 ISSN:2302-8149
e-ISSN:2540-9719

Penerapan *Content-based Filtering* pada Sistem Rekomendasi Tempat Cuci Sepatu di Kabupaten Klaten

Implementation of Content-based Filtering in a Shoe Cleaning Service Recommendation System in Klaten Regency

[1]Wahyu Nugroho*, [2]Arif Nur Rohman, [3]Ika Nur Fajri
[1,2,3]Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas AMIKOM Yogyakarta
[1,2,3]Jl.Ring Road Utara, Ngringin, Condongcatur, Kec. Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta, 55281
e-mail: wahyunugroho57@students.amikom.ac.id, arifrahman@amikom.ac.id, fajri@amikom.ac.id

(***received*:** 14 September 2026, ***revised*:** 28 October 2026, ***accepted*:** 29 October 2026)

Abstrak

Usaha jasa cuci sepatu di Kabupaten Klaten berkembang cukup pesat, tercemin dari 98 usaha yang teridentifikasi melalui data Google Maps dengan variasi layanan mulai dari cuci sepatu, semir, dry cleaning, hingga laundry sepatu dan tas. Banyaknya pilihan ini justru menyulitkan masyarakat dalam menentukan tempat yang sesuai kebutuhan, karena selama ini pencarian masih dilakukan secara manual dengan membandingkan rating dan ulasan satu per satu, tanpa ada sistem yang secara khusus mengelompokkan usaha berdasarkan kemiripan karakteristik layanannya. Masalah ini diperparah oleh kondisi data yang tidak konsisten, seperti kesalahan pelabelan kategori dan nilai rating yang kurang informatif. Penelitian ini bertujuan membangun sistem rekomendasi tempat cuci sepatu di Klaten menggunakan pendekatan content-based filtering. Metode yang digunakan meliputi pengumpulan data dari Google Maps, preprocessing untuk membersihkan dan menstandarkan atribut, ekstraksi fitur dengan pembobotan TF-IDF pada nama dan kategori usaha, perhitungan kemiripan antar usaha memakai cosine similarity, hingga perangkingan top-N rekomendasi yang diimplementasikan dalam antarmuka pencarian sederhana. Hasil penelitian menunjukan bahwa sistem mampu menghasilkan rekomendasi yang sesuai secara kategori, dengan rata-rata precision@5 berbasis kesesuaian kategori sebesar 72,9%, serta 44,9% usaha mencapai precision@5 sempurna (1,0) pada lima rekomendasi teratasnya. Hasil ini juga mengindikasikan bahwa kualitas data mentah hasil scraping kemungkinan turut memengaruhi batas kinerja sistem rekomendasi berbasis konten pada domain jasa lokal seperti ini, meskipun pengaruh tersebut belum di uji secara statistik interferensial.

Kata kunci: content-based filtering, cosine similarity, cuci sepatu, klaten, sistem rekomendasi, TF-IDF

Abstract

Shoe cleaning services in Klaten Regency have experienced considerable growth, with 98 businesses identified through Google Maps data, offering a variety of services ranging from shoe washing and polishing to dry cleaning and shoe and bag laundry. Despite the wide range of available options, consumers often face difficulties in identifying service providers that match their specific needs. This is because service searches are still conducted manually by comparing ratings and reviews individually, without a dedicated system for grouping businesses based on similarities in their service characteristics. This problem is further compounded by data inconsistencies, including incorrectly labeled business categories and rating values that provide limited discriminatory information. This study aims to develop a shoe cleaning service recommendation system in Klaten Regency using a content-based filtering approach. The methodology involves collecting data from Google Maps, preprocessing the data to clean and standardize attributes, extracting features through Term Frequency–Inverse Document Frequency (TF-IDF) weighting applied to business names and categories, calculating similarities between businesses using cosine similarity, and ranking the top-N recommendations through a simple search interface. The results indicate that the system can generate category-relevant recommendations, achieving an average category-based Precision@5 of 72.9%. Furthermore, 44.9% of the businesses achieved a perfect Precision@5 score (1.0) across their five

 ISSN:2302-8149
e-ISSN:2540-9719

highest-ranked recommendations. These findings also suggest that the quality of raw scraped data may influence the performance limitations of content-based recommendation systems in local service domains. However, this relationship has not yet been statistically tested using inferential analysis.
Keywords: *content-based filtering, cosine similarity, klaten, recommendation system, shoe cleaning, TF-IDF*

1. Pendahuluan

Usaha jasa cuci sepatu di Indonesia mengalami pertumbuhan dalam beberapa tahun terakhir seiring meningkatnya kesadaran Masyarakat terhadap perawatan alas kaki, yang kini dipandang tidak hanya sebagai kebutuhan fungsional tetapi juga bagian dari gaya hidup [20]. Sepatu kini digunakan untuk beragam kegiatan sehari-hari, sehingga semakin banyak sepatu yang dimiliki dan semakin besar pula kebutuhan perawatannya [24]. Perkembangan serupa dapat diamati di Kabupaten Klaten. Berdasarkan penelusuran data Lokasi usaha pada Google Maps yang dilakukan dalam penelitian ini, tercatat 98 penyedia jasa cuci sepatu yang tersebar di berbagai kecamatan, dengan variasi layanan mulai dari cuci sepatu, semir, dry cleaning, hingga laundry sepatu dan tas. Sebagian usaha juga menawarkan layanan tambahan seperti reparasi dan kustomisasi sepatu. Banyak pilihan ini memberikan alternatif yang luas bagi konsumen, tetapi sekaligus menyulitkan konsumen dalam menentukan usaha yang paling sesuai dengan kebutuhannya, terlebih Ketika persaingan antarpenyedia jasa semakin ketat [20], [24].

Pelaku usaha jasa cuci sepatu memanfaatkan media sosial seperti Instagram untuk promosi [20], [24], sedangkan calon konsumen, berdasarkan pengamatan awal penulis, umumnya mencari informasi melalui Google Maps dan membandingkan rating serta ulasan satu persatu secara manual. Sistem rekomendasi dapat membantu konsumen memilih item yang sesuai dengan kebutuhannya, dan salah satu pendekatan yang banyak digunakan content-based filtering (CBF), yaitu merekomendasikan item yang atributnya mirip dengan item acuan atau profil preferensi pengguna [18]. Keunggulan CBF adalah tidak bergantung pada data penelitian dari pengguna lain dan mampu merekomendasikan item baru yang belum pernah dinilai [18], sehingga pendekatan ini dipandang sesuai untuk domain yang tidak memiliki riwayat interaksi pengguna tetapi memiliki atribut item yang dapat dipresentasikan sebagai fitur. CBF telah diterapkan pada berbagai domain, antara lain rekomendasi buku [1], produk perawatan kulit [2], destinasi wisata [4],[7],[10],[14], wisata kuliner [8], [16], serta UMKM kuliner berbasis data Google Maps [11]. Meskipun demikian, kinerja CBF bergantung pada kecukupan dan kualitas deskripsi atribut item [18].

Di sisi lain, data usaha jasa lokal hasil scraping Google Maps memiliki sejumlah persoalan kualitas.
Pada data penelitian ini ditemukan satu entri dengan kesalahan input kategori, 14 dari 98 usaha (14,3%) tidak memiliki jumlah ulasan, dan 68 dari 98 usaha (69,4%) tidak mencantumkan tautan website. Kategori usaha juga ditulis secara beragam untuk makna yang serupa, sehingga 18 kategori awal perlu distandarkan menjadi 14 kategori. Selain itu, 82 dari98 usaha (83,7) memiliki nilai rating yang identik, yaitu nilai tertinggi pada data, sehingga rating tidak dapat membedakan satu usaha dari usaha lainnya dan kurang informatif sebagai dasar perhitungan kemiripan. Kondisi ini sejalan dengan kecenderungan rating daring yang terkonsentrasi pada nilai tinggi sehingga rata-rata rating menjadi indikator kualitas yang kurang andal [19]. Kajian sistematis terhadap sistem rekomendasi juga menempatkan persoalan data sebagai salah satu faktor yang membatasi keandalan sistem [3]. Dengan kondisi tersebut, pencarian manual atau filter sederhana berbasis nama dan lokasi belum memanfaatkan kemiripan karakteristik layanan antarusaha secara sistematis.

Berdasarkan uraian tersebut, rumusan masalah penelitian ini adalah: (1) bagaimana membangun sistem rekomendasi tempat cuci sepatu di Kabupaten Klaten dengan pendekatan content-based filtering yang memanfaatkan fitur teks gabungan nama usaha dan kategori layanan; (2) bagaimana kinerja sistem tersebut jika diukur menggunakan precision@5 berbasis kesesuaian kategori; dan (3) bagaimana karakteristik kualitas data hasil scraping Google Maps serta implikasinya terhadap kinerja sistem. Fitur yang digunakan dibatasi pada nama usaha dan kategori. Rating tidak digunakan sebagai fitur karena 83,7% usaha memiliki nilai rating yang identic sedangkan jumlah ulasan dan tautan website tidak digunakan karena tidak tersedia pada Sebagian usaha, sebagaimanana diuraikan diatas.

