

Implementasi Metode *K-Means* Dalam Analisis Teks Aduan Pelayanan Masyarakat

Mareta Putri Wardhana¹

¹ Program Studi Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional “Veteran” Jawa Timur
Email: maretawardhana@gmail.com

Abstrak

Aduan masyarakat terhadap pelayanan publik merupakan sumber informasi penting dalam evaluasi kualitas layanan. Aduan umumnya disampaikan dalam bentuk teks tidak terstruktur, yang mana hal ini menyulitkan identifikasi pola permasalahan utama secara cepat di tengah volume data yang besar. Penelitian ini bertujuan untuk mengimplementasikan metode *K-Means* dalam analisis teks aduan pelayanan masyarakat untuk mengelompokkan aduan berdasarkan kemiripan isi. Tahapan penelitian meliputi *text preprocessing* berupa *cleaning*, *tokenizing*, *normalization*, *stopword removal*, dan *stemming*, kemudian dilakukan pembobotan kata menggunakan TF-IDF untuk mengubah data teks menjadi representasi numerik. Proses klusterisasi dilakukan menggunakan metode *K-Means* dengan penentuan jumlah klaster optimal menggunakan *Variance Ratio Criterion* (VRC). Evaluasi kualitas klaster dilakukan menggunakan *silhouette coefficient* dan *Index Davies-Bouldin* (IDB) untuk mengukur tingkat kekompakan dan pemisahan antar klaster. Hasil penelitian menunjukkan bahwa metode *K-Means* menghasilkan 3 klaster utama dengan nilai *silhouette coefficient* sebesar 0,49 dan nilai IDB sebesar 3,47. Klaster yang terbentuk mampu merepresentasikan pola aduan masyarakat berdasarkan topik keluhan yang dominan. Visualisasi *wordcloud* menunjukkan perbedaan karakteristik aduan pada setiap klaster, sedangkan analisis emosi mengungkap bahwa aduan masyarakat secara keseluruhan didominasi oleh emosi sedih. Implementasi metode *K-Means* pada analisis teks aduan masyarakat diharapkan dapat membantu proses pengelompokan data secara otomatis dan mendukung perumusan strategi peningkatan kualitas layanan publik.

Kata kunci: Aduan masyarakat, *K-Means*, Klusterisasi, *Text mining*, TF-IDF

Abstract

Public complaints regarding public services are an important source of information for evaluating service quality. Complaints are generally submitted in the form of unstructured text, making it difficult to identify major problem patterns quickly within large volumes of data. This study aims to implement the *K-Means* method in analyzing public service complaint texts to group complaints based on content similarity. The research stages include text preprocessing consisting of *cleaning*, *tokenizing*, *normalization*, *stopword removal*, and *stemming*, followed by TF-IDF weighting to transform text data into numerical representations. The clustering process was carried out using the *K-Means* method with the optimal number of clusters determined using the *Variance Ratio Criterion* (VRC). Cluster quality evaluation was conducted using the *silhouette coefficient* and *Davies-Bouldin Index* (DBI) to measure cluster compactness and separation. The results show that the *K-Means* method produced 3 main clusters with a *silhouette coefficient* value of 0.49 and a DBI value of 3.47. The resulting clusters were able to represent patterns of public complaints based on dominant complaint topics. Wordcloud visualization showed differences in complaint characteristics within each cluster, while emotion analysis revealed that public complaints were predominantly dominated by sadness. The implementation of the *K-Means* method in complaint text analysis is expected to support automatic data grouping and assist in formulating strategies to improve public service quality more effectively.

Keywords: Clustering, *K-Means*, Public complaints, *Text mining*, TF-IDF

I. PENDAHULUAN

Analisis klaster merupakan salah satu teknik multivariat yang bertujuan untuk mengelompokkan objek berdasarkan tingkat kemiripan karakteristik yang dimilikinya [1]. Teknik ini memungkinkan identifikasi pola atau struktur tersembunyi dalam data tanpa memerlukan label atau kategori awal. Dalam konteks pengolahan data teks, analisis klaster sering dikombinasikan dengan pendekatan *text mining* untuk mengekstraksi informasi dan pola tersembunyi dari kumpulan dokumen tidak terstruktur.

Pada pelayanan publik, khususnya di instansi pemerintah, aduan masyarakat menjadi salah satu sumber informasi penting dalam evaluasi kualitas layanan. Aduan masyarakat umumnya disampaikan dalam bentuk teks bebas yang tidak terstruktur sehingga menyulitkan proses identifikasi pola permasalahan

apabila dilakukan secara manual, terutama ketika jumlah data terus meningkat. Oleh karena itu, diperlukan pendekatan berbasis komputasi yang mampu mengelompokkan aduan masyarakat secara otomatis berdasarkan kemiripan isi teks.

