IMPLEMENTASI METODE WARD DALAM ANALISIS TEKS ADUAN PELAYANAN MASYARAKAT (STUDI KASUS: DISPENDUKCAPIL KOTA SURABAYA)

Mareta Putri Wardhana, Kartika Maulida Hindrayani, Amri Muhaimin
Sains Data, Universitas Pembangunan Nasional "Veteran" Jawa Timur
Jl. Raya Rungkut Madya, Gunung Anyar, Surabaya, Jawa Timur, Indonesia
kartika.maulida.ds@upnjatim.ac.id

ABSTRAK

Aduan masyarakat terhadap pelayanan publik merupakan sumber informasi penting dalam evaluasi kualitas layanan, namun umumnya disampaikan dalam bentuk teks tidak terstruktur yang sulit dianalisis secara manual. Ketiadaan pengelompokan otomatis menyulitkan identifikasi pola permasalahan utama secara cepat di tengah volume data yang besar. Penelitian ini bertujuan mengelompokkan aduan masyarakat di Dispendukcapil Kota Surabaya berdasarkan kemiripan teks menggunakan metode *Agglomerative Hierarchical Clustering* dengan pendekatan Ward. Metode Ward dipilih karena kemampuannya meminimalkan variansi dalam klaster untuk menghasilkan kelompok yang homogen. Tahapan penelitian meliputi *text preprocessing*, pembobotan kata menggunakan *Term Frequency–Inverse Document Frequency* (TF-IDF), serta klasterisasi menggunakan jarak Euclidean. Berdasarkan analisis dendrogram, penelitian ini menghasilkan dua klaster optimal dengan nilai *Silhouette Coefficient* sebesar 0,7603 dan *Davies Bouldin Index* 1,0283. Nilai tersebut mengindikasikan struktur klaster yang baik dengan tingkat kohesi dan separasi yang tinggi. Visualisasi *wordcloud* menunjukkan perbedaan karakteristik aduan pada setiap klaster, sedangkan analisis emosi mengungkap bahwa aduan masyarakat secara keseluruhan didominasi oleh emosi sedih. Hasil penelitian ini diharapkan dapat membantu instansi terkait dalam mengidentifikasi pola permasalahan pelayanan dan mendukung perumusan strategi peningkatan kualitas layanan publik secara lebih efektif.

Kata kunci : *Text Mining, TF-IDF, Ward's Method, Klasterisasi Teks, Aduan Masyarakat, Analisis Emosi*

1. 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. Pada data teks, analisis klaster umumnya dikombinasikan dengan pendekatan *Text Mining*, yaitu proses analisis teks untuk mengekstraksi informasi dan pola tersembunyi dari kumpulan dokumen dalam jumlah besar [2]. *Text Mining* melibatkan penerapan teknik *Natural Language Processing* (NLP), statistik, dan *machine learning* untuk mengubah data teks menjadi representasi numerik yang dapat dianalisis secara komputasional. Pendekatan ini banyak digunakan dalam berbagai bidang, seperti analisis sentimen, pengelompokan dokumen, dan analisis opini publik.

Dalam pelayanan publik, khususnya pada instansi pemerintah, data aduan masyarakat merupakan salah satu sumber informasi penting untuk mengevaluasi kualitas layanan. Namun, di Dispendukcapil Kota Surabaya, permasalahan utama yang dihadapi adalah aduan tersebut umumnya disampaikan dalam bentuk teks bebas yang tidak terstruktur. Hal ini menyulitkan proses analisis apabila dilakukan secara manual, terutama dalam mengidentifikasi tren masalah yang serupa di tengah volume data yang besar. Tanpa adanya sistem pengelompokan otomatis, pola permasalahan utama sulit diidentifikasi secara sistematis, sehingga tindak lanjut terhadap aduan menjadi kurang optimal.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mengimplementasikan metode klasterisasi teks pada data aduan masyarakat di Dispendukcapil Kota Surabaya. Melalui proses ini, diharapkan diperoleh gambaran mengenai pola permasalahan utama yang sering dihadapi masyarakat secara objektif dan sistematis, yang nantinya dapat menjadi dasar bagi pengembangan strategi peningkatan kualitas pelayanan publik.

Untuk mencapai tujuan tersebut, metode yang digunakan dalam penyelesaian masalah adalah menggunakan metode *Agglomerative Hierarchical Clustering* (AHC) dengan pendekatan metode Ward. Metode Ward atau *Ward's Method* merupakan metode klasterisasi hierarkis dengan pendekatan *bottom-up*, di mana setiap objek awalnya dianggap sebagai satu klaster tersendiri, kemudian digabungkan secara bertahap berdasarkan kriteria tertentu [3]. Keunggulan utama metode Ward terletak pada kemampuannya dalam meminimalkan peningkatan variansi total dalam setiap klaster, sehingga menghasilkan kelompok yang lebih homogen dan kompak [4]. Karakteristik ini menjadikan Metode Ward sesuai untuk analisis data teks yang memiliki variasi kosakata dan konteks yang tinggi.

