Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1320−1331
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8145

Analisis Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan Sentimen Ulasan Aplikasi E-Commerce

R Immanuel Giovanni Italiano Pecaro[1], Galet Guntoro Setiaji[1], Ahmad Rifa'i[2,*]

[1] Fakultas Teknologi Informasi dan Komunikasi, Prodi Teknik Informatika, Universitas Semarang, Semarang, Indonesia
[2] Fakultas Teknologi Informasi dan Komunikasi, Prodi Sistem Informasi, Universitas Semarang, Semarang, Indonesia
Email: [1]12a11.immanuel.giovanni@gmail.com, [2]gallet@usm.ac.id, [3,*]rifai@usm.ac.id
Email Penulis Korespondensi: rifai@usm.ac.id
Submitted: **31/07/2025**; Accepted: **04/09/2025**; Published**: 05/09/2025**

Abstrak−Teknologi digital yang semakin berkembang telah mendorong pertumbuhan pesat aplikasi *E-Commerce*, yang ditandai dengan banyaknya aplikasi sejenis yang tersedia di *Google Play Store*. Hal ini menyebabkan meningkatnya kebutuhan masyarakat terhadap kemudahan dalam berbelanja secara *online*, sekaligus menunjukkan persaingan yang semakin ketat di ranah jual beli *online*. Ulasan pada aplikasi *E-Commerce* di *Google Play Store* seringkali menjadi dasar pengguna *Google Play Store* untuk mengunduh aplikasi tersebut. Dengan adanya ulasan pada aplikasi tersebut maka beberapa pengguna *Google Play Store* mulai mempertimbangkan dan menilai apakah aplikasi tersebut layak untuk diunduh atau tidak. *Shopee* merupakan salah satu aplikasi *E-Commerce* terbesar yang ada di Indonesia, dan saat ini memiliki lebih dari 15 juta data rating dan ulasan di *Google Play Store*. Penelitian ini bertujuan untuk membandingkan performa algoritma *K-Means* dan *K-Medoids* dalam mengelompokkan data numerik dari ulasan aplikasi tersebut**.** Pengelompokan dilakukan menggunakan teknik *Clustering* berdasarkan dua variabel numerik, yaitu *score* dan *thumbsupcount*, untuk memberikan gambaran awal mengenai kecenderungan opini pengguna terhadap aplikasi tersebut. Dataset ini diambil dari ulasan di *Google Play Store* pada bulan Desember tahun 2024 sebanyak 500 data. Hasil *Davies-Bouldin Index* dari penelitian tersebut didapatkan bahwa *K-Means* lebih unggul dari *K-Medoids* dengan perbandingan 0,457 berbanding 0,803.

Kata Kunci: Clustering; E-Commerce; K-Means; K-Medoids; Ulasan

Abstract−The growing digital technology has driven the rapid growth of E-Commerce applications, which is characterized by the number of similar applications available on the Google Play Store. This phenomenon has led to an increase in people's need for convenience in online shopping, as well as showing increasingly fierce competition in the digital marketplace. Reviews on E-Commerce applications in the Google Play Store often serve as a basis for users to decide whether to download an application. These reviews provide valuable insights, allowing users to assess whether the application is worth downloading. Shopee, one of the largest E-Commerce applications in Indonesia, currently has more than 15 million ratings and reviews on the Google Play Store. This study aims to compare the performance of the K-Means and K-Medoids algorithms in clustering numerical data from application reviews. Clustering was performed using the Clustering technique based on two numerical variables, namely score and thumbsupcount, to provide an initial overview of user opinion trends regarding the application. The dataset, consisting of 500 reviews, was collected from the Google Play Store in December 2024. The results Davies-Bouldin Index of the study indicate that K-Means outperforms K-Medoids, with a comparison score of 0.457 to 0.803.

Keywords: Clustering; E-Commerce; K-Means; K-Medoids; Reviews

1. PENDAHULUAN

Pertumbuhan pesat industri *e-commerce* di Indonesia telah membawa dampak signifikan terhadap perilaku konsumsi masyarakat. *Platform - platform* seperti *Shopee*, *Tokopedia*, *Bukalapak*, dan *TikTok Shop* terus bersaing dalam menawarkan kemudahan, efisiensi, dan berbagai promosi menarik yang mampu menarik perhatian konsumen dari berbagai lapisan masyarakat [1]. Di antara berbagai platform tersebut, *Shopee* menonjol sebagai salah satu aplikasi belanja daring dengan jumlah unduhan tertinggi di Indonesia, didorong oleh strategi pemasaran agresif dan antarmuka pengguna yang mudah digunakan. Seiring dengan peningkatan jumlah pengguna, volume ulasan yang ditinggalkan oleh konsumen di *platform* distribusi aplikasi seperti *Google Play Store* juga mengalami lonjakan signifikan. Ulasan-ulasan ini tidak hanya mencerminkan kepuasan atau ketidakpuasan pengguna, tetapi juga memengaruhi keputusan calon pengguna lain serta menjadi masukan penting bagi pengembang aplikasi dalam proses peningkatan layanan [2].

Namun, banyaknya ulasan yang tersedia secara *online* juga menimbulkan tantangan baru dalam hal pengolahan data. Sebuah aplikasi populer seperti *Shopee* dapat menerima ribuan hingga jutaan ulasan, yang jika dianalisis secara manual tentu akan sangat memakan waktu dan sumber daya. Meskipun pendekatan berbasis *Natural Language Processing* (NLP) dapat membantu dalam analisis sentimen berbasis teks, proses ini masih relatif kompleks dan membutuhkan daya komputasi tinggi [3]. Oleh karena itu, dibutuhkan solusi alternatif yang lebih efisien dan dapat dieksekusi dengan cepat, seperti pemanfaatan teknik *clustering* berdasarkan data numerik yang tersedia secara langsung di ulasan, yakni *score* dan jumlah *ThumbsUp*. Metode ini memungkinkan pengelompokan persepsi pengguna tanpa harus menganalisis isi teks secara mendalam, sehingga proses analisis menjadi lebih sederhana dan ringan.

Penelitian ini bertujuan untuk menganalisis sentimen pengguna aplikasi *Shopee* dengan menggunakan pendekatan *clustering* berdasarkan atribut numerik, yaitu *score* dan jumlah *ThumbsUp* yang diberikan oleh pengguna lain. Pendekatan ini dinilai efektif dalam mengidentifikasi kelompok pengguna dengan kecenderungan sentimen tertentu secara tidak langsung. Dua algoritma *clustering* yang digunakan dalam penelitian ini adalah *K-Means* dan *K-Medoids*, yang akan dibandingkan performanya dalam proses pengelompokan data numerik tersebut. Untuk

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1320−1331
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8145

mengevaluasi hasil *clustering*, digunakan metrik *Davies-Bouldin Index* (DBI) yang mengukur validitas *cluster* berdasarkan tingkat kedekatan antar anggota *cluster* dan pemisahan antar *cluster* [4].

K-Means merupakan algoritma *clustering* yang populer karena kesederhanaannya dan efisiensi prosesnya dalam menangani data berukuran besar. Namun, algoritma ini memiliki kelemahan utama yaitu sensitivitas terhadap outlier, karena menggunakan rata-rata sebagai pusat *cluster*. Sementara itu, *K-Medoids* menggunakan data aktual sebagai pusat *cluster*, sehingga lebih tahan terhadap outlier dan cenderung menghasilkan *cluster* yang lebih stabil [5]. Dengan membandingkan kedua algoritma ini menggunakan metrik DBI, diharapkan penelitian ini dapat memberikan gambaran yang lebih komprehensif mengenai efektivitas masing-masing metode dalam pengelompokan ulasan pengguna aplikasi *Shopee* berdasarkan data numerik.