 ISSN:2302-8149
e-ISSN:2540-9719

Tujuan penelitian ini adalah membangun sistem rekomendasi tempat cuci sepatu di Kabupaten Klaten berbasis content-based filtering dan mengevaluasi kinerjanya pada data hasil scraping Google Maps. Secara rinci, tujuan tersebut meliputi: (1) membangun sistem rekomendasi menggunakan pembobotan TF-IDF dan perhitungan cosine similarity pada gabungan nama usaha dan kategori; (2) mengevaluasi kinerja sistem menggunakan precision@5 berbasis kesesuaian kategori; dan (3) mengidentifikasi permasalahan kualitas data hasil scraping beserta implikasinya terhadap kinerja sistem. Sistem ini diharapkan membantu pengguna memperoleh pilihan tempat cuci sepatu yang relevan tanpa menelusuri puluhan listing satu per satu, serta menjadi rujukan awal bagi pengembangan sistem rekomendasi serupa pada sektor usaha jasa local lainnya.

2. Tinjauan Literatur

Penelitian tentang sistem rekomendasi berbasis content-based filtering (CBF) telah banyak dilakukan di Indonesia, dengan pilihan fitur dan metode yang umumnya menyesuaikan data yang tersedia pada masing-masing domain. CBF telah diterapkan pada berbagai domain, mulai dari rekomendasi buku menggunakan data katalog perpusatakaan kampus [1], produk skincare di marketplace [2], film Indonesia [6], hingga menu minuman pada cafe [12]. Sebaran domain ini menunjukkan bahwa CBF dapat diterapkan pada berbagai jenis item selama atributnya dapat direpresentasikan sebagai fitur yang dapat di bandingkan [18].

Secara konseptual, CBF merekomendasikan item yang atributnya mirip dengan item acuan atau dengan profil preferensi pengguna. Berbeda dengan collaborative filtering, misalnya pendekatan berbasis item pada rekomendasi wisata kuliner di Yogyakarta [13], CBF tidak memerlukan penilaian dari pengguna lain, tetapi bergantung pada kecukupan dan kualitas deskripsi atribut item [18]. Atribut berbentuk teks lazim diubah menjadi vektor numerik dengan Term Frequency-Inverse Document Frequency (TF-IDF). Bobot sebuah term pada suatu dokumen merupakan hasil kali frekuensi term termasuk pada dokumen (TF) dengan kebaikan proporsi dokumen yang memuatnya (IDF), sehingga term yang sering muncul pada satu dokumen tetapi jarang di seluruh korpus memperoleh bobot lebih tinggi [21]. Kemiripan antardokumen kemudian dihitung dengan cosine similarity, yaitu konsius sudut antar dua vektor. Karena yang dibandingan adalah arah vektor dan bukan panjangnya, ukuran ini tidak dipengaruhi panjang teks, dan nilainya berada antara 0 dan 1 untuk bobot TF-IDF yang tidak negative [22].

Pemilihan CBF dengan TF-IDF dan cosine similarity pada penelitian ini dilandasi beberapa alasan. Data usaha cuci sepatu hasil scraping Google Maps tidak di sertai Riwayat interaksi pengguna, sehingga pendekatan yang tergantung pada penilaian pengguna lain tidak dapat diterapkan secara langsung[18]. Atribut yang tersedia untuk setiap usaha, yaitu nama dan kategori, berbentuk teks pendek yang dapat direpresentasikan dengan TF-IDF tanpa memperlukan data latih yang besar. Selain itu, kombinasi TF-IDF dan cosine similarity bersifat transaparan sehingga prosesnya mudah dijelaskan, direplikasi, dan dibandingkan dengan studi lain yang memaikai kombinasi serupa [1], [14], [15], [17].

Pada sektor pariwisata dan kuliner, sumber serta kekayaan data yang dipakai berbeda-beda. Aljihadu [16] mengukur kemiripan antartempat kuliner di Gunungkidul dengan cosine similarity berdasarkan nama dan deskripsi tempat dari Google Maps. Arum dkk. [11] mengambil data 50 UMKM kuliner Surakarta dari Google Maps melalui web scraping, lalu menggabungkan kategori, menu andalan, kata kunci ulasan, dan fasilitas tanpa memakai rating. Penelitian lain memakai data instansi, yaitu data Dinas Pariwisata Kabupaten Pekalongan [14] dan Aceh Tengah [15], atau data yang memuat riwayat rating pengguna, seperti 75 destinasi dan 3.000 rating dari 100 pengguna pada penelitian di Bali [4]. Suryadi dkk. [10] memanfaatkan kategori, deskripsi, rating, dan ulasan wisatawan untuk merekomendasikan wisata alam di Gunungkidul, sedangkan penelitian wisata kuliner Madura [8] dan wisata Kabupaten Rembang [7] mengandalkan preferensi pengguna. Dengan demikian, Google Maps menjadi sumber data yang mudah diakses pada Sebagian studi [11], [16], tetapi kelengkapan informasinya bergantung pada pemilik usaha dan pengulas, sebagaimana dicatat oleh [11].

Dari sisi metode, beberapa studi yang ditinjau [1], [11], [14], [15], [17] memakai kombinasi TF-IDF dan cosine similarity, namun berbeda dalam cakupan fitur dan cara membentuk query. Salsabilah dan Utomo [14] hanya memakai kolom kategori dengan delapan nilai; pada contoh yang di tampilkan,

 ISSN:2302-8149
e-ISSN:2540-9719

skor kemiripan antartempat wisata hanya bernilai 0,0 atau 1,0, yang mengindikasikan bahwa fitur kategori saja menghasilkan skor yang nyaris biner. Penelitian [11] memperkaya fitur dengan menu, kata kunci ulasan, dan fasilitas serta menambahkan tahap stopword removal dan stemming, sedangkan [16] bertumpu pada nama dan deskripsi. Pada [15], query dibentuk dari maksimal lima atribut yang dipilih pengguna, sementara [14] dan penelitian ini memakai item acuan, yaitu tempat atau usaha yang dipilih pengguna, sebagai query. Penelitian [17] menerapkan kombinasi yang sama pada data resep berdasarkan bahan makanan, sedangkan [4] menggabungkan CBF dengan collaborative filtering yang dilatih pada riwayat rating pengguna, sehingga pendekatannya tidak dapat dipakai pada data listing usaha yang tidak memiliki riwayat interaksi.

Hasil yang dilaporkan sulit dibandingkan secara langsung karena ukuran yang dipakai berbeda-beda: RMSE sebesar 0,2427 dan loss sebesar 0,0589 pada [4], akurasi rata-rata 90% pada [1], skor kemiripan 0,949 pada satu query contoh di [11], serta pengujian black box yang menilai kesesuaian fungsi sistem pada [15]. Penelitian [14] menyajikan contoh daftar rekomendasi tanpa metrik kuantitatif. Arum dkk. [11] sendiri menyarankan pengujian Precision, Recall, MAP, atau NDCG pada penelitian lanjutan, karena simulasi mereka baru memakai korpus empat dokumen dari 50 data yang terkumpul. Dari sisi data, [14] dan [15] bersumber dari instansi resmi dan bukan hasil scraping, sedangkan [11] mencatat potensi bias akibat variasi kelengkapan informasi dari pemilik usaha maupun pengulas Google Maps.

Jika dibandingkan, studi-studi tersebut berbeda pada tiga dimensi yang saling berkaitan. Dari sisi sumber data, studi yang datanya dikurasi instansi [14], [15] atau memuat Riwayat rating pengguna [4] tidak menghadapi persoalan kelengkapan atribut seperti pada data hasil scraping Google Maps [11], [16]. Dari sis fitur, studi yang memakai atribut teks lebih kaya, seperti menu, ulasan, dan fasilitas [11] atau deskripsi tempat [10], [16], memiliki lebih banyak term pembeda antaritem, sedangkan studi yang hanya memakai kategori [14] menghasilkan skor kemiripan yang nyaris biner. Dari sisi evaluasi, studi yang ditinjau menilai hasilnya melalui contoh keluaran, pengujian fungsional, simulasi satu query, atau metrik yang tidak mengukur relevansi daftar top-N [4], [11], [14], [15], sehingga hasilnya tidak dapat dibandingkan secara langsung. Ketiga dimensi ini mengindikasikan bahwa kinerja CBF pada studi yang ditinjau berkaitan erat dengan sumber dan kekayaan data yang dipakai. Penelitian ini berbeda pada kondisi yang belum dikaji pada studi yang ditinjau, yaitu data hasil scraping dengan stribut teks pendek berupa nama dan kategori, tanpa Riwayat interaksi pengguna, dan dievaluasi dengan metrik relevansi top-N.