Tahapan teknis penyelesaian masalah ini diawali dengan mengubah data teks ke dalam bentuk numerik. Penelitian ini menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF) sebagai teknik

pembobotan kata. TF-IDF memberikan bobot pada setiap term berdasarkan frekuensi kemunculannya dalam dokumen dan tingkat kepentingannya terhadap keseluruhan dokumen, sehingga mampu merepresentasikan karakteristik dokumen secara kuantitatif [12]. Representasi TF-IDF ini selanjutnya digunakan sebagai dasar perhitungan jarak antar dokumen dalam proses klasterisasi menggunakan metode Ward dengan metrik jarak *Euclidean*. Kemudian, untuk memastikan validitas struktur klaster yang terbentuk, dilakukan evaluasi menggunakan *Silhouette Coefficient* dan *Davies Bouldin Index* (DBI) guna menentukan kualitas separasi dan kepadatan klaster.

Dengan menerapkan Metode Ward dalam analisis teks aduan, penelitian ini diharapkan dapat memberikan kontribusi nyata dalam penerapan *text mining* untuk pengelolaan aduan masyarakat. Hasil klasterisasi yang lebih homogen dan mudah diinterpretasikan diharapkan dapat membantu instansi terkait dalam mengidentifikasi tema permasalahan utama secara akurat, serta menjadi landasan strategis dalam pengambilan keputusan perbaikan layanan.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Penelitian terkait penerapan *Text Mining* dan analisis klaster pada data teks telah banyak dilakukan dalam berbagai konteks, termasuk pengelompokan aduan masyarakat dan data layanan publik. Penelitian-penelitian tersebut menunjukkan bahwa pendekatan klasterisasi mampu mengungkap pola dan tema tersembunyi dalam data teks tidak terstruktur.

Selanjutnya, penelitian oleh [5] menerapkan analisis klaster hierarkis menggunakan *Ward's Method* dan *Complete Linkage* pada data layanan kesehatan di Kota Makassar. Hasil penelitian menunjukkan bahwa metode Ward menghasilkan klaster yang lebih stabil dan homogen dibandingkan metode *Complete Linkage*. Nilai *Davies Bouldin Index* (DBI) yang rendah menunjukkan bahwa klaster yang dihasilkan memiliki pemisahan yang baik dan tingkat kekompakan yang tinggi. Temuan ini menunjukkan bahwa metode Ward efektif dalam meminimalkan variansi dalam klaster sehingga sesuai untuk data dengan karakteristik kompleks.

Penelitian lain oleh [6] menerapkan analisis klaster hierarkis dengan pendekatan *Ward's Method* untuk membangun taksonomi ancaman insider berbasis data teks. Penelitian ini menghasilkan lima klaster utama dengan nilai koefisien aglomerasi yang tinggi, yang menunjukkan validitas klaster yang baik. Hasil penelitian ini memperkuat bahwa metode Ward mampu menghasilkan struktur klaster yang interpretatif dan konsisten pada data teks.

Selain itu, penelitian oleh [7] menerapkan *Text Mining* untuk pengelompokan dokumen berita pada data teks tidak terstruktur. Hasil penelitian menunjukkan bahwa proses *preprocessing* teks dan pembobotan kata berperan penting dalam meningkatkan kualitas klaster yang dihasilkan. Penelitian ini menekankan bahwa keberhasilan analisis klaster pada data teks sangat dipengaruhi oleh tahapan awal pengolahan data.

2.2. *Text Mining*

Text Mining merupakan istilah umum yang digunakan untuk menganalisis dan mengolah data teks yang bersifat semi-terstruktur maupun tidak terstruktur [8]. Tujuan utama *Text Mining* adalah mengekstraksi informasi serta pola tersembunyi dari data teks dalam jumlah besar melalui penerapan teknik *Natural Language Processing* (NLP), statistik, dan *machine learning*. Metode ini banyak dimanfaatkan dalam berbagai aplikasi, seperti analisis sentimen untuk mengidentifikasi emosi atau opini dalam teks [9], pengelompokan dokumen berdasarkan kesamaan topik [10], serta ekstraksi informasi guna mendukung pengambilan keputusan strategis [11]. Mengingat data teks umumnya tidak terstruktur, diperlukan tahapan text preprocessing sebelum analisis lebih lanjut dilakukan. Dalam penelitian ini, *Text Mining* digunakan untuk mengekstraksi informasi dari aduan pelayanan masyarakat berbentuk teks sehingga dapat dianalisis dan dikelompokkan menggunakan metode klasterisasi.

2.3. Pembobotan TF-IDF

Pembobotan kata *Term Frequency–Inverse Document Frequency* (TF-IDF), digunakan untuk mengonversi data teks menjadi representasi numerik. Metode ini menggabungkan dua komponen utama, yaitu *Term Frequency* (TF) yang menunjukkan frekuensi kemunculan suatu kata dalam dokumen tertentu, dan *Inverse Document Frequency* (IDF). menunjukkan tingkat keumuman kata tersebut berdasarkan kemunculannya pada seluruh dokumen yang dianalisis [12].

