Identifikasi Pola dan Jenis Hotel Menggunakan Algoritma DBSCAN

Ikhwan Wahyudin
Fakultas Informatika
Universitas Telkom
Bandung
ikhwanwahyudin@student.telkomuniversity.ac.id

Helmi Muzakki Kurnianto
Fakultas Informatika
Universitas Telkom
Bandung
helmimuzakkik@student.telkomuniversity.ac.id

***Abstract*—Industri perhotelan di Indonesia terus berkembang , hal ini menyebabkan kemunculan hotel yang sangat bervariasi dengan jumlah banyak. Keanekaragaman hotel yang tinggi, hal ini membuat para pengguna atau wisatawan kesulitan dalam menemukan dan mencari hotel yang sesuai dengan kebutuhan mereka. Dengan adanya teknologi dan memanfaatkan klasterisasi dapat mengatasi tantangan akan hal ini. Salah satu algoritma yang digunakan untuk memecahkan permasalahan ini adalah algoritma DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Algoritma dapat mengidentifikasi pola dan jenis hotel berdasarkan kesamaan fitur-fitur pada setiap hotel tanpa memerlukan jumlah klaster yang ditentukan. Maka dengan ini, pengguna atau wisatawan dapat dengan mudah menemukan pilihan hotel yang cocok sesuai dengan kebutuhannya.**

Tahapan yang dilakukan untuk mencapai tujuan tersebut, hal pertama yang dilakukan mencari dan mengumpulkan data, dengan data yang dipakai berupa nama hotel, fasilitas hotel, deskripsi hotel, rating hotel, dan kota tempat hotel itu berada, setelah data terkumpul maka dilakukan pembersihan data dan vektorisasi kata. Selanjutnya penerapan dan pelatihan model klasterisasi dengan algoritma DBSCAN. Setelah itu evaluasi model untuk mengevaluasi hasil dari klasterisasi dan melakukan visualisasi serta analisis untuk melihat karakteristik dan pola setiap klaster yang dihasilkan. Hasil dari visualisasi dan analisis dapat membantu memecahkan masalah serta membantu pengguna dalam menentukan hotel sesuai dengan kebutuhan mereka.

Keywords—DBSCAN, Hotel, Klasterisasi

I. PENDAHULUAN

I.1. Latar Belakang

Industri perhotelan di Indonesia terus mengalami pertumbuhan yang signifikan seiring dengan perkembangan pariwisata dan ekonomi. Fenomena ini menyebabkan munculnya beragam jenis hotel dengan jumlah yang terus bertambah. Keanekaragaman ini, meskipun menggembirakan, juga membawa tantangan tersendiri bagi para pengguna atau wisatawan dalam mencari dan memilih akomodasi yang sesuai dengan kebutuhan dan preferensi mereka.

Di tengah banyaknya pilihan hotel yang tersedia, seringkali sulit bagi wisatawan untuk menavigasi dan menemukan opsi yang paling sesuai. Informasi yang terbatas atau kurangnya pemahaman tentang jenis-jenis hotel yang tersedia dapat memperumit proses pengambilan keputusan. Sebagai akibatnya, pengalaman menginap di hotel bisa jadi tidak sesuai dengan harapan atau kebutuhan wisatawan, yang pada gilirannya dapat mempengaruhi keseluruhan pengalaman liburan atau perjalanan bisnis mereka.

I.2. Rumusan Masalah

Dalam menghadapi tantangan ini, penting untuk mengembangkan solusi yang memanfaatkan kemajuan teknologi untuk meningkatkan pengalaman pengguna. Salah satu cara yang efektif adalah dengan menggunakan metode klasterisasi untuk mengelompokkan jenis-jenis hotel berdasarkan karakteristik dan fitur-fitur tertentu yang dimilikinya.

Kondisi ini menjadi semakin kompleks mengingat jumlah dan variasi hotel yang terus bertambah. Oleh karena itu, diperlukan pendekatan yang lebih adaptif dan fleksibel dalam mengidentifikasi pola-pola dan jenis-jenis hotel tanpa harus memiliki pengetahuan sebelumnya tentang jumlah klaster yang dihasilkan.

I.3. Tujuan

Tujuan dari penelitian ini adalah untuk melakukan klasterisasi dari hotel berdasarkan informasi data text dan data numerik yang dimiliki oleh berbagai hotel menggunakan algoritma DBSCAN.