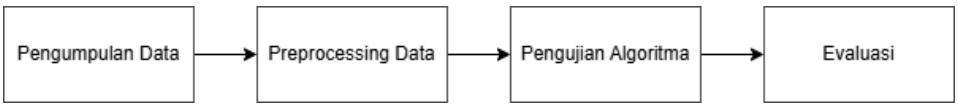
Manfaat dari pendekatan ini sangat relevan dan aplikatif bagi pengembang. Pengelompokan berdasarkan *score* rendah dan jumlah *ThumbsUp* tinggi dapat membantu mengidentifikasi masukan negatif yang ternyata mendapatkan banyak persetujuan dari pengguna lain. Artinya, keluhan tersebut bukan kasus individual, melainkan mencerminkan persepsi kolektif terhadap kekurangan tertentu dalam aplikasi. Sebagai contoh, jika suatu fitur sering dikritik namun disukai oleh banyak pengguna melalui *ThumbsUp*, hal ini bisa menjadi prioritas perbaikan yang lebih penting. Sebaliknya, *score* tinggi dengan banyak *ThumbsUp* dapat menunjukkan elemen aplikasi yang sangat diapresiasi pengguna, seperti kecepatan layanan atau kemudahan navigasi [6]. Dengan demikian, hasil *clustering* ini dapat membantu pengembang dalam pengambilan keputusan strategis untuk meningkatkan kualitas dan kepuasan pengguna secara keseluruhan.

Penelitian ini juga mengkaji sejumlah studi sebelumnya terkait penerapan teknik *clustering* dalam ruang lingkup *e-commerce*. Arsyad dan Sulastri, misalnya, menerapkan algoritma *K-Means* dan *K-Medoids* untuk mengelompokkan ulasan pengguna aplikasi *Blibli.com* berdasarkan analisis teks [7]. Meskipun metode tersebut memberikan hasil yang cukup informatif, penelitian tersebut belum memanfaatkan data numerik seperti *score* dan *ThumbsUp* sebagai dasar *clustering*. Selain itu, Mohemad et al. membandingkan performa *K-Means* dan *K-Medoids* dalam pengelompokan dokumen kriminal berbahasa Melayu, yang juga masih jauh dari konteks *e-commerce* [8]. Oleh karena itu, penelitian ini memberikan kontribusi baru dengan menggabungkan pendekatan numerik dan fokus pada aplikasi marketplace *Shopee*, yang belum banyak diteliti dengan cara serupa.

Dataset dalam penelitian ini diperoleh melalui proses *scraping* dari *Google Play Store*, yang menghasilkan 500 data ulasan pengguna aplikasi *Shopee*. Setiap data hanya terdiri dari dua atribut numerik yaitu *score* dan jumlah *ThumbsUp*, tanpa menyertakan konten teks dari ulasan itu sendiri. Pendekatan ini dipilih untuk menjaga efisiensi dalam pengolahan data serta menghindari kompleksitas analisis berbasis teks. Studi oleh Danureksa et al. menunjukkan bahwa pendekatan berbasis numerik seperti ini cukup efektif dalam kasus pengelompokan data supplier di Shopee menggunakan algoritma *K-Means* [9]. Selain itu, pemilihan metrik *Davies-Bouldin Index* juga didukung oleh studi Aruadi et al., yang menggunakannya untuk mengevaluasi pengelompokan data gizi berdasarkan provinsi di Indonesia, dan terbukti mampu mengukur kualitas *cluster* dengan baik [10].

2. METODOLOGI PENELITIAN

Berdasarkan pada pendahuluan di atas, metodologi penelitian ini dibagi menjadi empat tahapan, yaitu: Pengumpulan Data, *Preprocessing* Data, Pengujian Algoritma, dan Evaluasi [11]. Setiap tahapan memiliki fungsi yang saling berkaitan untuk memperoleh hasil yang akurat dalam penelitian ini. Secara umum, alur proses penelitian digambarkan pada Gambar 1 berikut:



Gambar 1. Metodologi Penelitian

Pada Gambar 1, tahapan diawali dengan pengumpulan data dari sumber ulasan pengguna aplikasi *Shopee*. Data yang telah diperoleh kemudian diproses melalui tahap *preprocessing* untuk menyiapkan data mentah menjadi layak untuk dianalisis. Tahap berikutnya adalah pengujian algoritma klasifikasi sentimen, dan diakhiri dengan evaluasi performa model untuk mengetahui akurasi serta efektivitas metode yang digunakan.

2.1 Pengumpulan Data

Pada tahapan pengumpulan data, *Dataset* diambil dari ulasan aplikasi *com.Shopee.id* di *Google Play Store* pada bulan Desember 2024. Data diperoleh dengan metode *scraping* menggunakan bahasa pemrograman *Python* untuk mengumpulkan 500 ulasan pengguna aplikasi tersebut dalam format CSV [12]. Data yang dikumpulkan mencakup berbagai kolom, namun hanya kolom *username*, *score*, dan *thumbsupcount* yang digunakan dalam penelitian ini.

2.2 Preprocessing Data

Pada tahap *Preprocessing Data*, dilakukan persiapan dataset mentah yang diperoleh dalam format CSV agar siap digunakan dalam proses pengelompokan menggunakan algoritma *K-Means* dan *K-Medoids*. *Dataset* awal

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1320−1331
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8145

mengandung beberapa kolom, namun hanya tiga kolom yang diambil untuk dianalisis, yaitu kolom *username*, *score*, dan *thumbsupcount*. Dari ketiga kolom tersebut, hanya dua kolom numerik *score* dan *thumbsupcount* yang digunakan sebagai fitur utama dalam pengelompokan.

Karena rentang nilai *score* dan *thumbsupcount* berbeda, dilakukan proses normalisasi menggunakan metode *Min-Max Scaling* agar nilai pada kedua fitur tersebut berada dalam rentang 0 hingga 1. Normalisasi ini bertujuan untuk menyamakan skala fitur sehingga algoritma clustering dapat berjalan optimal tanpa bias terhadap nilai fitur yang memiliki rentang lebih besar.

Langkah selanjutnya adalah normalisasi nilai pada kolom *score* dan *thumbsupcount* menggunakan metode *Min-Max Scaling* dengan cara mengubah nilai asli pada kolom (X) ke dalam skala baru antara 0 hingga 1. Proses ini menggunakan acuan nilai terkecil (Xmin) dan nilai terbesar (Xmax) pada kolom tersebut. Hasil normalisasi kemudian dinyatakan sebagai X', yaitu nilai yang sudah dipetakan agar berada dalam rentang 0 sampai 1 dengan rumus sebagai berikut:

$$X' = \frac{X - Xmin}{Xmax - Xmin} \tag{1}$$

Namun, hasil dari proses normalisasi ini dapat menghasilkan nilai nol ketika nilai asli X sama dengan Xmin. Nilai nol ini dapat menjadi masalah bagi beberapa algoritma *clustering* seperti *K-Means* atau *K-Medoids*, karena algoritma tersebut cenderung tidak bekerja optimal saat terdapat data berdimensi nol (*sparse data*). Untuk mengatasi hal ini, dilakukan penanganan nilai nol menggunakan fungsi tambahan [13].

Nilai hasil normalisasi yang sama dengan nol (0) akan diganti dengan nilai kecil acak menggunakan fungsi *IF* dan *RANDBETWEEN* di *Microsoft Excel*. Fungsi *IF* digunakan sebagai logika pengecekan sparsity, yaitu untuk mendeteksi apakah hasil normalisasi bernilai nol. Bila ditemukan, maka akan dihasilkan angka acak dalam rentang 0.01 hingga 0.09, yang berasal dari distribusi nilai hasil normalisasi lain yang bukan nol. Hal ini dilakukan untuk menghindari sparsity (nilai 0) dalam data, menjaga agar nilai pengganti tetap berada dalam skala yang wajar dan tidak merusak interpretasi data, tetap mempertahankan proporsi data secara keseluruhan.
Rumus yang digunakan:

$$Y = \begin{cases} \frac{randbetween(1,9)}{100}, & jika\ X' = 0 \\ X', & jika\ X' \neq 0 \end{cases} \tag{2}$$

Dimana X' adalah nilai hasil normalisasi awal, Y adalah nilai hasil normalisasi setelah penggantian nilai nol, Fungsi *RANDBETWEEN*(1,9) menghasilkan angka acak dari 1 sampai 9, lalu dibagi 100 sehingga menghasilkan angka antara 0,01 hingga 0,09.

