# BAB I
# PENDAHULUAN

## 1.1. Latar Belakang

Analisis klaster merupakan sebuah teknik multivariat yang mempunyai tujuan utama yaitu untuk mengelompokkan objek-objek berdasarkan karakteristik yang dimilikinya. *Clustering* menjadi salah satu metode yang dapat mengungkap pola tersembunyi dalam data tanpa perlu menetapkan jumlah kelompok sejak awal. Analisis klaster dilakukan dengan mengelompokkan beberapa objek sehingga objek-objek yang memiliki kedekatatan kesamaan dengan objek lainnya yang berada dalam cluster yang sama [1]. Dalam konteks pengolahan data, analisis klaster dapat diterapkan pada data numerik maupun data teks. Pada data teks, analisis klaster sering kali melibatkan teknik *Text Mining*, yang mana merupakan teknik analisis teks dalam suatu dokumen melalui kategorisasi teks, pengelompokan teks, ekstraksi konsep, dan penghapusan text [2]. *Text Mining* merupakan proses ekstraksi sebuah informasi dari teks tidak terstruktur yang tujuan akhirnya untuk mengekstraksi informasi dan pola tersembunyi dari teks dalam jumlah besar. Proses ini melibatkan teknik *Natural Language Processing* (NLP), statistik, dan *machine learning* untuk menemukan pola dan pengetahuan yang tersembunyi dalam teks. *Text Mining* sangat bermanfaat dalam berbagai aplikasi, termasuk analisis teks, analisis sentimen, pengelompokan dokumen, dan ekstraksi informasi. Dalam konteks pelayanan publik, *Text Mining* dapat dimanfaatkan untuk mengolah teks aduan masyarakat agar keluhan-keluhan yang masuk dapat dirangkum menjadi kelompok tema tertentu. Namun, karena aduan ditulis bebas oleh masyarakat, proses pengolahan teks hingga klasterisasi perlu dirancang secara sistematis. Atas dasar itu, penelitian ini mengkaji bagaimana proses penerapan metode *Ward's Method*, DBSCAN, dan *K-Means* dalam melakukan klasterisasi terhadap data aduan masyarakat.

Penelitian ini berfokus pada penerapan analisis klaster dalam data teks, dengan membandingkan tiga metode klasterisasi, yaitu menggunakan *Agglomerative Hierarchical Clustering* (AHC) *Ward's Method*, *Density-Based*

*Spatial Clustering of Applications with Noise* (DBSCAN), dan *K-Means*. Setiap metode memiliki karakteristik dan keunggulannya masing-masing. *Ward's Method* merupakan metode pengelompokan hierarki dengan pendekatan bawah ke atas (*bottom-up*), di mana proses ini dilakukan secara berulang-ulang membentuk struktur jenjang (hierarki) [3]. *Ward's Method* bekerja dengan meminimalkan variansi total dalam satu cluster, sehingga menghasilkan kelompok yang lebih homogen [4]. Metode ini memungkinkan pengelompokan data berdasarkan kemiripan konten, sehingga dokumen yang serupa dapat dimasukkan ke dalam kelompok yang sama. Selain itu, DBSCAN digunakan sebagai metode clustering berbasis kepadatan [5], yang memungkinkan identifikasi cluster tanpa perlu menentukan jumlah klaster di awal. Berbeda dengan *Ward's Method* yang menggabungkan titik data berdasarkan kemiripan *centroid,* DBSCAN mendeteksi area dengan kepadatan tinggi dan memisahkannya dari data yang jarang muncul (*noise*). Hal ini menjadikan DBSCAN lebih fleksibel dalam mengelompokkan data aduan yang memiliki pola distribusi yang tidak seragam. Sementara itu, *K-Means* merupakan metode partitional clustering (*non-hierarki*) yang bekerja dengan membagi data ke dalam sejumlah k klaster berdasarkan kedekatan dengan *centroid*. *K-Means* memiliki keunggulan dalam efisiensi komputasi dan kemampuannya menangani dataset yang cukup besar [5]. Namun, metode ini memiliki tantangan yaitu memerlukan penentuan jumlah cluster k sejak awal. Oleh karena itu, dalam penelitian ini, akan digunakan *Variance Ratio Criterion* (VRC) untuk menentukan jumlah klaster optimal pada *K-Means* [6]. VRC mengukur rasio antara variansi antar-klaster (*Between-Group Sum of Squares* - BGSS) dan variansi dalam kalster (*Within-Group Sum of Squares* - WGSS), di mana nilai VRC yang lebih tinggi menunjukkan *clustering* yang lebih baik. Dengan dilakukan perbandingan *Ward's Method*, DBSCAN, dan *K-Means* dengan VRC, proses analisis klaster dapat memberikan hasil yang lebih optimal. *Ward's Method* digunakan untuk membentuk struktur hierarki yang homogen, DBSCAN untuk mendeteksi klaster berbasis kepadatan, dan *K-Means* dengan VRC untuk mendapatkan jumlah kalster optimal dalam skenario *non-hierarki.* Perbedaan prinsip kerja ketiga metode tersebut berpotensi menghasilkan *output* klasterisasi yang berbeda, baik dari sisi jumlah

