Nusantara Journal of Multidisciplinary Science
Vol. 1, No. 8, Maret 2024
Hal 611-621
E-ISSN : 3024-8752
P-ISSN : 3024-8744
Site : https://jurnal.intekom.id/index.php/njms

Perbandingan Metode *Clustering* pada Kelompok Dokumen Teks Untuk Menentukan Nilai Ambang Batas Similaritas

Scoriny Noor Hasanah[1]
[1]Fakultas Teknik, Universitas Tanjungpura, Pontianak, Indonesia

Article Info

Article history:

Received Maret 14, 2023
Revised Maret 14, 2023
Accepted Maret 14, 2023

Kata Kunci :

K-Means
K-Medoids
Dokumen
Jaccard Similarity

Keywords:

K-Means
K-Medoids
Document
Jaccard Similarity

ABSTRAK

Dokumen teks adalah bentuk informasi tertulis yang dapat dibaca oleh komputer dan dapat dibaca oleh manusia. Jumlah dokumen teks yang dihasilkan terus meningkat seiring dengan pertumbuhan data digital. Untuk menemukan pola atau hubungan antara dokumen teks dalam hal ini, sangat penting untuk dapat mengelompokkannya.Salah satu tantangan dalam pemrosesan dokumen teks adalah menentukan nilai similaritas minimal yang diperlukan agar dokumen termasuk dalam satu kelompok. Pengelompokan dokumen dapat digunakan untuk mencari nilai ambang batas. Nilai ambang batas menunjukkan batas data yang dapat hadir dalam anggota *cluster*. Metode pengelompokan yang akan digunakan adalah *K-Means* dan *K-Medoids*.Penelitian ini bertujuan untuk membandingkan metode *cluster* kemudian mencari nilai ambang batas pada kelompok dokumen menggunakan jaccard similarity. Hasil dari penelitian ini adalah berdasarkan nilai silhouette coefficient algoritma *K-Means* lebih unggul dari pada algoritma *K-Medoids* untuk pengelompokan dokumen teks dengan nilai silhouette *K-Means* sebesar 0.0304 sedangkan untuk *K-Medoids* menghasilkan nilai silhouette sebesar 0.226. Diperoleh ambang nilai similaritas yang dihitung menggunakan jaccard similarity sebesar 0.07 berdasarkan pengelompokan *K-Means*.

ABSTRACT

Text documents are a form of written information that can be read by computers and can be read by humans. The number of text documents generated continues to increase along with the growth of digital data. To find patterns or relationships between text documents in this case, it is very important to be able to cluster them. One of the challenges in text document processing is determining the minimum similarity value required for documents to belong to a group. Document clustering can be used to find a threshold value. The threshold value indicates the limit of data that can be present in a cluster member. The clustering methods to be used are K-Means and K-Medoids. This study aims to compare cluster methods and then find the threshold value on document groups using jaccard similarity. The results of this study are based on the silhouette coefficient value of the K-Means algorithm is superior to the K-Medoids algorithm for clustering text documents with a K-Means silhouette value of 0.0304 while for K-Medoids it produces a silhouette value of 0.226. The similarity value threshold calculated using Jaccard similarity is 0.07 based on K-Means clustering.

This is an open access article under the CC BY license.

Corresponding Author:

Scoriny Noor Hasanah
Fakultas Teknik, Universitas Tanjungpura,
Pontiannak, Indonesia
Email: scorinynoor@student.untan.ac.id

1. PENDAHULUAN

Dokumen teks adalah bentuk informasi tertulis yang dapat dibaca oleh komputer dan dapat dibaca oleh manusia. Jumlah dokumen teks yang dihasilkan terus meningkat seiring dengan pertumbuhan data digital. Untuk menemukan pola atau hubungan antara dokumen teks dalam hal ini, sangat penting untuk dapat mengelompokkannya. Teknik pengorganisasian dokumen ke dalam kelompok-kelompok berdasarkan kesamaan topik atau substansi dikenal sebagai pengelompokan dokumen [1].