Perhitungan nilai bobot akhir TF-IDF dihitung sebagaimana ditunjukkan pada persamaan berikut:

$$TF - IDF_{t,d} = TF_{t,d} \times IDF_t \quad (1)$$

Keterangan:

$TF - IDF_{t,d}$: nilai bobot akhir
$TF_{t,d}$: banyaknya kata t pada dokumen d
IDF_t : nilai *Inverse Document Frequency*

Hasil perhitungan TF-IDF selanjutnya disusun dalam bentuk matriks vektor numerik, yang kemudian digunakan sebagai input pada proses klasterisasi teks.

2.4. Metode Ward

Metode Ward atau *Ward's Method* merupakan pendekatan *Agglomerative Hierarchical Clustering* (AHC) yang bertujuan minimalkan variasi dalam klaster (*within-cluster variance*) berdasarkan peningkatan *Sum of Squared Errors* (SSE) terkecil [13]. Dengan pendekatan ini, metode Ward cenderung menghasilkan klaster yang lebih homogen dan kompak.

Secara umum, Ward menggunakan fungsi objektif yang optimal sebagai dasar untuk memilih pasangan klaster [14]. Pada tahapan algoritma Ward dilakukan dengan memberikan inisialisasi klaster Setiap objek data atau dokumen dianggap sebagai satu klaster tersendiri (*singleton cluster*). Jika terdapat n dokumen, maka terbentuk n klaster awal. Selanjutnya dilakukan perhitungan jarak antar objek menggunakan *euclidean distance*, yang dirumuskan pada persamaan berikut:

$$d(i,j) = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \cdots + \left(x_{in} - x_{jn}\right)^2} \quad (2)$$

Keterangan:

$d(i,j)$: jarak antara data i dan j
x_i : koordinat x untuk data i
x_j : koordinat x untuk data j

Selanjutnya dilakukan perhitungan peningkatan SSE antar klaster Untuk setiap pasangan klaster, dihitung peningkatan nilai *Sum of Squared Errors* (SSE) apabila kedua klaster digabungkan, dengan persamaan:

$$SSE = \sum_{j=1}^{p}\left(\sum_{i=1}^{n} x_{ij}^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}\right)^2\right) \quad (3)$$

Keterangan:

SSE : *sum of square error* / jumlah kuadrat kesalahan
x_{ij} : nilai untuk objek i dalam klaster j
p : jumlah variabel yang diukur
n : jumlah objek dalam klaster yang terbentuk

Penggabungan klaster dilakukan secara iteratif pada dua klaster yang menghasilkan peningkatan *Sum of Squared Errors* (SSE) paling kecil, diikuti pembaruan matriks jarak hingga seluruh objek tergabung menjadi satu. Proses tersebut kemudian divisualisasikan dalam bentuk dendrogram, guna menggambarkan tahapan penggabungan klaster serta membantu dalam penentuan jumlah klaster yang optimal.

2.5. *Silhoutte Coefficient*

Silhouette Coefficient adalah suatu metrik yang menghitung sejauh mana data dalam sebuah kelompok memiliki kesamaan, dan dihitung secara individual untuk setiap objek dalam kelompok. Nilai *Silhouette Coefficient* berkisar antara -1 hingga 1 [15], yang mana semakin nilainya mendekati angka 1, maka semakin tinggi kualitas pengelompokkan pada satu kelompok tersebut. Untuk menghitung *Silhouette Coefficient* maka diperlukan persamaan sebagai berikut:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i)),(b(i))} \quad (4)$$

Keterangan:

$S(i)$: nilai *Silhouette Coeffcient* pada objek ke-i
(i) : *object* yang diteliti
$a(i)$: rata-rata jarak antar angota i dalam klaster yang sama
$b(i)$: nilai minimal jarak rata-rata antara item i dengan *object* pada klaster lain

2.6. *Davies Bouldin Index*

Davies Bouldin Index (DBI) *validity test* adalah metode untuk mengukur seberapa baik hasil clustering dengan menilai dissimilarity antara objek dalam klaster. DBI bertujuan untuk memaksimalkan jarak antara klaster satu sama lain (*separation value*) dan memperkirakan jarak antara titik-titik dalam satu klaster (*compactness value)* [5]. Untuk menghitung DBI maka diperlukan rumus sebagai berikut:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} R_{ij} \quad (5)$$

Keterangan:

DBI : *Davies Bouldin Index*
k : jumlah klaster
R_{ij} : rasio antara kompaksi dan separasi antara klaster i dan j

3. METODE PENELITIAN

3.1. Pengumpulan Data

Penelitian ini menggunakan data sekunder mengenai aduan yang tercatat dalam sistem atau database internal Dinas Kependudukan dan Pencatatan Sipil (Dispendukcapil) Kota Surabaya selama periode Januari 2023 hingga Desember 2024 (2 tahun) dengan jumlah data berjumlah 1.266 aduan.

