KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)
Terakreditasi Nomor 204/E/KPT/2022 | Vol. 5, No. 4, Oktober (2024), pp. 1636-1646

Sistem Rekomendasi Hotel Dengan Ektraksi Fitur Deskripsi Menggunakan Metode *Text Mining* dan *Content Based Filtering*

Asshiddiq[1], Lily Ayu Wulandhari[2]
[1,2]Program Studi Magister Teknik Informatika, BINUS Graduate Program, Universitas Bina Nusantara, Jakarta, Indonesia
E-mail: asshiddiq@binus.ac.id[1], lili.wulandhari@binus.ac.id[2]

Abstract

Recommendation systems have become a crucial component in various digital applications to help users find relevant products or services based on their preferences. In the context of tourism and hospitality, recommendation systems facilitate users in selecting hotels that suit their needs and preferences. An effective approach to building a recommendation system is by using Content Based Filtering techniques. This research aims to develop a hotel recommendation model that can address the cold start problem, a situation where the recommendation system struggles to provide accurate suggestions to new users or for new items that do not yet have many interactions. By using text-mining methods, hotel descriptions and amenities are extracted into important features, which are then used by measurement methods to calculate similarity scores between features to generate relevant and accurate recommendations for users. Two similarity score measurement methods compared in this study are Cosine Similarity and RBF Kernel. The similarity score measurement were conducted using 20 hotels, where each hotel selected 10 recommended hotels with the highest similarity scores. The results showed that the RBF Kernel method outperformed with an accuracy percentage of 99.8279 %. Meanwhile, the Cosine Similarity method had a slightly lower accuracy percentage of 99,8187 %.

Keywords: *Recommendation Systems, Content Based Filtering, Text Mining, Cold Start Problem, Cosine Similarity, RBF Kernel, Similarity Score*

Abstrak

Dalam konteks pariwisata dan perhotelan, sistem rekomendasi memfasilitasi pengguna dalam memilih hotel yang sesuai dengan kebutuhan dan preferensi mereka. Pendekatan efektif untuk membangun sistem rekomendasi adalah dengan menggunakan teknik Content Based Filtering. Penelitian ini bertujuan untuk mengembangkan model rekomendasi hotel yang dapat mengatasi masalah cold start, yaitu situasi di mana sistem rekomendasi kesulitan memberikan saran yang akurat kepada pengguna baru atau untuk item baru yang belum memiliki banyak interaksi. Dengan menggunakan metode text-mining, deskripsi hotel dan fasilitas diekstraksi menjadi fitur penting, yang kemudian digunakan oleh metode pengukuran untuk menghitung skor kesamaan antara fitur guna menghasilkan rekomendasi yang relevan dan akurat bagi pengguna. Dua metode pengukuran skor kesamaan yang dibandingkan dalam studi ini adalah Cosine Similarity dan RBF Kernel. Pengukuran skor kesamaan dilakukan menggunakan 20 hotel, di mana masing-masing hotel memilih 10 hotel rekomendasi dengan skor kesamaan tertinggi. Hasilnya menunjukkan bahwa metode RBF Kernel menunjukkan performa lebih baik dengan persentase akurasi 99,8279%. Sementara itu, metode Cosine Similarity memiliki persentase akurasi sedikit lebih rendah, yaitu 99,8187%.

Kata kunci: *Sistem Rekomendasi, Content Based Filtering, Text Mining, Masalah Cold Start, Cosine Similarity, RBF Kernel, Skor Kesamaan.*

ISSN: 2720-992X

1. Pendahuluan

Perkembangan teknologi informasi yang pesat menghadirkan berbagai inovasi penawaran suatu produk atau jasa. Salah satu inovasi yang mengalami evolusi dari konvensional menjadi moderen ada pada sektor pariwisata [1]. Dengan adanya berbagai macam hotel berbintang maupun non bintang yang ada di suatu destinasi wisata, tentunya masyarakat yang ingin melakukan liburan akan kebingungan dalam memilih hotel sesuai dengan kriteria maupun fasilitas yang diminati. OTA menyediakan macam-macam pilihan hotel yang dapat disesuaikan dengan kebutuhan dan keinganan masyarakat. Segala informasi mengenai lokasi, fasilitas, dan keunggulan hotel yang meliputi penjelasan mengenai kamar, restoran, kolam renang, area parkir, dan lain-lain dijelaskan dalam kolom deskripsi pada suatu *website* OTA. Sistem rekomendasi memiliki peran penting dalam perkembangan teknologi dikarenakan peningkatan produk dan layanan pada *platform online* untuk memberikan rekomendasi kepada pengguna berdasarkan preferensi mereka dengan mudah [2]. Sistem rekomendasi merupakan sebuah algoritma yang bertujuan untuk memberikan rekomendasi produk yang relevan kepada pengguna sesuai dengan preferensi mereka berdasarkan interaksi pengguna dengan produk-produk yang ada.

