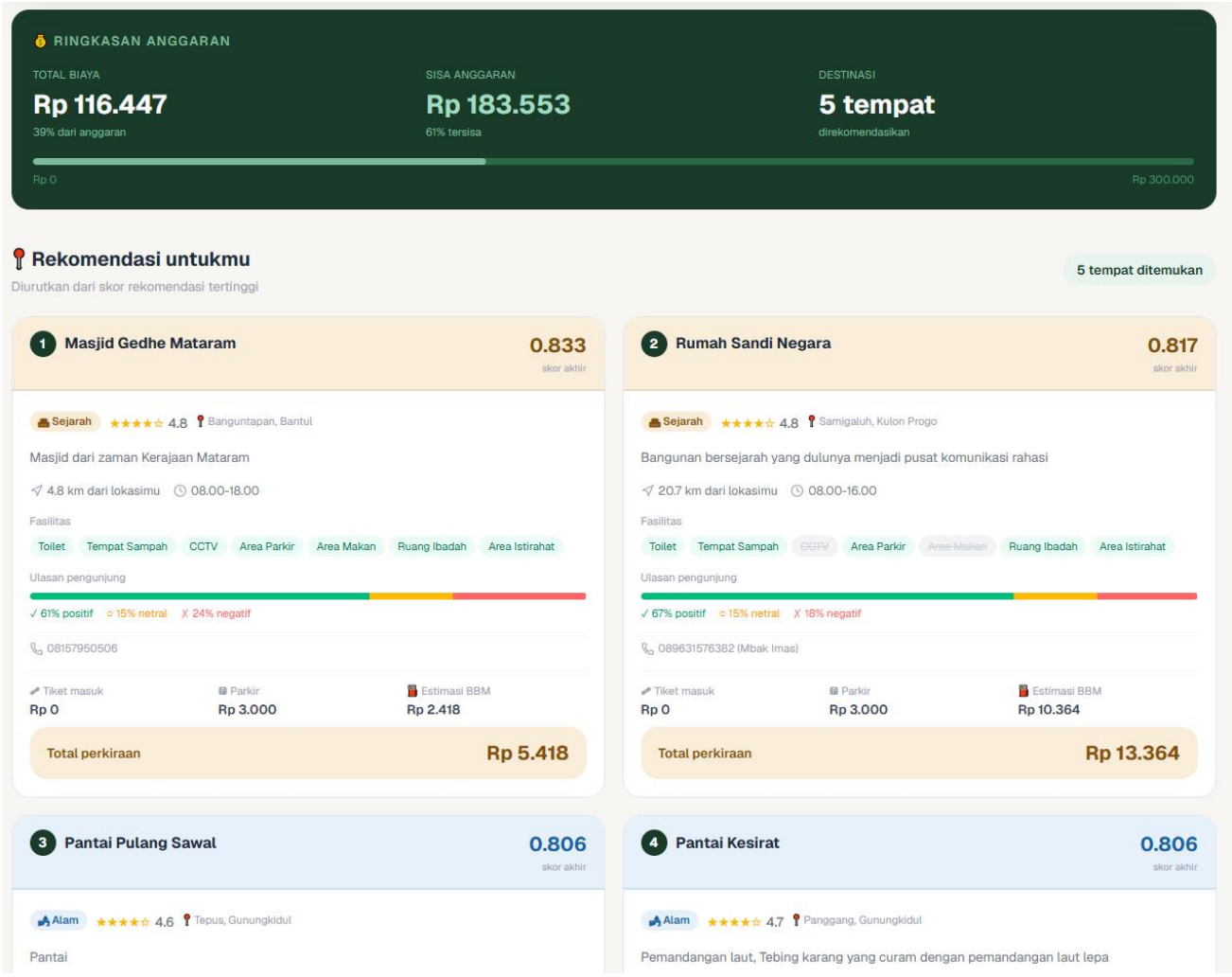
Gambar 5.45 Pilihan *Website* Sistem Rekomendasi

Pengguna juga dapat menentukan jumlah maksimum destinasi yang ingin ditampilkan, mulai dari satu hingga sepuluh destinasi. Parameter ini berfungsi membatasi banyaknya hasil rekomendasi yang akan ditampilkan oleh sistem. Setelah seluruh preferensi diisi, pengguna dapat menekan tombol Cari Rekomendasi untuk menjalankan proses rekomendasi. Sistem kemudian menggabungkan hasil Rule-Based Filtering, Content-Based Filtering, dan proses seleksi menggunakan algoritma Greedy untuk menghasilkan daftar destinasi wisata yang paling sesuai dengan preferensi pengguna.