3.2. *Preprocessing* Data

Kemudian pada proses ini, data yang telah kita kumpulkan akan masuk ke tahap *preprocessing* data. Tahapan ini diperlukan karena data teks mentah umumnya mengandung noise, inkonsistensi, serta elemen yang tidak relevan untuk proses analisis [16]. Melalui text preprocessing, format teks disesuaikan agar dapat diproses lebih lanjut pada tahap analisis berikutnya [17]. Secara umum, *preprocessing* data yang memiliki beberapa tahapan sebagai berikut:

a. *Cleaning*
   Merupakan tahap awal preprocessing yang bertujuan untuk membersihkan teks dari karakter yang tidak diperlukan, sehingga teks menjadi lebih mudah dianalisis [18]. Pada tahapan ini akan dilakukan penghapusan untuk nama orang, kemudian semua huruf akan diubah menjadi huruf *lowercase*, dan membuang karakter-karakter tidak penting seperti tanda baca, misalnya titik (.), koma (,), tanda tanya (?), tanda hubung (-), kemudian angka, dan juga simbol, seperti @, #. Kemudian, sisa karakter n di awal kata yang berasal dari proses baris baru (\n) dan *backslash* (\).
b. *Tokenizing*
   Merupakan tahap pemotongan string input berdasarkan setiap kata untuk kemudian dapat dianalisis [19]. Pada tahapan ini dilakukan proses

pemecahan kalimat menjadi sebuah kata-kata individu atau token yang berpaku pada spasi.

c. *Normalization*
Pada tahapan normalisasi, dilakukan proses penyelarasan kata untuk mengatasi variasi penulisan yang tidak baku atau penggunaan singkatan dalam teks aduan masyarakat sesuai dengan kaidah Bahasa Indonesia yang mengacu pada Kamus Besar Bahasa Indonesia (KBBI) [18]. Contohnya, kata singkatan seperti "sgt" dinormalisasi menjadi "sangat", "tdk" menjadi "tidak", dan "sdh" menjadi "sudah".

d. *Stopword Removal*
Pada tahapan ini, dilakukan proses penghapusan kata-kata umum yang sering muncul tetapi tidak bermakna penting atau tidak memberikan informasi signifikan dengan tujuan untuk mengurangi dimensi data [18]. Kata-kata tersebut meliputi kata hubung (konjungsi), kata ganti orang, dan kata kepemilikan, seperti "dan", "atau", "yang", "saya", "untuk", "ke", "karena".

e. *Stemming*
Pada tahapan ini setiap kata akan menjadi bentuk kata dasar dengan menghilangkan imbuhan awal maupun akhir, guna meningkatkan efektivitas proses analisis teks [18]. Contohnya "prosesnya" menjadi "proses", "bimbingannya" menjadi "bimbing", "bantuan" menjadi "bantu".

3.3. Pembobotan TF-IDF

Setelah tahap *preprocessing* selesai, langkah berikutnya adalah melakukan pembobotan kata menggunakan perhitungan TF-IDF. Berikut ini merupakan tahapan proses pembobotan kata (TF-IDF) pada penelitian ini:

a. Menghitung *Term Frequency* (TF) untuk setiap kata.
b. Menghitung *Inverse Document Frequency* (IDF) untuk setiap kata.
c. Setelah mendapatkan nilai TF dan juga nilai IDF, selanjutnya dilakukan perhitungan bobot akhir kata. Skor TF-IDF dihitung dengan mengalikan TF dengan IDF. Oleh karena itu, jika nilai TF untuk suatu kata adalah 0 (karena kata tersebut tidak muncul dalam dokumen), maka skor TF-IDF untuk kata tersebut akan menjadi 0, meskipun nilai IDF-nya tinggi. Dengan kata lain, jika sebuah kata tidak ada di dokumen, maka secara otomatis dianggap tidak memiliki kontribusi dalam representasi dokumen tersebut.
d. Membangun matriks TF-IDF untuk seluruh dokumen. Hasil perhitungan TF-IDF diubah menjadi matriks fitur berbentuk tabel yang merepresentasikan dokumen sebagai vektor numerik. Hasil dari matriks TF-IDF selanjutnya dapat dilanjutkan untuk dilakukan Pemodelan data.

3.4. Ward

Setelah mendapatkan TF-IDF, selanjutnya dilakukan pemodelan menggunakan metode Ward untuk melakukan klasterisasi. Berikut ini merupakan tahapan Langkah analisis metode Ward pada penelitian ini:

a. Setiap dokumen hasil pembobotan TF-IDF diasumsikan sebagai satu klaster awal.
b. Jarak antar klaster dihitung menggunakan *Euclidean Distance* untuk mengukur tingkat kemiripan antar vektor dokumen.
c. Untuk setiap kemungkinan penggabungan klaster, dihitung peningkatan *Sum of Squared Errors* (SSE) sebagai ukuran perubahan homogenitas klaster.
d. Dua klaster dengan peningkatan SSE paling kecil dipilih dan digabungkan menjadi satu klaster baru.
e. Setelah penggabungan, jarak antar klaster diperbarui untuk merefleksikan struktur klaster terbaru.
f. Langkah perhitungan jarak, evaluasi SSE, dan penggabungan klaster diulang hingga seluruh dokumen tergabung atau jumlah klaster tertentu tercapai.
g. Jumlah klaster optimal ditentukan menggunakan dendrogram, yang digunakan untuk mengidentifikasi titik potong terbaik sebagai pemisah antar klaster.

