# Perbandingan Penerapan *Text Mining* Pada Aplikasi Wargaku Surabaya Untuk Aduan Masyarakat Terhadap Satpol Pp Kota Surabaya Dengan Algoritma *K-Means Clustering* Dan *Fuzzy C-Means Clustering*

Ica Amilatul Kholidah[1], Aries Dwi Indriyanti[2]

[1,2] Jurusan Teknik Informatika/Program Studi S1 Sistem Informasi, Universitas Negeri Surabaya
[1]icaamilatul.21012@mhs.unesa.ac.id
[2]ariesdwi@unesa.ac.id

***Abstrak*— Di era digital, teknologi memiliki peran penting dalam memfasilitasi komunikasi antara pemerintah dan masyarakat, salah satunya melalui layanan pengaduan publik. Aplikasi WargaKu Surabaya merupakan inovasi dari Pemerintah Kota Surabaya yang memudahkan masyarakat dalam menyampaikan aduan, kritik, dan saran secara langsung. Salah satu instansi yang menerima laporan dari aplikasi ini adalah Satpol PP Kota Surabaya. Aduan yang masuk sangat beragam dan jumlahnya cukup besar. Namun, data tersebut umumnya berbentuk teks tidak terstruktur, serta banyak aduan yang berulang atau membahas topik yang sama. Kondisi ini menunjukkan bahwa penanganan aduan oleh Satpol PP belum berjalan optimal, sehingga diperlukan pengelolaan data yang lebih efektif.**

**Penelitian ini bertujuan untuk membandingkan kinerja algoritma K-Means dan Fuzzy C-Means dalam mengelompokkan data aduan masyarakat. Tujuannya adalah untuk membantu pemerintah memetakan isu-isu yang paling sering muncul agar dapat diprioritaskan penanganannya. Data yang digunakan merupakan aduan masyarakat kepada Satpol PP melalui aplikasi WargaKu Surabaya, sebanyak 500 data yang dikumpulkan dari Januari 2023 hingga April 2024.**

**Hasil penelitian menunjukkan bahwa algoritma K-Means memiliki performa lebih baik dalam membentuk klaster, dengan nilai Silhouette sebesar 0.5048 dibandingkan Fuzzy C-Means sebesar 0.5000. Selain itu, K-Means lebih cepat dan menghasilkan pemisahan klaster yang lebih jelas, sementara Fuzzy C-Means lebih sensitif terhadap data outlier.**

***Kata Kunci*— *Text Mining*, Aduan Masyarakat, *K-Means*, *Fuzzy C-Means*, *Clustering*.**

## I. PENDAHULUAN

Dalam era digital saat ini, teknologi memainkan peran penting dalam memfasilitasi interaksi antara pemerintah dan masyarakat. Layanan pengaduan masyarakat menjadi salah satu alat yang vital dalam meningkatkan transparansi, responsivitas, dan kualitas pelayanan publik. Menurut Wulandari dkk. (2024) pengaduan masyarakat merupakan suatu sumber informasi yang sangat penting untuk memperbaiki kesalahan yang mungkin terjadi, selain itu secara konsisten dapat menjaga dan meningkatkan pelayanan yang dihasilkan agar sesuai dengan standar yang telah ditetapkan.

Mengutip dari laman Suara Surabaya Pemerintah Kota Surabaya telah berupaya meningkatkan kualitas pelayanan publik mereka melalui pengembangan aplikasi WargaKu Surabaya, sebuah aplikasi berbasis android yang diluncurkan oleh Pemerintah Kota Surabaya pada tanggal 22 Maret 2021. Melalui aplikasi ini, warga Surabaya dapat menyampaikan aduan, kritik, saran, permohonan informasi, atau apresiasi kepada Pemerintah Kota Surabaya. Aplikasi ini akan menerima dan meneruskan berbagai aduan-aduan masyarakat ke dinas-dinas terkait untuk di tindak lanjuti. Sejak diluncurkan pada bulan Maret 2021 hingga akhir Desember 2021, aplikasi WargaKu menerima 11.316 pengaduan yang diajukan melalui aplikasi tersebut.

Dalam pengelolaan pengaduan masyarakat untuk peningkatan layanan publik yang lebih responsif dan efektif dapat menggunakan *text mining* dengan algoritma *K-Means Clustering* dan *Fuzzy C-Means Clustering*. *Text mining* adalah proses mengubah data teks tidak terstruktur menjadi informasi yang dapat dimanfaatkan. Penelitian ini membandingkan algoritma *K-Means* dan *Fuzzy C-Means* (FCM) dalam proses *clustering* data keluhan masyarakat. *K-Means* merupakan algoritma *hard clustering* (membagi data ke satu *cluster*) yang umum digunakan dalam pengelompokan data, sedangkan FCM dipilih karena mampu menangani ketidakpastian dalam data teks dengan memberikan derajat keanggotaan ganda sehingga satu data bisa dimasukkan kedalam beberapa *cluster* (*soft clustering*).

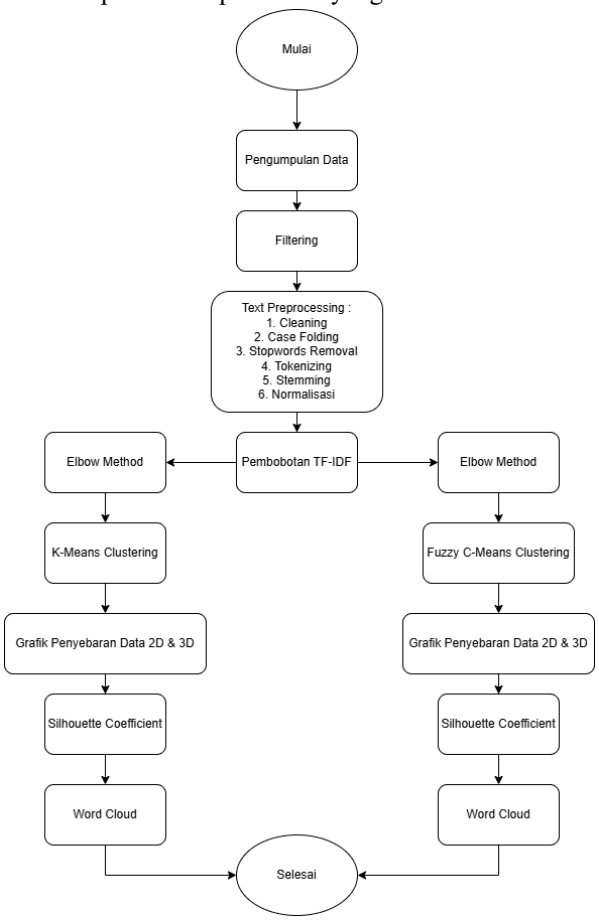
Satuan Polisi Pamong Praja (Satpol PP) Kota Surabaya sebagai salah satu perangkat daerah juga menerima dan menangani aduan dari aplikasi WargaKu Surabaya yang ditujukan untuk Satpol PP Kota Surabaya dengan berbagai macam topik.