Ditampilkan juga bagian Cara Kerja Sistem yang menjelaskan alur penggunaan website secara ringkas, yaitu: (1) pengguna memasukkan preferensi wisata, (2) sistem

memproses data dan menghitung skor rekomendasi berdasarkan berbagai atribut destinasi, dan (3) sistem menampilkan hasil rekomendasi beserta rincian biaya perjalanan. Pada bagian paling bawah halaman juga disajikan informasi mengenai jenis wisata yang tersedia beserta jumlah destinasi pada masing-masing kategori. Informasi tersebut memberikan gambaran kepada pengguna mengenai cakupan data yang dimiliki sistem sebelum proses rekomendasi dilakukan.

Setelah pengguna mengisi seluruh referensi dan menekan tombaol Cari Rekomendasi didapatkan Gambar 5.46, hasil rekomendasi diawali dengan Ringkasan Anggaran yang menampilkan informasi mengenai total estimasi biaya perjalanan, sisa anggaran, serta jumlah destinasi yang berhasil direkomendasikan. Selain itu, ditampilkan juga indikator visual berupa *progress bar* yang menunjukkan proporsi anggaran yang telah digunakan terhadap total anggaran yang dimasukkan pengguna. Ringkasan ini bertujuan memberikan gambaran secara cepat mengenai efisiensi penggunaan anggaran sebelum pengguna melihat rincian setiap destinasi.



Gambar 5.46 Hasil Sistem Rekomendasi

Kemudian ditampilkan daftar destinasi yang telah diurutkan berdasarkan skor rekomendasi (Hybrid *Score*) dari nilai tertinggi ke nilai terendah. Pengurutan tersebut hasil proses seleksi menggunakan algoritma Greedy, sehingga destinasi yang ditampilkan merupakan pilihan terbaik sesuai preferensi pengguna dengan tetap mempertimbangkan batas anggaran yang tersedia. Pada setiap destinasi yang direkomendasikan menampilkan informasi yang lengkap mengenai destinasi wisata, meliputi nama destinasi, kategori wisata, rating, lokasi, deskripsi singkat, jarak dari lokasi pengguna, jam operasional, dan skor rekomendasi yang diperoleh. Selain itu, sistem juga menampilkan informasi fasilitas yang tersedia, seperti toilet, tempat sampah, CCTV, area parkir, area makan, ruang ibadah, dan area istirahat sehingga pengguna dapat mempertimbangkan kelengkapan fasilitas sebelum menentukan tujuan wisata. Sebagai tambahan, setiap destinasi dilengkapi dengan hasil analisis sentimen yang ditampilkan dalam bentuk persentase ulasan positif, netral, dan negatif. Informasi tersebut memberikan gambaran mengenai persepsi pengunjung berdasarkan hasil analisis sentimen terhadap ulasan wisata, sehingga pengguna tidak hanya

mempertimbangkan biaya dan fasilitas, tetapi juga pengalaman yang dirasakan oleh pengunjung sebelumnya.

Hasil rekomendasi menunjukkan bahwa sistem berhasil menggabungkan informasi karakteristik destinasi dan persepsi pengguna dalam menghasilkan rekomendasi wisata yang relevan. Selain menampilkan nilai skor rekomendasi, sistem juga menampilkan informasi tambahan seperti rating destinasi, kategori atau jenis wisata, fasilitas yang tersedia, daya tarik utama destinasi, jarak wisata dengan lokasi pengguna, jam operasional, kategori sentimen, sosial media, tiket masuk, biaya parkir, dan estimasi biaya perjalanan. Serta dibagian bawah, diberikan estimasi biaya perjalanan yang terdiri atas biaya tiket masuk, biaya parkir, serta estimasi biaya bahan bakar (BBM). Seluruh komponen biaya tersebut dijumlahkan menjadi total perkiraan biaya untuk masing-masing destinasi. Informasi tersebut dapat membantu pengguna dalam memperkirakan pengeluaran yang diperlukan apabila mengunjungi destinasi tersebut.