3.5. Evaluasi Model

Pada tahapan ini dilakukan evaluasi model dengan tujuan untuk mengetahui bagaimana kualitas dan validitas klaster yang terbentuk berdasarkan hasil pengelompokan data. Pada penelitian ini, peneliti menggunakan evaluasi *Silhouette Coefficient* (SC) dan dan *Davies Bouldin Index* (DBI). Nilai SC berkisar antara -1 hingga 1, yang mana nilai mendekati 1 menunjukkan bahwa objek telah terkelompok dengan baik dalam klasternya, yang berarti klasterisasi memiliki kualitas yang baik. Sementara, semakin kecilnya nilai DBI, semakin baik hasil klasterisasi. Nilai DBI yang rendah menunjukkan bahwa klaster lebih kompak dan memiliki pemisahan yang jelas, sehingga dapat diartikan bahwa pengelompokan yang dilakukan sudah optimal.

4. HASIL DAN PEMBAHASAN

4.1. Pengumumpulan Data

Data yang digunakan pada penelitian ini merupakan data keluhan dari masyarakat terkait pelayanan yang dari Dispendukcapil Kota Surabaya, selama periode bulan Januari 2023 hingga bulan Desember 2024 (2 tahun) dengan jumlah data berjumlah 1.266 aduan.

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

4.2. *Preprocessing*

Proses preprocessing bertujuan untuk menghilangkan unsur-unsur yang tidak relevan, menyeragamkan format teks, serta mengekstraksi kata-kata yang memiliki makna penting agar dapat merepresentasikan isi aduan secara lebih akurat. Tahapan *preprocessing* yang dilakukan meliputi *cleaning*, *tokenizing*, *normalization*, *stopword removal*, dan *stemming* dengan memanfaatkan *library* Sastrawi yang dirancang khusus untuk pengolahan teks berbahasa Indonesia. Tabel 3 merupakan contoh hasil *Preprocessing*.

Tabel 2. Hasil *Preprocessing*

Proses	Hasil
Data Mentah	Selamat Pagi, \n\\saya ingin bertanya bagaimana prosesnya untuk pembuatan akta perkawinan? apakah bisa melalui website atau hrs datang ke kantor kecamatan/kelurahan? mohon bantuan dan bimbingannya secara terperinci langkah- langkahnya, karena saya tidak paham. Terima kasih.
Cleaning	selamat pagi saya ingin bertanya bagaimana prosesnya untuk pembuatan akta perkawinan apakah bisa melalui website atau hrs datang ke kantor kecamatan kelurahan mohon bantuan dan bimbingannya secara terperinci langkah langkahnya karena saya tidak paham terima kasih
Tokenizing	['selamat', 'pagi', 'saya', 'ingin', 'bertanya', 'bagaimana', 'prosesnya', 'untuk', 'pembuatan', 'akta', 'perkawinan', 'apakah', 'bisa', 'melalui', 'website', 'atau', 'hrs', 'datang', 'ke', 'kantor', 'kecamatan', 'kelurahan', 'mohon', 'bantuan', 'dan', 'bimbingannya', 'secara', 'terperinci', 'langkah', 'langkahnya', 'karena', 'saya', 'tidak', 'paham', 'terima', 'kasih']
Normalization	['selamat', 'pagi', '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'
Stopword Removal	['selamat', 'pagi', 'prosesnya', 'pembuatan', 'akta', 'perkawinan', 'website', 'kantor', 'kecamatan', 'kelurahan', 'mohon', 'bantuan', 'bimbingannya', 'terperinci', 'langkah', 'langkahnya', 'paham', 'terima', 'kasih']
Stemming	selamat pagi proses buat akta kawin website kantor camat kelurahan mohon bantu bimbing perinci langkah langkah paham terima kasih