Laporan yang masuk melalui aplikasi WargaKu Surabaya khususnya kepada Satpol PP Kota Surabaya sangat banyak dan beragam. Data ini sering kali berbentuk teks tidak terstruktur. Beberapa tantangan utama dalam pengelolaan pengaduan masyarakat antara lain banyaknya aduan yang disampaikan

oleh masyarakat seringkali berisi topik yang sama dan aduan yang sama dilakukan berulang-ulang karena belum adanya solusi terhadap aduan yang mereka sampaikan. Keadaan ini menunjukkan bahwa banyaknya aduan masyarakat yang tidak tertangani dengan optimal oleh Satpol PP Kota Surabaya sehingga sehingga perlu dilakukan pengelolaan aduan. Penelitian ini akan berfokus pada *clustering* aduan yang disampaikan oleh masyarakat pada aplikasi WargaKu Surabaya dengan tujuan untuk prioritisasi masalah yang memungkinkan pemerintah untuk mengidentifikasi isu-isu yang paling sering dilaporkan dan yang memerlukan perhatian segera, serta menyediakan layanan yang lebih responsif dan sesuai dengan kebutuhan dan harapan warga.

## II. METODOLOGI PENELITIAN

Metode yang digunakan dalam penelitian ini adalah metode penelitian kuantitatif, yang merupakan pendekatan penelitian yang objektif yang melibatkan pengumpulan dan analisis data menggunakan teknik statistik. Penelitian akan berfokus pada aduan Masyarakat terhadap Satuan Polisi Pamong Praja Kota Surabaya yang disampaikan melalui aplikasi WargaKu Surabaya. Pada penelitian ini terdapat beberapa tahapan yang harus dilalui guna mencapai tujuan penelitian. Gambar 1 berikut merupakan alur penelitian yang dilakukan.




Gbr. 1 Alur Penelitian

### A. *Data Set*

Dataset yang digunakan dalam penelitian ini dikumpulkan dengan menggunakan metode pengumpulan data sekunder. Data sekunder diperoleh dari dokumentasi aduan masyarakat yang disampaikan melalui aplikasi WargaKu Surabaya, yang secara khusus ditujukan kepada Dinas Satuan Polisi Pamong Praja (Satpol PP) Kota Surabaya. Data tersebut diperoleh dalam bentuk file Excel yang berisi 500 data keluhan masyarakat selama periode Januari 2023 hingga April 2024.

### B. *Filtering*

*Filtering* adalah menghapus data yang duplikat atau ganda sehingga hanya tersisa satu. Pada awal pengumpulan data diperoleh sebanyak 500 data, kemudian setelah dilakukan filtering diperoleh hasil menjadi 500 data yang berarti tidak terdapat data duplikat atau ganda.

### C. *Text Preprocessing*

*Text Preprocessing* digunakan untuk membersihkan dan mempersiapkan data agar dapat digunakan dalam analisis yang lebih lanjut. Berikut adalah beberapa tahapan *text preprocessing* yang dilakukan :

1. *Cleaning*
   *Cleaning* adalah proses untuk membersihkan tanda baca dan kata-kata yang tidak diperlukan untuk mengurangi *noise* (data yang tidak konsisten) (Syafii, 2023).
2. Case Folding
   *Case folding* adalah proses mengubah semua huruf dalam teks menjadi huruf kecil untuk menghilangkan perbedaan antara huruf besar dan huruf kecil dalam analisis teks.
3. Stopwords Removal
   *Stopword Removal* melakukan pemilihan kata-kata yang tidak penting yang ada dalam dokumen seperti 'dan', 'atau', 'tetapi', dan 'di', karena kata-kata tersebut muncul sangat sering namun tidak membantu dalam menentukan konteks dokumen (Rifaldi dkk., 2023).
4. Tokenizing
   *Tokenizing* bertujuan untuk memecah teks menjadi unit-unit yang lebih kecil, biasanya kata atau frasa, yang disebut sebagai *token*. *Tokenization* memungkinkan setiap kata dalam teks dapat dianalisis secara individual.
5. Stemming
   *Stemming* adalah proses mengidentifikasi dan mengubah kata yang memiliki imbuhan menjadi bentuk dasar memiliki imbuhan menjadi bentuk dasar (Puspitasari & Indriyanti, 2024).
6. Normalisasi
   Tahap Normalisasi adalah proses memperbarui kata yang keliru eja ataupun kata-kata tidak baku kedalam bentuk baku memakai kamus normalisasi.

### D. *Pembobotan TF-IDF*

TF-IDF *(Term Frequency-Inverse Document Frequency)* merupakan metode yang digunakan dalam pengolahan teks

untuk menghitung bobot kata dalam dokumen. TF-IDF menggabungkan dua metrik, yaitu *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF), yang dihitung secara terpisah dan kemudian dikalikan untuk memberikan bobot akhir bagi setiap kata. TF-IDF memberikan bobot yang lebih tinggi untuk kata-kata yang sering muncul dalam dokumen yang jarang muncul dalam dokumen-dokumen lain. Sehingga memungkinkan analisis yang lebih efektif dalam mengidentifikasi kata-kata yang paling penting dalam dokumen.

### E. *K-Means Clustering*

1. Elbow Method

   *Elbow Method* akan menentukan jumlah *cluster* optimal dengan melihat presentase setiap *cluster* yang akan membentuk siku atau *elbow* pada suatu titik tertentu (Maori & Evanita, 2023).