Langkah ini dipilih karena dalam konteks dataset yang kecil dan terbatas, pendekatan ini dapat menghindari ketidakseimbangan atau kegagalan algoritma yang disebabkan oleh nilai nol. Meskipun penggunaan angka acak bukan merupakan praktik normalisasi yang umum di penelitian berskala besar, dalam konteks penelitian ini langkah tersebut diambil sebagai solusi praktis untuk menjaga kualitas input data dalam proses analisis lebih lanjut [14].

Terakhir, proses normalisasi ini dilakukan dengan menggunakan *Microsoft Excel* karena sifat dataset yang relatif kecil dan untuk kemudahan pemrosesan. Sedangkan untuk penelitian lanjutan atau dataset berskala besar, disarankan menggunakan bahasa pemrograman seperti *Python* atau *tools* statistik yang lebih mendukung replikasi dan akurasi pengolahan data.

2.3 Pengujian Algoritma

Tahap pengujian algoritma ini bertujuan untuk mengevaluasi dan membandingkan kinerja algoritma *K-Means* dan *K-Medoids* dalam melakukan analisis sentimen terhadap 500 ulasan pengguna aplikasi *Shopee*, dengan memanfaatkan *tool Rapid Miner*. *Dataset* yang akan digunakan telah melalui proses *preprocessing* yang meliputi pemilihan kolom dan normalisasi data. Dalam pengujian ini, fokus utama adalah pada nilai *Davies-Bouldin Index* (DBI) yang dihasilkan oleh masing-masing algoritma [15].

2.3.1 Prosedur Pengujian

Pengujian dilakukan dengan langkah-langkah berikut:

a. Persiapan Data: Dataset yang telah di-*preprocessed* dalam format CSV diunggah ke *Rapid Miner*.
b. Implementasi Algoritma: Algoritma *K-Means* dan *K-Medoids* diimplementasikan menggunakan operator yang tersedia di *Rapid Miner*.
c. Konfigurasi Parameter: Parameter algoritma, seperti jumlah *cluster* (*k*), dikonfigurasi. Penentuan nilai *k* dilakukan dengan menjalankan nilai k=3 dan *max runs* = 10, lalu mengamati nilai DBI yang dihasilkan.
d. Evaluasi DBI: Nilai DBI dievaluasi untuk masing-masing algoritma dengan menggunakan nilai k = 3. DBI digunakan untuk menentukan kualitas *cluster*, di mana nilai DBI yang lebih rendah menunjukkan pengelompokan yang lebih baik.
e. Perbandingan: Hasil DBI dari *K-Means* dan *K-Medoids* dibandingkan untuk menentukan algoritma mana yang memberikan hasil pengelompokan yang lebih baik.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1320−1331
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8145

2.3.2 *K-Means*

Algoritma *K-Means* merupakan salah satu pendekatan *clustering* yang bekerja dengan membagi sekumpulan data ke dalam sejumlah grup (*k cluster*), di mana masing-masing data dikaitkan dengan pusat *cluster* terdekat. Titik pusat (*centroid*) dari tiap *cluster* ditentukan berdasarkan rata-rata dari data yang termasuk di dalamnya. Proses ini dimulai dengan memilih jumlah *cluster* serta posisi awal *centroid* secara acak, lalu diikuti dengan penghitungan jarak antar data dan pusat menggunakan rumus *Euclidean*. *Cluster* dan pusatnya akan diperbarui secara bertahap sampai posisi *centroid* stabil atau jumlah iterasi tertentu telah tercapai [16][17][18].
Implementasi di *Rapid Miner*:

a. Dalam RapidMiner, algoritma *K-Means* diimplementasikan menggunakan operator '*K-Means*'.
b. Operator ini membutuhkan beberapa parameter, termasuk jumlah *cluster* (k) dan metrik jarak yang digunakan (dalam kasus ini, *Euclidean Distance*).
c. Proses dimulai dengan inisialisasi k pusat *cluster* secara acak atau menggunakan metode inisialisasi lainnya yang tersedia di *Rapid Miner*.
d. Selanjutnya, setiap data ditetapkan ke *cluster* terdekat berdasarkan jarak *Euclidean* ke pusat *cluster*.
e. Setelah semua data ditetapkan, pusat *cluster* dihitung ulang sebagai rata-rata dari semua data dalam *cluster*.
f. Proses penetapan dan perhitungan ulang ini diulang hingga konvergen, yaitu ketika pusat *cluster* tidak lagi berubah secara signifikan atau mencapai jumlah iterasi maksimum yang ditentukan.

2.3.3 *K-Medoids*

Algoritma *K-Medoids* memiliki kemiripan konsep dengan *K-Means*, tetapi berbeda pada cara menentukan pusat *cluster*. Alih-alih menggunakan rata-rata (*mean*), *K-Medoids* memilih salah satu data aktual dalam *cluster* sebagai representasi pusat atau *medoid*. Pendekatan ini membuat *K-Medoids* lebih tangguh terhadap kehadiran data *outlier* dan *noise* dibandingkan *K-Means*, karena pusat *cluster* selalu berupa data nyata dalam himpunan tersebut [19][20].
Implementasi di *Rapid Miner*:

a. Dalam *Rapid Miner*, algoritma *K-Medoids* dapat diimplementasikan menggunakan operator '*K-Medoids*' atau operator serupa yang tersedia.
b. Operator ini juga membutuhkan parameter seperti jumlah *cluster* (k) dan metrik jarak.
c. Proses dimulai dengan pemilihan k data secara acak sebagai *medoid* awal.
d. Selanjutnya, setiap data ditetapkan ke *cluster* terdekat berdasarkan jarak ke *medoid*.
e. Setelah semua data ditetapkan, *medoid* baru dipilih dalam setiap *cluster*.
f. *Medoid* baru dipilih sebagai data yang meminimalkan total jarak ke semua data lain dalam *cluster*.
g. Proses penetapan dan pemilihan *medoid* ini diulang hingga tidak ada perubahan pada *medoid*.

2.3.4 *Davies-Bouldin Index* (DBI)

DBI digunakan sebagai metrik utama untuk mengevaluasi kualitas pengelompokan [21]. DBI mengukur *average similarity* setiap *cluster* dengan *cluster* yang paling mirip dengannya [22]. Dengan rendahnya nilai DBI, semakin baik hasil pengelompokan, karena menunjukkan bahwa *cluster-cluster* tersebut lebih terpisah dan homogen .

2.3.5 Hasil yang Diharapkan

Dalam pengujian, diharapkan dapat ditemukan nilai *k* yang optimal untuk masing-masing algoritma, serta mengetahui algoritma mana yang menghasilkan nilai DBI yang lebih rendah [23]. Algoritma dengan nilai DBI terendah dianggap lebih efektif dalam mengelompokkan sentimen ulasan aplikasi *e-commerce* dalam *dataset* yang digunakan.

2.3.6 *Tools* yang Digunakan

Rapid Miner dipilih sebagai *tool* utama dalam penelitian ini untuk menjalankan alur kerja *clustering* secara visual tanpa perlu pemrograman manual. Melalui *Rapid Miner*, peneliti dapat menyiapkan data numerik, melakukan preprocessing, menjalankan algoritma *K-Means* dan *K-Medoids*, hingga menghitung *Davies-Bouldin Index* (DBI) secara otomatis [24].

3. HASIL DAN PEMBAHASAN

Pada bagian ini, data yang diperoleh dari *Google Play Store* sebanyak 500 ulasan aplikasi *Shopee* akan diproses untuk keperluan analisis. Dari *dataset* tersebut, hanya dua kolom yang akan digunakan, yaitu kolom *Score* yang berisi nilai rating dari pengguna serta kolom *ThumbsUpCount* yang menunjukkan jumlah tanda suka pada ulasan. Kedua variabel ini menjadi fokus dalam pengolahan data untuk membandingkan hasil *clustering* menggunakan algoritma *K-Means* dan *K-Medoids* pada sentimen pengguna.