Berdasarkan uraian di atas, maka perlu dikembangkan suatu model sistem rekomendasi hotel untuk membantu pengguna memilih sesuai preferensi. Model ini menggunakan data fitur dari daftar fasilitas dan deskripsi hotel, dengan metode *text-mining* untuk ekstraksi fitur dari deskripsi tersebut. Pendekatan yang digunakan dalam penelitian ini adalah *vectorization* berbasis *frequency*, yaitu *Term Frequency* (TF). Pemilihan TF didasarkan pada kemampuannya untuk menangkap seberapa sering kata atau frasa muncul dalam deskripsi hotel. Dengan menggunakan TF, sistem dapat memahami konteks dan kepentingan fitur tertentu berdasarkan frekuensi kemunculannya [3]. Hal ini sangat relevan dalam konteks rekomendasi, di mana fitur-fitur yang sering muncul dalam deskripsi dapat dianggap lebih penting dan memberikan informasi yang lebih kaya tentang karakteristik hotel [4]. Penggunaan TF memungkinkan sistem untuk mengidentifikasi kata-kata kunci yang paling sering muncul, yang kemudian digunakan untuk membangun profil fitur dari setiap hotel [5]. Kemudian untuk algoritma *content-based filtering* digunakan untuk menghasilkan rekomendasi yang efektif dan mengatasi *cold start problem*. Skor kemiripan antar fitur diukur menggunakan metode *Cosine Similarity* dan *RBF Kernel*.

Hasil akhirnya adalah untuk menemukan model rekomendasi dengan kinerja terbaik, di mana model ini dikembangkan dan diintegrasikan dalam sistem rekomendasi hotel yang dirancang untuk memberikan rekomendasi yang relevan kepada pengguna berdasarkan data yang telah diproses. Proses integrasi mencakup ekstraksi fitur dari deskripsi hotel, penerapan model untuk menghitung skor kemiripan, dan penyajian hasil rekomendasi dalam bentuk daftar 10 hotel terbaik yang paling sesuai dengan preferensi pengguna.

2. Metodologi Penelitian

Pada penelitian sebelumnya yang dilakukan oleh Wahyudi dan rekan pada tahun 2020 mengembangkan sistem rekomendasi menggunakan *content-based filtering* yang mencoba merekomendasikan item serupa dengan yang telah disukai pengguna di masa lalu. Sistem ini menghitung rating kategori hotel berdasarkan kota dan menggabungkan dua fitur utama, yaitu kategori dan kota dari hotel yang dipilih. Hasil evaluasi menunjukkan bahwa sistem ini memiliki nilai precision sebesar 88,9%, akurasi 85%, dan recall 80%, yang menunjukkan peningkatan dari penelitian serupa dengan akurasi 61,37% [6].

Melanjutkan upaya ini, pada tahun 2022 Shah dan Jacob mengembangkan sistem rekomendasi hotel berbasis ulasan pelanggan menggunakan metode *content-based filtering*. Sistem ini memanfaatkan data ulasan dari *dataset "515K Hotel Reviews*

Data in Europe" yang diambil dari Kaggle, mencakup ulasan hotel dalam bahasa Inggris dari tahun 2015 hingga 2017. Algoritma yang digunakan meliputi teknik *Natural Language Processing* seperti *word2vec* dan TF-IDF untuk ekstraksi fitur. Ulasan pelanggan digunakan untuk menghitung *cosine similarity* yang kemudian diterapkan dalam model rekomendasi untuk memberikan daftar 10 hotel terbaik berdasarkan lokasi pengguna. Eksperimen menunjukkan bahwa model ini dapat memberikan rekomendasi yang akurat dan relevan. Model ini diimplementasikan dalam bentuk sistem rekomendasi yang menampilkan rekomendasi hotel beserta geolokasinya pada peta, meskipun belum direalisasikan dalam bentuk aplikasi *web* atau mesin pencari [7].

Sejalan dengan penelitian tersebut, Melyani dan rekan pada tahun 2022 yang bertujuan untuk membuat sistem rekomendasi hotel menggunakan pendekatan *content-based filtering* untuk hotel-hotel di Yogyakarta pada situs Nusatrip. Sistem ini dibangun untuk membantu calon penghuni hotel memilih hotel sesuai preferensi mereka dan membantu perusahaan meningkatkan pemesanan kamar melalui situs web *Online Travel Agent (*OTA*)*. Metode yang digunakan meliputi pembobotan data teks menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) dan pengukuran kemiripan dokumen menggunakan *cosine similarity*. Data yang digunakan adalah deskripsi dari masing-masing hotel. Hasil uji coba dengan hotel Good Karma Yogyakarta menunjukkan bahwa sistem berhasil merekomendasikan 10 hotel yang mirip dengan nilai *cosine similarity* tertinggi sebesar 0.9567 hingga 0.8715. Implementasi sistem rekomendasi ini dilakukan melalui *website* sederhana menggunakan *Flask* dan *Heroku*, memudahkan pengguna mencari hotel yang serupa dengan hotel yang dipilih sebelumnya [8].