Pengelompokan dokumen dapat digunakan untuk mencari nilai ambang batas. Nilai ambang batas menunjukkan batas data yang dapat hadir dalam anggota *cluster* [2]. Salah satu tantangan dalam pemrosesan dokumen teks adalah menentukan nilai similaritas minimal yang diperlukan agar dokumen termasuk dalam satu kelompok. Nilai similaritas ini dapat berbeda-beda tergantung pada tujuan analisis dan jenis dokumen yang sedang dipelajari. Pada penelitian kali ini akan menentukan nilai ambang batas dari kelompokan dokumen dengan menggunakan algortima *K-Means* dan *K-Medoids*. *K-Means* adalah teknik analisis data yang menerapkan pemodelan data unsupervised. Ini adalah salah satu metode yang digunakan untuk melakukan partisi atau pengelompokan pada setiap jenis data [3]. Untuk *K-Medoids* juga merupakan salah satu algoritma yang dapat digunakan untuk pengelompokan data menggunakan objek-objek dalam sebuah kelompok objek untuk menggambarkan sebuah klaster [4].

Beberapa penelitian telah dilakukan dengan menerapkan algoritma *K-Means* atau *K-Medoids*. Dalam penelitian [5] menggunakan algoritma, *K-Means*, *K-Medoids*, Fuzzy C-Means untuk pengelompokan buku menghasilkan nilai silhoutte sebesar 0.28 untuk *K-Means*, sebesar 0.30 untuk K-Meodids dan 0,26 untuk Fuzzy C-Means. Selanjutnya dalam penelitian [6] dalam mengelompokan indeks pembangunann mahasiswa, menghasilakan nilai DBI sebesar 108.18 untuk *K-Means*, DBI sebesar 0.956 untuk Hirearichale dan DBI sebasar 1.253 untuk *K-Medoids*.

Untuk menghitung nilai similaritas antar dokumen akan dihitung menggunakan Jaccard similarity. Jaccard similarity adalah metode yang digunakan untuk mengukur kesamaan antara dua himpunan berdasarkan elemen yang dibagikan di antara keduanya. Dalam konteks dokumen teks, Jaccard similarity dapat digunakan untuk mengukur sejauh mana dua dokumen memiliki kata-kata yang sama atau serupa. Semakin tinggi nilai Jaccard similarity antara dua dokumen, semakin mirip kedua dokumen tersebut [7]. Berdasarkan uraian tersebut penelitian kali akan mencari nilai ambang batas pada kelompok dokumen teks yang akan digunakan sebagai nilai minimal yang akan digunakan sebagai nilai rekomendasi untuk menyatakan kedua dokumen similar.

2. METODE

2.1 *K-Means*

K-Means adalah salah satu algoritma pengelompokan data yang bertujuan untuk membagi data yang dimiliki menjadi dua kelompok atau lebih. Pada algoritma ini, data dibagi menjadi beberapa kelompok berdasarkan kesamaan karaktersistik, data yang memiliki karakteristik sama masuk dalam satu kelompok dan data yang berkarakteristik berbeda dikelompokan ke dalam kelompok lain [8].

Adapun langkah-langkah untuk algoritma *K-Means* sebagai berikut[9]:

1. menentukan jumlah *cluster* k
2. ambing sampel data sebanyak jumlah *cluster* secara acak sebagai centroid
3. hitung jarak antara data dengan pusat *cluster* (centroid) dengan menggunakan rumus jarak Euclidean pada persamaan (1) berikut:

$$D(m,n) = \sqrt{\sum_{k=1}^{p}(x_{ik-}X_{jk})^2} \qquad (1)$$

Dimana D (m, n) adalah jarak data ke m ke pusat *cluster* n, x_{ik} adalah titik dari objek i pada ukuran k, x_{jk} merupakan titik dari objek j pada ukuran k.

4. Hitung kembali centroid dengan keanggotaan *cluster* yang baru
5. Jika pusat *cluster* tidak berubah maka proses *cluster* telah selesai, jika belum maka ulangi langkah 3 hingga pusat *cluster* tidak berubah lagi.

2.2 *K-Medoids*

K-medoids adalah algoritma partisi untuk mengelompokkan data yang mengelompokkan n objek ke dalam sebuah *cluster*. Untuk menghasilkan sebuah klaster, pendekatan ini menggunakan objek dalam sekumpulan objek untuk menggambarkan sebuah *cluster* data [10]. Berikut ini adalah langkah-langkah yang terlibat dalam algoritma *K-Means* [11]:

1. menentukan jumlah *cluster* yang ada (k).
2. Menentukan pusat asli setiap *cluster* secara acak.
3. Menghitung jarak euclidean dengan persamaan (1)
4. Pilih objek data secara acak dari setiap *cluster* untuk menerima medoid baru; ini menghasilkan total cost yang dihasilkan dalam penghitungan jarak, yang diperoleh dengan menghitung nilai terendah yang ditemukan di setiap *cluster*.