3.1 Hasil Penelitian

Pada tahap ini, sebanyak 500 data ulasan aplikasi *Shopee* yang telah diperoleh melalui teknik *scraping* akan digunakan sebagai bahan analisis. Data mentah tersebut terdiri dari berbagai informasi terkait rating dan interaksi pengguna yang

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1320−1331
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8145

akan diolah lebih lanjut untuk membandingkan performa dua algoritma *clustering*, yaitu *K-Means* dan *K-Medoids*. Sebelum dilakukan pemrosesan lebih lanjut, data mentah ini dapat dilihat secara rinci pada Tabel 1 di bawah, yang menampilkan kolom-kolom utama yang menjadi fokus analisis:

Tabel 1. Data mentah

reviewId	User Name	User Image	Content	Score	Thumbs Up Count	review Created Version	At	Reply Content	Replied At	appVersion
130cb160-a061-4314-920e-bf7dd71f7d76	Chikka Risa	https://play-lh.googleusercontent.com/a/ACg8oc...	Aplikasi bagus, sayang di pengiriman lamaaaaaa...	1	20	3.39.15	2024-12-07 08:04:32	hi kak, maaf ya buat kendala pesanan kakak, ha...	2024-12-07 10:15:54	3.39.15
4e3ae6df-adb4-4b82-9954-11202eb3fad7	Dewi Nurlaela	https://play-lh.googleusercontent.com/a-/ALV-U..	Sangat muaskan belanja di *Shopee*. selain lengk...	5	40	3.39.15	2024-12-07 07:04:42	hi kak, makasih buat review bintang 5 nya, yuk...	2024-12-07 09:26:49	3.39.15
…	…	…	…	…	…	…	…	…	…	…
d894da97-f7ce-44c4-9e72-c96081093126	Rahmat Hidayat	https://play-lh.googleusercontent.com/a-/ALV-U...	Pembayaran sudah dilakukan namun status pembel...	1	0	3.37.31	2024-11-23 10:06:11	Hai kak, maaf buat kamu ga nyaman terkait kend...	2024-11-23 11:18:46	3.37.31
35d29fdc-1104-421b-81fe-4e7419af38d8	Toraya Gaming	https://play-lh.googleusercontent.com/a-/ALV-U..	Pengiriman lama terutama expedisi *Shopee* expre...	1	3	3.37.31	2024-11-15 21:21:57	Haii kak, maaf ya udh bikin km kecewa terkait ...	2024-11-15 22:13:26	3.37.31

Pada Tabel 1 ditampilkan data mentah yang belum melalui proses pengolahan lebih lanjut. Dari data tersebut, terlihat bahwa terdapat banyak kolom dengan beragam jenis informasi yang direkam, mulai dari identitas pengguna hingga detail interaksi pada ulasan. Mengingat tujuan utama analisis ini adalah untuk melakukan *clustering* berdasarkan sentimen ulasan, maka perlu dilakukan seleksi kolom yang relevan agar fokus analisis dapat terjaga dan hasilnya lebih akurat. Proses pemilihan kolom ini bertujuan untuk mengurangi kompleksitas data serta memastikan bahwa hanya atribut yang berkontribusi langsung terhadap analisis sentimen yang akan digunakan.

Dalam hal ini, kolom yang dipilih adalah *UserName*, *Score*, dan *ThumbsUpCount*. Kolom *UserName* berfungsi sebagai identitas unik untuk setiap pengulas sehingga dapat memudahkan pengelolaan data dan mencegah duplikasi. Sementara itu, kolom *Score* merepresentasikan rating numerik yang diberikan oleh pengguna, yang merupakan parameter utama untuk mengklasifikasikan sentimen dalam penelitian ini. Kolom *ThumbsUpCount* menunjukkan jumlah tanda suka (*likes*) yang diterima oleh ulasan tersebut dari pengguna lain, yang dapat menjadi indikator popularitas atau relevansi ulasan. Pemilihan ketiga kolom ini juga didasarkan pada pertimbangan bahwa kombinasi antara nilai rating dan jumlah thumbs up dapat memberikan gambaran yang lebih lengkap mengenai persepsi pengguna terhadap aplikasi. Untuk melakukan seleksi pada kolom mana yang akan diambil, disini kolom yang akan diambil adalah kolom *UserName*, *Score*, dan *ThumbsUpCount* dapat dilihat pada Gambar 2 dibawah ini:

```
Data=data[["userName","score","thumbsUpCount"]]
```

Gambar 2. Pengambilan Kolom

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1320−1331
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8145

Seperti yang ditunjukkan pada Gambar 2, proses pengambilan kolom dilakukan dengan menggunakan sintaks pemrograman *Python* yang memanggil kolom *userName*, *score*, dan *thumbsUpCount* dari keseluruhan dataset. Proses ini bertujuan untuk menyaring data agar hanya menyisakan informasi yang benar-benar relevan dengan tujuan analisis, yaitu mengevaluasi sentimen pengguna terhadap aplikasi *Shopee* berdasarkan rating numerik yang mereka berikan serta jumlah pengguna lain yang menyetujui ulasan tersebut.Pemilihan ketiga kolom ini didasarkan pada pertimbangan bahwa *userName* dapat merepresentasikan identitas unik pengguna, *score* mencerminkan penilaian numerik mereka terhadap aplikasi, dan *thumbsUpCount* menggambarkan seberapa bermanfaat ulasan tersebut menurut pengguna lainnya. Dengan menyederhanakan data hanya ke kolom-kolom tersebut, proses analisis menjadi lebih fokus, efisien, dan sesuai dengan tujuan pengujian algoritma *clustering*. Setelah kolom-kolom tersebut berhasil diekstraksi dari dataset awal, data yang telah difilter siap untuk digunakan dalam proses pemodelan. Hasil dari data yang telah diseleksi ini dapat dilihat pada Tabel 2 berikut:

Tabel 2. Data yang dipakai

User Name	Score	ThumbsUpCount
Chikka Risa	1	20
Dewi Nurlaela	5	40
…	…	…
Rahmat Hidayat	1	0
Toraya Gaming	1	3

Setelah data yang relevan berhasil disaring dan ditampilkan pada Tabel 2, langkah selanjutnya adalah melakukan proses normalisasi untuk menyamakan skala antar variabel. Normalisasi ini bertujuan agar algoritma *clustering* seperti *K-Means* dan *K-Medoids* dapat bekerja secara optimal tanpa bias terhadap skala data yang berbeda. Teknik yang digunakan dalam penelitian ini adalah *Min-Max Scaling*, yang diterapkan melalui rumus di Microsoft Excel. Metode ini mengubah nilai asli dalam kolom menjadi rentang antara 0 hingga 1, sehingga distribusi nilai menjadi lebih seragam dan proporsional.

Namun, dalam proses normalisasi ditemukan beberapa nilai nol pada kolom tertentu, khususnya pada *ThumbsUpCount*. Nilai nol ini berpotensi mengganggu perhitungan jarak dalam algoritma *clustering* karena dapat menyebabkan kesalahan interpretasi dalam pengelompokan data. Oleh karena itu, dilakukan penyesuaian lanjutan dengan menerapkan fungsi *IF* dan *RANDBETWEEN* pada Excel untuk mengganti nilai nol dengan angka acak dalam rentang tertentu. Pendekatan ini digunakan agar tetap menjaga proporsi data dan menghindari kekosongan nilai yang dapat mengganggu proses analisis.

Hasil dari normalisasi dan penyesuaian data ini akan membentuk dataset baru yang telah siap untuk digunakan dalam proses *clustering*. *Dataset* tersebut menyajikan nilai *score* dan *thumbsUpCount* dalam bentuk terstandarisasi, yang memungkinkan perbandingan algoritma *K-Means* dan *K-Medoids* dilakukan secara adil dan konsisten pada tahapan selanjutnya.