Di bidang hiburan, Sinha dan Sharma pada tahun 2023 mengembangkan sistem rekomendasi film *content-based* yang dipersonalisasi. Sistem ini menggunakan atribut film seperti *genre*, aktor, sutradara, dan sinopsis untuk memberikan rekomendasi yang akurat dan relevan. Metodologi yang digunakan melibatkan preprocessing data untuk membersihkan dan mengubah data film mentah, vektorisasi teks untuk mengubah deskripsi film menjadi vektor numerik, dan penggunaan cosine similarity untuk mengukur kesamaan antar film. Hasil penelitian menunjukkan bahwa sistem berhasil menghasilkan rekomendasi film yang akurat dan dipersonalisasi, meningkatkan pengalaman menonton film pengguna. Meskipun demikian, sistem ini memiliki beberapa keterbatasan seperti kurangnya pemahaman semantik dan masalah skalabilitas [9].

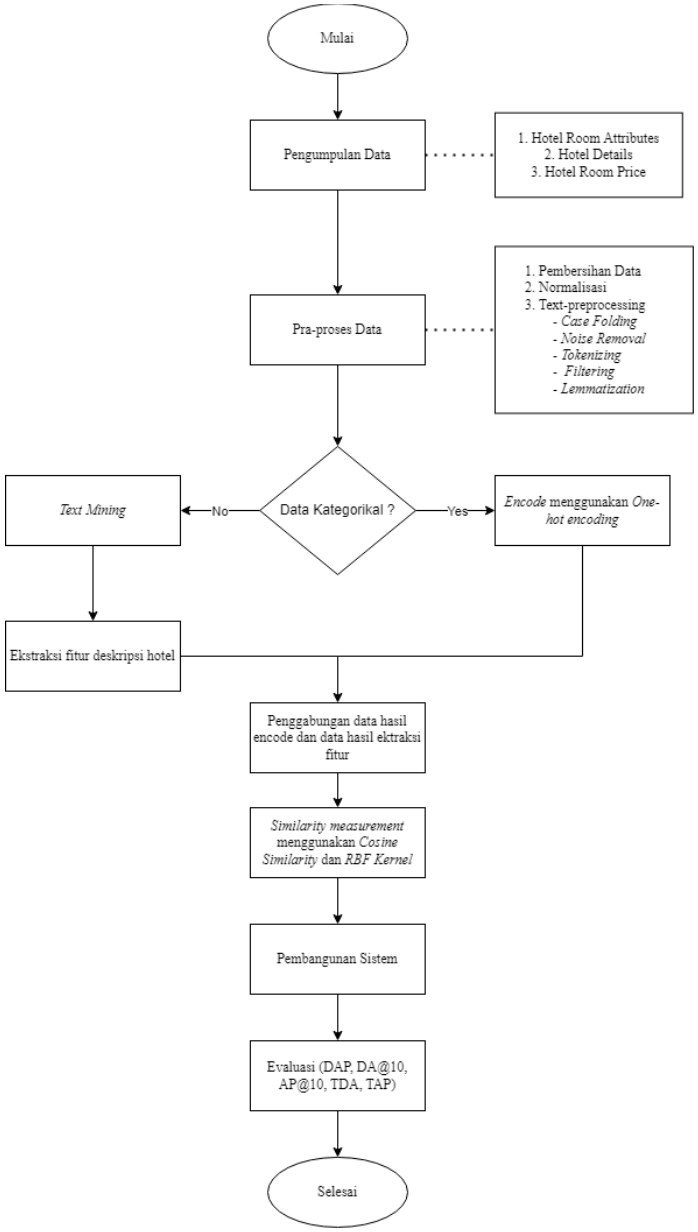
Berdasarkan penelitian-penelitian yang telah dilakukan sebelumnya, metode *Content Based Filtering* (CBF) dan *text mining* dipilih karena keunggulannya dalam mengatasi masalah *cold start* dan kemampuan untuk memberikan rekomendasi yang relevan berdasarkan konten deskriptif [10]. Penelitian sebelumnya menunjukkan bahwa metode ini efektif dalam berbagai domain, seperti rekomendasi hotel dan rekomendasi film.

