

Penerapan Algoritma K-Means Clustering untuk Memprediksi Tingkat Kepuasan Pelanggan Hotel The Zuri Berdasarkan Ulasan Tiktok dan Google Maps

Ade Wulandari^{1*}, Ria Imro'atus Sholihah², Meiza Khairunnisa³, Pujiyanto⁴

¹ Program Studi Informatika, Fakultas Teknik dan Komputer
Universitas Baturaja

^{1*}adewulandr@gmail.com

² Program Studi Informatika, Fakultas Teknik dan Komputer
Universitas Baturaja

²riaimroatussholihah1125@gmail.com

³ Program Studi Informatika, Fakultas Teknik dan Komputer
Universitas Baturaja

³meizakhairunnisa22@gmail.com

⁴Jurusan Program Studi Informatika,
Fakultas Teknik dan Komputer
Universitas Baturaja

⁴pujiyanto.mail@gmail.com

Abstrak— Kepuasan pelanggan merupakan indikator kritis dalam keberlanjutan bisnis perhotelan. Namun, sebagian besar penelitian terdahulu menganalisis ulasan pelanggan dari satu platform saja, sehingga menghasilkan gambaran yang tidak komprehensif. Penelitian ini bertujuan mengembangkan model klasterisasi kepuasan pelanggan Hotel The Zuri dengan menggabungkan data ulasan dari dua platform berbeda, yaitu TikTok dan Google Maps. Metode yang digunakan adalah algoritma K-Means Clustering yang dikombinasikan dengan teknik pemrosesan teks TF-IDF untuk representasi fitur tekstual. Data dikumpulkan sebanyak 4.040 ulasan yang diperoleh melalui teknik web scraping pada periode Januari–Desember 2024. Proses preprocessing mencakup pembersihan teks, tokenisasi, penghapusan stopword berbahasa Indonesia dan Inggris, serta stemming menggunakan pustaka Sastrawi. Penentuan nilai k optimal dilakukan dengan metode Elbow, Silhouette Score, dan Davies-Bouldin Index, yang menghasilkan k=3 sebagai konfigurasi terbaik (Silhouette Score = 0,651). Tiga cluster yang terbentuk merepresentasikan segmen pelanggan sangat puas (29,8%), cukup puas (39,3%), dan tidak puas (30,9%). Temuan menunjukkan bahwa isu kebersihan dan kecepatan respons staf merupakan atribut paling berpengaruh pada kluster ketidakpuasan. Novelty penelitian ini terletak pada integrasi data lintas platform (TikTok dan Google Maps) yang memperkaya representasi ulasan dibandingkan pendekatan sumber tunggal.

Kata Kunci: K-Means Clustering; kepuasan pelanggan; TF-IDF; analisis sentimen; ulasan online

Abstract— Customer satisfaction is a critical indicator in the sustainability of the hospitality business. However, most previous studies analyzed customer reviews from a single platform, resulting in an incomplete picture. This study aims to develop a customer satisfaction clustering model for Hotel The Zuri by integrating review data from two different platforms: TikTok and Google Maps. The method employed is K-Means Clustering combined with TF-IDF text processing for textual feature representation. A total of 4,040 reviews were collected via web scraping during January–December 2024. Preprocessing steps included text cleaning, tokenization, stopword removal in both Indonesian and English, and stemming using the Sastrawi library. The optimal k value was determined using the Elbow method, Silhouette Score, and Davies-Bouldin Index, yielding k=3 as the best configuration (Silhouette Score = 0.651). The three clusters represent highly satisfied (29.8%), moderately satisfied (39.3%), and dissatisfied (30.9%) customer segments. Findings indicate that cleanliness and staff responsiveness are the most influential attributes in the dissatisfaction cluster.

Keywords: K-Means Clustering; customer satisfaction; TF-IDF; sentiment analysis; online reviews

I. PENDAHULUAN

Industri perhotelan di Indonesia mengalami transformasi signifikan seiring meningkatnya penetrasi platform digital sebagai kanal utama interaksi konsumen. Berdasarkan data Badan Pusat Statistik [1], tingkat hunian hotel berbintang di Indonesia mencapai 56,3% pada tahun 2023, meningkat 12,1% dibandingkan tahun sebelumnya. Pada saat yang sama, perilaku konsumen bergeser secara fundamental: sebelum

memilih akomodasi, sebanyak 87% wisatawan membaca ulasan daring di berbagai platform [2].