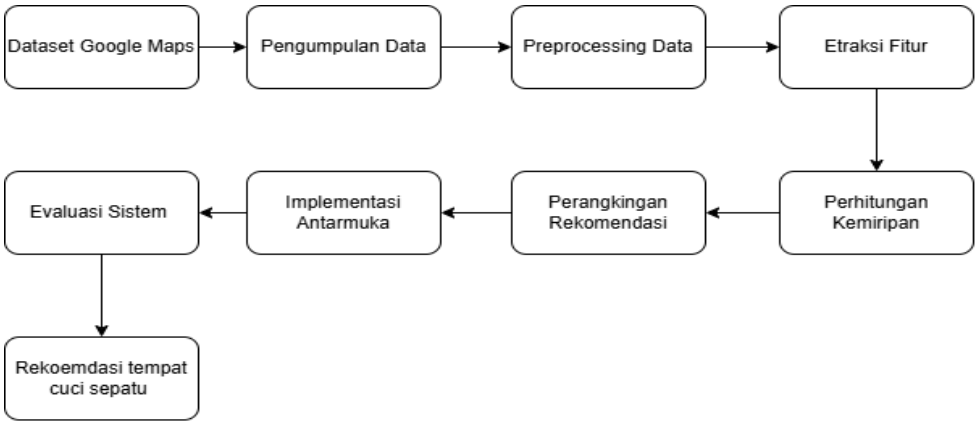
Tinjauan sistematis terhadap sistem rekomendasi menunjukan bahwa keterbatasan data dan variasi teknik evaluasi masih menjadi tantangan umum [3]. Pada studi CBF yang ditinjau, persoalan kualitas data umumnya muncul sebagai tahap praproses atau catatan keterbatasan, bukan sebagai fokus kajian. Penelitian [14] mengisi nilai kosong dengan placeholder dan menambahkan kolom rating secara manual dari ulasan Google, sedangkan [11] menyebut variasi kelengkapan informasi sebagai sumber bias pada saran penelitian lanjutan. Pada studi yang ditinjau, belum ditemukan penelitian yang secara khusus menelaah kualitas data mentah hasil scraping listing usaha mikro, misalnya rating yang terkonsentrasi pada satu nilai, kategori yang tidak konsisten, dan kolom kosong, beserta implikasinya terhadap kinerja rekomendasi.

Dari perbandingan tersebut terdapat tiga kesenjangan pada studi yang ditinjau. Pertama, fitur teks pada studi tersebut bersumber dari data instansi [14], [15], atribut yang lebih kaya seperti menu, ulasan, dan fasilitas [11], atau nama dan deskripsi tempat [16], sedangkan data listing usaha jasa lokal hasil scraping minim teks deskriptif dan didominasi atribut terstruktur yang tidak konsisten. Penelitian [14] paling mendekati kondisi ini karena hanya memakai kategori, tatapi datanya berasal dari instansi dan hasilnya disajikan sebagai contoh keluaran. Kedua, di antara studi yang ditinjau [1], [2], [4], [5], [6], [7], [8], [9], [10], [11], [12], [14], [15], [16], [17], belum ditemukan penelitian CBF dengan objek jasa perawatan sepatu, padahal satu kabupaten saja memiliki puluhan usaha (98 usaha pada data penelitian ini). Ketiga, pada sebagian studi yang ditinjau [4], [11], [14], [15], evaluasi tidak memakai metrik relevansi daftar top-N seperti precision@N. Berdasarkan hal tersebut, penelitian ini membangun sistem rekomendasi CBF untuk usaha cuci sepatu di Kabupaten Klaten dari data hasil scraping Google Maps, mendokumentasikan kualitas data mentahnya, dan mengevaluasi hasilnya dengan precision@5

 ISSN:2302-8149
e-ISSN:2540-9719

3. Metode Penelitian

Penelitian ini memakai pendekatan kuantitatif deskriptif dalam kerangka design and development research [23]. Design and development merupakan strategi penelitian secara keseluruhan: peneliti merancang dan membangun suatu produk, yaitu sistem rekomendasi tempat cuci sepatu, kemudian mengevaluasi kinerjannya. Pendekatan kuantitatif deskriptif berperan pada analisis data di dalam strategi tersebut, yakni menggambarkan karakteristik data hasil scraping dan mengukur kinerja sistem melalui hitungan, persentase, dan rata-rata precision@5 tanpa pengujian hipotesis inferensial. Dengan demikian, design and development menentukan apa yang dikerjakan (merancang, membangun, dan mengevaluasi sistem), sedangkan pendekatan kuantitatif deskriptif menentukan cara data dan hasil evaluasi dianalisis serta dilaporkan. Algoritma utama yang dipakai adalah content-based filtering dengan TF-IDF sebagai pembobot term dan cosine similarity sebagai ukuran kemiripan antarusaha cuci sepatu. Obyek penelitian adalah usaha jasa cuci sepatu di Kabupaten Klaten, Jawa Tengah, dengan ruang lingkup dibatasi pada usaha yang terdaftar dan dapat diakses melalui Google Maps. Penelitian dilakukan dalam tujuh tahap sebagai ditampilkan pada Gambar 1, yang dijabarkan pada sub-bab berikut.



Gambar 1 Alur metode penelitian

3.1 Pengumpulan Data

Data dikumpulkan secara daring melalui web scraping pada Google Maps menggunakan ekstensi peramban Instant Data Scraper dengan kata kunci pencarian “Tempat cuci sepatu di klaten”. Pengambilan data dilakukan pada 9 September 2026, sehingga data mencerminkan kondisi listing pada tanggal tersebut. Atribut yang diambil meliputi nama usaha, kategori, rating, jumlah ulasan, alamat, status dan jam operasional, nomor telepon, tautan website jika tersedia, serta tautan Google Maps, sehingga setiap entri membawa 10 atribut mentah. Data disimpan dalam berkas CSV. Pendekatan pengumpulan data dari listing Google Maps ini juga dipakai pada penelitian sejenis di sektor kuliner [11], [16].

Kreteria inklusi yang diterapkan adalah listing yang muncul pada hasil pencarian dengan kata kunci tersebut dan beralamat di wilayah administratif Kabupaten Klaten. Tidak diterapkan kriteria eksklusi berdasarkan kategori, sehingga seluruh 98 entri yang diperoleh dipertahankan. Akibatnya, dataset memuat listing dengan kategori yang tidak secara langsung merupakan usaha jasa cuci sepatu, yaitu Corporate office (2 entri), Janitorial service (3 entri), dan Shoe store (6 entri), yang bersama-sama berjumlah 11 dari 98 entri (11,2%). Entri tersebut dipertahankan agar data mencerminkan hasil pencarian yang dihadapi pengguna pada Google Maps, dan keberadaanya menjadi keterbatasan penelitian karena entri tersebut dapat muncul sebagai rekomendasi.

Secara operasional, kategori usaha merujuk pada jenis layanan yang tercatum pada listing Google Maps dan, bersama nama usaha, menjadi fitur konten utama dalam CBF. Rating adalah nilai rata-rata penilaian pengguna terhadap suatu usaha, sedangkan jumlah ulasan menunjukan banyaknya pengguna yang memberikan penilaian.

3.2 Preprocessing Data

Data mentah berjumlah 98 entri diproses melalui empat langkah berurutan, dan setiap langkah tidak ada entri yang dikeluarkan. Langkah pertama adalah koreksi kesalahan input kategori: satu entri

 ISSN:2302-8149
e-ISSN:2540-9719

yang kolom kategorinya berisis alamat dikoreksi mengikuti konteks nama usahanya (awal 98, dipertahankan 98, dikeluarkan 0). Langkah kedua adalah standardisasi kategori yang bermakna serupa, yang menurunkan jumlah kategori unik dari 18 menjadi 14 (awal 98, dipertahankan 98, dikeluarkan 0). Langkah ketiga adalah penanganan nilai kosong pada kolom jumlah_ulasan (14 entri) dan telepon (8 entri) tanpa menghapus entri: pada basis data aplikasi, jumlah_ulasan yang kosong disimpan sebagai 0 dan telepon yang kosong dibiarkan kosong (awal 98, dipertahankan 98, dikeluarkan 0). Langkah keempat adalah pembersihan teks pada nama usaha dan kategori, yang meliputi case folding, penghapusan karakter khusus, dan tokenisasi, untuk membentuk kolom clean_features, yaitu gabungan nama usaha dan kategori yang mejadi masukan ekstraksi fitur TF-IDF (awal 98, dipertahankan 98, dikeluarkan 0). Korpus akhir dengan demikian terdiri atas 98 dokumen.

Selain itu, nilai rating distandardisasi dengan z-score pada persamaan (1). Standardisasi ini hanya berfungsi sebagai langkah diagnostik untuk memeriksa sebaran rating, bukan sebagai masukan perhitungan kemiripan.

$$x' = \frac{x - \mu}{\sigma} \quad (1)$$

Dengan x adalah nilai rating asli, μ adalah rata-rata rating, σ adalah simpangan baku, dan x′ adalah nilai hasil standardisasi. Rating hasil standardisasi tidak dimasukkan ke dalam clean_features dan tidak dipakai dalam perhitungan kemiripan. Hasil pemeriksaan ini menjadi dasar keputusan untuk tidak memakai rating sebagai fitur dan dilaporkan pada bagian Hasil dan Pembahasan.