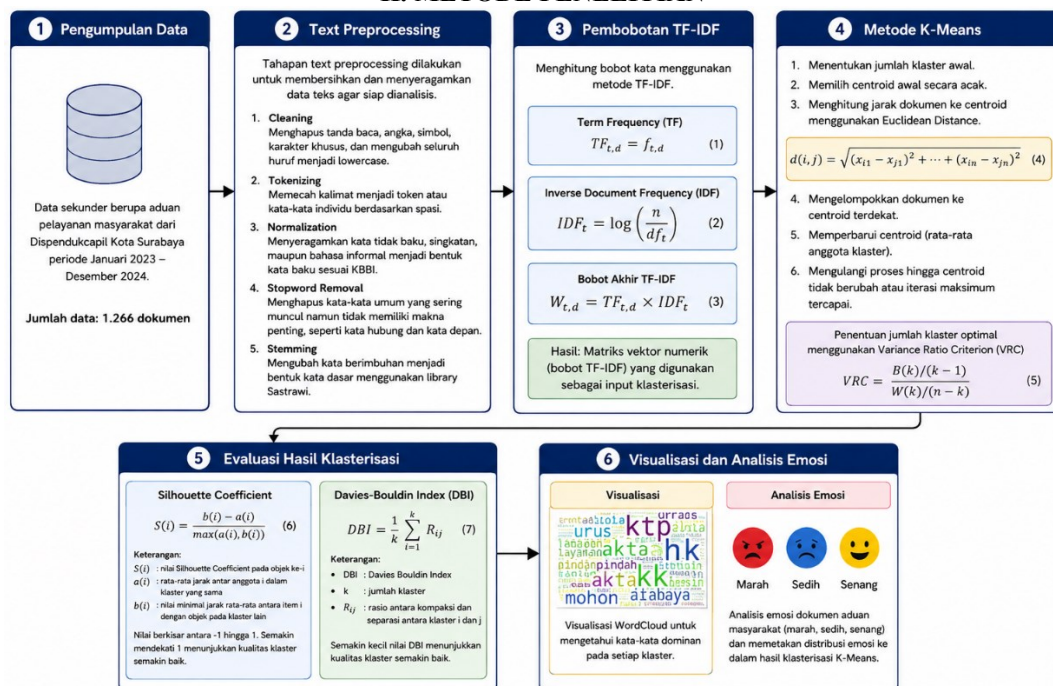
Salah satu metode klasterisasi yang banyak digunakan dalam analisis data teks adalah *K-Means*. Metode *K-Means* merupakan algoritma klasterisasi berbasis centroid yang bekerja dengan membagi data ke dalam sejumlah klaster berdasarkan kedekatan jarak terhadap pusat klaster (*centroid*). Metode ini memiliki keunggulan dalam proses komputasi yang relatif cepat dan sederhana sehingga cocok digunakan pada data berukuran besar [2]. Namun, *K-Means* memiliki kelemahan dalam menentukan jumlah klaster optimal secara manual yang dapat memengaruhi kualitas hasil klasterisasi.

Dalam analisis teks, sebelum proses klasterisasi dilakukan, data teks perlu diubah ke dalam bentuk numerik menggunakan metode pembobotan kata. Penelitian ini menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk merepresentasikan dokumen dalam bentuk vektor numerik. TF-IDF memberikan bobot pada kata berdasarkan frekuensi kemunculan kata dalam dokumen serta tingkat kepentingannya terhadap keseluruhan dokumen [3]. Representasi TF-IDF kemudian digunakan sebagai dasar dalam proses klasterisasi menggunakan metode *K-Means*.

Penelitian sebelumnya oleh [4] menunjukkan bahwa metode *K-Means* mampu menghasilkan pengelompokan data teks yang cukup baik pada data berita tidak terstruktur dengan nilai Purity sebesar 0,83. Selain itu, penelitian oleh [5] menunjukkan bahwa penggunaan *Variance Ratio Criterion* (VRC) dapat membantu menentukan jumlah klaster optimal secara lebih objektif pada metode *K-Means*. Penelitian lain oleh [6] menerapkan *text mining* dan klasterisasi pada data aduan masyarakat untuk mengidentifikasi pola utama permasalahan pelayanan publik.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mengimplementasikan metode *K-Means* dalam analisis teks aduan pelayanan masyarakat di Dinas Kependudukan dan Pencatatan Sipil (Dispendukcapil) Kota Surabaya. Penelitian ini menggunakan tahapan *text preprocessing*, pembobotan TF-IDF, penentuan jumlah klaster optimal menggunakan *Variance Ratio Criterion* (VRC), serta evaluasi menggunakan *silhouette coefficient* dan *Davies-Bouldin Index* (DBI). Selain itu, penelitian ini juga melakukan visualisasi *wordcloud* dan analisis emosi untuk mengetahui karakteristik aduan pada masing-masing klaster. Hasil penelitian diharapkan dapat membantu instansi terkait dalam mengidentifikasi pola permasalahan pelayanan masyarakat dan mendukung peningkatan kualitas layanan publik secara lebih efektif.

II. METODE PENELITIAN



Gambar 1. Alur Penelitian

Penelitian ini bertujuan untuk mengimplementasikan metode *K-Means* dalam analisis teks aduan pelayanan masyarakat di Dispendukcapil Kota Surabaya. Metode penelitian dapat dilihat pada Gambar 1 yang digunakan terdiri dari beberapa tahapan utama, yaitu pengumpulan data, *text preprocessing*, pembobotan TF-IDF, proses klasterisasi menggunakan metode *K-Means*, evaluasi hasil klasterisasi, hingga visualisasi dan analisis emosi.