   *Metode elbow* di visualisasikan dalam bentuk grafik untuk mengetahui lebih jelas siku yang terbentuk. Tujuan dari metode *elbow* untuk memilih nilai k yang kecil dan masih memiliki nilai *withinss* yang rendah.
2. K-Means Clustering

   Algoritma *K-Means* didefinisikan sebagai metode *Unsupervised Learning* yang memiliki proses iteratif, di mana dataset dikelompokkan ke dalam *k* jumlah klaster atau subkelompok yang telah ditentukan sebelumnya dan tidak saling tumpang tindih (Susanti & Wibawa, 2024).

   *K-Means Clustering* akan mengelompokkan data ke dalam *cluster* yang memiliki karakteristik yang sama atau paling mirip.
3. Grafik Penyebaran Data

   Grafik penyebaran data berfungsi untuk memvisualisasikan hasil pengelompokan data (*clustering*) yang telah dilakukan. Visualisasi ini memberikan gambaran bagaimana data dikelompokkan ke dalam *cluster* berdasarkan pola atau fitur tertentu.

   Dalam penelitian ini di bentuk grafik penyebaran data menggunakan dua dimensi dan tiga dimensi. Pada grafik penyebaran data dua dimensi menggunakan *Principal component analysis* (PCA) untuk mereduksi dimensi hasil pembobotan TF-IDF menjadi dua dimensi (2D), sehingga dapat divisualisasikan dalam bentuk plot sebaran data.
4. Silhouette Coeffcient

   *Silhouette Coefficient* dapat digunakan untuk menilai kualitas kluster yang dihasilkan oleh *Elbow Method* dengan cara menghitung nilai *Silhouette Coefficient* untuk setiap data point dalam kluster.

   Nilai *Silhouette Coefficient* berkisar antara -1 hingga 1, di mana nilai yang mendekati 1 menunjukkan bahwa objek berada di klaster yang tepat, dan sebaliknya.
5. Word Cloud

   *Word Cloud* adalah metode visualisasi data yang digunakan untuk menampilkan kata-kata atau frase yang paling sering muncul dalam suatu kumpulan teks.

   Dengan demikian, dapat dilihat tema-tema atau pola-pola yang dominan dalam setiap *cluster*, sehingga dapat membantu dalam mengidentifikasi masalah-masalah yang paling sering muncul.

### F. *Fuzzy C-Means Clustering*

1. Elbow Method

   *Elbow Method* akan menentukan jumlah *cluster* optimal dengan melihat presentase setiap *cluster* yang akan membentuk siku atau *elbow* pada suatu titik tertentu (Maori & Evanita, 2023).

   *Metode elbow* di visualisasikan dalam bentuk grafik untuk mengetahui lebih jelas siku yang terbentuk. Tujuan dari metode *elbow* untuk memilih nilai k yang kecil dan masih memiliki nilai *withinss* yang rendah.
2. Fuzzy C-Means Clustering

   Algoritma *Fuzzy C-Means Clustering* melakukan *clustering* terhadap aduan masyarakat dengan mengelompokkan data ke dalam beberapa klaster berdasarkan derajat keanggotaan atau "*membership degree*" dari setiap data terhadap klaster tertentu. *Fuzzy C-Means Clustering* merepresentasikan setiap aduan sebagai vektor kata kunci menggunakan teknik TF-IDF.

   FCM digunakan sebagai model *clustering* pada *fuzzy* sehingga data bisa menjadi anggota dari keseluruhan kelas atau *cluster* yang dapat terbentuk dari derajat atau tingkat keanggotaan yang berbeda antara 0 sampai 1 (Maulidiyah & Nuryana, 2022).
3. Grafik Penyebaran Data

   Grafik penyebaran data berfungsi untuk memvisualisasikan hasil pengelompokan data (*clustering*) yang telah dilakukan. Visualisasi ini memberikan gambaran bagaimana data dikelompokkan ke dalam *cluster* berdasarkan pola atau fitur tertentu.

   Dalam penelitian ini di bentuk grafik penyebaran data menggunakan dua dimensi dan tiga dimensi. Pada grafik penyebaran data dua dimensi menggunakan *Principal component analysis* (PCA) untuk mereduksi dimensi hasil pembobotan TF-IDF menjadi dua dimensi (2D), sehingga dapat divisualisasikan dalam bentuk plot sebaran data.
4. Silhouette Coeffcient

   *Silhouette Coefficient* dapat digunakan untuk menilai kualitas kluster yang dihasilkan oleh *Elbow Method* dengan cara menghitung nilai *Silhouette Coefficient* untuk setiap data point dalam kluster.

   Nilai *Silhouette Coefficient* berkisar antara -1 hingga 1, di mana nilai yang mendekati 1 menunjukkan bahwa objek berada di klaster yang tepat, dan sebaliknya.
5. Word Cloud

   *Word Cloud* adalah metode visualisasi data yang digunakan untuk menampilkan kata-kata atau frase yang paling sering muncul dalam suatu kumpulan teks.

   Dengan demikian, dapat dilihat tema-tema atau pola-pola yang dominan dalam setiap *cluster*, sehingga

dapat membantu dalam mengidentifikasi masalah-masalah yang paling sering muncul.

## III. *HASIL DAN PEMBAHASAN*

### A. *Pembobotan TF-IDF*

Penerapan Term Frequency – Inverse Document Frequency digunakan untuk menunjukkan seberapa penting suatu kata (TF-IDF *score*) didalam dokumen tersebut. Nilai mendekati 1 berarti kata tersebut sangat penting terhadap dokumen tersebut, sedangkan nilai 0 berarti kata tidak muncul atau tidak relevan di dokumen tersebut. Berikut adalah tabel hasil pembobotan menggunakan TF-IDF.