3.3 Ekstraksi Fitur

Kolom clean_features diubah menjadi representasi numerik dengan pembobotan TF-IDF (Term Frequency-Inverse Document Frequency), yang juga dipakai pada penelitian CBF berbasis teks pendek maupun panjang [1], [14], [15], [17]. Pembobotan mengikuti perilaku bawaan kelas TfidfVectorizer pada pustaka scikit-learn, yaitu smoothed IDF dan normalisasi L2, sehingga formula yang berlaku adalah sebagai berikut. Frekuensi term tf(t,d) adalah banyaknya kemunculan term t pada dokumen d, sedangkan IDF dihitung dengan smoothing menggunakan persamaan (2) dan bobot awal term dihitung dengan persamaan (3):

$$\text{idf}(t) = \ln\left(\frac{1+N}{1+\text{df}(t)}\right) + 1 \quad (2)$$

$$w(t,d) = \text{tf}(t,d) \cdot \text{idf}(t) \quad (3)$$

Bobot w(t,d) kemudian dinormalisasi dengan norma L2 pada setiap dokumen menggunakan persamaan (4):

$$tfidf(t,d) = \frac{w(t,d)}{\sqrt{\sum_{t'}(w(t',d))^2}} \quad (4)$$

Dalam formula tersebut N adalah jumlah dokumen (N = 98), df(t) adalah jumlah dokumen yang memuat term t, dan t′ adalah indeks seluruh term pada dokumen d. Hasilnya berupa matriks TF-IDF berukuran 98 × n, dengan n adalah jumlah term unik pada korpus clean_features, dan setiap baris merepresentasikan satu usaha sebagai vektor berpanjang satu. Bobot yang dilaporkan pada Tabel 3 merupakan nilai setelah normalisasi L2.

3.4 Perhitungan Kemiripan

Kemiripan antarusaha dihitung dengan cosine similarity menggunakan persamaan (5), yang juga dipakai pada penelitian CBF sejenis [11], [14], [15], [17], dengan formula:

$$sim(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i{}^2} \cdot \sqrt{\sum_{i=1}^{n} B_i{}^2}} \quad (5)$$

Dengan A dan B adalah vektor TF-IDF dari dua usaha yang dibandingkan, i adalah indeks term, dan n adalah jumlah term unik. Karena setiap vektor telah dinormalisasi L2, penyebut bernilai 1 sehingga cosine similarity setara dengan hasil kali dalam kedua vektor. Nilainya berkisar 0 sampai 1 karena bobot TF-IDF tidak negatif: semakin mendekati 1, semakin mirip kedua usaha tersebut. Hasil tahap ini berupa matriks kemiripan berukuran 98 x 98.

3.5 Perangkaian Rekomendasi

 ISSN:2302-8149
e-ISSN:2540-9719

Dari metriks kemiripan yang terbentuk, sistem mengambil baris milik usaha acuan, mengurutkannya secara descending, lalu mengambil N=5 usaha dengan skor tertinggi (top-5) sebagai rekomendasi. Secara formal, proses ini di notasikan pada persamaan (6):

$$R_N(u) = \text{Top} - \text{N}(\{sim(u,v) \mid v \in U, v \neq u\}) \quad (6)$$

Dengan u adalah usaha acuan dan U adalah himpunan 98 usaha dalam dataset, cos(u,v) adalah skor cosine similarity antara usaha u dan v menurut persamaan (5), dan R_N (u) adalah daftar N = 5 usaha dengan skor kemiripan tertinggi terhadap u. Pendekatan Top-N semacam ini juga dipakai pada penelitian rekomendasi wisata lain [7], [14].

Pendekatan item-to-item berbeda dengan permasalahan yang melatarbelakangi penelitian, yaitu pencarian usaha berdasarkan kebutuhan pengguna, karena pengguna tidak selalu memiliki usaha acuan. Antarmuka memang menyediakan pencarian nama usaha dan filter kategori, tetapi kedua fitur ini hanya membantu pengguna menemukan usaha acuan dan tidak menghasilkan rekomendasi berbasis konten dari kata kunci. Dengan demikian, sistem saat ini melayani pengguna yang sudah mengetahui satu usaha dan ingin melihat alternatif sejenis, tetapi belum mendukung rekomendasi dari kebutuhan yang dinyatakan sebagai kata kunci bebas. Fitur tersebut dapat ditambahkan pada pengembangan lanjutan dengan mengubah kata kunci pengguna mejadi vektor TF-IDF menggunakan kosakata yang sama, lalu menghitung kemiripannya terhadap seluruh usaha [22].

3.6 Implementasi Antarmuka

Hasil perangkingan diimplementasikan sebagai aplikasi web sederhana. Backend dibangun dengan bahasa pemograman PHP tanpa kerangka kerja (framework) dan basis data MySQL yang diakses melalui ekstensi mysqli, sedangkan frontend memakai HTML, CSS, dan JavaScript. Pembobotan TF-IDF, perhitungan cosine similarity, dan pemeringkatan Top-N ditulis langsung sebagai fungsi PHP dengan rumus yang setara dengan TfidfVectorizer pada pustaka scikit-learn dengan pengaturan bawaan. Halaman utama menampilkan daftar usaha dalam bentuk kartu (kategori, nama, alamat, dan jumlah ulasan) yang dapat dicari berdasarkan nama usaha dan difilter berdasarkan kategori. Setelah pengguna memiliki satu usaha acuan, halaman detail menampilkan alamat, status dan jam buka, telepon, jumlah ulasan, serta tautan Google Maps dan website, diikuti tabel lima usaha dengan skor kemiripan tertinggi (nomor, nama, kategori, dan skor kemiripan) beserta nilai precision@5 usaha acuan tersebut. Aplikasi juga menyediakan halaman evaluasi yang merangkum precision@5 seluruh usaha dan halaman pencarian usaha terdekat berdasarkan lokasi pengguna (nama, kategori, jarak, dan alamat). Jarak pada halaman tersebut dihitung dengan formula Haversine pada persamaan (7):

$$d = 2r \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\varphi}{2}\right) + \cos\varphi_1 \cos\varphi_2 \sin^2\left(\frac{\Delta\lambda}{2}\right)}\right) \quad (7)$$

3.7 Evaluasi Sistem

Kinerja sistem diuji melalui dua pendekatan: pengujian fungsional (black box testing) untuk memastikan setiap fitur antarmuka berjalan sesuai harapan, dan pengujian kualitas rekomendasi menggunakan precision@N dengan N=5 (precision@5), yaitu proporsi usaha relevan di antara lima rekomendasi teratas. Karena penelitian ini tidak memiliki data riwayat interaksi pengguna, relevansi didefinisikan secara proksi: usaha v diangap relevan terhadap usaha acuan u apabila kategorinya, setelah standardisasi, sama dengan kategori usaha acuan, sebagaimana dirumuskan pada persamaan (8):

$$Precision@N(u) = \frac{1}{N}\sum_{i=1}^{N} rel(u, v_i) \quad (8)$$

Dengan c(·) menyatakan kategori usaha dan v_i usaha pada peringkat ke-i dalam daftar Top-N. Precision@N untuk satu usaha acuan dihitung dengan Persamaan (9):

$$Precision@N = \frac{1}{|U|}\sum_{u \in U} Precision@N(u) \quad (9)$$

 ISSN:2302-8149
e-ISSN:2540-9719

Dengan Rel(u) adalah himpunan usaha yang relevan terhadap u, n(·) menyatakan jumlah anggota suatu himpunan, dan $R_N(u)$ adalah daftar Top-N pada Persamaan (6). Evaluasi dilakukan dengan menjadikan setiap dari 98 usaha sebagai usaha acuan, masing-masing dengan 97 kandidat, sehingga rata-rata precision@N pada seluruh usaha dihitung dengan Persamaan (10):

$$\overline{\text{Precision@}N} = \frac{1}{n(U)} \sum_{u \in U} \text{Precision@}N(u) \quad (10)$$

Dengan n(U) = 98 adalah jumlah usaha dalam dataset. Karena relevansi ditentukan oleh kesamaan kategori, usaha acuan yang kategorinya tidak dimiliki usaha lain memperoleh precision@5 sebesar 0, dan usaha acuan yang kategorinya beranggotakan k <6 usaha memiliki nilai maksimum teoretis (k-1)/5. Pengukuran ini hanya menilai kesesuaian kategori dan tidak menilai kecocokan rekomendasi dengan kebutuhan pengguna, sedangkan metrik yang memerlukan penilaian relevansi dari pengguna tidak dapat dihitung. Hasil evaluasi ini menjadi dasar untuk menilai kecocokan pendekatan content-based filtering pada data usaha jasa lokal semacam ini, sebagaimana dibahas pada bagian Hasil dan Pembahasan.

4. Hasil dan Pembahasan

4.1 Hasil Pengumpulan Data

Proses scraping pada Google Maps dengan kata kunci “tempat cuci sepatu di klaten” behasil mengumpulkan 98 entri usaha yang tersebar di berbagai kecamatan di Kabupaten Klaten. Setiap entri membawa 10 atribut mentah sebagaimana dirangkum pada Tabel 1.