2.1. Pengumpulan Data

Penelitian ini menggunakan data sekunder berupa aduan pelayanan masyarakat yang diperoleh dari database internal Dispendukcapil Kota Surabaya selama periode Januari 2023 hingga Desember 2024. Total data yang digunakan sebanyak 1.266 dokumen aduan masyarakat dalam bentuk teks tidak terstruktur. Data aduan mencakup berbagai layanan administrasi kependudukan seperti pengurusan KTP, Kartu Keluarga, akta kelahiran, pindah domisili, serta kendala layanan administrasi lainnya.

2.2. Text Preprocessing

Tahapan *text preprocessing* dilakukan untuk meningkatkan kualitas data teks sebelum dilakukan proses analisis. Data teks mentah umumnya masih mengandung karakter yang tidak relevan, kata tidak baku, serta variasi penulisan yang dapat memengaruhi hasil klasterisasi [7]. Oleh karena itu, dilakukan beberapa tahapan *preprocessing* sebagai berikut:

- Cleaning*: Tahap *cleaning* dilakukan dengan menghapus tanda baca, angka, simbol, karakter khusus, dan mengubah seluruh huruf menjadi huruf kecil (*lowercase*). Proses ini bertujuan untuk mengurangi noise pada data teks [8].
- Tokenizing*: Tahap tokenisasi dilakukan dengan memecah kalimat menjadi token atau kata-kata individu berdasarkan spasi sehingga setiap kata dapat dianalisis lebih lanjut [9].
- Normalization*: Tahap normalisasi dilakukan untuk menyeragamkan kata tidak baku, singkatan, maupun bahasa informal menjadi bentuk kata baku sesuai Kamus Besar Bahasa Indonesia (KBBI) [8].
- Stopword removal*: Tahap *stopword removal* dilakukan untuk menghapus kata-kata umum yang sering muncul namun tidak memiliki makna penting dalam proses analisis, seperti kata hubung dan kata depan [8].
- Stemming*: Tahap *stemming* dilakukan untuk mengubah kata berimbuhan menjadi bentuk kata dasar menggunakan *library* Sastrawi sehingga variasi kata yang memiliki makna sama dapat disatukan [8].

2.3. Pembobotan TF-IDF

Setelah proses *preprocessing* selesai dilakukan, tahapan berikutnya adalah pembobotan kata menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Metode TF-IDF bekerja dengan menggabungkan nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). TF digunakan untuk mengukur frekuensi kemunculan kata dalam suatu dokumen, sedangkan IDF digunakan untuk mengukur tingkat kepentingan kata terhadap keseluruhan dokumen [3]. Nilai *Term Frequency* (TF) dihitung menggunakan Persamaan [1].

$$TF_{t,d} = f_{t,d} \quad [1]$$

Keterangan:

$TF_{t,d}$: frekuensi kemunculan *term* t pada dokumen d

$f_{t,d}$: jumlah kemunculan *term* t pada dokumen d

Selanjutnya, nilai *Inverse Document Frequency* (IDF) dihitung menggunakan Persamaan [2].

$$IDF_t = \log \frac{n}{df_t} \quad [2]$$

Keterangan:

IDF_t : nilai *Inverse Document Frequency* (IDF)

n : banyaknya dokumen

df_t : banyaknya dokumen yang memiliki kata t

Bobot akhir TF-IDF dihitung menggunakan Persamaan [3].

$$W_{t,d} = TF_{t,d} \times IDF_t \quad [3]$$

Keterangan:

$W_{t,d}$: bobot TF-IDF term t pada dokumen d

Hasil pembobotan TF-IDF kemudian disusun dalam bentuk matriks vektor numerik yang digunakan sebagai input pada proses klusterisasi menggunakan metode *K-Means*.

2.4. Metode *K-Means*

Metode *K-Means* merupakan algoritma klusterisasi berbasis partisi yang digunakan untuk mengelompokkan dokumen berdasarkan tingkat kemiripan isi teks[10]. Metode ini bekerja dengan membagi data ke dalam sejumlah klaster berdasarkan kedekatan terhadap pusat klaster (centroid). Tahapan metode *K-Means* pada penelitian ini meliputi:

1. Menentukan jumlah klaster awal.
2. Memilih *centroid* awal secara acak.
3. Menghitung jarak antar dokumen dengan centroid menggunakan *euclidean distance*.
Perhitungan *euclidean distance* dilakukan menggunakan Persamaan [4].

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + \dots + (x_{in} - x_{jn})^2} \quad [4]$$

Keterangan:

$d(i, j)$: jarak antara data i dan j

x_i : koordinat x untuk data i

x_j : koordinat x untuk data j

4. Mengelompokkan dokumen ke *centroid* terdekat.
5. Memperbarui centroid berdasarkan rata-rata anggota klaster.
6. Mengulangi proses hingga centroid tidak mengalami perubahan atau iterasi maksimum tercapai.

Penentuan jumlah klaster optimal dilakukan menggunakan *Variance Ratio Criterion* (VRC). Metode ini digunakan untuk membandingkan beberapa jumlah klaster dan memilih jumlah klaster dengan nilai VRC tertinggi. Nilai VRC dihitung menggunakan Persamaan [5].

$$VRC = \frac{B(k)/(k-1)}{W(k)/(n-k)} \quad [5]$$

Keterangan:

$B(k)$: variasi antar klaster

$W(k)$: variasi dalam klaster

k : jumlah klaster

n : jumlah data

2.5. Evaluasi Hasil Klusterisasi

Evaluasi hasil klusterisasi dilakukan menggunakan *silhouette coefficient* dan *Davies-Bouldin Index* (DBI). Perhitungan *silhouette coefficient* dilakukan menggunakan Persamaan [6].