5) Menentukan nilai simpangan dengan menghitung nilai jarak baru dikurangi dengan besarnya nilai jarak lama. Jika simpangan< 0 maka akan membentuk suatu objek berukuran sama dengan data *cluster*, sehingga membentuk setiap k atau objek baru pada medoid, dan mengulangi langkah 2 – 4 hingga tidak ada perubahan pada medoid yang dihasilkan, sehingga diperoleh *cluster* dan anggota di setiap *cluster*.

2.3 TF-IDF

Term Frekuensi - Inverse Document Frekuensi (TF-IDF) merupakan teknik pembobotan kata dengan menghitung nilai Term Frekuensi dan menghitung kemunculan kata di seluruh kumpulan dokumen teks [12]. Jumlah kemunculan sebuah kata dalam suatu dokumen tertentu dikenal dengan istilah Term frekuensi, semakin sering suatu kata muncul, semakin tinggi frekuensinya. Banyaknya dokumen dengan istilah berdasarkan seluruh dokumen dalam koleksinya dikenal dengan invers frekuensi dokumen [6].Untuk menghitung TF-IDF menggunakan persamaan (2):

$w_{td} = tf_{td} * idf$ (2)

$w_{td} = tf_{td} * \log(\frac{N}{df_t})$ (3)

Keterangan:

w_{td}= bobot kata terhadap dokumen

tf_{td}= jumlah kemunculan kata dalam dokumen

N= jumlah semua dokumen

df_{ft}= jumlah dokumen yang mengandung kata

2.4 Preprocssing

Preprocessing adalah langkah yang dilakukan untuk menempatkan data dalam format yang sesuai untuk digunakan oleh sistem atau algoritma, sehingga dapat digunakan. langkah yang akan dilakukan dalam proses preprocessing adalan *case folding*, *removing punctuation* dan stopword removal [13].

Case folding mengubah huruf kapital menjadi huruf kecil. *Removing punctuation* adalah proses untuk menghilangkan tanda baca atau simbol. Tokenization adalah proses memisahkan kalimat menjadi kata-kata. Stopword removal adalah proses membuang kata yang dianggap tidak memiliki makna [14].

2.3. Jaccard Similarity

Jaccard Similarity digunakan untuk membandingkan dokumen dan menentukan seberapa mirip dua hal atau teks satu sama lain. Untuk persamaan jaccard dapat dilihat pada persamaan 4 ([7]:

Jaccard (x, y) = $\frac{|X \cap Y|}{|X \cup Y|}$ (4)

Dimana X adalah dokumen 1 dan Y adalah dokumen 2.

2.4 Silhouette Coeffecient

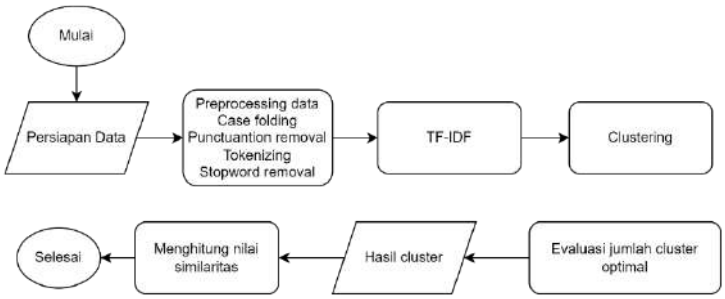
Silhouette Coefficient adalah salah satu metode untuk mengevaluasi atau mengukur kualitas dari sebuah *cluster* yang terbentuk. Untuk menghitung nilai Silhouette Coefficient menggunakan persamaan berikut [15]:

$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}$ (5)

dimana a_i nilai rata-rata objek ke-i dengan objek lainnya yang berada pada satu *cluster*, dan b_i nilai rata-rata objek ke-i dengan objek lainnya yang berada pada *cluster* yang berbeda. Nilai *silhouette coefficient* berkisar antara -1 hingga +1. Objek dengan nilai +1 ditetapkan dengan benar ke dalam satu *cluster* objek dengan nilai 0 berada di antara dua *cluster* atau memiliki posisi yang tidak jelas; dan objek dengan nilai -1 seharusnya ditetapkan ke dalam *cluster* yang berbeda.