BAB VI
PENUTUP

6.1 Kesimpulan

1. Berdasarkan hasil penelitian yang telah dilakukan, data destinasi wisata, fasilitas pendukung, dan informasi operasional di Daerah Istimewa Yogyakarta (DIY) berhasil diperoleh dan diintegrasikan dari kantor dinas langsung dan berbagai platform digital untuk membangun database yang mendukung sistem rekomendasi wisata berbasis Hybrid Filtering. Database yang dibangun mencakup informasi destinasi wisata, ringkasan wisata, titik koordinat, rating, harga tiket, harga parkir, fasilitas, serta informasi operasional yang selanjutnya digunakan sebagai data pendukung dalam proses rekomendasi wisata.
2. Analisis sentimen terhadap komentar wisatawan yang diperoleh dari Google Maps berhasil dilakukan menggunakan algoritma Support Vector Machine (SVM) dan Multinomial Naive Bayes (MNB). Sebelum proses klasifikasi dilakukan tahapan *preprocessing* teks, pelabelan sentimen menggunakan pendekatan Lexicon, serta penyeimbangan data menggunakan Random Under Sampling. Berdasarkan hasil perbandingan model, SVM dengan kernel Radial Basis Function (RBF) memberikan performa terbaik dibandingkan SVM kernel Linear dan Multinomial Naive Bayes, dengan nilai akurasi sebesar 72,65% dan F1-*macro* sebesar 66,47%, sehingga dipilih sebagai model analisis sentimen yang digunakan dalam penelitian ini.
3. Sistem rekomendasi wisata berhasil dibangun menggunakan pendekatan Hybrid Filtering yang menggabungkan Rule-Based Filtering dan Content-Based Filtering. Rule-Based Filtering digunakan untuk menyaring destinasi berdasarkan preferensi pengguna, seperti anggaran perjalanan, kategori wisata, lokasi, kendaraan, harga tiket, harga parkir, serta jumlah destinasi yang diinginkan. Selanjutnya, Content-Based Filtering menghitung tingkat kemiripan antar destinasi menggunakan representasi TF-IDF dan cosine similarity. Hasil analisis sentimen dari model SVM kernel RBF kemudian diintegrasikan sebagai salah satu komponen penilaian dalam perhitungan skor rekomendasi (*hybrid score*). Berdasarkan nilai tersebut, sistem menghasilkan peringkat destinasi sehingga mampu memberikan rekomendasi

wisata yang relevan, sesuai preferensi pengguna, serta mempertimbangkan persepsi pengunjung berdasarkan ulasan yang tersedia.

6.2 Saran

Berdasarkan penelitian yang telah dilakukan, ada beberapa saran yang dapat diterapkan untuk penelitian selanjutnya, yaitu:

1. Memperluas cakupan data destinasi wisata dalam penelitian, berdasarkan data hasil inventarisasi terdapat 346 destinasi wisata dan hanya 315 destinasi yang digunakan dikarenakan beberapa destinasi telah tutup permanen, tidak tersedia pada Google Maps, dan memiliki jumlah komentar yang tidak memenuhi kriteria analisis sentimen. Sehingga, penelitian selanjutnya dapat memperluas cakupan data agar representasi destinasi wisata menjadi lebih lengkap.
2. Menggunakan lebih banyak sumber data ulasan wisatawan. Penelitian ini hanya menggunakan ulasan dari Google Maps, sehingga pada penelitian selanjutnya dapat menambahkan sumber data lain seperti TikTok, Instagram, TripAdvisor, maupun platform digital lainnya untuk memperoleh informasi sentimen yang lebih beragam dan representatif.
3. Mengembangkan metode analisis sentimen menggunakan pendekatan *deep learning* atau *transformer-based models*, seperti Long Short-Term Memory (LSTM), BERT, maupun IndoBERT, untuk membandingkan performanya dengan metode Support Vector Machine (SVM) dan Multinomial Naive Bayes (MNB) yang digunakan pada penelitian ini. Pendekatan tersebut diharapkan mampu memahami konteks bahasa Indonesia dengan lebih baik sehingga menghasilkan klasifikasi sentimen yang lebih akurat.
4. Melakukan evaluasi sistem rekomendasi menggunakan data interaksi atau penilaian pengguna (*user feedback*) sehingga metrik evaluasi khusus sistem rekomendasi, seperti Precision@K, Recall@K, maupun Mean Average Precision (MAP), dapat diterapkan untuk mengukur kualitas rekomendasi secara lebih objektif.