Penelitian ini berkontribusi dengan melakukan ekstraksi fitur dari deskripsi hotel, yang merupakan aspek penting namun jarang digunakan dalam penelitian sebelumnya. Deskripsi hotel mengandung informasi yang kaya mengenai fasilitas, lokasi, dan layanan yang ditawarkan, yang dapat menjadi indikator penting dalam proses rekomendasi. Dengan menggunakan metode *text mining*, fitur-fitur penting seperti ukuran kamar, jenis tempat tidur, dan fasilitas tambahan dapat diekstraksi dari teks deskripsi ini. Proses ekstraksi fitur ini melibatkan langkah-langkah *text preprocessing* seperti *tokenization, case folding, filtering* dan *lemmatization*, serta penggunaan algoritma seperti TF untuk mengidentifikasi kata-kata kunci yang sering muncul. Selain itu, penelitian ini membandingkan dua metode pengukuran

kesamaan, yaitu *Cosine Similarity* dan *RBF Kernel*, untuk menghitung skor kemiripan antar fitur yang diekstraksi.

2.1. Tahapan Penelitian

Tahapan ini dirancang sesuai dengan tujuan yang ditetapkan pada kerangka pikir, yaitu untuk mengembangkan model untuk sistem rekomendasi hotel, yang dapat membantu pengguna dalam menentukan pilihan hotel disuatu tempat wisata, sesuai dengan preferensi mereka. Berikut tahapan penelitian yang dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

Proses dimulai dengan pengumpulan data dari tiga *dataset* yang digunakan yaitu *Hotel Room Attributes, Hotel Details,* dan *Hotel Room Price* yang diperoleh dari *Kaggle*. Data yang telah dikumpulkan kemudian dipra-proses melalui beberapa langkah, termasuk pembersihan data untuk menghilangkan data yang tidak dibutuhkan contohnya seperti *id, zipcode, url, curr,* dan *source*. Kemudian juga menghapus seluruh baris yang miliki data

kosong, menghapus data yang duplikat, dan melakukan normalisasi data yang bertipe numerik seperti *onsiterate*, hal ini untuk memastikan data numerik berada dalam rentang yang wajar dan tidak ada nilai ekstrim yang dapat memengaruhi analisis. Setelah pra-proses data selesai, dilakukan identifikasi apakah data bersifat kategorikal atau tidak. Jika data bersifat kategorikal, data tersebut di-*encode* menggunakan *one-hot encoding* untuk mengubahnya menjadi format numerik. Namun, jika data tidak bersifat kategorikal, data tersebut akan melalui proses *text mining* yang lebih lanjut, meliputi langkah-langkah *text-preprocessing* seperti *case folding* untuk mengubah semua huruf menjadi huruf kecil, *noise removal* untuk menghapus karakter atau simbol yang tidak relevan, *tokenization* untuk memecah teks menjadi kata-kata atau unit-unit kecil, *filtering* untuk menghapus kata-kata umum yang tidak memiliki makna penting, dan *lemmatization* untuk mengubah kata-kata menjadi bentuk dasarnya.

Dalam proses ekstraksi fitur menggunakan *Term Frequency* (TF), dimana frekuensi kemunculan setiap kata dalam teks dihitung. Dimana kata-kata yang memiliki frekuensi kemunculan tertinggi (*Top-N*) akan dijadikan fitur dan untuk kata-kata diluar *Top-N* akan dilakukan *masking* berdasarkan entitasnya, contohnya pada *roomammenities*, seluruh kata selain Top-N akan dimasking menjadi *OtherAmmenities*. Data hasil *encode* dan hasil ekstraksi fitur kemudian digabungkan untuk membentuk dataset yang lengkap. Langkah selanjutnya adalah pengukuran skor kemiripan antar fitur menggunakan dua metode, yaitu *Cosine Similarity* dan *RBF Kernel*.

Hasil pengukuran skor kemiripan antar fitur ini digunakan untuk membangun model sistem rekomendasi yang dapat memprediksi dan memberikan rekomendasi hotel yang sesuai dengan preferensi pengguna. Evaluasi sistem dilakukan menggunakan 5 metrik, yaitu:

a) *Difference Average Per Hotel* (DAP): Metrik ini mengukur rata-rata perbedaan antar fasilitas hotel per hotel. Formula untuk metrik ini dapat dilihat pada persamaan 1:

$$DAP = \frac{1}{m}\sum_{j=1}^{m} |x_{ij-}x_{1j}| \quad (1)$$

b) *Difference Average at 10 Hotels* (DA@10): Metrik ini mengukur rata-rata perbedaan fasilitas pada 10 hotel teratas yang dianalisis. Formula untuk metrik ini dapat dilihat pada persamaan 2:

$$DA@10 = \frac{1}{10}\sum_{i=2}^{11} DAP_i \quad (2)$$

c) *Accuracy Percentage at 10 Hotels* (AP@10): Metrik ini menunjukkan persentase rata-rata perbedaan pada 10 hotel teratas. Formula untuk metrik ini dapat dilihat pada persamaan 3:

$$AP@10 = \left(\frac{100 - DA@10}{100}\right) \times 100\% \quad (3)$$

d) *Total Difference Average* (TDA): Metrik ini mengukur rata-rata akhir dari perbedaan fasilitas di semua hotel. Formula untuk metrik ini pada persamaan 4:

$$TDA = \frac{1}{n}\left(\frac{1}{10}\sum_{i=1}^{n} F_i\right) \quad (4)$$

e) *Total Accuracy Percentage* (TAP): Metrik ini menunjukkan persentase rata-rata akhir dari perbedaan fasilitas di semua hotel. Formula untuk metrik ini pada persamaan 5:

$$TAP = \left(\frac{100 - TDA}{100}\right) \times 100\% \quad (5)$$

Keterangan:

x_{ij} = Nilai fasilitas ke-j untuk hotel ke-i.
x_{1j} = Nilai fasilitas ke-j untuk hotel pertama *(input)*.
m = Jumlah total fitur.
DAP_i = Nilai rata-rata perbedaan fasilitas untuk hotel ke-i.
F_i = Nilai akhir perbedaan fasilitas untuk hotel ke-i.
n = Jumlah hotel *input*.

2.2. Evaluasi

Untuk mengevaluasi sistem rekomendasi hotel, akan digunakan 20 *base hotel* yang dipilih berdasarkan *rating*, yaitu hotel dengan rating bintang 3 hingga 5, serta kelengkapan deskripsi, di mana hotel yang memiliki deskripsi lengkap dan rinci mengenai fasilitas dan layanan yang disediakan. Proses pemilihan ini memastikan bahwa *base hotel* yang digunakan dalam sistem rekomendasi memiliki data yang kaya dan relevan, seperti jumlah dan detail fitur deskripsi fasilitas kamar (misalnya, *Air conditioning, Free Wi-Fi in all rooms!, TV, dan Heating*), jenis kamar (misalnya, *Single Room*, dan *Double Room, Suite*), dan jenis layanan yang ditawarkan (misalnya, *Free breakfast, Room service*, dan *Executive lounge access*). Dengan demikian, hotel-hotel tersebut memberikan informasi yang komprehensif, memungkinkan sistem untuk menghasilkan rekomendasi yang lebih akurat dan sesuai dengan preferensi pengguna.

Setiap hotel dalam sampel ini kemudian akan diberikan 10 rekomendasi berdasarkan urutan *similarity score* dari yang tertinggi hingga terendah. Evaluasi akan dilakukan menggunakan 5 metrik yang telah dijelaskan sebelumnya pada sub bab 3.1 yaitu *Difference Average Per Hotel* (DAP), *Difference Average at 10 Hotels* (DA@10), *Accuracy Percentage at 10 Hotels* (AP@10), *Total Difference Average* (TDA), dan *Total Accuracy Percentage* (TAP).

Hasil dari pengukuran ini kemudian dievaluasi dan dianalisis performanya sesuai dengan distribusi nilai evaluasi performa. Analisis ini mencakup perbandingan nilai rata-rata, variasi, serta persentase kesesuaian prediksi untuk memastikan bahwa sistem rekomendasi bekerja dengan optimal dalam berbagai skenario. Dengan demikian, diharapkan sistem ini mampu memberikan rekomendasi hotel yang relevan dan akurat bagi pengguna.

3. Hasil dan Pembahasan

Dataset pertama yang akan dianalisis adalah *Hotel Room Attributes*, dimana terdapat 165,873 data dengan 3 fitur awal yaitu *Room Ammenities*, *Room Type* dan *Rate Description*, untuk sampel data dapat dilihat pada Tabel 1.

Fitur *Room Ammenities*, dan *Rate Description* terdapat 4,819 data yang hilang. Nantinya entri yang memiliki data yang hilang akan dihapus. Langkah ini dilakukan untuk memastikan bahwa analisis dan perhitungan skor kemiripan yang akan dilakukan tidak terpengaruh oleh ketidaksempurnaan data. Sedangkan pada fitur *Room Type* tidak terdapat data yang hilang. Pada fitur *Room Ammenities*, nantinya akan dilakukan *data preprocessing*, untuk mendapatkan informasi terkait sebaran data *ammenities*. Dari sebaran data fasilitas, memiliki variasi yang cukup banyak, sehingga nantinya akan diambil 30 ammenities dengan frekuensi kemunculan tertinggi, dan sisanya dikategorikan pada data *"OtherAmmenities"*.