II. STUDI LITERATUR.

II.1. Term Frequency-Inverse Document Frequency

Pada Proses ini dilakukan pembobotan kata dengan TF-IDF merupakan proses pembobotan dengan menilai signifikansi istilah tertentu dalam dokumen tertentu. TF-IDF dibangun dari komponen TF dan IDF.[1] TF merupakan pengukuran frekuensi istilah dalam dokumen tertentu sedangkan IDF merupakan perhitungan berapa banyak dokumen dalam korpus termasuk frasa yang ditentukan. Rumus TF-IDF adalah sebagai berikut :

$$TF(t,d) = \frac{\text{Jumlah kemunculan t dalam dokumen d}}{\text{Total jumlah kata dalam dokumen d}}$$

$$IDF(t,D) = \frac{\text{Total Jumlah Dokumen}}{\text{Jumlah dokumen d yang mengandung kata t}}$$

$$TF - IDF = TF * IDF$$

II.2. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Algoritma *Density-based Spatial Clustering of Application with Noise* (DBSCAN) merupakan metode klasterisasi yang mengelompokkan objek ke dalam klaster dengan mengelompokkan area area dengan kepadatan (*density-based*) tinggi ke dalam suatu klaster dan yang memiliki kepadatan rendah akan dianggap sebagai *noise* oleh DBSCAN. Algoritma ini dapat menemukan klaster dengan bentuk apa pun pada satu kondisi kepadatan.[2]
Berikut Langkah Algoritma DBSCAN

- Inisialisasi parameter input MinPts dan Eps. Inisiasi parameter yang optimum menggunakan fungsi *kNNdistplot* dari *packages dbscan* dengan

menghitung jarak rata-rata untuk setiap data ke k tetangga terdekatnya (*nearest neighbors*).

- Menentukan titik awal secara acak.
- Menghitung nilai ε atau semua jarak *density reachable* terhadap titik awal menggunakan rumus *Euclidean distance* berikut.

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2}$$

- Jika titik yang memenuhi Eps lebih dari MinPts maka titik awal merupakan core point dan terbentuk sebuah cluster.
- Jika amatan titik awal adalah *border points* dan tidak ada amatan yang *density-reachable* dengan amatan titik awal, maka proses dilanjutkan ke titik yang lain.

II.3. Silhouette Score

Silhouette Score ini digunakan untuk mengukur dan mengevaluasi hasil dari klasterisasi. *Silhouette Score* memberikan skor untuk setiap titik data, yang mencerminkan seberapa baik titik tersebut cocok dengan klaster. Skor ini berkisar dari -1 hingga 1, di mana nilai yang lebih tinggi menunjukkan bahwa titik tersebut lebih dekat ke klaster yang telah ditugaskan daripada klaster lainnya.[2]

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

II.4. Klasterisasi

Clustering atau klasterisasi adalah metode pengelompokan data. Menurut Tan, 2006 *clustering* adalah sebuah proses untuk mengelompokan data ke dalam beberapa *cluster* atau kelompok sehingga data dalam satu *cluster* memiliki tingkat kemiripan yang maksimum dan data antar *cluster* memiliki kemiripan yang minimum.[3]

III.METODE

III.1. Pengumpulan Data

Penelitian ini menggunakan dataset informasi hotel yang mencakup informasi kota,negara,nama hotel, rating hotel, alamat, aktraksi,fasilitas hotel, nomor telepon, map,kode pos dan deskripsi hotel berasal dari Hotels Dataset (kaggle.com) .

III.2. Praposessing Data

1. Filtering Data

   Melakukan filter data hanya hotel yang berada di negara Indonesia.

2. Split Data dan Menghapus Kolom tidak diperlukan

   Melakukan pembagian data pada kolom kolom tertentu.

   Pada kolom map berisi informasi latitude dan longtitude, maka nilai pada map akan dipisahkan dan terbagi menjadi dua kolom baru yaitu latitude dan longtitude.

   Pada Kolom Deskripsi akan dilakukan split karena didalam kolom deskripsi terdapat informasi hotel meliputi dining, room , renovation, checkin intruction, dan special intruction.

3. Menangani nilai null

   Pada proses ini dilakukan pengecekan nilai null dan mengisi nilai dengan kata ‘Nothing’ menandakan bahwa hotel tersebut tidak memiliki informasi menegnai atribut tersebut (hanya berlaku untuk kolom memiliki infomasi teks) selain data text akan dilakukan penghapusan, dan mengisi dengan 0 untuk kolom pincode menunjukan bahwa hotel tersebut tidak mencantumkan informasi pincode.

4. Membersihan tag html

   Melakukan pembersihan tag html pada kolom kolom yang memiliki tag html.

5. *Case Folding*

   Pengubahan semua data text menjadi huruf kecil.

6. Penghapusan tanda baca, karakter khusus dan kata km/mi.

   Menghapus semau tanda baca dan karakter khusus yang terdapat pada data text.