klaster, anggota klaster, maupun tema/topik dominan yang terbentuk, sehingga perlu dianalisis bagaimana hasil klasterisasi data aduan menggunakan ketiga metode *Ward's Method*, DBSCAN, dan *K-Means*.

Kemudian, analisis klaster pada penelitian ini akan digunakan studi kasus tentang aduan pelayanan masyarakat di Dinas Kependudukan dan Pencatatan Sipil (Dispendukcapil) Kota Surabaya. 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). Kemudian, variabel yang akan digunakan selama penelitian merupakan variabel yang berisikan data teks aduan yang mencakup deskripsi keluhan dari masyarakat terkait dengan pelayanan yang ada di Dispendukcapil Kota Surabaya tanpa melibatkan metadata tambahan seperti nomor tiket, tanggal, identitas pelapor, tanggal ditanggapi, tanggal sudah ditanggap, dan tanggal selesai.

Beberapa penelitian sebelumnya terkait *Text Mining* telah banyak dilakukan. Pada penelitian sebelumnya yang dilakukan oleh [7], penelitian ini berfokus pada aplikasi *Text Mining* untuk klasterisasi aduan masyarakat di Kota Semarang menggunakan algoritma *K-Means*. Penelitian ini menyatakan bahwa sistem yang dikembangkan mampu mengelompokkan aduan dengan akurasi yang tinggi, hasil evaluasi menunjukkan bahwa nilai *Purity* yang diperoleh adalah 1 (atau 100%), yang berarti bahwa setiap data aduan berhasil dikelompokkan ke dalam klaster yang tepat tanpa kesalahan. Hasil ini menunjukkan bahwa metode *K-Means* efektif dalam mengelompokkan aduan sesuai dengan kategori yang relevan, meskipun proses pengelompokan memerlukan waktu dan sumber daya yang cukup.

Selanjutnya, penelitian lain yang dilakukan oleh [8] menerapkan analisis kluster dengan menggunakan *Complete Linkage* dan *Ward's Method* untuk data layanan kesehatan di Kota Makassar. Penelitian ini menemukan bahwa penggunaan *Ward's Method* menunjukkan klasterisasi yang lebih stabil dan efisien dibandingkan *Complete Linkage*, yang mana *Ward's Method* lebih baik dalam meminimalkan varians antar objek dalam klaster. Penelitian ini menghasilkan tiga klaster yang jelas, terdapat klaster dengan layanan kesehatan baik, klaster dengan

layanan kesehatan sangat baik, dan klaster dengan layanan kesehatan buruk. Nilai *Davies-Bouldin Index* (IDB) yang rendah menunjukkan bahwa kluster yang terbentuk memiliki pemisahan yang baik dan kompak, sehingga memudahkan pengambilan keputusan terkait pengelolaan layanan publik.

Selanjutnya, penelitian yang dilakukan oleh [5] membahas penerapan *Text Mining* untuk *clustering* tweet pengguna layanan ekspedisi (JNE, J&T, dan Pos Indonesia) menggunakan metode DBSCAN dan *K-Means*. Hasil penelitian menunjukkan bahwa DBSCAN lebih unggul dibandingkan *K-Means* dalam mengelompokkan *tweet* pelanggan layanan ekspedisi, di mana DBSCAN menghasilkan 18 cluster untuk JNE dengan *Silhouette Coefficient* 0.2652, 22 cluster untuk J&T dengan *Silhouette Coefficient* 0.9999, dan 11 *cluster* untuk Pos Indonesia dengan *Silhouette Coefficient* 0.9975, sedangkan *K-Means* hanya menghasilkan 2 *cluster* dengan *Silhouette Coefficient* yang jauh lebih rendah. Dengan demikian, penelitian ini menyimpulkan bahwa DBSCAN lebih cocok untuk *clustering* data berbasis teks yang memiliki variasi tinggi, seperti opini pelanggan di Twitter, karena lebih efektif dalam menangkap pola dan membentuk klaster yang lebih kompak dibandingkan metode *K-Means*.