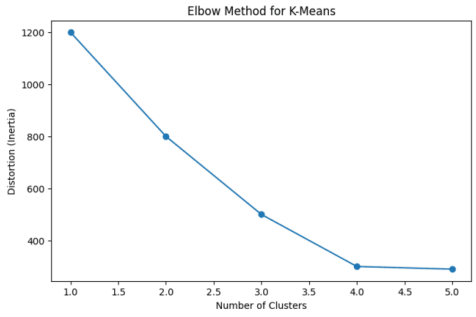
*TABEL I*
*HASIL PEMBOBOTAN TF-IDF*

| | jalan | jual | malam | pkl | tertib |
|---|---|---|---|---|---|
| 0 | 0.000000 | 0.0 | 1.000000 | 0.0 | 0.000000 |
| 1 | 0.000000 | 0.0 | 0.000000 | 0.0 | 0.000000 |
| 2 | 0.000000 | 0.0 | 0.000000 | 0.0 | 1.000000 |
| .. | ... | ... | ... | ... | ... |
| 498 | 0.000000 | 0.0 | 0.425021 | 0.0 | 0.905184 |
| 499 | 0.372636 | 0.0 | 0.819109 | 0.0 | 0.436122 |

### B. K-Means Clustering

1. Elbow Method

Dalam tahap ini, dilakukan proses penentuan jumlah klaster optimal menggunakan metode *Elbow method*. Untuk menentukan jumlah klaster optimal, dilakukan perulangan klaster dari k = 1 hingga k = 4 menggunakan algoritma *K-Means*. Pada setiap iterasi, dilakukan pelatihan model dan dihitung nilai inertia (distorsi), yang kemudian disimpan untuk divisualisasikan dalam bentuk grafik. Sumbu X menunjukkan jumlah klaster yang diuji, sedangkan sumbu Y menunjukkan nilai distorsi masing-masing klaster.




Gbr. 2 Hasil Elbow Method K-Means Clustering

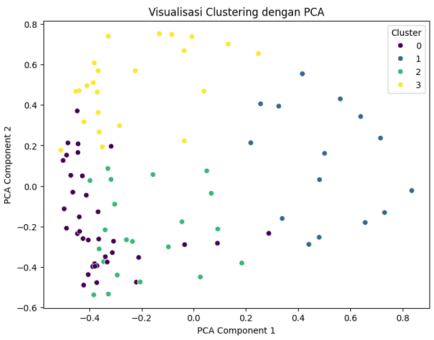
Titik optimal ditentukan berdasarkan bentuk grafik yang menyerupai siku (elbow), yaitu saat penurunan inertia mulai melambat. Dalam percobaan ini, titik elbow terlihat pada 4 *cluster*, yang menjadi indikasi awal jumlah klaster optimal yang akan digunakan dalam *clustering*.

2. K-Means Clustering

Algoritma ini termasuk dalam kategori unsupervised learning yang bertujuan untuk membagi data ke dalam sejumlah kelompok (klaster) berdasarkan kemiripan fitur antar data. Dalam implementasinya, digunakan parameter n_*cluster*s = 4 sebagai contoh pengujian awal, dengan nilai random_state = 42 untuk memastikan hasil klaster yang konsisten. Dari hasil *K-Means Clustering* dengan jumlah 4 kluster, pada *cluster* 1 berisi 74 data, *cluster* 2 berisi 111 data, *cluster* 3 berisi 215 data, dan *cluster* 4 berisi 100 data. Dari hasil *K-Means Clustering* dengan jumlah 4 *cluster*, diperoleh bahwa *cluster* ke-3 merupakan *cluster* paling dominan diantara *cluster-cluster* lain. Kemudian dari hasil analisis, ditemukan sebanyak 19 data yang termasuk outlier karena jaraknya signifikan lebih jauh dari pusat *cluster*, menunjukkan bahwa data tersebut menyimpang dari pola umum dalam kelompoknya.

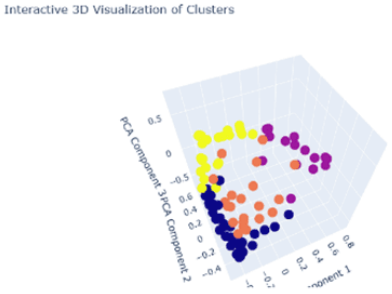
3. Grafik Penyebaran Data

Langkah selanjutnya adalah melakukan visualisasi penyebaran data untuk melihat bagaimana data dikelompokkan berdasarkan hasil klaster.




Gbr. 3 Grafik Penyebaran Data 2D K-Means

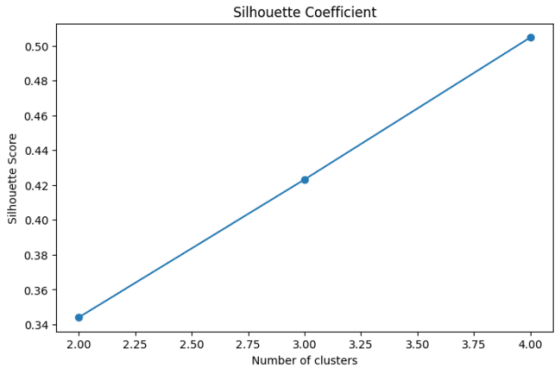
Setiap titik pada grafik merepresentasikan satu data keluhan masyarakat, sedangkan warna titik menunjukkan klaster tempat data tersebut dikelompokkan. Pewarnaan dilakukan berdasarkan hasil prediksi klaster oleh model *K-Means* yang telah dilakukan sebelumnya. Untuk memperjelas struktur klaster yang terbentuk, dan memberikan nilai tambah dalam interpretasi hasil klasterisasi maka dibentuk visualisasi tiga dimensi, karena grafik dapat di putar dan di eksplorasi tampilan data secara interaktif untuk memahami pola distribusinya secara lebih mendalam.




Gbr. 4 Grafik Penyebaran Data 3D K-Means

4. Silhouette Coefficient

Langkah selanjutnya adalah mengevaluasi kualitas dari hasil klaster menggunakan Silhouette Coefficient. Metode ini digunakan untuk melihat seberapa baik data yang ada berada dalam klaster yang sesuai. *Silhouette coefficient* memiliki nilai antara -1 hingga 1. Semakin mendekati angka 1, maka semakin baik data tersebut berada dalam kelompoknya. Hasil penghitungan kemudian divisualisasikan dalam bentuk grafik untuk memudahkan analisis. Sumbu X menunjukkan jumlah klaster yang dicoba, sedangkan sumbu Y menunjukkan nilai *Silhouette coefficient* yang didapat.