2.5 Alur Penelitian

Pada Gambar 1 merupakan proses dari alur penelitian yang akan dilakukan pada penelitian ini.



Gambar 1. Alur Penelitian

Adapun untuk alur penelitian yang akan dilakukan akan diuraikan sebagai berikut:

2.5.1 Persiapan Data

Data yang akan digunakan yaitu berupa data abstak jurnal berbahasa Indonesia dengan tema Informatika sebanyak 500 abstrak. Tema informatika yang digunakan adalah Sistem Informasi Geografis, Jaringan Komputer, Sistem Pendukung Keputusan, Kecerdasan Buatan, dan Desain Perangkat Lunak. Pengumpulan data dilakukan dengan cara mengakses sumber jurnal seperti dari website Jurnal Informatika & Rekayasa, Jurnal Ilmiah Teknik, JATISI (Jurnal Teknik Informatika dan Sistem Informasi) dan mengunduh dokumen. Data yang telah diunduh simpan dalam satu file CSV.

2.5.2. Preprocessing

Pada tahap selanjutnya akan dilakukan preprocessing atau pembersihan data sepeti mengubah huruf kapital menjadi huruf kecil (*case folding*), penghapuhan tanda baca (*removing punctuation*), penghapusan stopword dan *tokenizing*.

2.5.3. TF-IDF

Sebelum dilakukan proses *clustering* makan akan dilakukan vektorisa yang mengubah data teks menjadi numerik menggunakan TF-IDF.

2.5.4. *Cluster*ing

Tahap selanjutnya setelah dilakukan proses vektorisa maka akan dilakukan pengelompokan menggunakan *K-Means* dan *K-Medoids*.

2.5.5 Evaluasi *Cluster*

Evaluasi *cluster* merupakan proses pengujian untuk menentukan jumlah *cluster* optimal dengan menggunakan silhouette coefficient. Dimana nilai silhouette yang optimal adalah nilai yang mendekati 1.

2.5.6. Perhitungan Nilai Similaritas

Setelah diperoleh jumlah *cluster* optimal maka akan dihitung nilai similaritas menggunakan *Jaccard Similarity* untuk setiap *cluster* yang terbentuk.

3. HASIL DAN PEMBAHASAN

3.1. Evaluasi *Cluster*

Evaluasi *cluster* dilakukan dengan menggunkan silhoutte coefficient. Evalusi ini dilakukan dengan menentukan jumlah *cluster* optimal yang dimulai dari 2 hingga ditemukan

kelompokan yang optimal yaitu nilai silhoutte yang mendekati 1 yang akan dipilih sebagai *cluster* optimal. Untuk hasil evaluasi *cluster* pada *K-Means* dapat dilihat pada Tabel 1.

Tabel 1. Evaluasi *cluster K-Means*

K	Rata-rata Silhouette Coefficient
K=2	0.0160
K=3	0.0202
K=4	0.0255
K=5	0.0304
K=6	0.0264
K=7	0.0221

Pada Tabel 1 merupakan hasil evaluasi *cluster* menggunakan silhoutte coefficient, nilai silhoutte yang optimal terdapat pada k=5 sebesar 0.0304 karena nilai tersebut yang paling mendekati 1. Maka k=5 dipilih sebegai jumlah *cluster* optimal pada algoritma *K-Means*.

Adapun untuk hasil evaluasi *cluster* pada *K-Medoids* dapat dilihat pada Tabel 2.

Tabel 2. Evaluasi *cluster K-Medoids*

K	Rata-rata Silhouette Coefficient
K=2	0.0083
K=3	0.0036
K=4	0.0100
K=5	0.0158
K=6	0.0128
K=7	0.0226

Pada Tabel 2 merupakan hasil dari evaluasi *cluster* pada *K-Medoids* menggunakan. Pada *K-Medoids* diperoleh nilai k optimal adalah k=7 sebesar 0.0226 nilai ini yang menunjukkan nilai yang paling optimal diantara jumlah *cluster* yang lainnya.