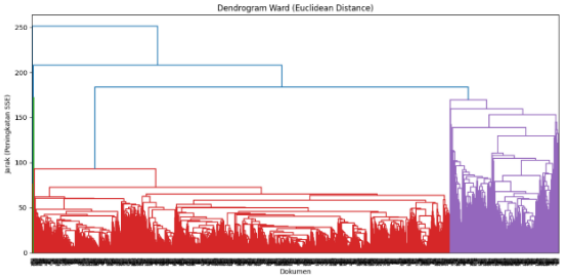
4.3. TF-IDF

Term	TF D1	TF D2	TF D3	DF	IDF	TF-IDF D1	TF-IDF D2	TF-IDF D3
aju	0.0	1.0	0.0	188	2.9000	0.0000	2.9000	0.0000
akta	1.0	0.0	0.0	202	2.8282	2.8282	0.0000	0.0000
amun	0.0	0.0	1.0	1	8.1365	0.0000	0.0000	8.1365
bantu	1.0	1.0	0.0	178	2.9547	2.9547	2.9547	0.0000
bimbing	1.0	0.0	0.0	2	7.4433	7.4433	0.0000	0.0000
buat	1.0	0.0	0.0	63	3.9933	3.9933	0.0000	0.0000
camat	1.0	0.0	0.0	83	3.7176	3.7176	0.0000	0.0000
cetak	0.0	0.0	2.0	126	3.3002	0.0000	0.0000	6.6004
desember	0.0	0.0	1.0	9	5.9393	0.0000	0.0000	5.9393
elektronik	0.0	1.0	0.0	31	4.7025	0.0000	4.7025	0.0000
kantor	1.0	0.0	1.0	30	4.7353	4.7353	0.0000	4.7353
kasih	1.0	0.0	0.0	187	2.9054	2.9054	0.0000	0.0000
kawin	1.0	0.0	0.0	48	4.2653	4.2653	0.0000	0.0000
ktp	0.0	2.0	3.0	344	2.2958	0.0000	4.5917	6.8875
langkah	2.0	0.0	0.0	11	5.7386	11.4772	0.0000	0.0000
lurah	1.0	0.0	1.0	244	2.6393	2.6393	0.0000	2.6393
mas	0.0	0.0	1.0	4	6.7502	0.0000	0.0000	6.7502
mohon	1.0	1.0	0.0	469	1.9859	1.9859	1.9859	0.0000
pagi	1.0	0.0	0.0	50	4.2245	4.2245	0.0000	0.0000
paham	1.0	0.0	0.0	6	6.3447	6.3447	0.0000	0.0000
perinci	1.0	0.0	0.0	1	8.1365	8.1365	0.0000	0.0000
proses	1.0	0.0	0.0	119	3.3574	3.3574	0.0000	0.0000
rusak	0.0	0.0	1.0	32	4.6707	0.0000	0.0000	4.6707
selamat	1.0	0.0	0.0	113	3.4091	3.4091	0.0000	0.0000
terima	1.0	0.0	0.0	197	2.8533	2.8533	0.0000	0.0000
website	1.0	0.0	0.0	17	5.3033	5.3033	0.0000	0.0000

Gambar 1. Hasil Proses TF-IDF

Berdasarkan hasil pembobotan TF-IDF yang ditunjukkan pada Gambar 1, setiap dokumen direpresentasikan dalam bentuk vektor numerik berdasarkan tingkat kepentingan kata. Nilai TF menggambarkan frekuensi kemunculan kata dalam dokumen, sedangkan nilai IDF menurunkan bobot kata yang sering muncul pada banyak dokumen karena dianggap kurang informatif. Kombinasi keduanya menghasilkan nilai TF-IDF yang menekankan kata-kata spesifik pada dokumen tertentu, seperti term “ktp” yang memiliki bobot tinggi pada dokumen tertentu. Sebaliknya, kata yang tidak muncul dalam dokumen memiliki nilai TF-IDF sebesar nol meskipun nilai IDF-nya tinggi. Representasi TF-IDF ini selanjutnya digunakan sebagai input utama dalam proses klasterisasi, sehingga pengelompokan dokumen dilakukan berdasarkan kemiripan bobot kata yang relevan.

4.4. Analisis Ward dan Evaluasi



Gambar 2. Dendogram Ward

Berdasarkan hasil klasterisasi hierarkis menggunakan metode Ward dengan jarak *euclidean*, pada Gambar 2, visualisasi dendrogram memperlihatkan lonjakan jarak penggabungan terbesar pada tahap tertentu, yang mengindikasikan titik pemotongan klaster yang optimal. Berdasarkan analisis tersebut, diperoleh jumlah klaster optimal pada metode Ward sebanyak dua klaster, seperti yang tertera pada Gambar 3. Selanjutnya jumlah klaster optimal tersebut digunakan sebagai dasar dalam proses berikutnya.

```
Jumlah klaster optimal (Ward): 2
```

Gambar 3. Klaster Optimal

Setelah jumlah klaster optimal ditentukan, proses klasterisasi Ward kembali dilakukan dengan menetapkan jumlah klaster sebanyak dua.

```
Silhouette Score : 0.7603
Davies-Bouldin Index : 1.0283
```

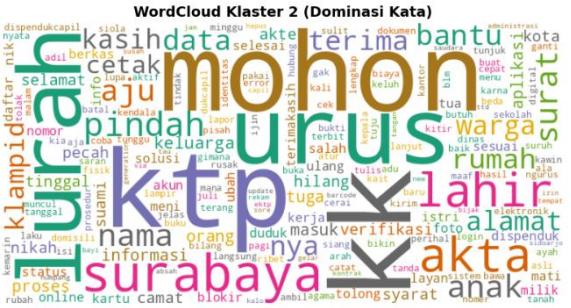
Gambar 4. Hasil Evaluasi *Silhouette Coefficient*