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad [6]$$

Keterangan:

$S(i)$: nilai *silhouette coefficient* pada objek ke- i

(i) : *object* yang diteliti

$a(i)$: rata-rata jarak antar anggota i dalam klaster yang sama

$b(i)$: nilai minimal jarak rata-rata antara item i dengan *object* pada klaster lain

silhouette coefficient digunakan untuk mengukur tingkat kohesi dan separasi antar klaster. Nilai *silhouette coefficient* berkisar antara -1 hingga 1 [11], di mana semakin mendekati angka 1 menunjukkan kualitas klaster yang semakin baik. Selain itu, evaluasi juga dilakukan menggunakan *Davies-Bouldin Index* (DBI) sebagaimana pada Persamaan [7].

$$DBI = \frac{1}{k} \sum_{i=1}^k R_{ij} \quad [7]$$

Keterangan:

DBI : *Davies Bouldin Index*

k : jumlah kluster

R_{ij} : rasio antara kompaksi dan separasi antara kluster i dan j

Sementara itu, *Davies-Bouldin Index* digunakan untuk mengukur tingkat compactness dan separation antar kluster. Semakin kecil nilai DBI, maka kualitas kluster yang dihasilkan semakin baik [6].

2.6. Visualisasi dan Analisis Emosi

Visualisasi hasil klusterisasi dilakukan menggunakan *wordcloud* untuk mengetahui kata-kata dominan pada setiap kluster. Visualisasi ini membantu mengidentifikasi karakteristik topik utama pada masing-masing kelompok aduan masyarakat. Selain itu, penelitian ini juga melakukan analisis emosi terhadap data aduan masyarakat. Analisis emosi dilakukan dengan mengelompokkan dokumen ke dalam kategori emosi marah, sedih, dan senang. Hasil analisis emosi kemudian dipetakan ke dalam hasil klusterisasi *K-Means* untuk mengetahui distribusi emosi pada masing-masing kluster.

III. HASIL DAN PEMBAHASAN

3.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data aduan pelayanan masyarakat di Dispendukcapil Kota Surabaya selama periode Januari 2023 hingga Desember 2024. Total data yang digunakan sebanyak 1.266 aduan dalam bentuk teks tidak terstruktur. Aduan masyarakat mencakup berbagai permasalahan administrasi kependudukan seperti pengurusan KTP, kartu keluarga, akta kelahiran, pindah domisili, serta kendala pelayanan administrasi lainnya.

Tabel 1. Data Aduan

No	Keluhan
1	Selamat Pagi, \n\saya ingin bertanya bagaimana prosesnya untuk pembuatan akta perkawinan? apakah bisa melalui website atau harus datang ke kantor kecamatan/kelurahan? mohon bantuan dan bimbingannya secara terperinci langkah- langkahnya, karena saya tidak paham. Terima kasih.
⋮	⋮
1266	Mohon bantuan untuk pengurusan pindah domisili dan ktp dari kota surabaya ke kabupaten gresik termasuk pecah kk

Tabel 1 menunjukkan bahwa data aduan masyarakat masih berbentuk teks bebas dan memiliki struktur yang tidak konsisten. Oleh karena itu, diperlukan tahapan *preprocessing* untuk membersihkan dan menyeragamkan data sebelum dilakukan proses analisis lebih lanjut.

3.2. Text Preprocessing

Tahapan *preprocessing* dilakukan untuk meningkatkan kualitas data teks agar lebih mudah dianalisis pada proses klusterisasi. Proses *preprocessing* meliputi *cleaning*, *tokenizing*, *normalization*, *stopword removal*, dan *stemming*.

Tabel 2. Cotoh Hasil *Text Preprocessing*

Tahapan	Hasil
Data Mentah	Mohon bantuan untuk pengurusan pindah domisili dan ktp dari kota surabaya ke kabupaten gresik termasuk pecah kk
<i>Cleaning</i>	mohon bantuan untuk pengurusan pindah domisili dan ktp dari kota surabaya ke kabupaten gresik termasuk pecah kk
<i>Tokenizing</i>	['mohon', 'bantuan', 'untuk', 'pengurusan', 'pindah', 'domisili', 'dan', 'ktp', 'dari', 'kota', 'surabaya', 'ke', 'kabupaten', 'gresik', 'termasuk', 'pecah', 'kk']
<i>Normalization</i>	['mohon', 'bantuan', 'untuk', 'pengurusan', 'pindah', 'domisili', 'dan', 'ktp', 'dari', 'kota', 'surabaya', 'ke', 'kabupaten', 'gresik', 'termasuk', 'pecah', 'kk']
<i>Stopword removal</i>	['mohon', 'bantuan', 'pengurusan', 'pindah', 'domisili', 'ktp', 'surabaya', 'kabupaten', 'gresik', 'pecah', 'kk']

Tahapan	Hasil
Stemming	mohon bantu urus pindah domisili ktp surabaya kabupaten gresik pecah kk

Berdasarkan Tabel 2, tahapan *preprocessing* berhasil menghilangkan kata-kata yang tidak relevan, menyeragamkan struktur teks, serta mengubah kata berimbuhan menjadi bentuk dasar. Hasil akhir *preprocessing* menghasilkan teks yang lebih ringkas dan representatif sehingga dapat meningkatkan kualitas analisis pada proses pembobotan TF-IDF dan klasterisasi menggunakan metode *K-Means*.