DAFTAR PUSTAKA

Aliady, H., Tuasikal, N. J., & Widodo, E. (2018). IMPLEMENTASI SUPPORT VECTOR MACHINE (SVM) DAN RANDOM FOREST PADA DIAGNOSIS KANKER PAYUDARA. *Seminar Nasional Teknologi Informasi dan Komunikasi 2018 (SENTIKA 2018)* (hal. 278-285). Yogyakarta: SENTIKA 2018.

Atmadji, E. S., Wabula, Y., Karsanti, H. T., & Kristopher. (2025). Penggunaan Naive Bayes Classifier pada Analisis Sentimen di Destinasi Pariwisata Super Prioritas Indonesia Berdasarkan Ulasan Pengguna. *Jurnal Sosioteknologi*, 226-239.

BAPPERIDA DIY. (2025, Desember 09). *Data Kinerja Dinas Pariwisata*. Diambil kembali dari Jogja Dataku: https://bapperida.jogjaprov.go.id/dataku/data_dasar/index/603-data-kinerja-dinas-pariwisata?id_skpd=49

BPS. (2019, Februari 1). *Jumlah kunjungan wisman ke Indonesia Desember 2018 mencapai 1,41 juta kunjungan*. Diambil kembali dari bps.go.id: https://www.bps.go.id/id/pressrelease/2019/02/01/1543/jumlah-kunjungan-wisman-ke-indonesia-desember-2018-mencapai-1-41-juta-kunjungan.html

BPS. (2020). *Perkembangan Jumlah Kendaraan Bermotor Menurut Jenis, 1949-2018*. Diambil kembali dari bps.go.id: https://www.bps.go.id/dynamictable/2016/02/09/1133/perkembangan-jumlah-kendaraan-bermotor-menurut-jenis-1949-2018.html

BPS. (2025, Desember 01). *Jumlah Kunjungan Wisatawan Mancanegara per bulan Menurut Kebangsaan (Kunjungan)*. Diambil kembali dari bps.go.id: https://www.bps.go.id/id/statistics-table/2/MTQ3MCMy/kunjungan-wisatawan-mancanegara-per-bulan-menurut-kebangsaan--kunjungan-.html

detikNews. (2025, Desember 24). *Berita: Jogja Diprediksi Padat, Komisi V DPR Sarankan Warga Bisa Wisata ke Kota Lain*. Diambil kembali dari news.detik.com: https://news.detik.com/berita/d-8276598/jogja-diprediksi-padat-komisi-v-dpr-sarankan-warga-bisa-wisata-ke-kota-lain

Dharmendra, I. K., Putra, I. M., & Atmojo, Y. P. (2024). Evaluasi Efektivitas SMOTE dan Random Under Sampling pada Klasifikasi Emosi Tweet. *Informatics for Educators And Professionals : Journal of Informatics*, 192-193.

Efendi, A. S. (2012). *SISTEM INFORMASI MANAJEMEN PARIWISATA DI DINAS PARIWISATA KOTA JEPARA.* Yogyakarta: DSpace UII.