3.2 Hasil *Cluster*ing

*Cluster*ing dilakukan dengan algoritma *K-Means* dan *K-Medoids*, adapun untuk hasil dari anggota *cluster* yang tersebut dari masing-masing algortima dapat dilihat pada Tabel 3 merupakan anggota *cluster K-Means* dan Tabel 4 merupakan anggota *cluster K-Medoids*.

Tabel 3. Anggota *cluster K-Means*

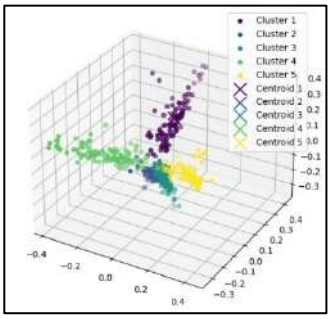
cluster	Jumlah Anggota
1	101
2	97
3	110
4	108
5	84

E-ISSN : 3024-8752
P-ISSN : 3024-8744

Tabel 4. Anggota *cluster K-Medoids*

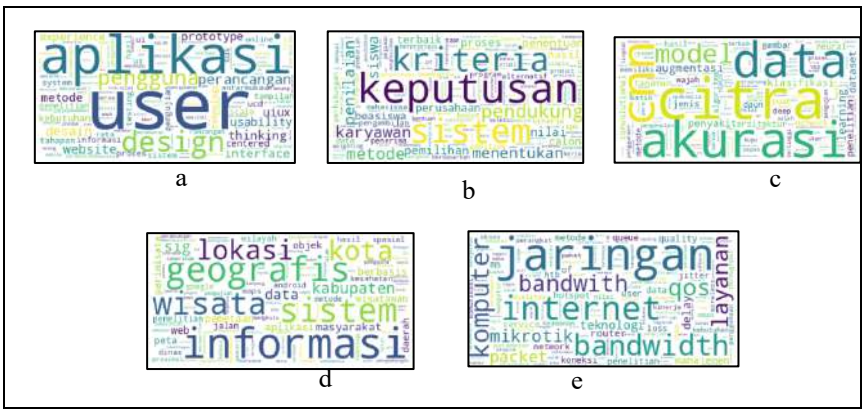
cluster	Jumlah Anggota
1	10
2	93
3	109
4	91
5	6
6	97
7	93

Pada tabel 3 merupakan anggota *cluster* yang terbentuk dari algortima *K-Means* untuk sebaran datanya cenderung seragam diatara setiap *cluster* yang terbentuk tetapi ada pada *cluster* 3 dan 4 memiliki anggota yang lebih banyak dari *cluster* lainnya. Berdasarakan Tabel 4 merupakan hasil *clustering* dari *K-Medoids* menghasilkan sebaran data yang lebih bervariasi dari pada *K-Means*. Pada Tabel 4 anggota *cluster* 1 dan 5 memiliki anggota yang lebih sedikit dari *cluster* yang lainnya. Adapun untuk hasil dari *clustering K-Means* dapat dilihat pada Gambar 2 dan untuk *Cluster*ing *K-Medoids* dapat dilibat pada tabel 4.



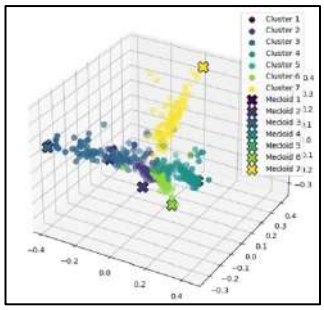
Gambar 2. *Cluster K-Means*

Pada Gambar 2 merupakan hasil dari sebaran data yang dihasilkan dari *clustering K-Means* yang memiliki sebaran data yang lebar dapat dilihat pada *cluster* 1 dan *cluster* 4. untuk *cluster* 2 lebih berkumpul dari pada *cluster* lainnya. *Cluster* 3 dan *cluster* 5 juga berkumpul tapi tidak sepadat seperti *cluster* 2. Adapun untuk informasi mengenai anggota yang terdapat pada setiap *cluster* dapat dilihat pada gambar 3 sebagai berikut.