Gbr. 5 Grafik Silhouette Coefficient K-Means

Silhouette *Score* for 2 *cluster*s: 0.3440
Silhouette *Score* for 3 *cluster*s: 0.4232
Silhouette *Score* for 4 *cluster*s: 0.5048

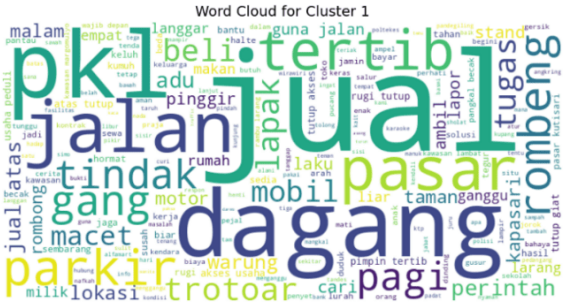
Nilai silhouette score tertinggi terdapat pada jumlah klaster sebanyak 4, yang menunjukkan pemisahan dan kohesi antar klaster yang paling optimal dibandingkan jumlah klaster lainnya. Hal ini juga konsisten dengan hasil elbow method yang menunjukkan titik optimal pada 4 klaster. Dengan demikian, dapat disimpulkan bahwa jumlah klaster yang paling tepat untuk data ini adalah sebanyak 4 klaster.

5. Word Cloud

Word Cloud ini bertujuan untuk menampilkan kata-kata yang paling sering muncul dalam masing-masing klaster, memberikan wawasan tentang elemen-elemen yang paling dominan dalam data.

**Word Cloud 1**

Word Cloud untuk Cluster 1 memperlihatkan bahwa klaster ini didominasi oleh aduan masyarakat yang berkaitan dengan aktivitas perdagangan informal yang dilakukan di ruang-ruang publik, khususnya oleh pedagang kaki lima (PKL).




Gbr. 6 Word Cloud Cluster 1 K-Means

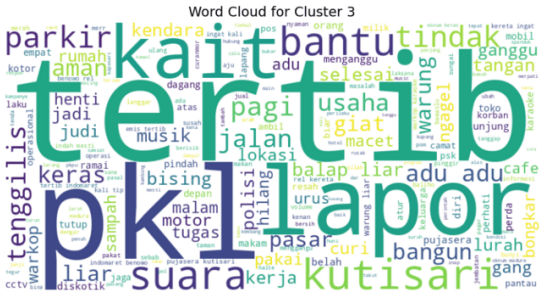
**Word Cloud 2**

Word Cloud untuk *Cluster* 2 menunjukkan bahwa aduan-aduan dalam klaster ini banyak berkaitan dengan aktivitas yang dianggap melanggar norma sosial dan ketertiban umum pada malam hari, khususnya di wilayah tertentu seperti area pemakaman atau tempat umum lainnya.


Gbr. 7 Word Cloud Cluster 2 K-Means

**Word Cloud 3**

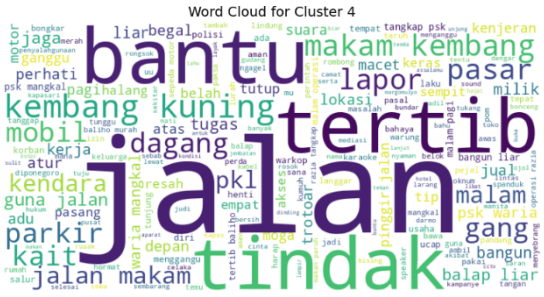
Word Cloud untuk *Cluster* 3 menggambarkan keluhan masyarakat mengenai pelanggaran ketertiban umum, baik berupa aktivitas PKL liar, gangguan suara, maupun penyalahgunaan ruang publik, yang memunculkan kebutuhan akan intervensi dan tindakan dari pihak berwenang.




Gbr. 8 Word Cloud Cluster 3 K-Means

**Word Cloud 4**

Word Cloud pada *Cluster* 4 menggambarkan aduan mengenai pelanggaran ketertiban umum di jalan dan area publik lainnya, termasuk gangguan sosial, dan pelanggaran lalu lintas.

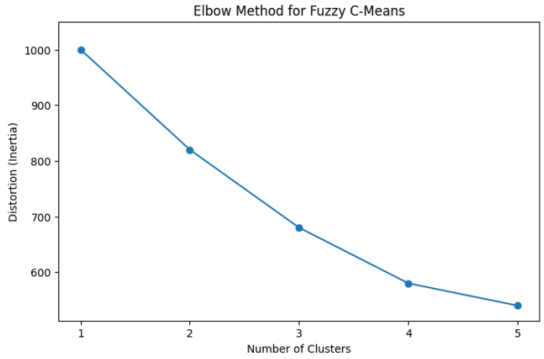



Gbr. 9 Word Cloud Cluster 4 K-Means

### C. *Fuzzy C-Means Clustering*

1. Elbow Method

Dalam tahap ini, dilakukan proses penentuan jumlah klaster optimal menggunakan metode *Elbow method*. Untuk menentukan jumlah klaster optimal, dilakukan perulangan klaster dari k = 1 hingga k = 4. Pada setiap iterasi, dilakukan pelatihan model dan dihitung nilai inertia (distorsi), yang kemudian disimpan untuk divisualisasikan dalam bentuk grafik. Sumbu X menunjukkan jumlah klaster yang diuji, sedangkan sumbu Y menunjukkan nilai distorsi masing-masing klaster.




Gbr. 10 Hasil Elbow Method Fuzzy C-Means

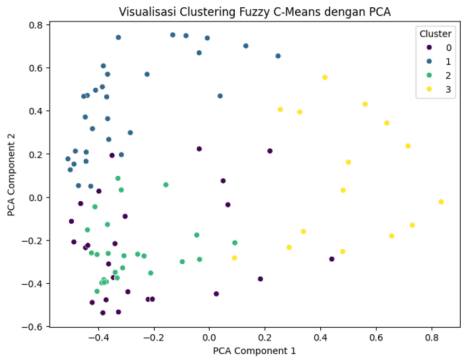
Titik optimal ditentukan berdasarkan bentuk grafik yang menyerupai siku (elbow), yaitu saat penurunan inertia mulai melambat. Dalam percobaan ini, titik elbow terlihat pada 4 *cluster*, yang menjadi indikasi awal jumlah klaster optimal yang akan digunakan dalam *clustering*.