Tabel 3. Data setelah normalisasi

User Name	Score	Thumbs Up Count
Chikka Risa	0,0400	0,0047
Dewi Nurlaela	1,0000	0,0094
…	…	…
Rahmat Hidayat	0,0500	0,0400
Toraya Gaming	0,0200	0,0007

Pada Tabel 3 menampilkan hasil akhir dari proses normalisasi yang telah dilakukan terhadap dua variabel utama, yakni *Score* dan *ThumbsUpCount*. Nilai-nilai pada kolom tersebut kini telah berada dalam skala antara 0 hingga 1 sebagai hasil dari penerapan teknik *Min-Max Scaling*. Data yang telah dinormalisasi ini selanjutnya akan digunakan dalam proses *cluster*isasi menggunakan algoritma *K-Means* dan *K-Medoids*. Akan tetapi, sebelum dilakukan implementasi penuh ke dalam perangkat lunak *Rapid Miner*, penulis terlebih dahulu melakukan simulasi dan verifikasi melalui perhitungan manual menggunakan *Microsoft Excel*, khususnya untuk algoritma *K-Means*. Langkah ini diambil untuk memastikan bahwa pemahaman terhadap proses kerja algoritma bersifat menyeluruh serta untuk mengevaluasi apakah hasil akhir dari *software* selaras dengan hasil *cluster*isasi yang diperoleh secara manual. Dalam proses manual ini, penulis memulai dengan menentukan *centroid* awal secara acak, kemudian menghitung jarak *Euclidean* antara masing-masing data dengan *centroid* tersebut. Setelah itu, dilakukan proses iteratif berupa pengelompokan ulang dan pembaruan nilai *centroid* hingga terjadi konvergensi, yaitu kondisi di mana perubahan posisi *centroid* dari iterasi sebelumnya sangat kecil atau tidak signifikan lagi. Proses ini membantu memperjelas bagaimana data yang berbeda dikelompokkan berdasarkan kemiripan jarak, dan bagaimana *centroid* secara dinamis menyesuaikan diri terhadap distribusi data. Setelah pembentukan *cluster* manual selesai, dilakukan perbandingan hasilnya dengan output dari *Rapid Miner*. Untuk mengevaluasi kualitas masing-masing hasil *cluster*, digunakan metrik *Davies-Bouldin Index (DBI)*. Nilai DBI yang lebih rendah menandakan bahwa *cluster* yang terbentuk memiliki jarak antar *cluster* yang cukup besar dan kepadatan dalam *cluster* yang relatif tinggi, yang merupakan indikator dari

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1320−1331
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8145

klasterisasi yang baik. Melalui tahapan ini, penulis dapat menilai secara lebih obyektif efektivitas dari metode yang digunakan dan sekaligus memastikan bahwa proses klasterisasi tidak hanya bergantung pada perangkat lunak, tetapi juga dapat dipahami dan direplikasi secara manual apabila dibutuhkan.

3.1.1 Deskripsi Data Akhir

Setelah melalui tahapan normalisasi yang telah dijelaskan diatas, diperoleh data akhir yang siap untuk digunakan dalam proses clustering. Dataset akhir terdiri dari 500 data atau baris, dengan dua atribut utama yaitu:

a. *score:* menggambarkan skor penilaian pengguna terhadap aplikasi (bernilai 1 sampai 5),
b. *thumbsupcount:* menunjukkan jumlah pengguna lain yang memberikan *thumbs up* atau dukungan terhadap ulasan tersebut.

Dua atribut numerik ini dipilih sebagai dasar dalam proses *clustering* karena dianggap mampu merepresentasikan sentimen pengguna secara kuantitatif tanpa melalui analisis teks yang kompleks. Analisis berbasis teks seringkali menghadapi tantangan berupa ambiguitas bahasa, perbedaan gaya penulisan, serta kebutuhan akan proses *preprocessing* yang cukup panjang, seperti tokenisasi, stemming, dan penghapusan *stopword*. Dengan menggunakan atribut numerik, proses pengolahan data dapat dilakukan dengan lebih sederhana, cepat, dan efisien, karena setiap nilai sudah dapat diukur secara langsung dalam bentuk angka. Selain itu, penggunaan atribut numerik memungkinkan penerapan algoritma berbasis jarak, seperti K-Means maupun K-Medoids, secara lebih optimal. Sebelum tahap pengelompokan dilakukan, data terlebih dahulu dinormalisasi agar seluruh nilai berada dalam rentang yang sebanding. Normalisasi ini penting untuk menghindari dominasi atribut tertentu yang memiliki skala lebih besar, sehingga hasil *clustering* yang diperoleh lebih akurat, objektif, serta mampu merefleksikan pola sesungguhnya dalam dataset.

3.1.2 Hasil Clustering *K-Means*

Setelah proses normalisasi data, langkah selanjutnya adalah melakukan penghitungan algoritma *K-Means* secara manual menggunakan *Microsoft Excel* dengan jumlah *cluster* sebanyak tiga (k=3). Penentuan *centroid* awal dilakukan secara acak dari tiga titik data yang berbeda dalam *dataset*. Proses iterasi dimulai dengan menghitung jarak *Euclidean* antara setiap data dengan masing-masing *centroid*, lalu data dikelompokkan ke dalam *cluster* dengan jarak terdekat. Untuk *centroid* awal dapat dilihat pada Tabel 4 berikut:

Tabel 4. *Centroid* Awal

Centroid	score	thumbsup
C1	0,0900	0,0162
C2	1,0000	0,0500
C3	0,5000	0,0200

3.1.2.1 Menghitung jarak ke *Centroid* (*Euclidean Distance*)

Contoh untuk data ke 1 (0.0900, 0.0162)

a. Jarak data C1

$$\sqrt{(0.0400 - 0.0900)^2} + \sqrt{(0.0047 - 0.0162)^2} = 0.0513$$

b. Jarak data C2

$$\sqrt{(0.0400 - 1.0000)^2} + \sqrt{(0.0047 - 0.0500)^2} = 0.9611$$

c. Jarak data C3

$$\sqrt{(0.0400 - 0.5000)^2} + \sqrt{(0.0047 - 0.0200)^2} = 0.4603$$

Karena data paling dekat ke C1, maka data ke-1 masuk ke *Cluster* C1. Selanjutnya untuk data ke 2 hingga data ke 500 dapat dilihat pada Tabel 5 berikut. Untuk langkah perhitungannya sesuaikan dengan data ke 1.

Tabel 5. Pengelompokan Centroid

User Name	Centroid 1	Centroid 2	Centroid 3	Centroid
Chikka Risa	0,0513	0,9611	0,4603	C1
Dewi Nurlaela	0,9100	0,0406	0,5001	C2
…	…	…	…	…
Rahmat Hidayat	0,0466	0,9501	0,4504	C1
Toraya Gaming	0,0717	0,9812	0,4804	C1

Selanjutnya, nilai *centroid* diperbarui dengan menghitung rata-rata dari seluruh atribut anggota *cluster* yang terbentuk pada tahap sebelumnya. Proses iterasi ini berlangsung secara berulang hingga posisi *centroid* mencapai kondisi stabil atau tidak mengalami perubahan yang signifikan. Pada hasil iterasi keempat, diperoleh *centroid* akhir

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1320−1331
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8145

yang merepresentasikan karakteristik utama dari masing-masing *cluster* yang terbentuk seperti pada tabel 6 dibawah ini.

Tabel 6. Centroid Akhir

C1	0,0884	0,0150
C2	0,9980	0,0564
C3	0,6227	0,0207

3.1.2.2 Menghitung DBI *K-Means*

Jumlah anggota dari masing-masing *cluster* yang terbentuk adalah sebanyak 124 data pada *cluster* 1, 266 data pada *cluster* 2, dan 110 data pada *cluster* 3. Setelah proses pengelompokan selesai, dilakukan tahap evaluasi untuk menilai kualitas hasil *clustering*. Evaluasi dilakukan menggunakan metode *Davies-Bouldin Index* (DBI), yang mempertimbangkan nilai dispersi dalam *cluster* (σ), jarak antar *centroid*, serta nilai R(i,j) pada setiap pasangan *cluster*. Semakin kecil nilai DBI, maka semakin baik kualitas pemisahan antar *cluster* karena menunjukkan bahwa data dalam satu *cluster* lebih homogen dan jarak antar *cluster* lebih besar. Berdasarkan perhitungan manual menggunakan *Microsoft Excel*, diperoleh nilai DBI sebesar 0,457. Nilai ini mengindikasikan bahwa hasil pengelompokan yang diperoleh cukup baik serta dapat diterima secara statistik. Perhitungan detail disajikan pada tabel 7 dibawah ini.