Hasil klasterisasi kemudian dievaluasi menggunakan *Silhouette Coefficient* dan *Davies Bouldin Index*, yang mana pada Gambar 4 didapatkan *Silhouette Coefficient* menghasilkan nilai sebesar 0,7603. Nilai ini menunjukkan bahwa struktur klaster yang terbentuk tergolong baik, dengan tingkat keseragaman data dalam klaster (kohesi) dan keterpisahan antar klaster (separasi) yang cukup jelas, sehingga hasil pengelompokan dapat dianggap *representative*. Kemudian, evaluasi menggunakan *Davies Bouldin Index* menghasilkan nilai 1,0283. Nilai DBI yang relatif rendah ini menegaskan bahwa klaster yang terbentuk memiliki penyebaran data yang *compact* (rapat) dengan separasi antar pusat klaster yang cukup jauh. Secara keseluruhan, hasil evaluasi menunjukkan bahwa metode yang digunakan mampu menghasilkan pengelompokan yang valid dan optimal.

4.5. Visualisasi

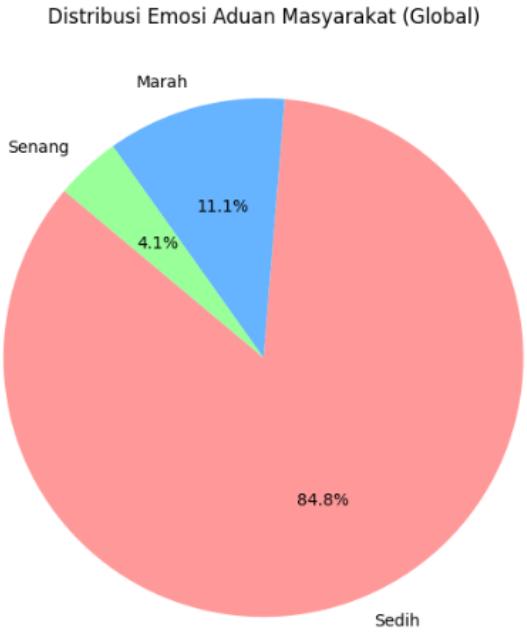
Wordcloud pada Klaster 1 menunjukkan dominasi kata-kata seperti akte lahir, bayi, nikah, duga, manipulasi, hasil, dan hukum, yang mengindikasikan aduan dengan konteks permasalahan administratif yang lebih spesifik, sensitif, dan memerlukan bantuan lanjutan berupa tindakan dari petugas.

Gambar 5. *Wordcloud* Klaster 1



Gambar 6. *Wordcloud* Klaster 2

Sementara itu, Klaster 2 didominasi oleh kata-kata seperti urus, mohon, surat, kk, anak, lurah, akta, pindah, dan terima kasih, yang menggambarkan aduan masyarakat yang berkaitan dengan permohonan layanan administrasi kependudukan serta proses pelayanan yang bersifat prosedural dan informatif.



Gambar 7. Distribusi Emosi aduan (Keseluruhan)

Kemudian, berdasarkan hasil analisis emosi secara keseluruhan, aduan masyarakat didominasi oleh emosi sedih sejumlah 84,4%, kemudian untuk emosi

senang sejumlah 4,1%, dan emosi marah sejumlah 11,1%. Dominasi emosi sedih menunjukkan bahwa mayoritas aduan berkaitan dengan kekecewaan, kebingungan, dan kesedihan terhadap pelayanan dan juga proses dalam mengurus administrasi kependudukan.

Tabel 3. Distribusi Emosi Aduan Dalam Ward

klaster	Marah	Sedih	Senang
1	0	2	0
2	140	1064	51

Selanjutnya, analisis emosi pada hasil klaster ward, menunjukkan bahwa Klaster 1 didominasi oleh emosi sedih sejumlah 2 aduan. Kemudian, Klaster 2 juga didominasi oleh emosi sedih, sejumlah 1064 aduan, kemudian marah sebanyak 140 aduan, dan senang sebanyak 51 aduan.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian, metode Ward mampu mengelompokkan aduan pelayanan masyarakat di Dispendukcapil Kota Surabaya secara efektif dengan jumlah klaster optimal sebanyak dua klaster yang diperoleh melalui analisis dendrogram. Hasil evaluasi menggunakan *Silhouette Coefficient* menghasilkan nilai sebesar 0,7603 yang menunjukkan bahwa klaster yang terbentuk memiliki tingkat kohesi dan separasi yang baik, sedangkan evaluasi menggunakan *Davies Bouldin Index* menghasilkan nilai sebesar 1,0283 yang mengindikasikan bahwa klaster bersifat cukup kompak dengan jarak antar klaster yang memadai. Visualisasi *wordcloud* memperlihatkan perbedaan karakteristik aduan pada setiap klaster, di mana satu klaster didominasi oleh aduan prosedural dan informatif terkait layanan administrasi kependudukan, sementara klaster lainnya menunjukkan aduan yang lebih spesifik, sensitif, dan memerlukan penanganan lanjutan. Selain itu, hasil analisis emosi menunjukkan bahwa aduan masyarakat secara keseluruhan didominasi oleh emosi sedih, yang mencerminkan adanya kekecewaan dan kebingungan masyarakat terhadap proses pelayanan Oleh karena itu, hasil klasterisasi ini dapat menjadi dasar bagi instansi terkait dalam mengidentifikasi pola permasalahan pelayanan serta menyusun strategi peningkatan kualitas layanan yang lebih tepat sasaran. Penelitian selanjutnya disarankan untuk memperluas variasi fitur, mengoptimalkan pemodelan emosi dengan pendekatan berbasis pembelajaran mesin, serta membandingkan hasil klasterisasi dengan data layanan aktual guna meningkatkan validitas dan pemanfaatan hasil analisis.