2. Fuzzy C-Means Clustering

Algoritma ini bersifat metode *soft clustering* yang memperbolehkan setiap data memiliki tingkat keanggotaan (*membership*) di lebih dari satu klaster. Dari hasil Fuzzy C-Means Clustering dengan jumlah 4 kluster, berikut adalah jumlah data pada setiap cluster. Pada cluster 1 berisi 99 data, cluster 2 berisi 102 data, cluster 3 berisi 197 data, dan cluster 4 berisi 102 data. Dari hasil Fuzzy C-Means Clustering dengan jumlah 4 cluster, diperoleh bahwa cluster ke-3 merupakan cluster paling dominan diantara cluster-cluster yang lain dengan jumlah 197 data. Pada algoritma *Fuzzy C-Means* (FCM), outlier diidentifikasi berdasarkan tingkat keanggotaan *fuzzy* tiap data pada *cluster*. Hasil analisis menunjukkan terdapat 167 data yang termasuk outlier karena memiliki tingkat keanggotaan rendah, menandakan data tersebut ambigu dan tidak jelas masuk ke cluster manapun secara tegas.

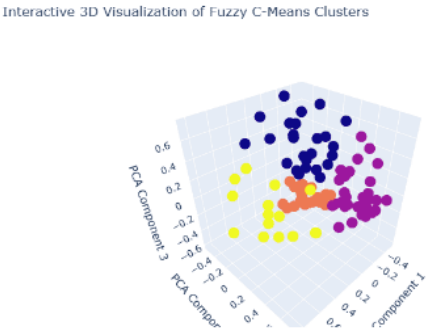
3. Grafik Penyebaran Data

Langkah selanjutnya adalah melakukan visualisasi penyebaran data untuk melihat bagaimana data dikelompokkan berdasarkan hasil klaster.




Gbr. 11 Grafik Penyebaran Data 2D Fuzzy C-Means

Setiap titik pada grafik merepresentasikan satu data keluhan masyarakat, sedangkan warna titik menunjukkan klaster tempat data tersebut dikelompokkan. Pewarnaan dilakukan berdasarkan hasil prediksi klaster oleh model *Fuzzy C-Means* yang telah dilakukan sebelumnya. Untuk memperjelas struktur klaster yang terbentuk, dan memberikan nilai tambah dalam interpretasi hasil klasterisasi maka dibentuk visualisasi tiga dimensi, karena grafik dapat di putar dan di eksplorasi tampilan data secara interaktif untuk memahami pola distribusinya secara lebih mendalam.

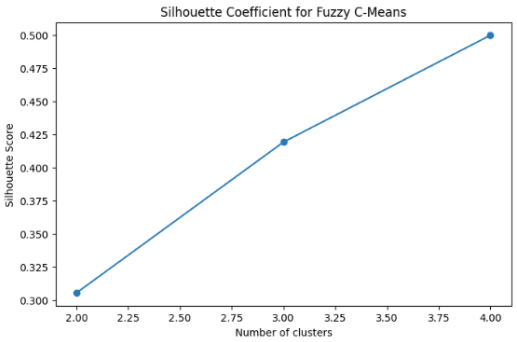



Gbr. 12 Grafik Penyebaran Data 3D Fuzzy C-Means

4. Silhouette Coefficient

Langkah selanjutnya adalah mengevaluasi kualitas dari hasil klaster menggunakan Silhouette Coefficient. Metode ini digunakan untuk melihat

seberapa baik data yang ada berada dalam klaster yang sesuai. *Silhouette coefficient* memiliki nilai antara -1 hingga 1. Semakin mendekati angka 1, maka semakin baik data tersebut berada dalam kelompoknya. Hasil penghitungan kemudian divisualisasikan dalam bentuk grafik untuk memudahkan analisis. Sumbu X menunjukkan jumlah klaster yang dicoba, sedangkan sumbu Y menunjukkan nilai *Silhouette coefficient* yang didapat.




Gbr. 13 Grafik Silhouette Coefficient Fuzzy C-Means

Silhouette Score for 2 clusters: 0.3055
Silhouette Score for 3 clusters: 0.4193
Silhouette Score for 4 clusters: 0.5000

Nilai silhouette score tertinggi terdapat pada jumlah klaster sebanyak 4, yang menunjukkan pemisahan dan kohesi antar klaster yang paling optimal dibandingkan jumlah klaster lainnya. Hal ini juga konsisten dengan hasil elbow method yang menunjukkan titik optimal pada 4 klaster. Dengan demikian, dapat disimpulkan bahwa jumlah klaster yang paling tepat untuk data ini adalah sebanyak 4 klaster.

5. Word Cloud

Word Cloud ini bertujuan untuk menampilkan kata-kata yang paling sering muncul dalam masing-masing klaster, memberikan wawasan tentang elemen-elemen yang paling dominan dalam data.

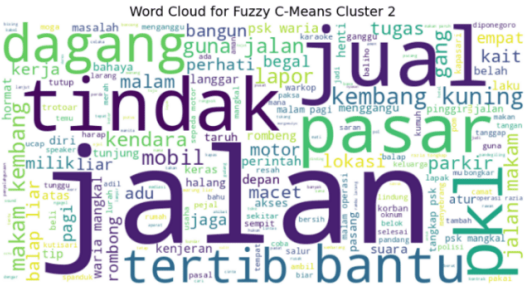
**Word Cloud 1**

Word Cloud untuk Cluster 1 merepresentasikan aduan masyarakat terkait pelanggaran penggunaan ruang publik oleh pedagang kaki lima dan parkir liar, yang dinilai mengganggu ketertiban, lalu lintas, serta kenyamanan lingkungan.


Gbr. 14 Word Cloud Cluster 1 Fuzzy C-Means

**Word Cloud 2**

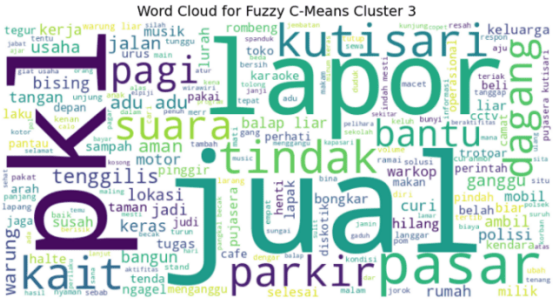
Berdasarkan *word cloud Cluster* 2 yang terbentuk, dapat disimpulkan bahwa topik utama dari aduan masyarakat dalam *cluster* ini berkaitan dengan ketertiban jalan dan pasar, termasuk permasalahan pedagang liar, parkir sembarangan, pelanggaran akses jalan, serta tuntutan terhadap Satpol PP untuk melakukan tindakan penertiban dan bantuan di lapangan.