Tabel 1. Kelengkapan atribut data mentah hasil scraping

Atribut	Jumlah Terisi	Jumlah Kosong
Nama_bisnis	98	0
Kategori	98	0
Rating	98	0
Jumlah_ulasan	84	14
Alamat	97	1
Status_buka	91	7
Jam_buka	83	15
Telepon	90	8
Link_website	30	68
Link_gmaps	98	0

Kolom link_website kosong pada 68 dari 98 usaha (69,4%), yang mengindikasikan bahwa sebagian besar usaha cuci sepatu di Klaten belum mencantumkan situs web pada listing Google Maps-nya. Kondisi data mentah semacam ini berbeda dari penelitian CBF di sektor pariwisata yang dapat dijadikan rujukan metodologis, yang datanya umumnya telah dikurasi oleh dinas pariwisata sehingga kelengkapan atributnya lebih terjaga [14], [15].

Selain link_website, dua atribut lain yang tidak lengkap adalah jumlah_ulasan (kosong pada 14 usaha) dan telepon (kosong pada 8 usaha). Keduannya ditangani tanpa menghapus entri: pada basis data aplikasi, jumlah_ulasan yang kosong disimpn sebagai 0, sedangkan telepon yang kosong dibiarkan kosong dan ditampilkan sebagai tanda “-” pada antarmuka. Kedua atribut tersebut tidak termasuk dalam clean_features,sehingga nilai kosongnya tidak mempengaruhi pembobotan TF-IDF, perhitungan kemiripan, maupun peringkat rekomendasi. Dampaknya terbatas pada infromasi yang ditampilkan pada halaman detail usaha. Khusu jumlah ulasan, nilai 0 pada 14 usaha tersebut menandakan data tidak tersedia dan belum tentu berarti usaha tidak pernah diulas.

Rating tercatat pada seluruh 98 usaha, sedangkan jumlah ulasan kosong pada 14 usaha. Kondisi ini tidak lazim karena listing Google Maps tanpa ulasan umumnya tidak menampilkan rating. Seluruh 14 usaha tersebut memiliki rating pada kelompok nilai tertinggi, yaitu kelompok yang sama dengan 82 usaha berating identik yang dibahas pada sub-bab berikutnya. Dengan demikian, rating pada usaha tanpa ulasan tersebut kemungkinan bukan penilaian pengguna, melainkan nilai bawaan atau hasil

 ISSN:2302-8149
e-ISSN:2540-9719

pengisian pada tahap pengambilan atau pembersihan data. Karena rating tidak digunakan dalam perhitungan kemiripan, kondisi ini tidak memengaruhi proses rekomendasi, tetapi memperkuat alasan untuk tidak menjadikan rating sebagai fitur.

4.2 Hasil Preprocessing Data

Ditemukan satu entri dengan kesalahan input kategori: usaha 'ToePuff 1 Cuci Sepatu & Tas' memiliki isi kolom kategori berupa alamat, yang dikoreksi menjadi Shoe repair shop mengikuti konteks nama usahanya. Setelah koreksi tersebut terdapat 18 kategori unik. Kategori yang bermakna serupa kemudian digabung, yaitu Laundry (9 usaha) dan Laundromat (2 usaha) ke Laundry service, Repair service (1 usaha) ke Shoe shining service, sehingga tersisa 14 kategori. Pemetaan kategori sebelum dan sesudah standardisasi beserta jumlah usahanya ditampilkan pada Tabel 2.

Tabel 2 Pemetaan dan distribusi kategori usaha sebelum dan sesudah standardisasi

Kategori setelah standardisasi	Kategori sebelum standardisasi (Jumlah Usaha)	Jumlah
Laundry service	Laundry service (20), Laundry (9), Laundromat (2)	32
Shoe repair shop	Shoe repair shop (28), Repair service (1), kategori berisi alamat (1)	30
Shoe shining service	Shoe shining service (11), Upholstery cleaning service (1)	12
Dry cleaner	Dry cleaner (7)	7
Shoe store	Shoe store (6)	6
Janitorial service	Janitorial service (3)	3
Corporate office	Corporate office (2)	2
Lainnya (7 kategori)	House cleaning service, Shipping service, Shopping mall, Business to business service, T-shirt store, Second hand store, Store (masing-masing 1)	7
Total (14 kategori)	18 kategori setelah koreksi	98

Temuan lain yang perlu dicatat datang dari kolom rating. Sebanyak 82 dari 98 usaha (83,7%) memiliki nilai rating identik, yaitu nilai tertinggi pada kolom tersebut, sedangkan 16 usaha lainnya terbagi merata pada dua nilai di bawahnya (masing-masing 8 usaha). Setelah standardisasi z-score, ketiga nilai tersebut menjadi 0,4126; -1,2893; dan -2,9911. Di antara 82 usaha berating identik, 14 usaha tidak memiliki jumlah ulasan, sebagaimana diuraikan pada sub-bab sebelumnya. Kondisi ini membuat kolom rating nyaris tidak informatif secara statistik untuk membedakan satu usaha dari usaha lain, sehingga penelitian ini memutuskan tidak menjadikan rating sebagai fitur utama perhitungan kemiripan. Keputusan ini selaras dengan penelitian rekomendasi kuliner UMKM di Surakarta yang sengaja tidak memakai rating sebagai fitur agar terhindar dari bias subjektivitas penilaian [11], dan sejalan dengan kecenderungan rating daring yang menumpuk pada nilai tinggi [19]. Temuan ini juga sejalan dengan catatan pada tinjauan sistematis mengenai pentingnya kualitas dan variasi data bagi keandalan sistem rekomendasi [3].

4.3 Hasil Ekstraksi Fitur

Proses pembobotan TF-IDF pada clean_features menghasilkan matriks berukuran 98 x 207, artinya terdapat 207 term unik dari seluruh korpus teks 98 usaha. Sebagai contoh untuk usaha 'Wareshoes Care (Cuci Sepatu Klaten)', lima term dengan bobot tertinggi di tampilkan pada Tabel 3.

Tabel 3 Top-5 Term TF-IDF untuk usaha 'wareshoes care'

Term	Bobot TF-IDF
Wareshoes	0,6223
Shining	0,3847
Klaten	0,3506

 ISSN:2302-8149
e-ISSN:2540-9719

Care	0,3506
cuci	0,2626

Pola bobot ini konsisten dengan prinsip dasar TF-IDF, yakni memberi bobot besar pada kata yang jarang muncul di seluruh korpus tapi representatif untuk satu dokumen [21]. Pada contoh ini, 'wareshoes' hanya muncul pada satu usaha (IDF 4,90) dan 'shining' pada 12 usaha (IDF 3,03), sedangkan 'cuci' muncul pada 33 usaha (33,7%; IDF 2,07). Bedanya, korpus teks pada penelitian ini jauh lebih pendek (nama usahaa dan kategori) dibandingkan studi yang memakai deskripsi tempat [10], [16], sehingga vocabulary yang terbentuk relatif kecil.

4.4 Hasil Perhitungan Kemiripan

Cosine similarity dihitung antarseluruh pasangan usaha, menghasilkan matriks kemiripan berukuran 98 x 98. Matriks ini memuat 9.506 pasangan non-diagonal (98 x 97). Angka tersebut dihitung sebagai pasangan berurutan, sehingga setiap pasangan usaha terhitung dua kali, yaitu sekali untuk tiap arah perbandingan, dengan skor yang sama karena matriksnya simetris. Jumlah pasangan unik adalah 4.753 (98 x 97 / 2). Rata-rata skor kemiripan adalah 0,0906 dengan skor tertinggi 0,8101, dan kedua nilai ini sama untuk kedua cara perhitungan karena simetris tersebut. Sebanyak 2.342 pasangan berurutan (24,6%), atau 1.171 pasangan unik, memiliki skor tepat 0, yang menunjukan tidak ada kata yang beririsan antar kedua usaha tersebut. Tingginya proporsi ini sejalan dengan gejala sparsity yang disinggung sebagai tentangan umum sistem rekomendasi berbasis konten [3], meskipun penyebab utamannya di sini adalah keterbatasan panjang teks, bukan volume data pengguna.

4.5 Hasil Perangkaian Rekomendasi

Sebagai contoh penerapan, Tabel 4 menampilkan Top-5 rekomendasi untuk usaha acuan "Wareshoes Care (Cuci Sepatu Klaten)", yang berkategori Shoe shining service. Kolom Relevan menandai apakah kategori rekomendasi sama dengan kategori usaha acuan.

Tabel 4 Top-5 Rekomendasi untuk usaha acuan 'wareshoes care' (shoe shining service)

Peringkat	Nama Usaha	Kategori	Skor	Relevan
1	Cuci Sepatu Klaten	Laundry service	0,4972	Tidak
2	Human Care Cuci Sepatu Klaten	Shoe repair shop	0,4293	Tidak
3	Djowi cuci sepatu Klaten	Laundry service	0,3431	Tidak
4	Bersih Shoes and Care	Shoe shining service	0,3418	Ya
5	Cuci Sepatu Klaten(Richgold)	Shoe repair shop	0,3253	Tidak

Hanya peringkat keempat (Bersih Shoes and Care) yang berkategori sama dengan usaha acuan, sehingga precision@5 untuk contoh ini adalah 1/5 = 0,2. Usaha pada peringkat 1, 2, 3, dan 5 memperoleh skor tinggi bukan karena kesamaan layanan, melainkan karena sama-sama memuat frasa "cuci sepatu Klaten" pada namannya, sehingga berbagi term cuci, sepatu, dan klaten dengan usaha acuan. Pola Top-N semacam ini konsistem dengan penelitian rekomendasi wisata lain yang membatasi keluaran pada item dengan skor kemiripan tertinggi [7], [14], hanya saja basis kemiripan disini bergeser dari deskripsi tempat ke pola penamaan usaha.