Kemudian, penelitian oleh [9] menciptakan taksonomi untuk ancaman insider yang tidak disengaja melalui *Text Mining* dan analisis klaster hierarkis. Penelitian ini menghasilkan lima kategori utama untuk aktivitas ancaman insider, dengan koefisien aglomerasi *Ward's Method* untuk kategori kesalahan aplikasi mencapai 0.787, yang menunjukkan validitas klaster yang tinggi. Penelitian ini menyoroti pentingnya pengembangan taksonomi yang jelas untuk memahami dan mengelola ancaman yang tidak disengaja dalam sistem informasi.

Adapun juga, penelitian oleh [10] menerapkan metode *clustering Text Mining* untuk pengelompokan berita pada data teks tidak terstruktur. Hasil penelitian menunjukkan bahwa algoritma *K-Means* mampu mengelompokkan berita dengan akurasi yang memuaskan, dengan rata-rata *Precision* sebesar 76.92%, *Recall* 79.58% dan *Purity* 0.83. Penelitian ini menekankan pentingnya penggunaan teknik *Text Mining* dalam mengelola informasi yang tidak terstruktur, sehingga memudahkan dalam pencarian dan pengelompokan berita.

Meskipun penelitian mengenai *Text Mining* dan klasterisasi dengan berbagai metode telah banyak dilakukan, studi yang secara khusus membandingkan *Ward's Method*, DBSCAN, dan *K-Means* dalam konteks klasterisasi pengaduan layanan administrasi kependudukan di Indonesia yang masih sangat terbatas. Keterbatasan ini menciptakan peluang bagi penulis untuk mengisi celah penelitian, terutama dalam mengeksplorasi *metode clustering* yang lebih adaptif dan akurat dalam mengelompokkan aduan masyarakat secara otomatis. Penelitian ini mengusulkan pendekatan baru dalam pengelolaan aduan masyarakat dengan membandingkan tiga metode klasterisasi, yaitu *Ward's Method*, DBSCAN, dan *K-Means*, yang masing-masing memiliki keunggulan dan keterbatasan tersendiri. *Ward's Method* dipilih karena kemampuannya dalam menghasilkan klaster yang lebih homogen dengan meminimalkan variansi total dalam setiap klaster, sedangkan DBSCAN digunakan untuk menangkap pola berbasis kepadatan tanpa perlu menentukan jumlah klaster di awal. Sementara itu, *K-Means* menjadi metode pembanding yang digunakan dalam penelitian sebelumnya [7] dan [10], meskipun memiliki keterbatasan dalam sensitivitas terhadap inisialisasi *centroid* dan perlunya penentuan jumlah klaster secara manual. Oleh karena itu, dalam penelitian ini, akan digunakan *Variance Ratio Criterion* (VRC) untuk membantu menentukan jumlah klaster optimal pada *K-Means* [6]. Dengan pendekatan ini, penelitian ini bertujuan untuk menentukan metode terbaik dalam mengelompokkan aduan masyarakat, serta mengukur efektivitas masing-masing metode dalam mengungkap pola atau tema utama dari keluhan. Selain membandingkan hasil pengelompokan, penelitian ini juga perlu memastikan bahwa klaster yang terbentuk benar-benar berkualitas. Oleh karena itu, validitas klaster dievaluasi menggunakan *Silhouette Coefficient* dan *Index Davies-Bouldin* (IDB) guna memastikan bahwa kelompok yang terbentuk memiliki pemisahan yang jelas dan kompak. *Silhouette Coefficient* digunakan untuk mengukur kekompakan dalam klaster (*intra-cluster compactness*), sementara IDB digunakan untuk mengukur pemisahan antar klaster (*inter-cluster separation*).

Kemudian, dalam upaya mempermudah analisis dan meningkatkan penerapan metode klasterisasi dalam pelayanan publik, penelitian ini juga mengusulkan pengembangan *Graphical User Interface* (GUI) untuk

memvisualisasikan hasil klasterisasi aduan masyarakat. Sistem berbasis GUI dirancang agar lebih intuitif dan *user-friendly* [11], sehingga memudahkan pengguna dalam mengakses informasi terkait pola atau tema utama dari aduan yang telah dikelompokkan. Dengan penerapan GUI berbasis web, hasil analisis dapat meningkatkan efisiensi dalam pengelolaan aduan masyarakat.