Gbr. 15 Word Cloud Cluster 2 Fuzzy C-Means

**Word Cloud 3**

Word Cloud untuk *Cluster* 3 menggambarkan keluhan masyarakat terhadap aktivitas pedagang liar, kebisingan, gangguan lalu lintas, dan pelanggaran ketertiban umum yang terjadi di ruang publik, khususnya di lingkungan permukiman.




Gbr. 16 Word Cloud Cluster 3 Fuzzy C-Means

**Word Cloud 4**

Word Cloud pada *Cluster* 4 menggambarkan aduan masyarakat mengenai gangguan ketertiban yang dominan terjadi pada malam hari, seperti kebisingan dan aktivitas sosial yang dianggap menyimpang seperti prostitusi di kembang kuning.


Gbr. 17 Word Cloud Cluster 4 Fuzzy C-Means

D. *Perbandingan K-Means Clustering dan Fuzzy C-Means Clustering*

Dalam penelitian ini, kedua algoritma *clustering* yaitu *K-Means* dan *Fuzzy C-Means* (FCM) dibandingkan berdasarkan beberapa aspek penting, yaitu nilai *Silhouette coefficient*, waktu *preprocessing*, proses pemisahan data, performa *clustering*, dan deteksi *outlier.*

*TABEL II*
*PERBANDINGAN HASIL K-MEANS DAN FUZZY C-MEANS*

| Perbandingan | *K-Means* | *Fuzzy C-Means* |
|---|---|---|
| Nilai *Silhouette Coeffcient* | Nilai Silhouette *Score* untuk 2, 3, dan 4 *cluster* berturut-turut adalah 0.3440, 0.4232, dan 0.5048. *K-Means* memiliki nilai Silhouette yang sedikit lebih tinggi pada 4 *cluster* dibandingkan dengan FCM | Nilai Silhouette *Score* untuk 2, 3, dan 4 *cluster* berturut-turut adalah 0.3055, 0.4193, dan 0.5000. |
| Kecepatan *Preprocessing* | *Preprocessing* dilakukan sebelum proses *clustering* dan tidak bergantung pada metode *clustering* yang dipilih. | *Preprocessing* dilakukan sebelum proses *clustering* dan tidak bergantung pada metode *clustering* yang dipilih. |
| Pemisahan Data | Pemisahan data pada *K-Means* bersifat *hard clustering*, dimana setiap data pasti masuk ke satu *cluster* tertentu. | Pada FCM, pemisahan data bersifat *soft clustering*, sehingga setiap data memiliki keanggotaan pada beberapa *cluster* dengan nilai yang berbeda-beda. |
| Performa *Clustering* | *K-Means* menunjukkan performa yang lebih baik dalam mengelompokkan kata-kata kunci dari data keluhan masyarakat. Hal ini terlihat dari representasi kata dalam *word cloud* yang lebih dominan dan tidak saling tumpang tindih antar klaster, sehingga memudahkan dalam mengidentifikasi topik utama pada masing-masing *cluster*. | Pada *Fuzzy C-Means*, hasil *word cloud* cenderung lebih tumpang tindih karena pendekatan *soft clustering* menyebabkan satu data dapat memiliki keanggotaan pada lebih dari satu *cluster*, sehingga beberapa kata muncul di lebih dari satu klaster dan mengurangi keunikan dari masing-masing visualisasi. |
| Deteksi *Outlier* | *K-Means* mendeteksi outlier berdasarkan jarak data ke centroid *cluster*, menghasilkan 19 data outlier karena letaknya jauh dari pusat *cluster*. | *Fuzzy C-Means* mengidentifikasi 167 data outlier berdasarkan nilai keanggotaan *fuzzy* rendah, yang menunjukkan adanya data ambigu atau *noise*. |

Berdasarkan hasil perbandingan, *algoritma K-Means* terbukti lebih unggul dalam pengelompokan aduan masyarakat terhadap Satpol PP Kota Surabaya. Hal ini ditunjukkan oleh nilai *Silhouette Score* yang lebih tinggi yaitu 0.5048 dibandingkan *Fuzzy C-Means* 0.5000, visualisasi *word cloud* yang lebih jelas dan tidak tumpang tindih antar klaster, serta jumlah *outlier* yang lebih sedikit yaitu 19 data dibandingkan 167 data pada *Fuzzy C-Menas*. Selain itu, *K-Means* juga lebih cepat dalam proses *clustering* karena menggunakan pendekatan *hard clustering* yang lebih sederhana. Oleh karena itu, *K-Means* lebih direkomendasikan untuk pengelompokan data aduan masyarakat yang memerlukan hasil yang tegas dan efisien.

## IV. KESIMPULAN

Berdasarkan penelitian yang dilakukan mengenai penerapan *text mining* pada aplikasi WargaKu Surabaya terhadap aduan masyarakat kepada Satpol PP Kota Surabaya menggunakan algoritma *K-Means Clustering* dan *Fuzzy C-Means Clustering* (FCM), dapat ditarik kesimpulan bahwa :

1. Berdasarkan hasil penelitian, baik algoritma *K-Means* maupun *Fuzzy C-Means* (FCM) mampu mengelompokkan data keluhan masyarakat terhadap Satpol PP Kota Surabaya dengan cukup baik. *K-Means* unggul dalam hal kecepatan proses, kejelasan pemisahan data (*hard clustering*), dan menghasilkan visualisasi *word cloud* yang lebih fokus dan representatif. Sedangkan *Fuzzy C-Means* menghasilkan klaster yang tumpang tindih akibat pendekatan *soft clustering*, serta lebih sensitif dalam mendeteksi outlier berdasarkan derajat keanggotaan rendah. Dari segi evaluasi, *K-Means* menghasilkan

nilai Silhouette yang sedikit lebih tinggi 0.5048 dibanding *Fuzzy C-Means* 0.5000 pada jumlah *cluster* terbaik, yaitu empat *cluster*. Hal ini menunjukkan bahwa *K-Means Clustering* menghasilkan klaster yang lebih terpisah dan konsisten dibanding *Fuzzy C-Means Clustering*. Namun, keduanya masih termasuk dalam struktur yang baik.