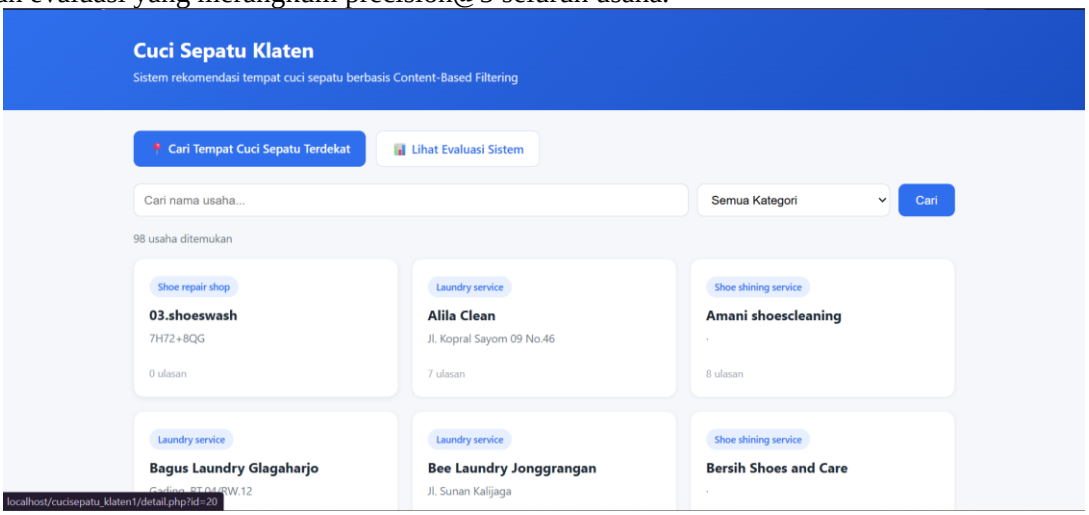
Contoh ini memperlihatkan keterbatasan pemeringkatan berbasis kemiripan teks: skor tertinggi mencerminkan kesamaan kata pada nama, belum tentu kesamaan jenis layanan. Term 'cuci','sepatu', dan 'klaten' muncul masing-masing pada 33 (33,7%), 39 (39,8%), dan 16 (16,3%) dari 98 usaha, sehingga IDF-nya rendah dan daya pembedanya terbatas. Namun, ketiganya muncul bersamaan pada nama sehingga kontribusi gabungan cukup besar untuk mengungguli satu term pembeda seperti 'shining'. Selain itu, kata pada nama kategori seperti 'service' (50 usaha, 51,0%) dan 'shoe' (48 usaha, 49,0%) menimbulkan kemiripan lintas kategori, misalnya antara Shoe shining service dan Laundry service pada peringkat pertama dan ketiga. Penggunaan domain-specific stopwords atau

 ISSN:2302-8149
e-ISSN:2540-9719

pembobotan terpisah antara nama usaha dan kategori dapat dipertimbangkan untuk mengatasi hal ini, dan keduanya dibahas pada sub-bab Pembahasan.

4.6 Hasil Implementasi Antarmuka

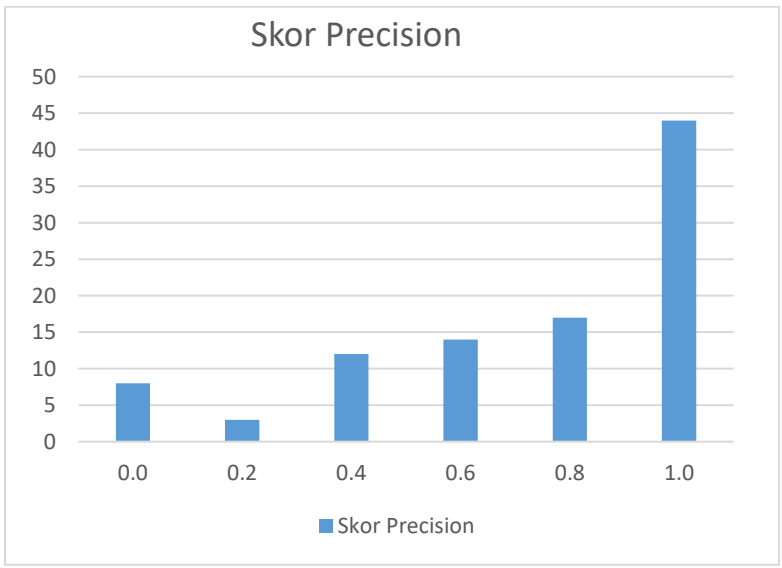
Hasil pemeringkatan diimlementasikan dalam bentuk aplikasi web sederhana. Halaman utama menampilkan daftar usaha dalam bentuk kartu (kategori, nama, alamat, dan jumlah ulasan) yang dapat dicari berdasarkan nama usaha dan difilter berdasarkan kategori. Setelah pengguna memilih satu usaha, halaman detail menampilkan alamat, status dan jam buka, telepon, jumlah ulasan, serta tautan Google Maps dan website, diikuti tabel lima usaha dengan kemiripan tertinggi (nomor, nama, kategori, dan skor kemiripan) beserta skor precision@5 usaha tersebut. Karena keterbatasan atribut kontak (69,4% usaha tidak mencantumkan link_website), antarmuka (Gambar 2) menyediakan alternatif kontak berupa nomor telepon dan tautan Google Maps. Aplikasi juga memiliki halaman pencarian usaha terdekat berdasarkan lokasi pengguna (nama, kategori, jarak, dan alamat) serta halaman evaluasi yang merangkum precision@5 seluruh usaha.



Gambar 2 Implementasi antarmuka

4.7 Hasil Evaluasi Sistem

Karena tidak tersedia data riwayat interaksi pengguna, evaluasi dilakukan dengan pendekatan proxy precision@5 berbasis kesesuaian kategori, sebagaimana didefinisikan pada persamaan (8) dan (9), yang dihitung untuk seluruh 98 usaha sebagaimana ditampilkan pada Gambar 3.



Gambar 3 Distribusi skor precision

 ISSN:2302-8149
e-ISSN:2540-9719

Rata-rata precision@5 keseluruhan adalah 0,7286 (72,9%). Sebanyak 44 dari 98 usaha (44,9%) mencapai precision@5 sempurna (1,0) dan 8 usaha (8,2%) memperoleh precision@5 sebesar 0, sedangkan usaha lainnya tersebar pada nilai 0,2 (3 usaha), 0,4 (12 usaha), 0,6 (14 usaha), dan 0,8 (17 usaha).

Delapan usaha dengan precision@5 nol tidak seluruhnya disebabkan oleh kategori unik. Tujuh diantaranya berkategori unik, yaitu House Cleaning service, Shipping service, Shopping mall, Business to business service, T-shrit store, Second hand store, dan Store (sesuai dengan tujuh kategori tunggal pada tabel 2), sehingga tidak ada usaha lain yang dapat dinilai relevan. Usaha kedelapan, Boogie Lab – Shoes Cleaning UII, justru berada pada kategori terbesar (Laundry service, 31 usaha) tetapi tetap memperoleh nilai nol. Kelima rekomendasi teratasnya berkategori lain (Shoe shining service, Shoe repair shop, dan Shoe store) karena kemiripan terbesar berasal dari kata ‘shoes’, ‘cleaning’, dan ‘uii’ pada nama, sedangkan usaha berkategori Laundry service pertama baru muncul pada peringkat keenam (Clean Laundry, skor 0,2003). Selisih satu usaha antara delapan usaha bernilai nol dan tujuh kategori tunggal dengan demikian berasal dari kesalahan peringkat, bukan dari ukuran kategori.

Karena relevansi ditentukan oleh kesamaan kategori, nilai yang dapat dicapai dibatasi oleh ukuran kategori. Apabila setiap usaha menerima lima rekomendasi yang seluruhnya berasal dari kategori sendiri, rata-rata maksimum precision@5 dihitung dengan persamaan (11):

$$\overline{\text{Precision@5}}_{\text{maks}} = \frac{1}{n(U)} \sum_{u \in U} \frac{\min(k_u - 1{,}5)}{5} \qquad (11)$$

Dengan k adalah jumlah usaha pada kategori usaha acuan u. Nilai maksimum tersebut adalah 0,894 (89,4%), karena tujuh usaha berkategori untuk tidak mungkin memperoleh rekomendasi relevan, sedangkan usaha pada kategori Junitorial service (3 usaha) dan Corporate office (2 usaha) paling tinggi mencapai 0,4 dan 0,2. Sebagai pembanding, nilai harapan precision@5 apabila lima usaha dipilih secara acak dihitung dengan persamaan (12):

$$E(\text{Precision@5}) = \frac{\sum_{c \in C} k_c(k_c - 1)}{n(U) \cdot (n(U) - 1)} \qquad (12)$$

Dengan C adalah himpunan 14 kategori dan k_c adalah jumlah usaha pada kategori c. Nilai harapan tersebut adalah 0,212 (21,2%). Capaian 72,9% berada di atas pembanding acak dan setara dengan 81,5% dari batas teoretis.