   Menghapus huruf km dan mi karena banyaknya kata yang muncul dalam data teks yang kurang memberikan informasi berpegaruh

7. Tokenisasi

   Memecah teks menjadi token token kata.

8. *Stopword Removal*

   Menghapus kata-kata umum yang tidak memberikan nilai informasi yang signifikan dan kata kata yang sering muncul dalam teks pada bahasa inggris.

9. Lematisasi

   Mengubah kata kata ke bentuk dasar dengan memperhatikan konteks dan bagian dari kalimat serta sesuai dengan kamus. Lematisasi yang digunakan untuk bahasa inggris.

10. Mengubah nilai rating hotel menjadi numerik

    Pengubahan nilai pada kolom rating hotel menjadi numerik. Dalam kolom rating hotel terdapat nilai OneStar, TwoStart, ThreeStart, FourStart, FiveStart dan AllStart mengubahnya menjadi rating 1 sampai 5.

III.3. Exploratory Data Analysis (EDA)

Dalam penelitian ini, EDA digunakan untuk mendapatkan wawasan awal dari dataset ulasan hotel yang digunakan. EDA membantu dalam memahami struktur data, menemukan pola, menemukan anomali, dan membangun hipotesis awal untuk analisis lebih lanjut. Berikut EDA yang dilakukan yaitu:

1. Visualisasi persebaran HotelRating
2. Visualisasi frekuensi kemunculan kata untuk data texts.

III.4. Term Frequency-Inverse Document Frequency

Pada Proses ini dilakukan pembobotan kata dengan TF-IDF dengan library scikit-learn pada python pada data

data teks yang akan dijadikan fitur input pada algoritma klasaterisasi.

III.5. Standarisasi Fitur Numerik

Selain dilakukan TF-IDF pada data teks, data numerik akan dilakukan standarisasi menggunakan bantuan library dari scikit learn

III.6. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Setelah dilakukannya standarisasi pada fitur numerik dan pembobotan kata pada data text dengan TF-IDF , selanjutnya dilakukan klasterisasi dengan data fitur input yang digunakan yaitu kolom room,attraction,dining, Renovaation, checkin Intruction, SpecialIntruction, Hotel Facilities, city code, hotel rating, latitude, longtitude, dan pin code.

Pada klasterisasi ini juga dilakukan beberapa kali percobaan parameter untuk mendapatkan nilai evaluasi yang cukup dan noise yang rendah. Dengan 50 kombinasi parameter, untuk eps 0,5 sampai 0,9 dan min sample 10,20,30,40,50,60,70,80,90,100.

III.7. Evaluasi

Setelah dilakukannya klasterisasi selanjutnya dilakukan evaluasi menggunakan silhouette score untuk mengukur seberapa baik model melakukan klasterisasi.

IV. HASIL DAN PEMBAHASAN

IV.1. Exploratory Data Analysis (EDA)

Berikut hasil EDA pada data hotel Indonesia :