3.3. Pembobotan TF-IDF

Setelah *preprocessing* selesai dilakukan, tahapan berikutnya adalah pembobotan kata menggunakan metode TF-IDF. Proses ini bertujuan untuk mengubah dokumen teks menjadi representasi numerik berdasarkan tingkat kepentingan kata dalam dokumen.

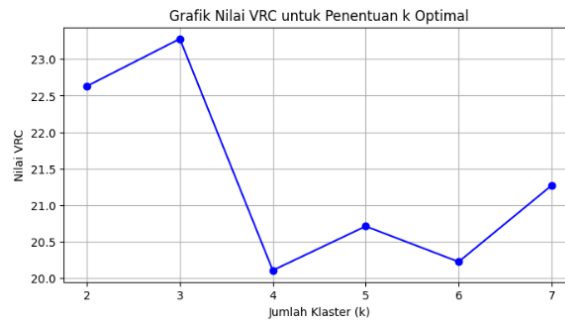
Tabel 3. Hasil Perhitungan TF-IDF

Term	TF			DF	IDF	TF-IDF		
	D1	D2	D3			D1	D2	D3
aju	0	1	0	188	2,9000	0,0000	2,9000	0,0000
akta	1	0	0	275	2,5197	2,5197	0,0000	0,0000
bantu	1	1	0	178	2,9547	2,9547	2,9547	0,0000
bimbing	1	0	0	2	7,4433	7,4433	0,0000	0,0000
buat	1	0	0	63	3,9933	3,9933	0,0000	0,0000
camat	1	0	0	83	3,7176	3,7176	0,0000	0,0000
cetak	0	0	2	126	3,3002	0,0000	0,0000	6,6004
desember	0	0	1	9	5,9393	0,0000	0,0000	5,9393
elektronik	0	1	0	31	4,7025	0,0000	4,7025	0,0000
kantor	1	0	1	30	4,7353	4,7353	0,0000	4,7353
kasih	1	0	0	187	2,9054	2,9054	0,0000	0,0000
kawin	1	0	0	48	4,2653	4,2653	0,0000	0,0000
ktp	0	2	3	344	2,2958	0,0000	4,5917	6,8875
langkah	2	0	0	11	5,7386	11,4772	0,0000	0,0000
lurah	1	0	1	244	2,6393	2,6393	0,0000	2,6393
mas	0	0	1	4	6,7502	0,0000	0,0000	6,7502
mohon	1	1	0	469	1,9859	1,9859	1,9859	0,0000
pagi	1	0	0	50	4,2245	4,2245	0,0000	0,0000
paham	1	0	0	6	6,3447	6,3447	0,0000	0,0000
perinci	1	0	0	1	8,1365	8,1365	0,0000	0,0000
proses	1	0	0	119	3,3574	3,3574	0,0000	0,0000
rusak	0	0	1	32	4,6707	0,0000	0,0000	4,6707
selamat	1	0	0	113	3,4091	3,4091	0,0000	0,0000
terima	1	0	0	197	2,8533	2,8533	0,0000	0,0000
website	1	0	0	17	5,3033	5,3033	0,0000	0,0000

Tabel 3 menunjukkan hasil representasi dokumen dalam bentuk matriks TF-IDF. Setiap dokumen direpresentasikan sebagai vektor numerik berdasarkan bobot kata yang muncul pada dokumen tersebut. Kata yang sering muncul pada dokumen tertentu namun jarang muncul pada dokumen lain akan memiliki nilai bobot lebih tinggi. Representasi numerik ini selanjutnya digunakan sebagai input pada proses klasterisasi menggunakan metode *K-Means*.

3.4. Hasil Klasterisasi *K-Means* dan Evaluasi Hasil Klasterisasi

Setelah dilakukan proses pembobotan TF-IDF, tahapan berikutnya adalah proses klasterisasi menggunakan metode *K-Means*. Penentuan jumlah klaster optimal dilakukan menggunakan *Variance Ratio Criterion* (VRC). Pengujian dilakukan dengan membandingkan beberapa jumlah klaster untuk memperoleh nilai VRC terbaik.



Gambar 2. Grafik K Optimal K-Means

Berdasarkan Gambar 2, nilai VRC tertinggi diperoleh pada jumlah kluster sebanyak tiga kluster. Nilai VRC mengalami peningkatan hingga jumlah kluster tertentu, kemudian mengalami penurunan pada jumlah kluster berikutnya. Nilai VRC tertinggi menunjukkan bahwa variasi antar kluster lebih besar dibandingkan variasi dalam kluster, sehingga menghasilkan pemisahan kluster yang lebih optimal. Oleh karena itu, jumlah kluster sebanyak tiga dipilih sebagai jumlah kluster terbaik pada penelitian ini.

Setelah diperoleh jumlah kluster optimal sebanyak tiga kluster berdasarkan pengujian *Variance Ratio Criterion* (VRC), langkah selanjutnya adalah melakukan proses klusterisasi menggunakan metode *K-Means* untuk menentukan anggota kluster pada setiap dokumen aduan masyarakat.