Perbandingan numerik langsung dengan penelitian terdahulu tidak dilakukan, karena studi yang ditinjau memakai data, metode, dan metrik evaluasi yang berbeda, misalnya RMSE pada [4] dan simulasi satu query pada [11], sehingga angka tersebut tidak sebanding dengan precision@5 pada penelitian ini. Sebagai gantinya, capaian sistem dibandingkan dengan dua pembanding internal yang dihitung pada data yang sama, yaitu nilai harapan pemilihan acak dan batas atas teoretis. Sejalan dengan temuan pada bagian preprocessing, hal evaluasi ini megindikasikan bahwa kualitas data turut memengaruhi kinerja sistem CBF. Dugaan ini sejalan dengan catatan pada tinjauan sistematis mengenai peran kualitas dataset dalam keandalan sistem rekomendasi [3].

4.8 Pembahasan

Penelitian ini menerapkan content-based filtering dengan TF-IDF dan cosine similarity pada 98 usaha jasa cuci sepatu di Kabupaten Klaten. Pada studi yang ditinjau, belum ditemukan penelitian CBF dengan obyek jasa perawatan sepatu, dan sebagaian di antaranya berfokus pada sektor pariwisata dan kuliner [4], [7], [8], [11], [14], [15], [16]. Hasil penelitian ini dapat dirangkum dalam tiga hal. Pertama, CBF diterapkan pada domain jasa perawatan sepatu dengan data yang hanya terdiri atas nama dan kategori usaha. Kedua, persoalan kualitas data hasil scraping listing Google Maps didokumentasikan, yaitu kesalahan input kategori, kategori yang tumpeng tindih, rating yang nyaris seragam, dan atribut yang tidak lengkap, beserta penanganannya. Ketiga, evaluasi kuantitatif dilaporkan berupa precision@5 berbasis kesesuaian kategori, disertai pembanding internal berupa nilai harapan pemilihan acak (21,2%) dan batas atas teoretris (89,4%).

Hasil evaluasi mengindikasikan bahwa sistem bekerja lebih baik daripada pemilihan acak, tetapi sumber kesalahan rekomendasi lebih terkait dengan data dan representasi fitur daripada perhitungan

 ISSN:2302-8149
e-ISSN:2540-9719

kemiripannya. Tujuh dari delapan usaha ber-precision@5 nol berkategori unik, yang tidak dapat dipenuhi oleh sistem mana pun dalam skema evaluasi ini. Pada contoh Wareshoes Care dan Boogie Lab – Shoes Cleaning UII, kesalahan berasal dari kemiripan kata yang tidak mencerminkan kesamaan layanan, yaitu frasa umum pada nama dan kata pada nama kategori yang dipakai bersama oleh banyak kategori. Selama itu, sebagian kategori bermakna tumpeng tindih, misalnya Laundry service, Shoe repair shop, dan Shoe shining service, sehingga rekomendasi yang dinilai tidak relevan oleh proxy kategori belum tentu tidak berguna bagi pengguna. Nilai precision@5 pada penelitian ini dengan demikian merupakan ukuran kesesuaian kategori, bukan ukuran kepuasan atau kecocokan menurut pengguna.

Sebagai uji eksploratif di luar rancangan utama, tiga term berdaya pembeda rendah (‘cuci’, ‘sepatu’, dan ‘klaten’) dibuang dari clean_features sebelum pembobotan TF-IDF, dengan prosedur evaluasi yang sama. Rata-rata precision@5 naik dari 0,7286 menjadi 0,7837, usaha dengan precision@5 sempurna naik dari 44 menjadi 50, usaha ber-precision@5 nol turun dari 8 menjadi 7, dan kelima rekomendasi Wareshoes Care seluruhnya berkategori Shoes shining service. Hasil ini mengindikasikan bahwa domain-specific stopwords dapat mengurangi dominasi frasa umum, tetapi sifatnya masih awal karena hanya satu konfigurasi yang dicoba pada satu dataset. Pembobotan terpisah antara nama usaha dan kategori belum diuji secara kuantitatif. Relevansi pada penelitian ini didefinisikan melalui kategori, sehingga menaikkan bobot kategori akan menaikkan precision@5 secara mekanis tanpa bukti bahwa kualitas rekomendasi benar-benar membaik. Opsi tersebut perlu dievaluasi dengan penilaian relevansi yang independent dari kategori.

Persoalan kualitas data pada penelitian ini, yaitu rating yang nyaris seragam pada 83,7% usaha, jumlah ulasan yang kosong pada 14 usaha, dan website yang kosong pada 69,4% usaha, membatasi atribut yang dapat dipakai sebagai fitur sehingga representasi hanya bertumpu pada nama dan kategori. Kondisi ini konsisten dengan temuan penelitian pada tinjauan sistematis mengenai peran kualitas data [3]. Selain itu, penerapan CBF pada sektor jasa lokal berskala mikro menghadapi tantangan yang berbeda dari penelitian di sektor pariwisata yang datanya bersumber dari instansi [14], [15], terutama karena teks deskriptifnya minim dan kualitas atribut hasil scraping-nya rendah. Penelitian ini bersifat deskriptif dan tidak menguji hubungan antara kualitas data dan kinerja secara statistic inferensial, sehingga kesimpulan tersebut masih berupa indikasi.

4.9 Keterbatasan Penelitian

Pengumpulan data hanya memakai satu kata kunci pencarian (“tempat cuci sepatu di katen”) dan satu sumber data (Google Maps) yang diambil dengan satu alat (Instant Data Scraper) pada satu waktu (9 September 2026). Usaha yang tidak muncul pada hasil pencarian kata kunci tersebut tidak tercakup, sedangkan listing yang muncul dapat memuat usaha yang bukan jasa cuci sepatu, yakni 11 entri (11,2%) berkategori Corporate office, Janitorial service, dan Shoe store yang tetap dipertahankan. Dampaknya, cakupan data dan kualitas rekomendasi bergantung pada hasil pencarian dan kelengkapan informasi yang diisi pengelola listing, dan entri yang bukan jasa cuci sepatu dapat muncul sebagai rekomendasi.

Jumlah data sebanyak 98 usaha dari satu kabupaten menghasilkan vocabulary yang kecil (207 term) dan memuat hasil peka terhadap sejumlah usaha tertentu. Tujuh usaha berkategori unik saja sudah membatasi rata-rata meksimum precision@5 pada 89,4%. Karena itu, hasil penelitian belum dapat digeneralisasikan ke wilayah lain atau jenis jasa lokal lain, dan penelitian ini tidak memakai pengujian statistik inferensial.

Sistem menerima satu usaha acuan sebagai masukan (item-to-item), sedangkan permasalahan penelitian berkaitan dengan pencarian usaha berdasarkan kebutuhan pengguna. Pengguna yang belum mengetahui satu usaha tidak dapat memperoleh rekomendasi berbasis konten. Evaluasi yang menjadikan setiap usaha sebagai acuan satu per satu belum menggambarkan pola kebutuhan pengguna yang sesungguhnya.

Relevansi hanya diukur melalui kesamaan kategori dan belum divalidasi oleh responden atau pengguna. Kategori pada Google Maps ditentukan oleh pengelola listing, dapat saling tumpang tindih, dan satu entri harus dikoreksi secara manual. Akibatnya, precision@5 sebesar 72,9% tidak dapat ditafsirkan sebagai tingkatan kepuasan pengguna. Penelitian lanjutan dapat menambah kata kunci dan

sumber data, melibatkan responden untuk menilai relevansi, serta menambahkan pencarian rekomendasi berbasis kata kunci.

5. Kesimpulan

Penelitian ini membangun sistem rekomendasi tempat cuci sepatu di Kabupaten Klaten dengan pendekatan content-based filtering dan mengevaluasinya menggunakan precision@5. Sistem dibangun dari 98 usaha hasil scraping Google Maps dengan pembobotan TF-IDF pada gabungan nama dan kategori usaha (207 term unik) serta cosine similarity sebagai ukuran kemiripan. Data mentah memuat kesalahan input kategori, kategori yang tumpeng tindih, rating yang identik pada 83,7% usaha sehingga tidak dipakai sebagai fitur, dan atribut yang tidak lengkap. Rata-rata precision@5 berbasis kesesuaian kategori mencapai 72,9% dengan 44 usaha (44,9%) memperoleh nilai sempurna, berada di atas nilai harapan pemilihan acak (21,2%) dan setara dengan 81,5% dari batas atas teoretis (89,4%). Hasil ini mengindikasikan bahwa sistem mampu menghasilkan rekomendasi yang sesuai secara kategori dan bahwa kesalahan rekomendasi lebih terkait dengan kualitas data dan representasi fitur daripada perhitungan kemiripannya, meskipun interpretasi ini masih bersifat indikatif karena tidak diuji secara statistic inferensial. Tujuan penelitian tercapai dengan catatan bahwa data berasal dari satu kata kunci, satu sumber, dan 98 usaha, masukan sistem berupa satu usaha acuan, serta relevansi diukur melalui kesamaan kategori tanpa validasi pengguna, sehingga hasilnya belum dapat digeneralisasikan dan belum menggambarkan kepuasan pengguna. Penelitian lanjutan dapat melakukan uji pengguna untuk menilai relevansi secara langsung, memakai metrik evaluasi lain seperti normalized Discounted Cumulative Gain (nDCG), mengembangkan pencarian rekomendasi berbasis kata kunci, menambahkan data ulasan tekstual dan domain-specific stopwords, serta menggabungkan content-based filtering dengan collaborative filtering.