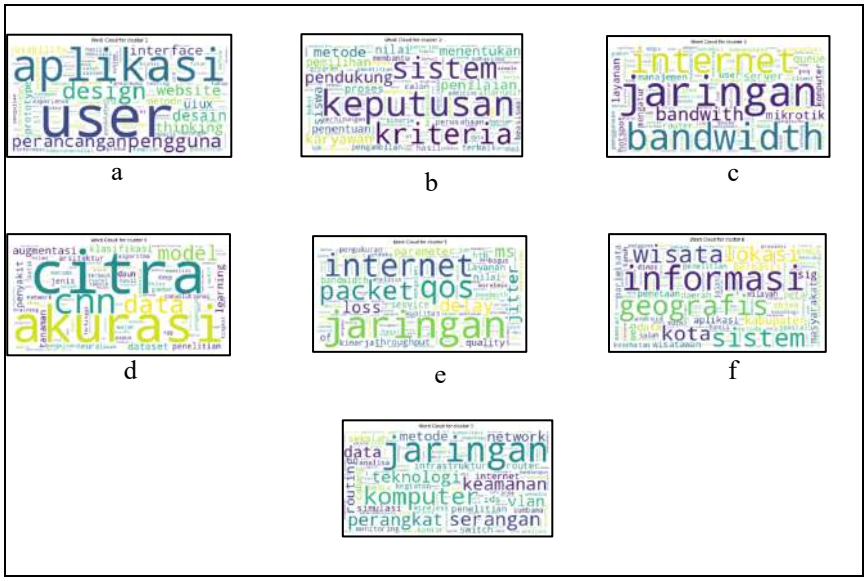
Gambar 3. Anggota *cluster K-Means*

Pada Gambar 3 merupakan anggota *cluster* yang dihasilkan, dimana pada gambar 3a merupakan *cluster* 1 yang menunjukan bahwa kata aplikasi, user, design, perancangan, yang sering muncul pada *cluster* ini. Dapat dikatakan bahwa *cluster* ini memiliki tema yang berkaitan tentan desain aplikasi. Pada gambar 3b merupakan *cluster* 2 dapat dilihat bahwa kata yang sering muncul adalah keputusan, kriteria, sistem, pendukung, metode. Sehingga dapat dikatakan bahwa *cluster* 2 ini berkaitan dengan tema sistem pendukung keputusan. Untuk Gambar 3c merupakan *cluster* 3 dapat dilihat bahwa kata yang sering muncul adalan akurasi, data, model, cnn, augmentasi, akurasi, klasifikasi. Sehingga dapat dikatakan bahwa *cluster* ini berkaitan dengan pemodelan atau klasifikasi data yang merupakan bidang dalam kecerdasan buatan. Untuk Gambar 3d merupakan *cluster* 4 menunjukan bahwa kata yang sering muncul adalah informasi, sistem, wisata, geografis. Dapat dikatakan bahwa pada *cluster* ini berkaitan dengan tema sistem informasi geografis dan untuk gambar 3e merupakan *cluster* 5 menunjukan bahwa kata yang sering muncul adalah jaringan, bandwidth, internet, komputer, miktoritik. Dapat dikatakan bahwa *cluster* ini berkaitan dengan tema jaringan komputer. Untuk hasil *clustering K-Medoids* dapat dilihat pada Gambar 4 sebegai berikut.



Gambar 4. *Cluster K-Medoids*

Pada Gambar 4 merupakan sebaran data yang dihasilkan dari *clustering* menggunakan *K-Medoids*. Hasil dari *clustering* ini memiliki sebaran yang luas, pembentukan kelompok tidak beraturan seperti pada *cluster* 1, *cluster* 2, *cluster* 3. Adapun untuk anggota *cluster* yang dihasilkan dari *clustering K-Medoids* dapat dilihat pada Gambar 5.



Gambar 5. Anggota *cluster K-Medoids*

Pada Gambar 5 merupakan anggota *cluster* yang dihasilkan dari *clustering K-Medoids*. Gambar 5a merupakan *cluster* 1 menunjukkan bahwa kata aplikasi, design, user, perancangan website yang sering muncul pada *cluster* ini. Pada *cluster* ini membahas tentang desain aplikasi. Pada Gambar 5b merupakan *cluster* 2 menunjukkan kata yang sering muncul adalah keputusan, sistem, kriteria, pendukung, metode. Pada *cluster* ini membahas tentang sistem pendukung keputusan. Pada gambar 5c merupakan *cluster* 3 dapat dilihat bahwa kata yang sering muncul adalah jaringan, bandwidth, mikrotik, internet, komputer. Pada *cluster* ini membahas mengenai jaringan komputer cluster ini membahas tema yang sama dengan *cluster* 5 dan 7, hal ini karena banyak variasi dalam penggunaan kata. Pada gambar 5d *cluster* 4 menunjukkan kata yang sering muncul adalah akurasi, citra, cnn, data, augmentasi, klasifikasi. Pada cluster ini membahas tentang klasifikasi data atau augmentasi yang berada pada bidang kecerdasan buatan. Pada Gambar 5f *cluster* 6 8 menunjukkan kata yang sering muncul adalah informasi, geografis, sistem, lokasi. Pada cluster ini membahas tema yang berkaitan dengan sistem informasi geografis.