Tabel 4. Output Pemodelan K-Means

Dokumen	Jarak Ke-			Kluster
	C0	C1	C2	
D1	26,16	105,71	21,34	2
D2	18,20	103,97	7,54	2
D3	21,64	103,69	14,81	2
D4	19,54	104,33	10,27	2
D5	30,59	106,88	28,56	2
D6	22,30	105,27	18,60	2
D7	63,93	122,23	65,55	0
D8	48,30	110,78	47,51	2
D9	24,71	106,01	21,96	2
D10	19,73	104,39	12,13	2

Berdasarkan Tabel 4, setiap dokumen dihitung jaraknya terhadap seluruh centroid menggunakan *euclidean distance*. Proses penentuan kluster dilakukan berdasarkan nilai jarak terkecil (*argmin*). Sebagai contoh, Dokumen D1 memiliki jarak terkecil terhadap centroid C2 sebesar 21,34 dibandingkan *centroid* lainnya, sehingga dokumen tersebut secara otomatis masuk ke dalam Kluster 2. Proses ini menunjukkan bahwa pengelompokan dokumen pada metode *K-Means* dilakukan berdasarkan kedekatan karakteristik antar dokumen terhadap centroid kluster.

Hasil klusterisasi kemudian dievaluasi menggunakan *silhouette coefficient* dan *Davies-Bouldin Index* (DBI).

Tabel 5. Evaluasi Model

Metode	Hasil Kluster	<i>Silhouette Coefficient</i>	<i>Index Davies-Bouldin</i>
K-Means	3 kluster	0,49	3,47

Berdasarkan hasil evaluasi Tabel 5, metode *K-Means* menghasilkan nilai *Silhouette Coefficient* sebesar 0,49. Nilai tersebut menunjukkan bahwa hasil klusterisasi memiliki tingkat kohesi dan separasi yang cukup baik. Nilai *silhouette coefficient* yang mendekati angka 1 menunjukkan bahwa data dalam suatu kluster memiliki tingkat kemiripan yang cukup tinggi dibandingkan dengan data pada kluster lainnya.

Selain menggunakan *silhouette coefficient*, evaluasi juga dilakukan menggunakan *Davies-Bouldin Index* (DBI). Hasil evaluasi menghasilkan nilai DBI sebesar 3,47. Nilai DBI yang relatif kecil menunjukkan bahwa kluster yang terbentuk cukup kompak dan memiliki pemisahan antar kluster yang cukup baik. Secara

keseluruhan, hasil evaluasi menunjukkan bahwa metode *K-Means* mampu melakukan pengelompokan aduan masyarakat berdasarkan kemiripan isi teks secara cukup optimal.

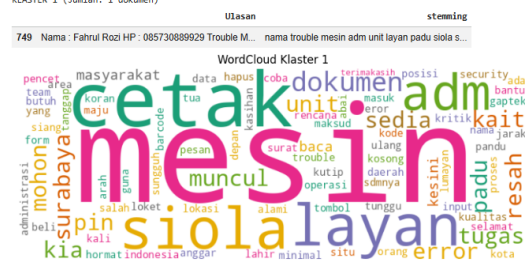
3.5. Visualisasi dan Analisi Emosi

Visualisasi dilakukan menggunakan *wordcloud* untuk mengetahui kata-kata dominan pada masing-masing klaster hasil *K-Means*. *Wordcloud* membantu menggambarkan karakteristik utama dari setiap kelompok aduan masyarakat berdasarkan frekuensi kemunculan kata.



Gambar 3. Wordcloud Klaster 0

Pada Gambar 3 merupakan visualisasi Klaster 0 yang berisikan total 103 dokumen. Klaster ini didominasi oleh kata-kata seperti “lurah”, “kk”, “alamat”, “surabaya”, dan “mohon”. Dominasi kata tersebut menunjukkan bahwa kelompok ini menampung aduan terkait layanan administrasi kependudukan di tingkat kewilayahan. Visualisasi ini menggambarkan keluhan masyarakat mengenai proses pindah datang, verifikasi tempat tinggal atau alamat rumah, serta pelayanan administrasi yang melibatkan persetujuan pejabat atau petugas kelurahan.



Gambar 4. Wordcloud Klaster 1

Pada Gambar 4 merupakan visualisasi Klaster 1 yang berisikan total 1 dokumen. Klaster ini didominasi oleh kata-kata seperti “cetak”, “mesin”, “siola”, dan “layan”. Dominasi kata tersebut menunjukkan bahwa kelompok ini menampung aduan terkait layanan administrasi kependudukan melalui mesin ADM di Unit Pelayanan Terpadu Siola Surabaya. Visualisasi ini menggambarkan keluhan masyarakat yang ingin mencetak dokumen kependudukan namun mengalami berbagai kendala, seperti tidak adanya petugas pendamping, fitur layanan pada mesin yang tidak lengkap, serta terjadinya gangguan atau *error* saat proses pencetakan, termasuk kendala PIN, dan proses cetak yang berulang kali gagal.