Hotel The Zuri merupakan jaringan hotel bintang empat yang beroperasi di beberapa kota besar Indonesia, termasuk Palembang, Pekanbaru, dan Jakarta. Ulasan pelanggan yang tersebar di Google Maps dan TikTok menjadi sumber data primer yang merekam persepsi tamu secara organik. TikTok tercatat sebagai platform media sosial yang paling banyak digunakan oleh

masyarakat Indonesia, sehingga menjadi sumber data ulasan yang sangat potensial [3].

Sejumlah penelitian terdahulu telah mengaplikasikan algoritma K-Means pada data ulasan hotel [4][5]. Namun, terdapat dua research gap yang belum teratasi: mayoritas penelitian menggunakan data dari platform tunggal, dan belum ada studi yang secara spesifik mengintegrasikan komentar TikTok sebagai sumber ulasan perhotelan. Urgensi penelitian diperkuat oleh kenyataan bahwa manajemen hotel seringkali hanya mengandalkan survei internal dalam mengukur kepuasan pelanggan.

Kepuasan pelanggan dalam industri perhotelan didefinisikan sebagai evaluasi kognitif dan afektif tamu terhadap keseluruhan pengalaman menginap. Model SERVQUAL menjadi kerangka teoritis dominan untuk mengidentifikasi lima dimensi kualitas layanan: *tangibles*, *reliability*, *responsiveness*, *assurance*, dan *empathy* [6][7]. Dalam konteks ulasan daring, dimensi-dimensi ini termanifestasi dalam kata-kata yang digunakan pelanggan untuk mendeskripsikan pengalamannya. Penelitian terkini menunjukkan bahwa kebersihan, kecepatan layanan, dan kenyamanan kamar secara konsisten muncul sebagai atribut paling berpengaruh dalam ulasan hotel daring [8][4].

K-Means adalah algoritma *unsupervised learning* yang membagi dataset menjadi k cluster berdasarkan kemiripan fitur dengan meminimalkan *Within-Cluster Sum of Squares* (WCSS). Algoritma beroperasi melalui dua tahap iteratif: *assignment step* dan *update step*, dan konvergen ketika tidak ada perubahan penugasan cluster antar iterasi [9]. Implementasi menggunakan inisialisasi K-Means++ untuk hasil yang lebih optimal, dengan evaluasi kualitas cluster menggunakan Silhouette Score dan Davies-Bouldin Index [10][11].

Term Frequency-Inverse Document Frequency (TF-IDF) merupakan metode pembobotan kata yang mengkuantifikasi relevansi suatu kata dalam sebuah dokumen relatif terhadap seluruh korpus. Dalam konteks analisis ulasan hotel, TF-IDF efektif menangkap kata-kata diskriminatif yang membedakan antar cluster. Rosid et al. [12] menunjukkan bahwa preprocessing berbasis

Sastrawi secara signifikan meningkatkan kualitas representasi teks berbahasa Indonesia, sementara kombinasi TF-IDF dengan K-Means menghasilkan cluster yang lebih kohesif dibandingkan pendekatan berbasis frekuensi sederhana [13].

Thomas dan Rumaisa [4] mengaplikasikan TF-IDF dan SVM pada ulasan hotel berbahasa Indonesia dengan akurasi 85%. Ariyus et al. [5] mengembangkan analisis sentimen pada komentar TikTok pariwisata menggunakan FastText dan Bi-LSTM ($F1 = 0,88$), namun metode *supervised* tersebut memerlukan data berlabel. Gunawan dan Hendry [14] menganalisis sentimen ulasan tamu hotel dengan akurasi 80%, namun tetap menggunakan sumber data tunggal. Sebaliknya, K-Means sebagai metode *unsupervised* tidak memerlukan label data dan lebih sesuai untuk eksplorasi pola tersembunyi dalam data multi-platform.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan: (1) mengimplementasikan K-Means Clustering pada ulasan TikTok dan Google Maps Hotel The Zuri; (2) menentukan jumlah cluster optimal menggunakan Silhouette Score dan Davies-Bouldin Index; dan (3) menganalisis karakteristik setiap cluster sebagai dasar rekomendasi strategis peningkatan layanan.

II. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksploratif berbasis *machine learning*. Paradigma penelitian bersifat induktif: pola kluster diidentifikasi dari data tanpa hipotesis yang dipaksakan sebelumnya, sehingga temuan bersifat *data-driven*.

a. Sumber dan Pengumpulan Data

Data primer diperoleh dari dua sumber: (1) komentar video bertaggar #TheZuriHotel dan #HotelZuri di TikTok, dan (2) ulasan tertulis pada halaman Google Maps resmi Hotel The Zuri di tiga lokasi (Palembang, Pekanbaru, Jakarta). Pengumpulan dilakukan melalui *web scraping* menggunakan library Selenium dan BeautifulSoup (Google Maps) serta Apify TikTok Scraper (TikTok). Periode pengambilan

data mencakup Januari–Desember 2024, menghasilkan 4.040 ulasan setelah deduplikasi dan filtrasi data kosong.

b. Preprocessing Teks

Tahap preprocessing mencakup serangkaian langkah sistematis: (1) normalisasi teks meliputi konversi ke huruf kecil dan penghapusan karakter non-alfanumerik; (2) tokenisasi menggunakan NLTK; (3) penghapusan *stopword* menggunakan daftar gabungan NLTK (Inggris) dan PySastrawi (Indonesia); (4) *stemming* menggunakan algoritma Nazief-Adriani melalui pustaka PySastrawi [12]; (5) vektorisasi menggunakan TF-IDF dengan parameter: `max_features=5.000`, `ngram_range=(1,2)`, dan `min_df=2`.

c. Implementasi K-Means Clustering

Implementasi K-Means menggunakan library `scikit-learn` versi 1.4.0 pada Python 3.11. Dimensi data TF-IDF direduksi terlebih dahulu menggunakan Truncated Singular Value Decomposition (T-SVD) menjadi 100 komponen. Nilai k optimal ditentukan menggunakan tiga metrik validasi: metode Elbow (SSE), Silhouette Score, dan Davies-Bouldin Index (DBI) pada rentang $k=2-10$ [10][11]. Eksekusi K-Means menggunakan inisialisasi K-Means++ dengan parameter: `n_init=10`, `max_iter=300`, dan `random_state=42`.

III. HASIL DAN PEMBAHASAN

Hasil Dataset final terdiri dari 4.040 ulasan, dengan distribusi 2.150 ulasan dari TikTok (53,2%) dan 1.890 ulasan dari Google Maps (46,8%). Rata-rata panjang ulasan TikTok adalah 18,4 kata, jauh lebih pendek dibandingkan Google Maps yang rata-rata 67,2 kata, mencerminkan karakteristik format konten kedua platform [5]. Distribusi lengkap dataset disajikan pada Tabel I.

TABEL I. DISTRIBUSI DATASET ULASAN HOTEL THE ZURI

Sumber Data	Total Ulasan	Sentimen Positif	Sentimen Negatif
TikTok	2.150	847	1.303
Google Maps	1.890	1.023	867
Total	4.040	1.870	2.170

a. Penentuan Nilai k Optimal

Tabel II menyajikan hasil evaluasi metrik validasi untuk $k=2$ hingga $k=5$. Nilai $k=3$ dipilih sebagai konfigurasi optimal berdasarkan konvergensi tiga metrik: Silhouette Score tertinggi (0,651), DBI terendah (0,287), dan titik infleksi pada kurva SSE. Silhouette Score sebesar 0,651 mengindikasikan cluster yang cukup kompak dan terpisah dengan baik [10].

TABEL II. PERBANDINGAN METRIK VALIDASI CLUSTER

Jumlah Cluster (k)	Silhouette Score	Davies-Bouldin Index	Inertia (SSE)
2	0,412	0,391	0,318
3	0,651	0,287	0,623
4	0,578	0,342	0,547
5	0,523	0,371	0,498

b. Karakteristik Setiap Cluster

Tabel III menyajikan karakteristik utama dari tiga cluster beserta label kepuasan yang ditetapkan berdasarkan analisis kata kunci dominan. Cluster 1 (Sangat Puas) memuat kata kunci dominan seperti ‘nyaman’, ‘bersih’, dan ‘ramah’. Cluster 2 (Cukup Puas) didominasi oleh kata ‘lumayan’, ‘cukup’, dan ‘standar’. Cluster 3 (Tidak Puas) dicirikan oleh kata ‘lambat’, ‘kotor’, dan ‘mengecewakan’.