Efendi, A. S. (2012, Maret 15). *SISTEM INFORMASI MANAJEMEN PARIWISATA DI DINAS PARIWISATA KOTA JEPARA.* Yogyakarta: UNIVERSITAS ISLAM INDONESIA. Diambil kembali dari DSpace UII: https://dspace.uii.ac.id/handle/123456789/36590

Ery Setiyawan Jullev Atmadji, Y. W. (2025). Penggunaan Naive Bayes Classifier pada Analisis Sentimen di Destinasi Pariwisata Super Prioritas Indonesia Berdasarkan Ulasan Pengguna. *Jurnal Sosioteknologi*, 226-239.

Fadhlan Hadaina, U. B. (2022). Implementasi Metode Multinomial Naïve Bayes Untuk Sentiment Analysis Terhadap Data Ulasan Produk Colearn Pada Google Play Store . *Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)*, 600-666.

Faizah, A. N., & Sancoko, S. D. (2024). Sistem Rekomendasi Wisata di Kabupaten Bantul Menggunakan Metode MOORA . *Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi*, 1905-1916.

Husna, S. O., Fajri, I. N., & Pristyanto, Y. (2024). SISTEM REKOMENDASI PARIWISATA GUNUNGKIDUL BERBASIS WEB MENGGUNAKAN METODE CONTENT-BASED FILTERING. *JURSISTEKNI (Jurnal Sistem Informasi dan Teknologi Informasi)*, 534-543.

KOMPAS. (2025, Desember 24). *Travel News: Libur Akhir Tahun, 7 Juta Wisatawan Diprediksi Berlibur ke Yogyakarta*. Diambil kembali dari kompas.com: https://travel.kompas.com/read/2025/12/24/070133527/libur-akhir-tahun-7-juta-wisatawan-diprediksi-berlibur-ke-yogyakarta

Muhammad Luthfi Fajri Martanto, W. I. (2024). Sentiment Analysis of M-Paspor App Reviews Using Multinomial Naive Bayes. *Journal of Logistics, Informatics and Service Science*, 311-326.

Osorio, J. H., Alvarez-Meza, A., Daza-Santacoloma, G., Orozco-Gutierrez, A., & Castellanos-Dominguez, G. (2021, 5 4). *Relevant information undersampling to support imbalanced data classification*. Diambil kembali dari ScienceDirect: https://doi.org/10.1016/j.neucom.2021.01.033

Raharjo, B. (2017). *Belajar Otodidak Flask (Framework Python Untuk Pengembangan Aplikasi Web).* Bandung: INFORMATIKA.

Scikit-learn Developers. (2026). *module Support Vector Machines*. Diambil kembali dari scikit learn: https://scikit-learn.org/stable/modules/svm.html

Sinaga, D. B., Sulistyanto, A., & Ichwan, H. (2026). Analisis Sentimen Komentar Youtube terhadap Kasus Dugaan Korupsi Migas Menggunakan Algoritma Support Vector Machine (SVM). *BLEND SAINS JURNAL TEKNIK*, 288-298.