2. Hasil *clustering* dan visualisasi Word Cloud dari masing-masing algoritma mampu menggambarkan tema-tema utama dalam aduan masyarakat. Pada algoritma *K-Means Clustering* klaster 1 berisi aduan mengenai aktivitas PKL atau perdagangan, klaster 2 gangguan suara dan prostitusi, klaster 3 aktivitas PKL di kutisari, dan klaster 4 mengenai jalan. Sedangkan pada *Fuzzy C-Means Clustering* klaster 1 berisi aduan mengenai gangguan di ruang publik oleh PKL dan parkir liar, klaster 2 mengenai pasar dan gangguan akses jalan, klaster 3 PKL, dan klaster 4 gangguan suara dan prostitusi di kembang kuning. Pada algoritma *K-Means cluster* dominan terletak pada *cluster* ke-3 dengan jumlah 215 data, sedangkan pada *Fuzzy C-Means* terletak pada *cluster* ke-3 dengan jumlah 197 data.

## V. SARAN

Berdasarkan dari penelitian ini, penulis memberikan saran untuk mengembangkan penelitian dalam masa yang akan datang sebagai berikut :

Untuk meningkatkan akurasi dan kualitas *clustering*, disarankan melakukan pengujian tambahan dengan indeks evaluasi lain seperti Davies-Bouldin Index atau Calinski-Harabasz *Score*.

Hasil *clustering* ini dapat dimanfaatkan oleh pihak Satpol PP Kota Surabaya untuk menyusun strategi penertiban yang lebih terfokus, seperti menetapkan prioritas waktu (misalnya malam hari) atau lokasi (seperti kawasan PKL, makam, dan taman kota) berdasarkan hasil klaster dominan.

Penelitian selanjutnya dapat dilakukan dengan menggunakan data aduan yang lebih besar dan mencakup periode waktu yang lebih panjang, serta eksplorasi algoritma *clustering* lainnya untuk membandingkan performa dan akurasi yang dihasilkan.

## UCAPAN TERIMA KASIH

Segala puji bagi Allah SWT yang telah memberikan rahmat dan petunjuk-Nya sehingga penulis berhasil menyelesaikan artikel yang berjudul ”Perbandingan Penerapan *Text Mining* pada Aplikasi Wargaku Surabaya untuk Aduan Masyarakat terhadap Satpol PP Kota Surabaya dengan Algoritma *K-Means Clustering* dan *Fuzzy C-Means Clustering*” dengan baik dan tepat waktu. Penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada pihak-pihak yang telah memberikan dukungan, bimbingan, dan motivasi selama proses penyelesaian artikel ini. Orang tua tercinta, atas doa, dukungan moral, dan materi yang tidak pernah putus selama masa studi hingga selesainya skripsi ini. Kepada Ibu Aries Dwi Indriyanti, S.Kom., M.Kom., sebagai Dosen Pembimbing, yang telah memberikan waktu, arahan, bimbingan, dan masukan yang sangat berharga dalam proses penelitian dan penulisan penlitian ini. Teman-teman terimakasih atas kebersamaan, diskusi, dan dukungan selama proses perkuliahan hingga penelitian ini dapat saya selesaikan dengan lancar sampai akhir.

## REFERENSI

[1] A. Rofpi and Tukiman, “Strategi Dinas Komunikasi dan Informatika dalam Mendukung Smart City Melalui Aplikasi Wargaku di Kota Surabaya,” *Journal of Governance Innovation*, vol. 6, no. 1, pp. 48–59, Mar. 2024, doi: 10.36636/jogiv.v6i1.4132.

[2] A. W. Wulandari, N. Widowati, and A. Subowo, “Manajemen Pengaduan Masyarakat Di Ombudsman Republik Indonesia Perwakilan Provinsi Jawa Tengah Di Kota Semarang,” *Journal of Public Policy and Management Review*, vol. 13, no. 2, pp. 1–10, 2024, doi: 10.14710/jppmr.v13i2.43800.

[3] D. Rifaldi, F. Abdul, and Herman, “Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet ‘Mental Health,’” *Decode: Jurnal Pendidikan Teknologi Informasi*, vol. 3, no. 2, pp. 161–171, Sep. 2023, doi: 10.51454/decode.v3i2.131.

[4] I. Mahmudi, A. D. Indriyanti, and I. Lazulfa, “PENERAPAN ALGORITMA K-MEANS CLUSTERING SEBAGAI STRATEGI PROMOSI PENERIMAAN MAHASISWA BARU PADA UNIVERSITAS HASYIM ASY’ARI JOMBANG,” *INOVATE*, vol. 04, 2020.

[5] I. N. S. N. Q. Maulidiyah and I. K. D. Nuryana, “Penerapan Fuzzy C-Means Untuk Manajemen Aplikasi Penyewaan Seni Reog Berbasis Android,” *Journal of Informatics and Computer Science*, vol. 03, 2022.

[6] M. D. E. Susanti and R. P. Wibawa, “Clustering Student Understanding Levels In Software Engineering Courses,” in *Proceedings of the 2023 Brawijaya International Conference (BIC 2023)*, 2024, pp. 693–703. doi: 10.2991/978-94-6463-525-6_76.

[7] M. I. Syafii, “Sentimen Analisis Pada Media Sosial Menggunakan Metode Naive Bayes Classifier (NBC),” *Jurna Teknologi Pintar*, vol. 3, no. 2, 2023

[8] N. A. Maori and Evanita, “METODE ELBOW DALAM OPTIMASI JUMLAH CLUSTER PADA K-MEANS CLUSTERING,” *Jurnal SIMETRIS*, vol. 14, 2023.

[9] R. Puspitasari and A. D. Indriyanti, “ANALISIS SENTIMEN OPINI PUBLIK TERHADAP KEBIJAKAN BARU SKRIPSI PADA MEDIA SOSIAL TWITTER MENGGUNAKAN METODE NAÏVE BAYES,” *Journal of Emerging Information Systems and Business Intelligence*, vol. 05, 2024.

[10] W. S. U. Saragih, N. A. Hasibuan, and R. K. Hondroo, “Penerapan Text Mining Dengan Menggunakan Metode TF-IDF Untuk Menentukan Genre Dari Komik,” *Konferensi Nasional Teknologi Informasi dan Kompute*,

vol. 4, no. 1, pp. 191–199, 2020, doi: 10.30865/komik.v4i1.2679.