Tabel 1. Sampel data pada dataset *Hotel Room Attributes*

Id	Hotel code	Room Ammenities	Room Type	Rate Description
70138124	449128	Air conditioning, Carpeting, Closet, Free Wi-Fi	Single Room	Room size: 20 m²/215 ft², City view, Shower and bathtub, 1 double bed or 2 single beds

Pada fitur *Room Type*, juga memiliki sebaran data yang variatif, sehingga pada fitur ini hanya diambil 10 *room type*, dengan frekuensi kemunculan tertinggi, sedangkan sisanya dikategorikan pada data *"OtherRoomType"*. Kemudian pada fitur *Rate Description*, akan dilakukan *feature extraction*. Seperti yang ditunjukkan pada Tabel 1, awalnya data disimpan sebagai *string* tunggal dengan informasi yang dipisahkan oleh tanda koma (","), oleh karena itu fitur ini perlu dilakukan beberapa teknik *text-preprocessing* dengan tahapan seperti berikut:

a) Case Folding
Dilakukan untuk mengubah semua huruf dalam teks menjadi huruf kecil [11]: *"room size: 20 m²/215 ft², city view, shower and bathtub, 1 double bed or 2 single beds"*.

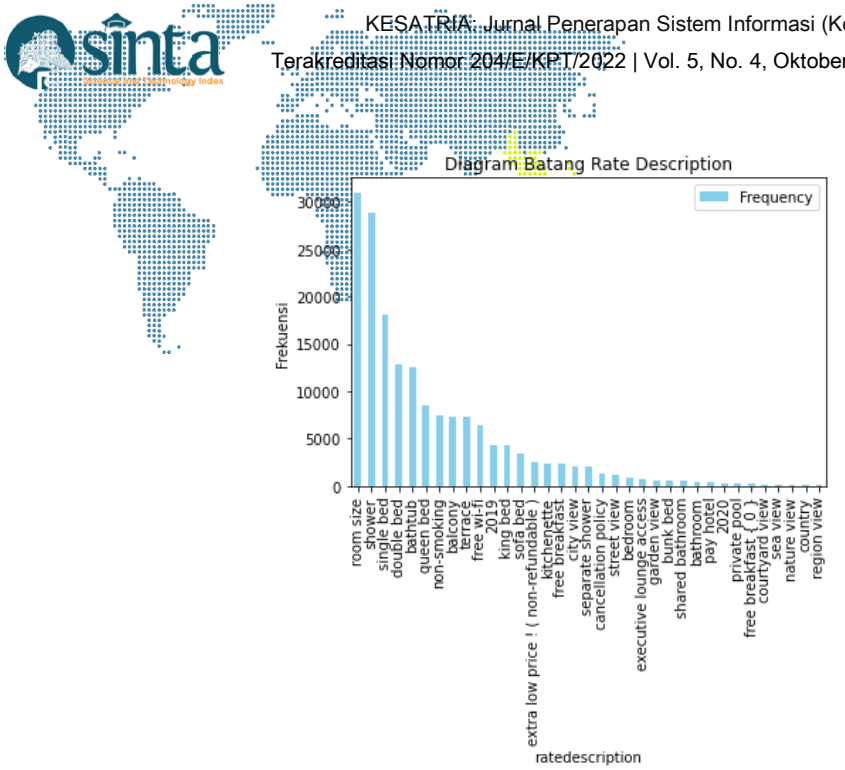
b) Tokenizing
Dilakukan untuk memisahkan teks menjadi unit-unit yang lebih kecil, seperti kata atau frasa, yang disebut token [12]. *String* yang telah menjadi huruf kecil semua kemudian dipecah menjadi token-token: *["room", "size:", "20", "m²/215", "ft²,", "city", "view,", "shower", "and", "bathtub,", "1", "double", "bed", "or", "2", "single", "beds"]*.

c) Filtering
Token-token umum yang tidak memberikan banyak informasi kontekstual dihapus: *["room", "size:", "20", "m²/215", "ft²,", "city", "view,", "shower", "bathtub,", "1", "double", "bed", "2", "single", "beds"]*.

d) Lemmatization
Setiap token diubah ke bentuk dasarnya: *["room", "size:", "20", "m²/215", "ft²,", "city", "view,", "shower", "bathtub,", "1", "double", "bed", "2", "single", "bed"]*.

e) Pattern Matching
Menggunakan *regular expression* untuk menemukan pasangan *"key: value"* pada entri. Seperti: *"room size: 20 m²/215 ft²"* cocok dengan pola ini dan diproses menjadi *tuple* seperti: *("room size", 20)*. Jika tidak cocok dengan *pattern*, maka akan diberikan nilai 1 , seperti: *"City view"*, maka akan menjadi *tuple* seperti: *("city view", 1)*.

Hasil akhir dari *text-preprocessing* ini adalah daftar pasangan *tuple* yang terstruktur, seperti: *[(room size, 20), (city view, 1), (shower, 1), (bathtub, 1), (double bed, 1), (single bed, 2)]*. Dimana nantinya nilai index 0 pada *tuple*, akan menjadi nama fitur, sedangkan nilai index 1 pada *tuple*, akan menjadi nilai untuk fitur tersebut. Setelah *text-preprocessing* selesai dilakukan, maka frekuensi kemunculan kata untuk menentukan 30 kata yang paling sering muncul dapat dihitung, yang dapat dilihat pada Gambar 2. Kata-kata inilah yang akan diekstrak menjadi fitur.