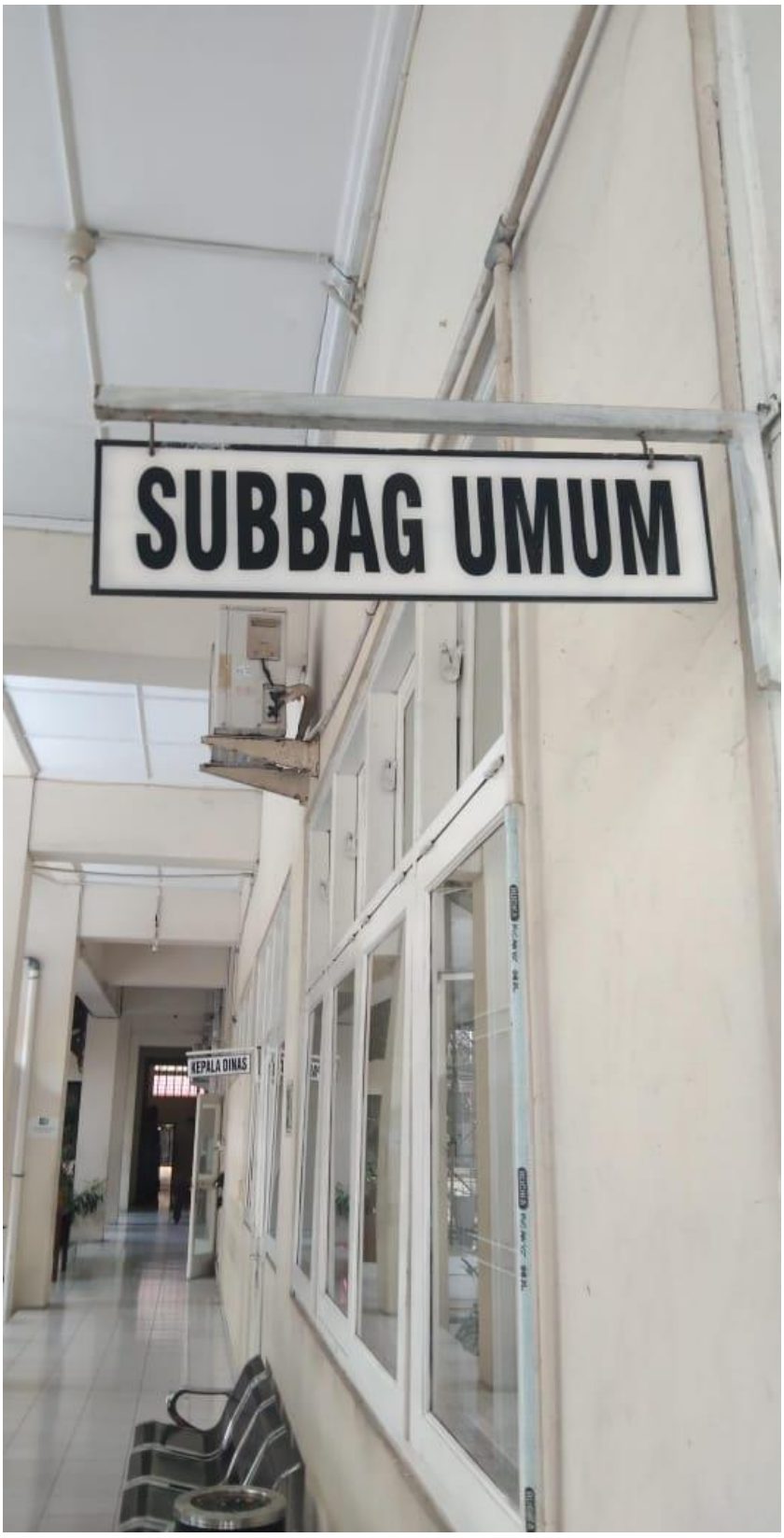
Tim Tugu Jogja. (2019, Juli 3). *Pemerintah DIY Targetkan Kunjungan Wisman Capai 600 Ribu*. Diambil kembali dari Kumparan.com: https://kumparan.com/tugujogja/pemerintah-diy-targetkan-kunjungan-wisman-capai-600-ribu-1rOVdRpBSMI/2

Wicaksono, A. (2020). NEW NORMAL PARIWISATA YOGYAKARTA. *Kepariwisataan : Jurnal Ilmiah* , 139-150.