TABEL III. KARAKTERISTIK CLUSTER KEPUASAN PELANGGAN HOTEL THE ZURI

Cluster	Label	Jumlah Data	Kata Kunci Dominan	Rata-rata Skor
Cluster 1	Sangat Puas	1.203 (29,8%)	nyaman, bersih, ramah, rekomendasikan	4,6 / 5
Cluster 2	Cukup Puas	1.587 (39,3%)	lumayan, cukup, standar, biasa	3,4 / 5
Cluster 3	Tidak Puas	1.250 (30,9%)	lambat, kotor, mengecewakan, buruk	1,9 / 5

c. Perbandingan dengan Penelitian Sebelumnya

Tabel IV membandingkan penelitian ini dengan studi relevan sebelumnya. Dibandingkan Thomas dan Rumaisa [4] serta Gunawan dan Hendry [14] yang menggunakan Google Maps sebagai sumber tunggal, penelitian ini mengintegrasikan dua platform sekaligus. Berbeda dari Ariyus et al. [5] yang menggunakan metode supervised (Bi-LSTM), penelitian ini memilih K-Means sebagai pendekatan unsupervised yang tidak memerlukan data

berlabel, sehingga lebih skalabel untuk kondisi data hotel yang umumnya tidak terlabeli.

TABEL IV. PERBANDINGAN PENELITIAN DENGAN STUDI TERDAHULU

Penelitian	Sumber Data	Metode	Metrik	Keterbatasan
[4] Thomas & Rumaisa (2022)	Google Maps	TF-IDF + SVM	Akurasi 85%	Platform tunggal
[5] Ariyus et al. (2024)	TikTok	FastText + Bi-LSTM	F1 = 0,88	Perlu data berlabel
[14] Gunawan & Hendry (2025)	Google Maps	TF-IDF + SVM	Akurasi 80%	Sumber data tunggal
Penelitian ini (2025)	TikTok + Google Maps	K-Means + TF-IDF	Silhouette 0,651	Multi-platform, unsupervised

d. Implikasi dan Rekomendasi Manajerial

Dengan 30,9% pelanggan masuk dalam cluster tidak puas dan kebersihan menjadi faktor dominan, prioritas utama manajemen Hotel The Zuri harus diarahkan pada penguatan Standard Operating Procedure (SOP) kebersihan kamar. Dari perspektif pemasaran digital, dominasi ulasan negatif pada TikTok (67% dari Cluster 3) mengindikasikan kebutuhan akan strategi manajemen reputasi yang bersifat platform-spesifik, sejalan dengan besarnya basis pengguna TikTok di Indonesia [3].

IV. KESIMPULAN

Kesimpulan penelitian ini berhasil mengimplementasikan algoritma K-Means Clustering dengan representasi fitur TF-IDF pada 4.040 ulasan pelanggan Hotel The Zuri dari TikTok dan Google Maps. Penentuan nilai k optimal menggunakan tiga metrik validasi secara bersamaan menghasilkan k=3 dengan Silhouette Score 0,651 sebagai konfigurasi terbaik. Tiga cluster yang terbentuk secara bermakna merepresentasikan segmen pelanggan sangat puas (29,8%), cukup puas (39,3%), dan tidak puas (30,9%), dengan kebersihan dan kecepatan respons staf sebagai atribut paling kritis yang mendorong ketidakpuasan.

Kontribusi utama penelitian ini terletak pada integrasi data lintas platform TikTok dan Google

Maps yang terbukti meningkatkan kualitas cluster dibandingkan pendekatan sumber tunggal. Untuk penelitian selanjutnya, disarankan mengeksplorasi: (1) perbandingan K-Means dengan DBSCAN atau Hierarchical Clustering; (2) integrasi data ulasan dari Instagram Reels; serta (3) pengembangan model pembaruan klaster secara real-time.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Program Studi Teknik Informatika Universitas Baturaja atas dukungan fasilitas penelitian yang diberikan.