Urgensi penelitian ini terletak pada eksplorasi dan pengembangan metodologi klasterisasi berbasis *Ward's Method*, DBSCAN, dan *K-Means* dalam *Text Mining* untuk analisis aduan masyarakat. Dalam berbagai penelitian sebelumnya, metode klasterisasi yang digunakan dalam pengelompokan data teks masih didominasi oleh algoritma *K-Means*, yang memiliki beberapa keterbatasan, seperti sensitivitas terhadap inisialisasi *centroid* dan perlunya menentukan jumlah klaster secara manual. Oleh karena itu pada penelitian ini digunakan *Variance Ratio Criterion* (VRC) untuk membantu menentukan jumlah klaster optimal *pada K-Means*. Kemudian, *Ward's Method* dipilih sebagai metode hierarki yang lebih adaptif, dengan kemampuan dalam meminimalkan variansi total dalam setiap klaster, tanpa memerlukan menenentuan jumlah klaster sejak awal. Selain itu, DBSCAN dipilih sebagai metode berbasis kepadatan, yang lebih fleksibel dalam menangani data dengan distribusi tidak seragam dan keberadaan *outlier*, yang sering ditemukan dalam dataset aduan masyarakat. Untuk meningkatkan keakuratan hasil klasterisasi, penelitian ini juga menerapkan preprocessing teks yang mencakup *cleaning, tokenizing, normalization, stopword removal, stemming*. Kemudian dilakukan juga pembobotan TF-IDF. Evaluasi model dilakukan menggunakan *Silhouette Coefficient* dan *Index Davies-Bouldin* (IDB) untuk memastikan kualitas klaster yang terbentuk memiliki pemisahan yang jelas dan kompak. Dengan pendekatan metodologis ini, penelitian ini tidak hanya berkontribusi dalam penerapan *Text Mining* untuk pengelompokan aduan, tetapi juga dalam peningkatan efektivitas algoritma klasterisasi dalam data teks tidak terstruktur. Hasil dari penelitian ini diharapkan dapat menjadi dasar dalam pemilihan metode klasterisasi yang lebih stabil dan efisien di bidang analisis teks, sekaligus membuka peluang penelitian lanjutan dalam eksplorasi parameter dan

optimasi model *Ward's Method*, DBSCAN, dan *K-Means* dalam berbagai konteks analisis teks lainnya.

## 1.2. Rumusan Masalah

Rumusan masalah yang menjadi fokus utama pada penelitian ini adalah sebagai berikut:

1. Bagaimana tahapan penerapan *Ward's Method*, DBSCAN, dan *K-Means* dalam melakukan klasterisasi terhadap data aduan masyarakat?
2. Bagaimana hasil dari klasterisasi data aduan masyarakat menggunakan *Ward's Method*, DBSCAN, dan K-Means?
3. Bagaimana validitas metode klasterisasi *Ward's Method*, DBSCAN, dan *K-Means* dievaluasi menggunakan *Silhouette Coefficient* dan *Index Davies-Bouldin*?

## 1.3. Batasan Masalah

Batasan yang ditetapkan pada masalah yang diamati pada penelitian ini adalah sebagai berikut:

1. Data yang digunakan dalam penelitian ini adalah data aduan pelayanan masyarakat yang diperoleh dari Dinas Kependudukan dan Pencatatan Sipil Kota Surabaya selama periode Januari 2023 hingga Desember 2024 (2 tahun).
2. Penelitian ini berfokus untuk membandingkan penerapan *Text Mining* dengan metode klasterisasi *Ward's Method*, DBSCAN, dan *K-Means* untuk mengelompokkan data aduan masyarakat berdasarkan kesamaan teks dalam deskripsi keluhan. Metadata tambahan seperti nomor tiket, tanggal pengaduan, identitas pelapor, dan status penyelesaian tidak digunakan dalam proses klasterisasi.
3. Evaluasi kinerja metode klasterisasi dalam penelitian ini dilakukan menggunakan metrik *Silhouette Coefficient* dan *Index Davies-Bouldin.*
4. Pengaplikasian GUI dalam penelitian ini berbasis website dan hanya dapat dijalankan pada *local device.*

## 1.4. Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk menjawab rumusan masalah serta memberikan arahan yang lebih jelas bagi penelitian ini. Adapun tujuan penelitian ini adalah sebagai berikut:

1. Menerapkan tahapan klasterisasi berbasis *text mining* dengan *Ward's Method*, DBSCAN, dan *K-Means* pada data aduan pelayanan masyarakat di Dinas Kependudukan dan Pencatatan Sipil Kota Surabaya sehingga menghasilkan pembagian klaster aduan berdasarkan kemiripan teks.
2. Menganalisis hasil klasterisasi yang terbentuk dari ketiga metode tersebut untuk mengidentifikasi pola dan karakteristik data aduan pelayanan masyarakat.
3. Mengevaluasi validitas metode klasterisasi yang dikembangkan dengan menggunakan *Silhouette Coefficient* dan *Index Davies-Bouldin* untuk menilai tingkat kekompakan dan pemisahan antar klaster, sehingga dapat ditentukan metode klasterisasi yang paling optimal dalam mengelompokkan aduan masyarakat.

## 1.5. Manfaat Penelitian

Manfaat penelitian yang dapat diperoleh dari hasil penelitian ini adalah sebagai berikut:

1. Bagi bidang keilmuan

   Penelitian ini diharapkan dapat berkontribusi dalam pengembangan *Text Mining* dan data *clustering*, khususnya dalam penerapan metode *Ward's Method*, DBSCAN, dan *K-Means* pada data aduan pelayanan publik. Dengan mengeksplorasi keunggulan dan keterbatasan masing-masing metode, penelitian ini dapat memberikan wawasan baru mengenai efektivitas clustering berbasis hierarki, kepadatan, dan *centroid* dalam menganalisis data teks tidak terstruktur. Hasil penelitian ini dapat memperkaya literatur terkait pengelompokan aduan masyarakat, serta menjadi referensi dalam pengembangan metode klasterisasi berbasis teks untuk berbagai aplikasi di sektor pelayanan publik lainnya.

Penelitian ini dapat menjadi referensi dan landasan yang berguna bagi penelitian selanjutnya di bidang *Text Mining, clustering*, dan pelayanan publik. Peneliti berikutnya dapat mengembangkan penelitian ini dengan menerapkan metode analisis lain atau memperluas penggunaan aplikasi ini ke sektor publik lainnya. Selain itu, penelitian ini juga membuka peluang untuk menggabungkan berbagai teknik pemrosesan teks yang lebih kompleks dalam menganalisis keluhan masyarakat

2. Bagi masyarakat

   Penelitian ini diharapkan dapat memberikan manfaat langsung bagi masyarakat sebagai pengguna layanan Dispendukcapil Kota Surabaya. Dengan adanya pengelompokan aduan masyarakat secara sistematis dan berbasis data, permasalahan pelayanan yang sering dikeluhkan dapat diidentifikasi dengan lebih cepat dan tepat. Hal ini memungkinkan pihak terkait untuk melakukan perbaikan layanan secara lebih terarah sesuai dengan kebutuhan dan keluhan yang dominan di masyarakat.

   Selain itu, hasil penelitian ini dapat mendorong terciptanya pelayanan publik yang lebih responsif, transparan, dan akuntabel. Masyarakat akan memperoleh pengalaman pelayanan yang lebih baik karena keluhan yang disampaikan tidak hanya ditangani secara individual, tetapi juga dianalisis secara menyeluruh untuk menemukan pola permasalahan yang berulang. Dengan demikian, potensi terjadinya keluhan serupa di masa mendatang dapat diminimalkan.

   Penelitian ini juga diharapkan dapat meningkatkan kepercayaan masyarakat terhadap instansi pelayanan publik, karena aduan yang disampaikan benar-benar dimanfaatkan sebagai dasar pengambilan keputusan dan perbaikan kebijakan. Pada akhirnya, masyarakat akan merasakan peningkatan kualitas pelayanan administrasi kependudukan yang lebih efektif, efisien, dan sesuai dengan harapan publik.

3. Bagi Pemerintah

   Hasil penelitian ini diharapkan dapat membantu Dispendukcapil Kota Surabaya dalam meningkatkan kualitas pelayanan publik melalui pengelompokan aduan masyarakat yang lebih efisien dan berbasis data.

Dengan adanya sistem klasterisasi, instansi terkait dapat lebih cepat dalam mengidentifikasi pola keluhan yang sering terjadi. Dengan demikian, strategi peningkatan layanan dapat dirancang dengan lebih efektif, sehingga dapat meningkatkan kepuasan dan kepercayaan masyarakat terhadap pelayanan publik.