6. Referensi

[1] D. Ridhwanullah, Y. K. Kumarahadi, and B. D. Raharja, "*Content-Based Filtering* pada Sistem Rekomendasi Buku Informatika," *Jurnal Ilmiah SINUS*, Vol. 22, No. 2, p. 57, Jul. 2024, DOI: 10.30646/sinus.v22i2.840.

[2] N. Azizah and A. F. Rozi, "Sistem Rekomendasi Produk *Somethinc* menggunakan Metode *Content-based Filtering*," *Jurnal Teknologi dan Sistem Informasi Bisnis*, Vol. 6, No. 3, pp. 461–468, Jul. 2024, DOI: 10.47233/jteksis.v6i3.1411.

[3] I. Saifudin and T. Widiyaningtyas, "*Systematic Literature Review on Recommender System: Approach, Problem, Evaluation Techniques, Datasets*," *IEEE Access*, Vol. 12, pp. 19827–19847, 2024, DOI: 10.1109/ACCESS.2024.3359274.

[4] R. Faurina and E. Sitanggang, "Implementasi Metode *Content-based Filtering* dan *Collaborative Filtering* pada Sistem Rekomendasi Wisata di Bali," *Techno.Com*, Vol. 22, No. 4, pp. 870–881, Nov. 2023, DOI: 10.33633/tc.v22i4.8556.

[5] D. Pratiwi, A. Asrianda, and L. Rosnita, "Penerapan Metode *Content-based Filtering* dalam Sistem Rekomendasi Objek Wisata di Aceh Tamiang," *Jurnal Ilmu Komputer dan Informatika*, Vol. 4, No. 2, pp. 85–96, Jan. 2025, DOI: 10.54082/jiki.169.

[6] D. T. Santoso, V. Atina, and D. Hartanti, "*Prototipe* Sistem Rekomendasi Film Indonesia menggunakan Pendekatan *Content based Filtering* dan Metode *Vector Space Model*," *Infotek: Jurnal Informatika dan Teknologi*, Vol. 7, No. 2, pp. 444–455, Jul. 2024, DOI: 10.29408/jit.v7i2.26083.

[7] S. Nurpila, H. A. Ismail, and S. A. Prakoso, "Rancang Bangun Rekomendasi Tempat Wisata di Kabupaten Rembang berbasis *Website* menggunakan Metode *Content based Filtering*," *INTECOMS: Journal of Information Technology and Computer Science*, Vol. 7, No. 4, pp. 1149–1158, Jul. 2024, DOI: 10.31539/intecoms.v7i4.10718.

[8] L. Cahyani, N. Sephiana, M. Tahir, and J. Aisyiah, "Sistem Rekomendasi Wisata Kuliner Madura menggunakan *Content based Filtering*," *Explore IT: Jurnal Keilmuan dan Aplikasi Teknik Informatika*, Vol. 16, No. 1, pp. 1–7, Jun. 2024, DOI: 10.35891/explorit.v16i1.5366.

[9] M. S. Negara and A. Z. Mardiansyah, "Implementasi Machine Learning dengan Metode *Collaborative Filtering* dan *Content-based Filtering* pada Aplikasi Mobile Travel (Bangkit Academy)," *JBegaTI*, Vol. 5, No. 1, pp. 126–136, Mar. 2024, DOI: 10.29303/jbegati.v5i1.1193.

[10] I. Suryadi, D. H. Gutama, D. Danianti, and A. S. Yazid, "Sistem Rekomendasi Tempat Wisata Alam di Gunung Kidul menggunakan Algoritma *Content based Filtering* berbasis *Web*," *JATI: Jurnal Mahasiswa Teknik Informatika*, Vol. 9, No. 2, pp. 3324–3331, Apr. 2025, DOI: 10.36040/jati.v9i2.13344.

[11] E. P. S. Arum, S. D. Larasati, A. P. Aprilia, A. Ichsan, and A. P. Mahendra, "Penerapan *Content-based Filtering* untuk Rekomendasi UMKM Kuliner di Kota Surakarta," in *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis (SENATIB)*, Surakarta, 2026, pp. 273–281. [Online]. Available: https://ojs.udb.ac.id/Senatib/article/view/6386

[12] K. Kosim and R. Prihandi, "*Recommendation System Algorithm Content-based Filtering Method to Provide Drink Menu Recommendations*," *Journal of Mathematics Instruction, Social Research and Opinion*, Vol. 2, No. 2, pp. 158–168, Jul. 2023, DOI: 10.58421/misro.v2i2.130.

[13] H. Hartatik, S. D. Nurhayati, and W. Widayani, "Sistem Rekomendasi Wisata Kuliner di Yogyakarta dengan Metode *Item-based Collaborative Filtering*," *Journal Automation Computer Information System*, Vol. 1, No. 2, pp. 55–63, Nov. 2021, DOI: 10.47134/jacis.v1i2.8.

[14] C. Salsabilla and D. W. Utomo, "*Pekalongan Regency Tourism Recommendation System with Content based Filtering*," *Sistemasi: Jurnal Sistem Informasi*, Vol. 14, No. 1, pp. 262–270, Jan. 2025, DOI: 10.32520/stmsi.v14i1.4839.

[15] S. Oyadila, D. Abdullah, and A. Razi, "Implementasi *Content-based Filtering* dengan *TF-IDF* dan *Cosine Similarity* untuk Sistem Rekomendasi Destinasi Wisata di Aceh Tengah," *RABIT: Jurnal Teknologi dan Sistem Informasi Univrab*, Vol. 10, No. 2, pp. 1329–1339, Jul. 2025, DOI: 10.36341/rabit.v10i2.6532.

[16] L. H. Aljihadu, "Sistem Rekomendasi Wisata Kuliner di Gunungkidul menggunakan Metode *Content based Filtering*," *Jurnal Informatika dan Teknik Elektro Terapan*, Vol. 13, No. 1, pp. 1435–1444, Jan. 2025, DOI: 10.23960/jitet.v13i1.5955.

[17] J. Wong and Yoannita, "Implementasi *TF-IDF* dan *Cosine Similarity* untuk Rekomendasi Resep Masakan Indonesia berdasarkan Bahan Makanan," *JATI: Jurnal Mahasiswa Teknik Informatika*, Vol. 10, No. 1, pp. 577–584, Jan. 2026, DOI: 10.36040/jati.v10i1.16816.

[18] P. Lops, M. de Gemmis, and G. Semeraro, "*Content-based Recommender Systems: State of the Art and Trends*," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Boston, MA, USA: Springer, 2011, pp. 73–105, DOI: 10.1007/978-0-387-85820-3_3.

[19] N. Hu, J. Zhang, and P. A. Pavlou, "*Overcoming the J-shaped Distribution of Product Reviews*," *Communications of the ACM*, Vol. 52, No. 10, pp. 144–147, Oct. 2009, DOI: 10.1145/1562764.1562800.

[20] M. Fadilah, Meliya, F. Q. Urdha, and I. Noviyanti, "Analisis Strategi Proaktif untuk meningkatkan Daya Saing Jasa Cuci Sepatu pada *Waves*. *Shoes laundry* di Pangkalpinang," *Jurnal Pendidikan Tambusai*, Vol. 9, No. 2, pp. 22466–22471, 2025, DOI: 10.31004/jptam.v9i2.30184.

ISSN:2302-8149
e-ISSN:2540-9719

[21] G. Salton and C. Buckley, "*Term-Weighting Approaches in Automatic Text Retrieval*," *Information Processing & Management*, Vol. 24, No. 5, pp. 513–523, 1988, DOI: 10.1016/0306-4573(88)90021-0.

[22] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge University Press, 2008, DOI: 10.1017/CBO9780511809071.

[23] R. C. Richey and J. D. Klein, "*Developmental Research Methods: Creating Knowledge from Instructional Design and Development Practice*," *Journal of Computing in Higher Education*, Vol. 16, No. 2, pp. 23–38, 2005, DOI: 10.1007/BF02961473.

[24] T. Ichwani, D. Kurniawati, M. Munira, A. Damayanti, T. Astuti, and S. M. Ani, "Pelatihan dan Pembinaan Mahasiswa Wirausaha Jasa Cuci Sepatu *SNWS (Shoes and Wash Service)*," *SULUH: Jurnal Abdimas*, Vol. 5, No. 2, pp. 184–190, Feb. 2024, DOI: 10.35814/suluh.v5i2.5203.