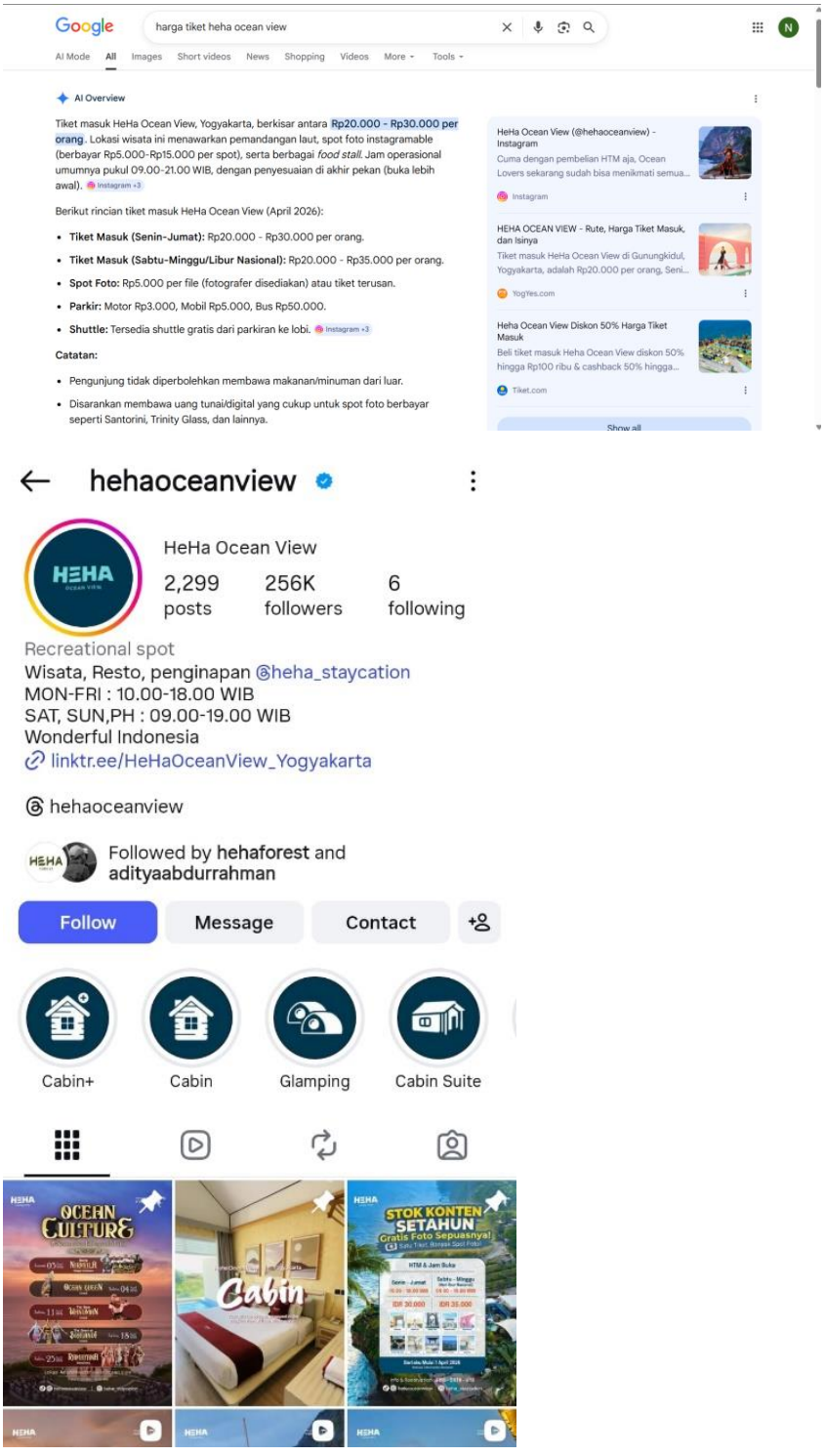
Zhu, Y. (2018). Comparison of Model Performance for Basic and Advanced Modeling Approaches to Crime Prediction. *Intelligent Information Management*, 123-132.