Tabel 7. Hitung SSW (σ) Tiap *Cluster*

SSW 1	0,0692
SSW 2	0,0594
SSW 3	0,1289

Penghitungan Jarak Antar *Centroid* (SSB)

$$d(C1,C2) = \sqrt{(0{,}0884 - 0{,}9980)^2} + \sqrt{(0{,}0150 - 0{,}0564)^2} = 0{,}9105$$

$$d(C1,C3) = \sqrt{(0{,}0884 - 0{,}6227)^2} + \sqrt{(0{,}0150 - 0{,}0207)^2} = 0{,}5343$$

$$d(C2,C3) = \sqrt{(0{,}9980 - 0{,}6227)^2} + \sqrt{(0{,}0564 - 0{,}0207)^2} = 0{,}3770$$

a. Hitung nilai R(i,j) untuk setiap pasangan *cluster*
Cluster 1:

$$R(1,2) = \frac{0{,}0692+0{,}0594}{0{,}9105} = 0{,}1413$$

$$R(1,3) = \frac{0{,}0692+0{,}1289}{0{,}5343} = 0{,}3707$$

Max R1=0,3707

Cluster 2:

$$R(2,1) = \frac{0{,}0594+0{,}0692}{0{,}9105} = 0{,}1413$$

$$R(2,3) = \frac{0{,}0594+0{,}1289}{0{,}3770} = 0{,}4995$$

Max R2=0,4995

Cluster 3:

$$R(3,1) = \frac{0{,}1289+0{,}0692}{0{,}5343} = 0{,}3707$$

$$R(3,2) = \frac{0{,}1289+0{,}0594}{0{,}3770} = 0{,}4995$$

Max R3=0,4995

b. Menghitung DBI

$$DBI = \frac{0{,}3707+0{,}4995+0{,}4995}{3} = \frac{1.3697}{3} = 0{,}457$$

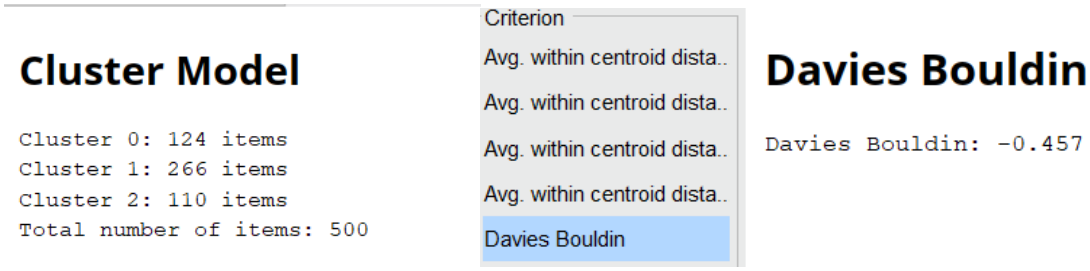
3.1.2.3 Implementasi Clustering *K-Means* dengan *Rapid Miner*

Proses *clustering* dengan algoritma *K-Means* menggunakan tiga *cluster* (k=3) dijalankan pada *software RapidMiner*. Hasilnya, data terbagi menjadi tiga kelompok dengan distribusi anggota sebagai berikut:
a. Cluster 1: 124 data

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1320−1331
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8145

b. Cluster 2: 266 data
c. Cluster 3: 110 data

Sesuai Gambar 3 Nilai *Davies-Bouldin Index* (DBI) yang dihasilkan dari proses clustering ini adalah sebesar 0.457, yang menunjukkan bahwa hasil pembagian cluster cukup baik. Nilai DBI yang rendah mengindikasikan bahwa antar-cluster memiliki jarak yang signifikan dan anggota dalam satu cluster memiliki kedekatan yang baik.



Gambar 3. Hasil Clustering *K-Means* pada *Rapid Miner*

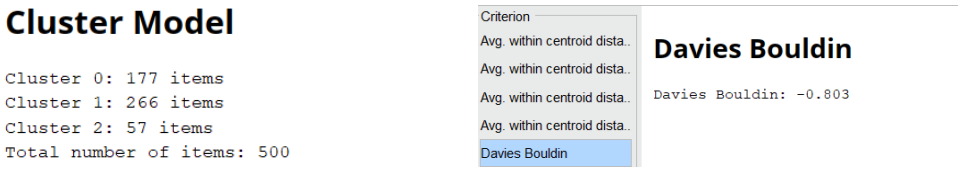
3.1.2.4 Interpretasi Makna Setiap Cluster

Setelah data terbagi dalam tiga cluster, dilakukan analisis terhadap karakteristik masing-masing kelompok berdasarkan nilai *score* dan *thumbsupcount* rata-rata:

a. Cluster 1: Berisi ulasan dengan score rendah namun memiliki jumlah thumbsup sedang. Ini dapat diartikan sebagai ulasan negatif yang relevan atau dipercaya oleh pengguna lain. Cluster ini merepresentasikan ketidakpuasan yang dianggap penting.
b. Cluster 2: Didominasi oleh data dengan score tinggi dan thumbsup tinggi. Ini menunjukkan ulasan yang sangat positif dan mendapat banyak dukungan dari pengguna lain. Cluster ini mencerminkan kepuasan pengguna tinggi.
c. Cluster 3: Memiliki score sedang dan thumbsup rendah, menunjukkan ulasan yang kemungkinan bersifat netral atau tidak terlalu berdampak bagi pengguna lain. Cluster ini menggambarkan respon rata-rata atau ulasan biasa saja.

3.1.3 Hasil Clustering *K-Medoids*

Setelah menghitung menggunakan algoritma *K-Means* dan mendapatkan hasil DBI 0,457. Penulis melakukan penghitungan menggunakan algoritma *K-Medoids* untuk melakukan perbandingan algoritma mana yang lebih baik dalam ulasan sentimen. Dalam penelitian ini, metode *K-Medoids clustering* diterapkan untuk mengelompokkan data menjadi tiga *cluster* berdasarkan *Euclidean Distance* sebagai metrik jarak. Hasil *clustering* menunjukkan bahwa *Cluster* 1 terdiri dari 177 objek, *Cluster* 2 terdiri dari 266 objek, dan *Cluster* 3 terdiri dari 57 objek. Evaluasi hasil *clustering* dilakukan menggunakan *Davies-Bouldin Index* (DBI), yang mengukur validitas *cluster* berdasarkan rasio antara dispersi dalam *cluster* dan jarak antar *cluster*. Nilai *Davies-Bouldin Index* (DBI) yang diperoleh dari hasil clustering *K-Medoids* adalah sebesar 0.803, yang jauh lebih tinggi dibandingkan hasil *K-Means*. Hal ini mengindikasikan bahwa pemisahan antar *cluster* pada metode *K-Medoids* kurang optimal, dengan kemungkinan bahwa batas antar *cluster* tidak terlalu jelas, atau terdapat banyak data yang saling tumpang tindih. Hasil *clustering* dapat dilihat pada Gambar 4 dibawah ini.



Gambar 4. Hasil Clustering *K-Medoids* pada *Rapid Miner*

Sayangnya, *Rapid Miner* tidak menyajikan secara eksplisit nilai medoid atau titik pusat dari masing-masing cluster dalam bentuk numerik yang mudah diakses, sehingga tidak dilakukan validasi manual seperti pada *K-Means*. Namun, hasil clustering tetap dapat dianalisis secara kualitatif berdasarkan karakteristik distribusi nilai *score* dan *thumbsupcount*.

3.1.3.1 Interpretasi *Cluster K-Medoids*

Analisis terhadap data dalam tiap cluster menunjukkan bahwa:

a. Cluster 1: Memiliki variasi score yang lebih lebar namun cenderung ke nilai rendah dan thumbsupcount rendah pula. Kemungkinan berisi ulasan negatif yang kurang mendapat perhatian dari pengguna lain.
b. Cluster 2: Konsisten menunjukkan nilai score tinggi dan jumlah thumbsupcount relatif besar, mengindikasikan ulasan positif dan relevan.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1320−1331
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8145

c. Cluster 3: Jumlah anggotanya paling sedikit, terdiri dari data dengan nilai score sedang dan thumbsupcount acak. Cluster ini merepresentasikan ulasan yang kurang menonjol baik dari segi kepuasan maupun perhatian pengguna lain.