IV.1.1. Visualisasi Fitur Checkin Intruction

Fig. 1. Visualisasi wordcloud checkin intruction

IV.1.2. Visualisasi Fitur Hotel Facilities

Fig. 2. Wordcloud fitur hotel facilities

IV.1.3. Visualisasi Fitur Room

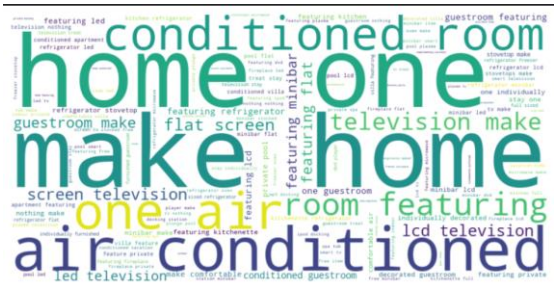

Fig. 3. Wordcloud fitur room

IV.1.4. Visualisasi Fitur Renovation

Fig. 4. Wordcloud fitur renovation

IV.1.5. Visualisasi Fitur Dining

Fig. 5. Wordcloud fitur Dining

IV.1.6. Visualisasi Fitur Special Intruction

Fig. 6. Wordcloud fitur Special Intruction

IV.1.7. Distribusi Hotel Rating

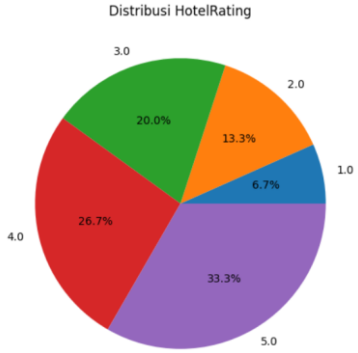

Fig. 7. Piechart fiitur Hotel Rating

IV.1.8.Kata Sering muncul untuk semua data text

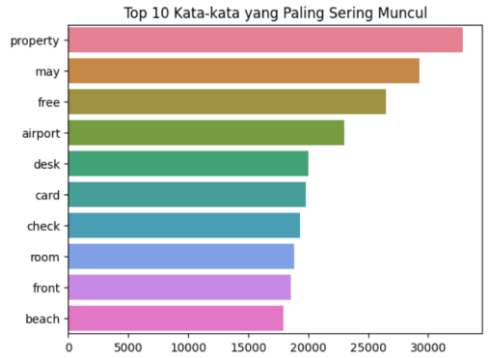

Fig. 8. Barplot kata yang sering muncul

IV.2. Hasil Clustering

Berdasarkan hasil percobaan klasterisasi dengan beberapa parameter didapatkan bahwa nilai silhouette score paling tinggi pada parameter 0.5 dengan min sample 30 namun memiliki noise yang sangat tinggi dan klaster yang terbentuk hanya 2 klaster. Berikut persebaran klasternya :

Fig. 9. Pesebaran klaster eps 0,5 dan minsample 30

Pada percobaan lain dengan parameter eps 0,9 dan min sample 20, klaster yang terbentuk cukup banyak yaitu 65 klaster dengan noise yang cukup rendah dibandingkan sekitar 3800 data namun nilai silhouette score tidak tinggi seperti parameter 0,5 min sample 30 tapi masih cukup baik. Berikut persebaran klasternya :

Fig. 10. Persebaran klaster dengan eps 0,9 min sample 20

Berikut grafik nilai Sihouttle Score untuk 50 kombinasi parameter :

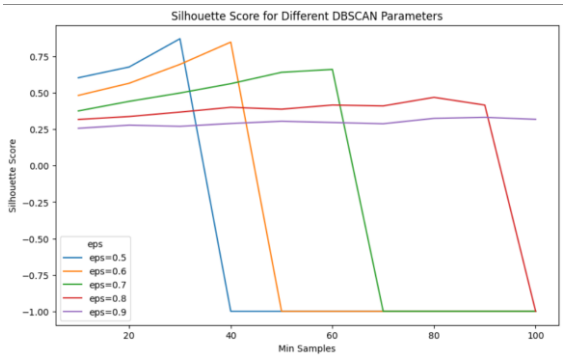

Fig. 11. Example of a figure caption. (*figure caption*)

V. KESIMPULAN DAN SARAN

Hasil eksperimen menunjukkan bahwa penggunaan algoritma DBSCAN dan pembobotan kata dengan TF-IDF mampu dalam mengklasterisasi hotel berdasarkan fitur numerik dan text namun memiliki beberapa kekurangan dimana pada beberapa parameter eps dan min sample menunjukan nilai matrix evaluasi kurang baik dan noise yang cukup banyak. Pada eksperimen ini dapat disimpulkan eps 0,9 dan min sample 20 dapat digunakan karena memiliki noise cukup rendah walaupun nilai matrix evaluasinya tidak terlalu tinggi dibandingkan eps 0,5 min sample 30.

Berdasarkan identifikasi tiap klaster yang telah dihasilkan oleh DSBCAN dapat disimpulkan bahwa tiap klaster memiliki kesamaan dan pembeda dengan klaster lain yaitu pada fitur cityName , Rating Hotel dan Attraction, namun pada fitur lain masih terdapat beberapa persamaan antar beberapa klaster.

Berdasarkan hasil yang diperolah, saran yang dapat dilakukan untuk penelitian selanjutnya agar dapat meningkatkan kualitas hasil diantaranya : Mengeksplorasi metode ektraksi fitur lain seperti word embedding atau Doc2Vec dan mencoba algoritma klasterisasi yang berbeda.

REFERENCES

[1] Mohammed, M. T., & Rashid, O. F. (2023). Document retrieval using term frequency inverse sentence frequency weighting scheme. Indonesian Journal of Electrical Engineering and Computer Science, 31(3), 1478-1485.

[2] Fauzan, A., Novianti, A., Ramadhani, R. R. M. A., & Adhiwibawa, M. A. S. (2022). Analysis of hotels spatial clustering in Bali: density-based spatial clustering of application noise (DBSCAN) algorithm approach. *EKSAKTA: Journal of Sciences and Data Analysis*, 25-38.

[3] Tan, P.N., Steinbach, M., Kumar, V. (2006) *Introduction to Data Mining*. Boston:Pearson Education.