3.3 Nilai Similaritas

Berdasarkan dari hasil *clustering* sebelumnya diperoleh *cluster* optimal berdasarkan nilai silhoutte yang dihasilkan maka algoritma *K-Means* merupakan algortma yang optimal dalam pengelompoka data, maka dari itu hasil dari pengelompokan dari *K-Means* akan dihitung nilai similaritasnya menggunakan Jaccard similarity. Adapun untuk hasilkan dapat dilihat pada Tabel 5.

Tabel 5. Jaccard Similarity

Jaccard similarity		
No	*K-MEANS*	
	Pasangan Dokumen	Nilai Similaritas Tertinggi
Cluster 1		
1	D303 & D308	0,44
2	D358 & D374	0,43
3	D353 & D378	0,40
Cluster 2		
1	D427 & D494	0,32
2	D477 & D497	0,29
3	D422 & D446	0,29
Cluster 3		
1	D202 & D204	0,32
2	D204 & D282	0,27
3	D212 & D244	0,25
Cluster 4		
1	D114 & D153	0,41
2	D432 & D480	0,34
3	D151 & D198	0,27
Cluster 5		
1	D45 & D84	0,33
2	D9 & D20	0,32
3	D17 & D93	0,23

Berdasarkan Tabel 5 merupakan nilai similaritas yang dihasilkan dari Jaccard similarity pada pengelompokan *K-Means*. Nilai similaritas yang dihasilan dari rentang 0.25-0.44 untuk menentukan nilai ambang batas dari masing-masing *cluster* tersebut menggunakan jarak euclidean. Adapun untuk nilai jarak rata-rata anggota *cluster* dapat dilihat pada Tabel 5.

Tabel 6. Jarak rata-rata anggota *cluster K-Means*

Cluster	Jarak rata-rata anggota *cluster*
1	0.94
2	0.95
3	0.96
4	0.94
5	0.93

Pada Tabel 6 merupakan nilai jarak rata-rata anggota *cluster K-Means* yang dihitung menggunakan jarak euclidean. Jarak euclidean dihitung antara setiap titik data dalam *cluster* dan centroid-nya. Jarak yang rendah menunjukkan bahwa *cluster* padat, dengan anggota *cluster* yang tidak terlalu tersebar jauh dari centroid. Berdasarkan hasil dari tabel 6 diperoleh bahwa *cluster* 5 merupakan *cluster* yang paling padat berdasarkan dari nilai rata-rata jarak euclidean sebesar 0.93 yang merupakan nilai paling rendah dari *cluster* lainnya artinya pada *cluster* ini anggota lebih dekat dengan centroidnya. Untuk nilai similaritas dari *cluster* 5 dapat dilihat pada Tabel 6.

Tabel 7. Nilai similaritas *cluster* 5

Pasangan Dokumen	Nilai Similaritas
D42 & D48	0.32
D9 & D20	0.31
D17 & D93	0.22
…	…
D52 & D81	0.049
D16 & D40	0.049
D31 & D70	0.048
Rata-rata	0.07

Pada Tabel 7 merupakan nilai similaritas yang dihasilkan dari *cluster* 5 yang merupakan *cluster* yang paling padat diantara *cluster* lainnya yang menghasilkan nilai ambang sebesar 0.07, nilai ini dapat digunakan sebagai nilai rekomendasi untuk menyatakan kedua dokumen mirip pada dokument abstrak jurnal sebanyak 500.

4. KESIMPULAN

Dari hasil penelitian yang dilakukan dapat disimpulkan sebagai berikut:

1. Berdasarkan nilai silhouette coefficient algoritma *K-Means* lebih unggul dari pada algoritma *K-Medoids* untuk pengelompokan dokumen teks dengan nilai silhouette *K-Means* sebesar 0.0304 sedangkan untuk *K-Medoids* menghasilkan nilai silhouette sebesar 0.226.