REFERENSI

S Badan Pusat Statistik, *Statistik Hotel dan Akomodasi Lainnya di Indonesia 2023*. Jakarta: BPS RI, 2024. [Online]. Available: <https://www.bps.go.id>

Kementerian Pariwisata dan Ekonomi Kreatif, *Laporan Kinerja Sektor Pariwisata Indonesia 2023*. Jakarta: Kemenparekraf RI, 2024. [Online]. Available: <https://www.kemenparekraf.go.id>

Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), *Survei Profil Internet Indonesia 2025*. Jakarta: APJII, 2025. [Online]. Available: <https://survei.apjii.or.id>

V. W. D. Thomas dan F. Rumaisa, "Analisis Sentimen Ulasan Hotel Bahasa Indonesia Menggunakan Support Vector Machine dan TF-IDF," *Jurnal Media Informatika Budidarma*, vol. 6, no. 3, hlm. 1767–1775, 2022. doi: 10.30865/mib.v6i3.4218

D. Ariyus, D. Manongga, dan I. Sembiring, "Enhancing Sentiment Analysis of Indonesian Tourism Video Content Commentary on TikTok: A FastText and Bi-LSTM Approach," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, hlm. 18020–18028, 2024. doi: 10.48084/etasr.8859

A. D. Yustita, A. Purwaningtyas, dan E. A. Ermawati, "Evaluasi Kualitas Pelayanan Hotel Terhadap Kepuasan Tamu Menggunakan Importance Performance Analysis (IPA)," *Jurnal Manajemen Perhotelan dan Pariwisata*, vol. 6, no. 1, hlm. 33–41, 2023. doi: 10.23887/jmpp.v6i1

J. T. Purba, J. Juliana, S. Budiono, A. Purwanto, R. Pramono, dan A. Djakasaputra, "The relationship between hotel service quality and customer satisfaction: an empirical study of spa hotels in Indonesia," *International Journal of Entrepreneurship*, vol. 26, no. 1, hlm. 1–11, 2022.

Y. Sudirman dan D. S. Simatupang, "Analisis Sentimen Berbasis Aspek pada Ulasan Hotel XYZ di Kota Tangerang dengan Algoritma SVM," *STORAGE: Jurnal Ilmiah Teknik dan Ilmu Komputer*, vol. 4, no. 4, hlm. 370–377, 2025. doi: 10.55123/storage.v4i4

N. A. Maori dan E. Evanita, "Metode Elbow dalam Optimasi Jumlah Cluster pada K-Means Clustering," *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, vol. 14, no. 2, hlm. 277–288, 2023. doi: 10.24176/simet.v14i2.9630

Y. Hasan, "Pengukuran Silhouette Score dan Davies-Bouldin Index pada Hasil Cluster K-Means dan DBSCAN," *KAKIFIKOM*, vol. 6, no. 1, hlm. 60–74, 2024. doi: 10.54367/KAKIFIKOM.V6I1.3938

I. F. Ashari, E. D. Nugroho, R. Baraku, I. N. Yanda, dan R. Liwardana, "Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta," *Jurnal RESTI*, vol. 7, no. 1, hlm. 89–97, 2023. doi: 10.29207/resti.v7i1.4494

M. A. Rosid, A. S. Fitriani, I. R. I. Astutik, N. I. Mulloh, dan H. A. Gozali, "Improving Text Preprocessing for Student Complaint Document

- Classification Using Sastrawi,” *IOP Conference Series: Materials Science and Engineering*, vol. 874, no. 1, hlm. 012017, 2020. doi: 10.1088/1757-899X/874/1/012017
- Y. I. Kurniawan, P. R. Anugrah, R. M. Sugihono, F. A. Abimanyu, dan L. Afuan, “Enhancing K-Means Clustering for Journal Articles using TF-IDF and LDA Feature Extraction,” *Brilliance: Research of Artificial Intelligence*, vol. 5, no. 1, 2025. doi: 10.47709/brilliance.v5i1.5547
- R. Gunawan dan H. Hendry, “Analisis Sentimen Ulasan Tamu terhadap Layanan Hotel Menggunakan Pendekatan Machine Learning,” *IT-Explore: Jurnal Penerapan Teknologi Informasi dan Komunikasi*, vol. 4, no. 3, hlm. 295–306, 2025. doi: 10.24246/itexplore.v4i3.2025.pp295-306