Gambar 2. 30 kata dengan frekuensi tertinggi dari fitur *ratedescription*

Kemudian, dataset selanjutnya yang akan dianalisis adalah *Hotel Details*. Dimana dari dataset ini akan digunakan 5 fitur yaitu *hotelname*, *city*, *country*, *propertytype*, dan *starrating*. Pemilihan fitur didasarkan pada relevansi dan kegunaannya dalam pembuatan sistem rekomendasi, sedangkan fitur lainnya lebih bersifat identifikasi atau metadata yang tidak memberikan informasi langsung yang dapat digunakan untuk pembuatan sistem rekomendasi, sampel data dapat dilihat pada Tabel 2. Pada dataset ini kelima fitur yang digunakan tidak terdapat data yang hilang. Data pada fitur *city*, dan *country* memiliki variasi yang cukup tinggi, sehingga akan diambil 10 data dengan frekuensi kemunculan tertinggi, sedangkan sisanya akan dikategorikan sebagai *'OtherCity'* dan *'OtherCountry'*. Untuk fitur *propertytype*, dikarenakan hanya memiliki 8 variasi, sehingga akan digunakan seluruh variasinya. Untuk fitur *starrating*, tidak terdapat data *outlier*, dimana rentang nilai pada fitur ini berkisar antara 2.0 hingga 4.0.

Tabel 2. Sampel data pada dataset Hotel Details

Hotel Name	City	Country	Property Type	Star Rating
Mediteran Bungalow	Omis	Croatia	Holiday parks	4

Terakhir, dataset yang akan dianalisis adalah *Hotel Room Price*. Dimana dari dataset ini akan digunakan 3 fitur yaitu *roomtype*, *onsiterate* dan *mealinclusiontype*. Ketiga fitur ini dipilih karena fitur-fitur ini dianggap paling relevan dan memiliki pengaruh cukup besar terhadap penetapan harga dan keputusan pengguna, sampel data dapat dilihat pada Tabel 3.

Seluruh fitur pada dataset ini tidak terdapat data yang hilang, fitur *roomtype* nantinya digunakan untuk menggabungkan dataset ini dengan dataset *Hotel Room Attributes*, yang juga memiliki fitur yang *roomtype*. Untuk fitur *onsiterate* terdapat data *outlier*, karena pada fitur ini terdapat data anomali yaitu bernilai jauh lebih tinggi dibandingkan dengan sebagian besar nilai lainnya dalam dataset. Nilai tersebut dianggap sebagai data yang hilang karena umumnya tidak ada *onsiterate* yang mencapai hingga di atas 10.000 atau

bahkan 15.000. *Outlier* ini menunjukkan adanya harga yang sangat tidak biasa dan jauh dari kisaran harga umum yang biasanya berada di bawah 2.500. Oleh karena itu, fitur *onsiterate* akan dilakukan normalisasi untuk mengurangi dampak *outlier* terhadap hasil analisis model.

Tabel 2. Sampel data pada dataset *Hotel Room Price*

roomtype	onsiterate	mealinclusiontype
Double Room	82.36	Free Breakfast

Setelah proses pembersihan data dan transformasi data, kemudian seluruh fitur yang tidak dibutuhkan untuk perhitungan *similarity score* dihapus. Sehingga total fitur yang ditemukan setelah melakukan praproses data menjadi 108 fitur, dimana untuk sampel data hasil praproses dapat dilihat pada Tabel 4.

Tabel 4. Sampel data hasil praproses

room size	shower	single bed	double bed	bathtub
22	1	0	0	0
15	1	1	0	0

Data hasil praproses kemudian digunakan untuk menghitung skor kemiripan antar fitur menggunakan metode *cosine similarity* dan *rbf kernel* dimana untuk sampel hasil menggunakan *input "Grouse & Claret by Marston's Inns"* dengan 3 hotel dengan skor kemiripan tertinggi dapat dilihat pada Tabel 5 dan Tabel 6.

Tabel 5. Sampel hasil rekomendasi menggunakan metode *cosine similarity*

Recommendation Hotel	Difference Average Per Hotel	Difference Average at 10 Hotel	Percentage at 10 Hotel
Central Hotel 21	0,0648	0,0806	99,9194
Hotel balladins la Roche-sur-Yon	0,0556		
Hotel Sandra	0,0741		

Tabel 6. Sampel hasil rekomendasi menggunakan metode *rbf kernel*

Recommendation Hotel	Difference Average Per Hotel	Difference Average at 10 Hotel	Percentage at 10 Hotel
Hotel balladins la Roche-sur-Yon	0,0556	0,0630	99,9370
Lefkoniko Bay	0,0556		
Magatzem 128	0,0648		

Hasil akhir dari rekomendasi hotel menggunakan metode *cosine similarity* yaitu nilai *Total Difference Average* adalah 0.1958, dan nilai *Total Accuracy Percentage* adalah 99.8187 %. Sedangkan untuk hasil rekomendasi hotel menggunakan metode *rbf kernel* yaitu nilai *Total Difference Average* adalah 0.1859, dan nilai *Total Accuracy Percentage* adalah 99.8279 %.