3.1.4 Perbandingan Kinerja Algoritma

Untuk menentukan algoritma yang memberikan hasil terbaik pada *dataset* ini, dilakukan perbandingan performa menggunakan metrik *Davies-Bouldin Index* (DBI). Nilai DBI dipilih karena mampu menggambarkan tingkat kedekatan antar anggota dalam satu *cluster* serta pemisahan antar *cluster*. Hasil perbandingan disajikan pada Tabel 8 berikut:

Tabel 8. Perbandingan performa berdasarkan nilai *Davies-Bouldin Index* (DBI)

	K-Means	*K-Medoids*
Cluster	C1 = 124 C2 = 266 C3 = 110	C1 = 177 C2 = 266 C3 = 57
DBI	0,457	0,803

Dari hasil tersebut, terlihat bahwa *K-Means* memiliki nilai DBI yang lebih rendah, yang berarti menghasilkan *cluster* dengan separasi yang lebih baik dan kompak dibandingkan *K-Medoids* pada data yang digunakan.

3.1 Pembahasan

Hasil penelitian ini menunjukkan bahwa algoritma *K-Means* menghasilkan nilai *Davies-Bouldin Index* (DBI) yang lebih rendah dibandingkan algoritma *K-Medoids*, yaitu sebesar 0,457 untuk *K-Means* dan 0,803 untuk *K-Medoids*. Nilai DBI yang lebih rendah menunjukkan bahwa *K-Means* mampu membentuk *cluster* yang lebih kompak dan memiliki jarak antar *cluster* yang baik, sehingga secara umum lebih optimal untuk digunakan dalam konteks dataset ini.

Keunggulan *K-Means* pada penelitian ini dapat dijelaskan dari karakteristik algoritmanya yang menggunakan rata-rata (*mean*) sebagai pusat *cluster*, sehingga bekerja dengan baik pada dataset numerik dan bersih yang tidak banyak mengandung outlier. Sementara itu, *K-Medoids* menggunakan representasi titik data aktual sebagai pusat cluster, dan lebih cocok diterapkan ketika terdapat *noise* atau nilai ekstrem. Namun, karena data dalam penelitian ini telah dinormalisasi dan relatif bersih dari outlier, maka keunggulan *K-Medoids* tidak terlalu terlihat.

3.2.1 Interpretasi Cluster sebagai Sentimen

Dengan menggunakan dua atribut numerik, yaitu *score* dan *thumbsupcount*, proses *clustering* dalam penelitian ini bertujuan untuk menangkap sentimen ulasan pengguna secara implisit. Nilai *score* merepresentasikan kepuasan pengguna terhadap aplikasi, sedangkan *thumbsupcount* menunjukkan validasi sosial atau relevansi ulasan bagi pengguna lain. Berdasarkan hasil clustering menggunakan *K-Means*, diperoleh tiga kelompok utama:

a. Cluster pertama terdiri dari ulasan dengan *score* rendah namun *thumbsupcount* tinggi, yang dapat diartikan sebagai ulasan negatif tetapi relevan dan dianggap penting oleh pengguna lain.
b. Cluster kedua berisi ulasan dengan score tinggi dan dukungan tinggi, yang mengindikasikan kepuasan yang kuat dan kredibel.
c. Cluster ketiga menunjukkan nilai score sedang dan dukungan rendah, yang dapat diartikan sebagai ulasan netral atau biasa saja.

Interpretasi ini menunjukkan bahwa pendekatan numerik terhadap sentimen dapat memberikan insight yang bermakna meskipun tidak menggunakan data teks secara langsung. Hal ini membuktikan bahwa data numerik juga dapat diolah untuk menggambarkan sentimen pengguna dengan metode clustering yang tepat.

Meskipun tidak dilakukan analisis teks atau *natural language processing*, pendekatan ini dapat dikatakan berhasil dalam menggambarkan struktur sentimen ulasan secara kuantitatif. Nilai score dan thumbsupcount telah terbukti mencerminkan dua sisi penting dari ulasan pengguna: kepuasan dan relevansi. Oleh karena itu, pembentukan cluster dari kombinasi kedua variabel ini memberikan alternatif ringan namun efektif untuk pengelompokan sentimen dalam skala besar.

3.2.2 Konsistensi dengan Penelitian Sebelumnya

Temuan dalam penelitian ini konsisten dengan berbagai penelitian sebelumnya, yang menunjukkan bahwa *K-Means* lebih unggul dalam pengelompokan data numerik yang telah dinormalisasi. Pada penjelasan pendahuluan diatas bahwa *K-Means* menghasilkan validitas DBI jauh lebih baik dibandingkan *K-Medoids*.

3.2.3 Implikasi Praktis bagi Pengembang Aplikasi *Shopee*

Dari sudut pandang implementasi, hasil clustering ini dapat memberikan informasi strategis bagi pengembang aplikasi Shopee. Beberapa implikasi praktis yang dapat diambil antara lain:

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1320−1331
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8145

a. Identifikasi Masalah: Cluster dengan nilai *score* rendah namun *thumbsupcount* tinggi dapat menjadi indikator adanya fitur aplikasi yang banyak dikritik namun dianggap penting. *Cluster* ini dapat menjadi prioritas utama untuk perbaikan.
b. Peningkatan Fitur Populer: *Cluster* dengan *score* tinggi dan *thumbsupcount* tinggi menunjukkan fitur atau aspek aplikasi yang sangat disukai pengguna. Ini dapat dijadikan landasan untuk memperkuat fitur unggulan dan menjaga kepuasan pengguna.
c. Evaluasi Fitur Netral: *Cluster* dengan *score* dan *thumbsupcount* sedang atau rendah bisa menjadi bahan refleksi. Fitur-fitur yang tidak menimbulkan kepuasan maupun ketidakpuasan signifikan mungkin perlu reformulasi atau penghapusan agar tidak membebani pengalaman pengguna.

Dengan pendekatan ini, pengembang dapat mengambil keputusan berbasis data untuk meningkatkan kualitas layanan dan pengalaman pengguna secara menyeluruh. Maka, hasil clustering bukan hanya alat analitik, melainkan juga pendukung strategi pengembangan produk yang berkelanjutan.

4. KESIMPULAN

Berdasarkan hasil penelitian yang penulis lakukan, penulis menyimpulkan bahwa pemanfaatan algoritma *K-Means* dan *K-Medoids* dalam membagi golongan data numerik yang berupa *score* dan *thumbsupcount* pada ulasan aplikasi *Shopee* di *Google Play Store* berhasil memberikan gambaran awal mengenai kecenderungan sentimen pengguna. Hasil penelitian yang berhasil membandingkan efektivitas yang diberikan oleh kedua algoritma adalah tujuan penelitian. Hal ini ditunjukkan menggunakan hasil evaluasi dengan metrik *Davies-Bouldin Index*, di mana algoritma *K-Means* dan *K-Medoids* mendapatkan masing-masing nilai 0,457 dan 0,803. Dengan demikian, didapatkan hasil bahwa algoritma *K-Means* memiliki performa yang lebih baik dalam pembentukan *cluster* yang kompak dan memiliki jarak terpisah. Pertanyaan penelitian ini telah berhasil dijawab oleh hasil tersebut dengan adanya solusi efektif terkait cara mana yang lebih efektif menggunakan metode klasterisasi pada data numerik ulasan pengguna aplikasi. Namun demikian, penelitian ini juga memiliki beberapa kekurangan, yaitu hanya terdapat 500 data ulasan yang digunakan menjadi hasil analisis ini tidak begitu merepresentasikan seluruh populasi pengguna *Shopee*. Selain itu, hasil analisis ini juga hanya menggunakan satu metrik evaluasi, yakni DBI sehingga belum sepenuhnya mampu menilai kualitas *cluster* secara mendalam. Selanjutnya untuk penelitian berikutnya, diharapkan untuk menggunakan data sebanyak dan beragam serta menambahkan metrik evaluasi standar yang lain seperti *Silhouette Score* sehingga analisis *cluster* yang didapat lebih dapat diuji coba secara modifikasi.