DAFTAR PUSTAKA

[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] Dessy Yulianti, Teguh Hermanto, and Meriska Defriani, "Analisis Clustering Donor Darah dengan Metode Agglomerative *Hierarchical clustering*," *Resolusi : Rekayasa Teknik Informatika dan Informasi*, vol. 3, no. 6, pp. 303–308, 2023, doi: https://doi.org/10.30865/resolusi.v3i6.977.

[4] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*. Wiley, 2011. doi: https://doi.org/10.1002/9780470977811.

[5] Muthahharah and A. Juhari, "A Cluster Analysis with Complete Linkage and Metode Wardfor Health Service Data in Makassar City", Jurnal Varian, vol. 4, no. 2, pp. 109 - 116, Apr. 2021.

[6] J. Baugher and Y. Qu, "Create the Taxonomy for Unintentional Insider Threat via *Text Mining* and Hierarchical clustering Analysis," European journal of electrical engineering and computer science, vol. 8, no. 2, pp. 36–49, Apr. 2024.

[7] 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.

[8] Riyantoko, P. A.., & Muhaimin, A (2023). A Simple Data Sentiment Analysis using Bjorka phenomenon on Twitter *Nusantara Science and Technology Proceedings*, *2023*(33), 330-336. https://doi.org/10.11594/nstp.2023.3353

[9] Rizky Ananta Pratama, "ANALISIS SENTIMEN KONSUMEN DENGAN TEKNIK *TEXT MINING*," *Jurnal Dunia Data*, vol. 1, no. 6, 2024, Accessed: Mar. 01, 2025. [Online]. Available: http://www.pusdansi.org/index.php/duniadata/article/view/109.

[10] S. M. Fani, R. Santoso, and S. Suparti, "Penerapan *Text Mining* untuk Melakukan Clustering Data Tweet Akun Blibli Pada Media Sosial Twitter Menggunakan *K-Means* Clustering," *Jurnal Gaussian*, vol. 10, no. 4, pp. 583-593, Dec. 2021. https://doi.org/10.14710/j.gauss.10.4.583-593.

[11] Genisia Pramestiloka Aulia, Tatik Widiharih, and Iut Tri Utami, "PENERAPAN *TEXT MINING* DAN FUZZY C-MEANS CLUSTERING UNTUK IDENTIFIKASI KELUHAN UTAMA PELANGGAN PDAM TIRTA MOEDAL KOTA SEMARANG," *Jurnal Gaussian*, vol. 12, no. 1, pp. 126–135, Dec. 2022, doi: https://doi.org/10.14710/j.gauss.12.1.126-135

[12] 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.

[13] I. Indra, N. Nur, Muh. Iqram, and Nurul Inayah, "Perbandingan *K-Means* dan Hierarchical Clustering dalam Pengelompokan Daerah Beresiko Stunting," *INOVTEK Polbeng - Seri Informatika*, vol. 8, no. 2, pp. 356–356, Nov. 2023, doi: https://doi.org/10.35314/isi.v8i2.3612.

[14] T. M. Fahrudin, P. A. Riyantoko, K. M. Hindrayani, and M. H. P. Swari, "Cluster Analysis of Hospital Inpatient Service Efficiency Based on BOR, BTO, TOI, AvLOS Indicators using Agglomerative Hierarchical Clustering", *Telematika*, vol. 18, no. 2, pp. 194–210, Oct. 2021.

[15] 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.

[16] 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.

[17] Raditya Rinandyaswara, Y. A. Sari, and Muhammad Tanzil Furqon, "Pembentukan Daftar Stopword Menggunakan Term Based Random Sampling Pada Analisis Sentimen Dengan Metode Naïve Bayes (Studi Kasus: Kuliah Daring Di Masa Pandemi)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 9, no. 4, pp. 717–717, Aug. 2022, doi: https://doi.org/10.25126/jtiik.2022934707

[18] S. J. Angelina, Arif, and Hafiz Muhardi, "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.

[19] 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.