Dari kedua metode pengukuran tersebut, dapat dilihat bahwa metode *RBF Kernel* lebih unggul dibandingkan *Cosine Similarity* dalam menghasilkan rekomendasi hotel. Metode *RBF Kernel* dapat mencapai performa yang baik dengan melakukan, *feature normalization*, dan *feature extraction*. *Feature normalization* dilakukan menggunakan *Standard Scaler*, sedangkan untuk *feature extraction* dilakukan dengan menggunakan teknik *text-mining* hasil dari *text-preprocessing*.

4. Kesimpulan

Metode pengukuran skor kemiripan antar fitur terbaik untuk menghasilkan rekomendasi hotel adalah *RBF Kernel* memiliki nilai *Total Accuracy Percentage* sebesar 99.8279 %, yang menunjukkan bahwa metode pengukuran tersebut sangat efektif dalam mengidentifikasi kesamaan antara berbagai fitur hotel. Keakuratan yang tinggi ini memastikan bahwa rekomendasi yang diberikan kepada pengguna sangat relevan dan sesuai dengan preferensi mereka. Selain itu sistem rekomendasi yang dikembangkan juga berhasil mengatasi *cold start problem.* Dengan menggunakan metode *text-mining* untuk ekstraksi fitur, sistem ini dapat memberikan rekomendasi yang relevan bahkan untuk pengguna baru atau item baru yang belum memiliki banyak interaksi. Hal ini menunjukkan bahwa sistem ini dapat tetap memberikan rekomendasi yang akurat meskipun terdapat keterbatasan data awal.

Daftar Pustaka

[1] S. S. Wachyuni and K. Wiweka, “Kepuasan Wisatawan dalam Penggunaan E-Commerce Agoda dalam Pemesanan Hotel,” Journal of Tourism Destination and Attraction, vol. 8, no. 1, pp. 61–70, 2020.

[2] L. Hendriyati, “Pengaruh online travel agent terhadap pemesanan kamar di hotel mutiara malioboro yogyakarta,” Media Wisata, vol. 17, no. 1, 2019.

[3] F. Shehzad, A. U. Rehman, K. Javed, K. A. Alnowibet, H. A. Babri, and H. T. Rauf, “Binned Term Count: An Alternative to Term Frequency for Text Categorization,” Mathematics, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:253401332.

[4] Z. Hussain, B. Mago, A. Khadim, and K. Amjad, “An Intelligent Data Analysis for Recommendation Systems Using Machine Learning,” 2023 International Conference on Business Analytics for Technology and Security (ICBATS), pp. 1–7, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:258738917.

[5] R. Ojino, L. Mich, and N. H. Mvungi, “Hotel room personalization via ontology and rule-based reasoning,” Int. J. Web Inf. Syst., vol. 18, pp. 369–387, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:251971957.

[6] K. Wahyudi, J. Latupapua, R. Chandra, and A. S. Girsang, “Hotel content-based recommendation system,” in Journal of Physics: Conference Series, IOP Publishing, 2020, p. 012017.

[7] H. Shah and L. Jacob, “Hotel Recommendation System Based on Customer’s Reviews Content Based Filtering Approach,” in Proceedings - 2022 4th International Conference on Advances in Computing, Communication Control and Networking, ICAC3N 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 222–226. doi: 10.1109/ICAC3N56670.2022.10074228.

[8] C. A. Melyani *et al.*, “Hotel Recommendation System with Content-Based Filtering Approach (Case Study: Hotel in Yogyakarta on Nusatrip Website),” 2022. [Online]. Available: www.unipasby.ac.id.

[9] S. Sinha and T. Sharma, “Content-Based Movie Recommendation System: An Enhanced Approach to Personalized Movie Recommendations,” International Journal of Innovative Research in Computer Science and Technology, vol. 11, no. 3, pp. 67–71, May 2023, doi: 10.55524/ijircst.2023.11.3.12.

[10] A. Patel, N. Shah, B. Parul, and K. S. Suthar, "Hotel Recommendation using Feature and Machine Learning Approaches: A Review," 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 1144–1149, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:257536341.

[11] M. F. Juna and M. Hayaty, "The observed preprocessing strategies for doing automatic text summarizing," Computer Science and Information Technologies, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:260585342.

[12] P. Prakrankamanant and E. Chuangsuwanich, "Tokenization-based data augmentation for text classification," 2022 19th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 1–6, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:251167768.