REFERENCES

[1] A. H. Hasugian, M. Fakhriza, and D. Zukhoiriyah, "Analisis Sentimen Pada Review Pengguna E-Commerce Menggunakan Algoritma Naïve Bayes," *J-SISKO TECH (Jurnal Teknol. Sist. Inf. dan Sist. Komput. TGD)*, vol. 6, no. 1, p. 98, Jan. 2023, doi: 10.53513/jsk.v6i1.7400.
[2] B. Z. Ramadhan, R. I. Adam, and I. Maulana, "Analisis Sentimen Ulasan pada Aplikasi E-Commerce dengan Menggunakan Algoritma Naïve Bayes," *J. Appl. Informatics Comput.*, vol. 6, no. 2, pp. 220–225, Dec. 2022, doi: 10.30871/jaic.v6i2.4725.
[3] S. Suswadi and M. Erkamim, "Sentiment Analysis of Shopee App Reviews Using Random Forest and Support Vector Machine," *Ilk. J. Ilm.*, vol. 15, no. 3, pp. 427–435, Dec. 2023, doi: 10.33096/ilkom.v15i3.1610.427-435.
[4] E. T. Ena Tasia and M. Afdal, "Perbandingan Algoritma *K-Means* Dan *K-Medoids* Untuk Clustering Daerah Rawan Banjir Di Kabupaten Rokan Hilir," *Indones. J. Inform. Res. Softw. Eng.*, vol. 3, no. 1, pp. 65–73, Mar. 2023, doi: 10.57152/ijirse.v3i1.523.
[5] N. A. S. Z. Abidin, R. D. Avila, A. Hermatyar, and R. Rismayani, "Perbandingan Algoritma *K-Means* dan *K-Medoids* untuk Pengelompokan Daerah Produksi Kakao," *J. Tek. Inform. dan Sist. Inf.*, vol. 8, no. 2, Aug. 2022, doi: 10.28932/jutisi.v8i2.4897.
[6] Laila Ali Putri, Mazayah Tsaqofah, Dea Syahfira Hasibuan, Hasti Fadillah, Maria Ulfa, and Mhd.Furqan, "Application of *K-Means* Clustering Algorithm for E-Commerce Data Analysis," *J. Artif. Intell. Eng. Appl.*, vol. 4, no. 3, pp. 2364–2367, Jun. 2025, doi: 10.59934/jaiea.v4i3.1170.
[7] M. R. H. Arsyad and Sulastri, "Klusterisasi Data Review Pengguna Aplikasi Marketplace Blibli.Com dengan Algoritma *K-Means* dan *K-Medoids*," *J. Tek. Inform. Unika Santo Thomas*, vol. 09, pp. 2657–1501, 2024, https://doi.org/10.54367/jtiust.v9i1.
[8] R. Mohemad, N. N. Mohd Muhait, N. M. Mohamad Noor, and Z. A. Othman, "Performance analysis in text clustering using *K-Means* and *K-Medoids* algorithms for Malay crime documents," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 5, p. 5014, Oct. 2022, doi: 10.11591/ijece.v12i5.pp5014-5026.
[9] M. M. Dinar Danureksa, R. Kurnawan, and Y. Ira Arie Wijaya, "Penerapan Algoritma *K-Means* untuk Optimasi Model Clustering Data Supplier di Aplikasi Shopee," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 1, pp. 1676–1684, Jan. 2025, doi: 10.36040/jati.v9i1.12723.
[10] A. R. Aruadi, F. Andreas, and B. N. Sari, "Analisis Klaster *K-Means* pada Data Rata-Rata Konsumsi Kalori dan Protein Menurut Provinsi dengan Metode Davies Bouldin Index," *J. Pendidik. dan Konseling*, vol. 4, no. 4, pp. 5652–5661, 2022.
[11] N. Amalia Putri, A. Srirahayu, and N. Arif Sudibyo, "Sentiment Analysis Towards the KitaLulus Application Using the Naive Bayes Method from Google Play Store Reviews," *J. Indones. Sos. Teknol.*, vol. 5, no. 10, pp. 4593–4603, Oct. 2024, doi: 10.59141/jist.v5i10.1244.
[12] E. R. Kaburuan and N. R. Setiawan, "Sentimen Analisis Review Aplikasi Digital Korlantas Pada Google Play Store Menggunakan Metode SVM," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 12, no. 1, pp. 105–116, Mar. 2023, doi:

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1320−1331
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8145

10.32736/sisfokom.v12i1.1614.
[13] P. P. Allorerung, A. Erna, M. Bagussahrir, and S. Alam, "Analisis Performa Normalisasi Data untuk Klasifikasi K-Nearest Neighbor pada Dataset Penyakit," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 9, no. 3, pp. 178–191, Sep. 2024, doi: 10.14421/jiska.2024.9.3.178-191.
[14] Maureen Tumalun, Melky Pangemanan, and Augustinus Robin Butarbutar, "Pembuatan Tabel Microsoft Excel Dan Rumus Fungsi Microsoft Excel Untuk Data Kesehatan," *J. Praba J. Rumpun Kesehat. Umum*, vol. 2, no. 2, pp. 36–41, Jun. 2024, doi: 10.62027/praba.v2i2.96.
[15] Y. A. Singgalen, "Pemilihan Metode dan Algoritma dalam Analisis Sentimen di Media Sosial : Sistematic Literature Review," *J. Inf. Syst. Informatics*, vol. 3, no. 2, pp. 278–302, Jun. 2021, doi: 10.33557/journalisi.v3i2.125.
[16] G. G. Setiaji, K. P. Gunata, and G. Setiarso, "Optimasi Clustering *K-Means* Menggunakan Algoritma Genetika Dengan Data View Dan Like Di Tiktok," *J. Transform.*, vol. 22, no. 2, pp. 115–120, Jan. 2025, doi: 10.26623/y2tedy77.
[17] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "*K-Means* clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (Ny).*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.
[18] E. U. Oti, M. O. Olusola, F. C. Eze, and S. U. Enogwe, "Comprehensive Review of *K-Means* Clustering Algorithms," *Int. J. Adv. Sci. Res. Eng.*, vol. 07, no. 08, pp. 64–69, 2021, doi: 10.31695/IJASRE.2021.34050.
[19] S. Z. Rindiawana, A. N. Rachman, and V. Purwayoga, "Optimasi Jumlah Cluster untuk Analisis Penjualan Barang Kosmetik Menggunakan *K-Medoids*," *J. SIistem dan Teknol. Inf.*, vol. 13, no. 1, pp. 148–165, 2025, doi: 10.26418/justin.v13i1.86637.
[20] A. E. Ezugwu *et al.*, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Eng. Appl. Artif. Intell.*, vol. 110, p. 104743, Apr. 2022, doi: 10.1016/j.engappai.2022.104743.
[21] N. Ain Kilo, M. R. Katili, and I. K. Hasan, "Perbandingan Metode *K-Means* dan *K-Medoids* Dengan Validitas Davies-Bouldin Indeks, Dunn Indeks dan Indeks Connectivity Pada Pengelompokkan Masyarakat Penerima Bantuan Langsung Tunai Under the licence CC BY-," *Res. Math. Nat. Sci*, vol. 4, no. 1, pp. 8–15, 2025, doi: 10.55657/rmns.v4i1.190.
[22] E. Muningsih, I. Maryani, and V. R. Handayani, "Penerapan Metode *K-Means* dan Optimasi Jumlah Cluster dengan Index Davies Bouldin untuk Clustering Propinsi Berdasarkan Potensi Desa," *J. Sains dan Manaj.*, vol. 9, no. 1, pp. 95–100, 2021, doi: 10.31294/evolusi.v9i1.10428.
[23] R. Siagian, P. Sirait, and A. Halim, "The Implementation of *K-Means* dan *K-Medoids* Algorithm for Customer Segmentation on E-commerce Data Transactions," *SISTEMASI*, vol. 11, no. 2, p. 260, May 2022, doi: 10.32520/stmsi.v11i2.1337.
[24] E. Prasetyaningrum and P. Susanti, "Perbandingan Algoritma *K-Means* Dan *K-Medoids* Untuk Pemetaan Hasil Produksi Buah-Buahan," *J. MEDIA Inform. BUDIDARMA*, vol. 7, no. 4, p. 1775, Oct. 2023, doi: 10.30865/mib.v7i4.6477.