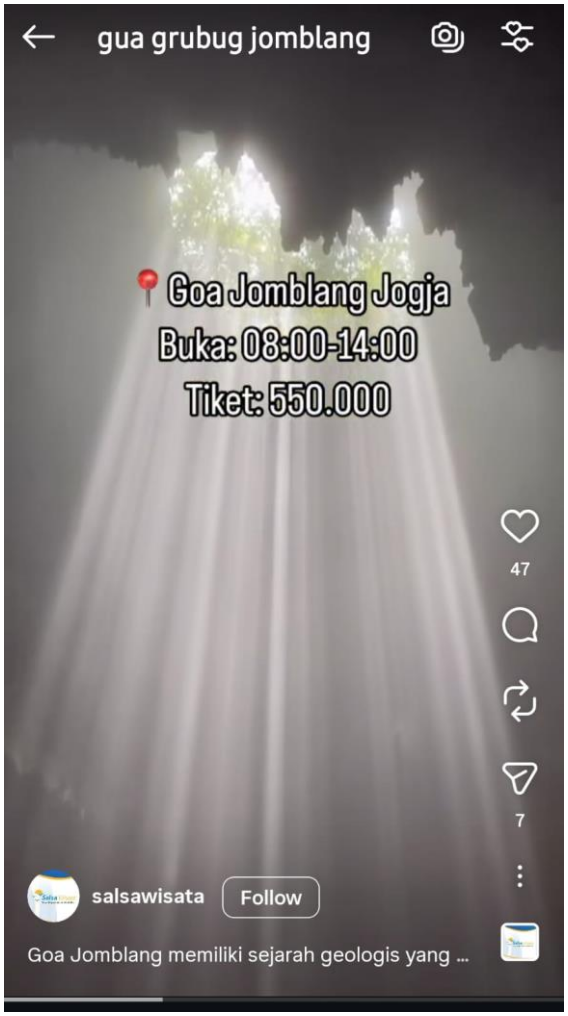
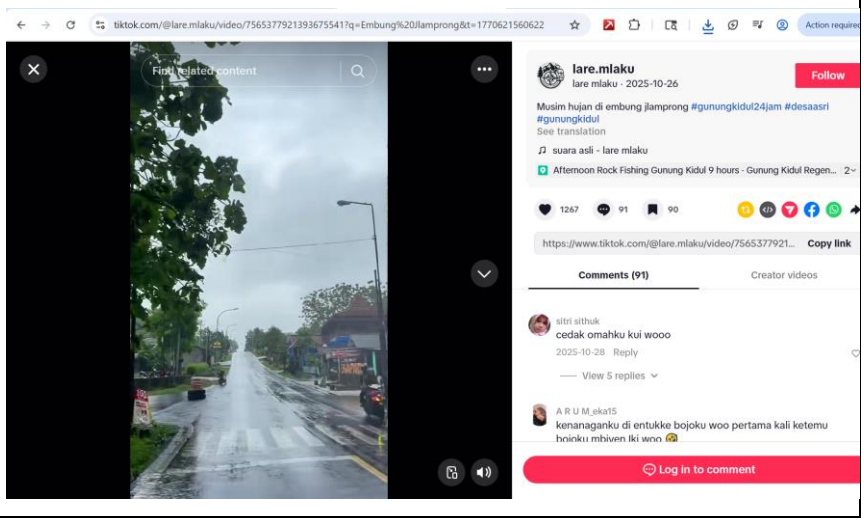
LAMPIRAN

Lampiran 1 Nama Ruangan Konsultasi Data Kantor Dinas Pariwisata DIY



Lampiran 2	Data, Hasil Analisis Sentimen, dan Website https://github.com/noell16/Lampiran_Skripsiii https://github.com/noell16/lampiran_website https://drive.google.com/drive/folders/1fT07k1p8M6EKLmGltu0Ae_Pt7q9mTtBa?usp=sharing

Lampiran 3	Cara Mencari Scaping Lainnya 

Lampiran 4	Cara Mencari Scaping Lainnya  



Taman Jamu Naturindo

+62 838-1713-9095

Call Catalog Share

Products SEE ALL



Wisata Edukasi Tanaman Obat

Tuesday 07:30 – 15:30

Wednesday 07:30 – 15:30



Wisata Edukasi Tanaman Obat

Tuesday 07:30 – 15:30

Wednesday 07:30 – 15:30

Thursday 07:30 – 15:30

Friday 07:30 – 15:30

Saturday 07:30 – 12:30

Sunday Closed

Monday 07:30 – 15:30

Medical & health

Taman Jamu NATURINDO Sendangsari, Secang, Sendangsari, Pengasih, Kulon Progo Regency, Special Region of Yogyakarta 55652, Indonesia



linktr.ee

Instagram

Wanagama Eco Edu Forest

Hutan Pendidikan, Pelatihan, Konservasi, & Ekowisata di Yogyakarta Kehutanan UGM

Reservation and Price List

Formulir Layanan Wanagama

Prosedur Kegiatan

Booklet Wanagama
Document

Lokasi

Wanagama Eco Edu Forest

Kantor Sekretariat