2. Diperoleh ambang nilai similaritas yang dihitung menggunakan jaccard similarity sebesar 0.07 berdasarkan pengelompokan *K-Means*.

REFERENSI

[1] A. Pujianti, "Implementasi Data Mining Menggunakan Metode K-Means Clustering Untuk Menentukan Status Kematian Bayi Di Jawa Barat," 2023.
[2] N. Khilmiyatul Ilmiyah, D. E. Ratnawati, and S. Anam, "Implementasi Gabungan Metode K-Means Learning Vector Quantization (LVQ) Untuk Klasifikasi Fungsi Senyawa Aktif Menggunakan Data SMILES," 2019. [Online]. Available: http://j-ptiik.ub.ac.id
[3] S. Fatimah and A. Usman, "Pengelompokkan Tingkat Pemahaman Guru PAUD Terhadap Pembelajaran Berbasis STEAM Menggunakan Metode X-Means Clustering," 2022.
[4] S. Bahri, D. Marisa Midyanti, and P. Korespondensi, "Penerapan Metode K-Medoids Untuk Pengelompokan Mahasiswa Berpotensi Drop Out Application Of K-Medoids Method For Dropout Potential Student Grouping," vol. 10, no. 1, pp. 165–172, 2023, doi: 10.25126/jtiik.2023106643.
[5] D. L. Karputri1 and W. Yustanti2, "Analisis Klastering Buku sebagai Evaluasi untuk Peningkatan Minat Baca Perpusatakaan SMAN 1 Grogol," 2022.
[6] E. Luthfi, A. Wahyu Wijayanto, and P. Statistika, "Analisis perbandingan metode hirearchical, k-means, dan k-medoids clustering dalam pengelompokkan indeks pembangunan manusia Indonesia," no. 4, pp. 761–773, 2021, [Online]. Available: http://journal.feb.unmul.ac.id/index.php/INOVASI
[7] P. Widiandana and I. Riadi, "Implementasi Metode Jaccard pada Analisis Investigasi Cyberbullying," masa berlaku mulai, vol. 1, no. 3, pp. 1046–1051, 2017.
[8] S. Dewi, S. Defit, and Y. Yuhandri, "Akurasi Pemetaan Kelompok Belajar Siswa Menuju Prestasi Menggunakan Metode K-Means," Jurnal Sistim Informasi dan Teknologi, pp. 28–33, Mar. 2021, doi: 10.37034/jsisfotek.v3i1.40.
[9] T. Hidayat, "Klasifikasi Data Jamaah Umroh Menggunakan Metode K-Means Clustering," Jurnal Sistim Informasi dan Teknologi, pp. 19–24, Feb. 2022, doi: 10.37034/jsisfotek.v4i1.115.
[10] J. Penerapan, T. Informasi, D. Komunikasi, G. B. Kaligis, and S. Yulianto, "It-Explore Analisa Perbandingan Algoritma K-Means, K-Medoids, Dan X-Means Untuk Pengelompokkan Kinerja Pegawai (Studi Kasus: Sekretariat DPRD Provinsi Sulawesi Utara)," 2022.
[11] F. Firzada and Y. Yuhandri, "Klasterisasi Tingkat Masa Studi Tepat Waktu Mahasiswa Menggunakan Algoritma K-Medoids," Jurnal Sistim Informasi dan Teknologi, pp. 162–168, Aug. 2021, doi: 10.37034/jsisfotek.v3i3.60.
[12] K. H. Manguri, R. N. Ramadhan, and P. R. Mohammed Amin, "Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks," Kurdistan Journal of Applied Research, pp. 54–65, May 2020, doi: 10.24017/covid.8.
[13] J. Khatib Sulaiman, M. Ikhsan, and R. R. Kurniawan, "Penerapan Text Mining pada Sistem Rekomendasi Pembimbing Skripsi Mahasiswa Menggunakan Algoritma Naïve Bayes Classifier," Indonesian Journal of Computer Science Attribution, vol. 12, no. 6, pp. 2023–4196.
[14] N. , & S. J. Feldman, The Text Mining Hand Book Advanced Approaches in Analyzing Unstructured Data. . United States of Amerika: Cambridge University Press., 2007.
[15] P. J. Rousseeue, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," 1987.