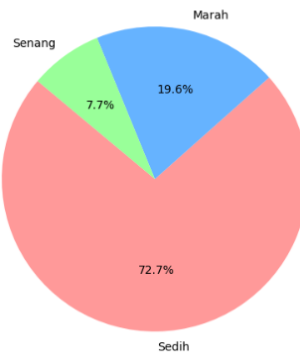


Gambar 5. Wordcloud Klaster 2

Pada Gambar 5 merupakan visualisasi Klaster 2 yang berisikan total 1.153 dokumen. Klaster ini didominasi oleh kata-kata seperti “ktp”, “kk”, “akta”, “lahir”, dan “mohon”. Dominasi kata tersebut menunjukkan bahwa kelompok ini menampung aduan terkait layanan administrasi kependudukan secara umum. Visualisasi ini menggambarkan keluhan dan permohonan masyarakat mengenai pengurusan dokumen identitas utama yang meliputi KTP, Kartu Keluarga, dan Akta Kelahiran.

Selain visualisasi *wordcloud*, penelitian ini juga melakukan analisis emosi terhadap seluruh data aduan masyarakat untuk mengetahui kecenderungan emosi yang muncul pada data pelayanan publik.

Distribusi Emosi Global (Seluruh Data)



Gambar 6. Distribusi Emosi aduan (Keseluruhan)

Berdasarkan hasil visualisasi distribusi emosi global terhadap seluruh data aduan pada Gambar 6, diperoleh bahwa mayoritas dokumen didominasi oleh emosi sedih dengan persentase sebesar 72,7%. Sementara itu, emosi marah tercatat sebesar 19,6%, dan emosi senang sebesar 7,7%. Komposisi tersebut menunjukkan bahwa sebagian besar aduan masyarakat mengandung nuansa ketidakpuasan, kekecewaan, kebingungan, atau permasalahan pelayanan yang belum terselesaikan. Dominasi emosi sedih mengindikasikan bahwa karakteristik utama aduan masyarakat cenderung berupa keluhan terhadap proses administrasi, keterlambatan pelayanan, maupun kendala teknis dalam pelayanan publik. Sementara itu, proporsi emosi marah sebesar 19,6% menunjukkan adanya masyarakat yang menyampaikan keluhan dengan ekspresi yang lebih tegas atau intens akibat pengalaman pelayanan yang dianggap kurang memuaskan. Adapun persentase emosi senang yang relatif rendah menunjukkan bahwa tingkat apresiasi atau kepuasan masyarakat terhadap pelayanan masih lebih kecil dibandingkan jumlah keluhan yang masuk.

Setelah dilakukan pelabelan emosi pada seluruh dataset, analisis dilanjutkan dengan memetakan distribusi emosi ke dalam hasil klusterisasi metode *K-Means*.

Tabel 6. Distribusi Emosi Aduan

Klaster	Emosi		
	Marah	Sedih	Senang
0	49	47	7
1	0	1	0
2	197	866	90

Berdasarkan Tabel 6, Klaster 0 menghasilkan 47 dokumen sedih, 49 dokumen marah, dan 7 dokumen senang. Klaster ini menunjukkan adanya keseimbangan antara emosi sedih dan marah, sehingga menggambarkan kelompok aduan dengan tingkat ketidakpuasan yang cukup tinggi terhadap pelayanan administrasi kependudukan.

Selanjutnya, Klaster 1 hanya menghasilkan 1 dokumen dengan emosi sedih. Jumlah dokumen yang sangat sedikit menunjukkan bahwa klaster ini memiliki anggota yang sangat terbatas dan kurang representatif terhadap keseluruhan data aduan masyarakat.

Sementara itu, Klaster 2 didominasi oleh emosi sedih sebanyak 866 dokumen, kemudian emosi marah sebanyak 197 dokumen, dan emosi senang sebanyak 90 dokumen. Dominasi emosi sedih pada Klaster 2 menunjukkan bahwa sebagian besar masyarakat mengalami kesulitan, kebingungan, atau kekecewaan dalam proses pelayanan administrasi kependudukan.

Namun demikian, jika dikaitkan dengan hasil evaluasi metode *K-Means* yang menghasilkan nilai *silhouette coefficient* sebesar 0,49 dan nilai *Davies-Bouldin Index* (DBI) sebesar 3,47, maka pembagian

klaster yang terlihat lebih “rapi” sebenarnya cenderung bersifat dipaksakan. Metode *K-Means* berusaha membagi data yang pada dasarnya cukup homogen ke dalam beberapa partisi berbeda, sehingga menyebabkan batas antar klaster menjadi kurang jelas dan memiliki tingkat overlap yang tinggi.

IV. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, tujuan penelitian untuk mengimplementasikan metode *K-Means* dalam analisis teks aduan pelayanan masyarakat di Dispendukcapil Kota Surabaya telah berhasil dicapai. Proses *text preprocessing* yang meliputi *cleaning*, *tokenizing*, *normalization*, *stopword removal*, dan *stemming* berhasil meningkatkan kualitas data teks sehingga dapat digunakan pada proses pembobotan TF-IDF dan klasterisasi.

Hasil penelitian menunjukkan bahwa metode *K-Means* mampu mengelompokkan data aduan masyarakat menjadi tiga klaster berdasarkan kemiripan isi teks. Berdasarkan visualisasi *wordcloud*, setiap klaster memiliki karakteristik topik yang berbeda. Selain itu, hasil evaluasi klasterisasi menghasilkan nilai *silhouette coefficient* sebesar 0,49 dan nilai *Davies-Bouldin Index* (DBI) sebesar 3,47. Nilai tersebut menunjukkan bahwa metode *K-Means* masih memiliki keterbatasan dalam memisahkan data yang cenderung homogen sehingga tingkat overlap antar klaster masih cukup tinggi.

Hasil analisis emosi menunjukkan bahwa mayoritas aduan masyarakat didominasi oleh emosi sedih sebesar 72,7%, yang menggambarkan adanya ketidakpuasan, kebingungan, dan berbagai kendala pelayanan yang dialami masyarakat dalam proses administrasi kependudukan.

Penelitian ini diharapkan dapat membantu instansi terkait dalam memahami pola aduan masyarakat serta mendukung evaluasi dan peningkatan kualitas pelayanan publik. Untuk penelitian selanjutnya, metode klasterisasi lain seperti *DBSCAN*, *Ward's Method*, atau *Fuzzy C-Means* dapat digunakan sebagai pembandingan untuk memperoleh hasil pengelompokan yang lebih optimal. Selain itu, penelitian selanjutnya dapat mengembangkan analisis emosi dengan kategori emosi yang lebih beragam serta menggunakan dataset yang lebih besar agar hasil analisis menjadi lebih representatif.

REFERENSI

- [1] W. Wardono, S. Sunarmi, dan M. R. Wirawan, “PENGELOMPOKKAN KABUPATEN/KOTA DI PROVINSI JAWA TENGAH BERDASARKAN INDIKATOR KESEJAHTERAAN DENGAN METODE KMEANS CLUSTER,” *EDUSAINTEK*, vol. 3, no. 0, 2019.
- [2] M. Afzali and S. Kumar, “Text Document Clustering: Issues and Challenges,” 2019 Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput., pp. 263–268, 2019.
- [3] P. A. Riyantoko, T. M. Fahrudin, D. A. Paceyta, Trimono, and T. D. Timur, “Analisis Sentimen Sederhana Menggunakan Algoritma LSTM dan BERT untuk Klasifikasi Data Spam dan Non-Spam,” *PROSIDING SEMINAR NASIONAL SAINS DATA*, vol. 2, pp. 103–111, 2022.
- [4] Nyoman Gede Yudiarta, Made Sudarma, and Wayan Gede Ariastina, “Penerapan Metode Clustering *Text mining* Untuk Pengelompokan Berita Pada Unstructured Textual Data,” *Majalah Ilmiah Teknologi Elektro*, vol. 17, no. 3, pp. 339–339, Dec. 2018.
- [5] C. Putra, K. Wahyu Megananda Kesuma Putra, K. Alit Arsani, A. Prabawa, and I. W. Kesumadana, “Gambaran Penyebab Infertilitas Pasangan Usia Subur di Rumah Sakit Kasih Ibu Denpasar Tahun 2021 - 2022,” *jiwp*, vol. 10, no. 12, pp. 371–385, Jun. 2024.
- [6] Muthahharah and A. Juhari, “A Cluster Analysis with Complete Linkage and *Ward's Method* for Health Service Data in Makassar City,” *Jurnal Varian*, vol. 4, no. 2, pp. 109 - 116, Apr. 2021.
- [7] Devi Putri Isnarwaty and Irhamah Irhamah, “Text Clustering pada Akun TWITTER Layanan Ekspedisi JNE, J&T, dan Pos Indonesia Menggunakan Metode Density-Based Spatial Clustering of Applications with Noise (DBSCAN) dan *K-Means*,” *Jurnal Sains dan Seni ITS*, vol. 8, no. 2, Feb. 2020, doi:<https://doi.org/10.12962/j23373520.v8i2.49094>.
- [8] S. J. Angelina, Arif, and Hafiz Muhandi, “Analisis Pengaruh Penerapan *Stopword removal* Pada Performa Klasifikasi Sentimen Tweet Bahasa Indonesia,” *Jurnal Aplikasi dan Riset Informatika*, vol. 1, no. 2, pp. 165–173, 2023, doi: <https://doi.org/10.26418/jari.v2i1.69680>.
- [9] T. M. Fahrudin, A. R. F. Sari, A. Lisanthoni, and A. A. D. Lestari, “ANALISIS SPEECH-TO-TEXT PADA VIDEO MENGANDUNG KATA KASAR DAN UJARAN KEBENCIAN DALAM CERAMAH AGAMA ISLAM MENGGUNAKAN INTERPRETASI AUDIENS DAN

-
- VISUALISASI WORD CLOUD,” *SKANIKA*, vol. 5, no. 2, pp. 190–202, Jul. 2022, doi: <https://doi.org/10.36080/skanika.v5i2.2942>.
- [10] Dita Afida, E. D. Udayanti, and Etika Kartikadarma, “Aplikasi *Text mining* untuk Klasterisasi Aduan Masyarakat Kota Semarang Menggunakan Algoritma *K-Means*,” *Jurnal Transformatika*, vol. 18, no. 2, pp. 215–224, 2021.
- [11] D. A. Prasetya, A. P. Sari, M. Idhom, and A. Lisanthoni, “Optimizing Clustering Analysis to Identify High-Potential Markets for Indonesian Tuber Exports,” *Indones. J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 1, pp. 113–122, 2025, doi: 10.35882/skzqbd57.