

BAB I

PENDAHULUAN

Bab ini membahas latar belakang, rumusan masalah, tujuan dan manfaat, ruang lingkup, serta sistematika penulisan skripsi mengenai klasterisasi judul berita menggunakan metode *K-Means*.

1.1. Latar Belakang

Era digital telah membawa pertumbuhan pesat dalam jumlah data informasi online, termasuk berita. Volume berita yang terus meningkat pesat membutuhkan teknologi yang canggih untuk pengelolaan dan analisis informasi secara efisien. Selama ini, pengelompokan kategori berita dilakukan dengan menentukan kategorinya satu per satu berdasarkan keputusan dan pengamatan editor berita, yang memerlukan proses manual yang memakan waktu dan tidak skalabel. Oleh karena itu, teknologi yang dapat mengotomatiskan pengelompokan berita menjadi sangat diperlukan untuk meningkatkan efisiensi dan akurasi dalam industri berita (Claudia, 2018)

Klasterisasi adalah teknik penting dalam *data mining* yang digunakan untuk mengelompokkan data berdasarkan kesamaan fitur. Metode klasterisasi ini membantu dalam mengidentifikasi struktur tersembunyi atau pola dalam data besar dengan cara yang terstruktur dan otomatis. Dalam konteks berita, klasterisasi dapat digunakan untuk mengelompokkan berita berdasarkan topik atau tema, memudahkan pengguna dalam menemukan informasi yang relevan (Chandra dkk, 2019).

Metode klasterisasi yang telah digunakan dalam klasterisasi judul berita sebelumnya adalah *Agglomerative Hierarchical Clustering (AHC)*. Penelitian menunjukkan bahwa *AHC* mampu mengidentifikasi topik berita namun kurang efisien untuk dataset besar karena metode ini tidak dapat menentukan jumlah klaster di awal proses klasterisasi dan sensitif terhadap outlier, yang mana metode ini menginisialisasi semua titik sebagai klaster tersendiri di awal proses klasterisasi (Sulaiman & Alquliyev, 2019). Kelemahan ini menjadikan *AHC* kurang ideal untuk skala data yang terus berkembang dalam industri berita.

Penelitian yang menggunakan *K-Means* dan *TF-IDF* menunjukkan bahwa metode ini dapat mengatasi beberapa kelemahan *AHC*. *K-Means* menawarkan pendekatan yang lebih fleksibel karena mampu menentukan jumlah kluster diawal, sementara *TF-IDF* meningkatkan representasi fitur teks dengan memberikan bobot lebih pada kata-kata yang jarang muncul di seluruh dokumen (Irawan & Suryani, 2019). Meskipun *K-Means* dengan *TF-IDF* memberikan hasil yang lebih baik dalam hal fleksibilitas, namun penelitian ini masih menggunakan representasi data berita lengkap yang mencakup seluruh artikel, sehingga memori yang diperlukan menjadi lebih besar, terutama untuk dataset yang besar.

Penelitian oleh Anita & Oliandi (2022), mengaplikasikan metode *K-Means* dengan metode *Elbow* dalam klasterisasi judul berita di Twitter. Metode *K-Means* yang digunakan menawarkan pendekatan yang efisien untuk pengelompokan berita dengan waktu komputasi yang relatif singkat. Metode *Elbow* digunakan untuk menentukan jumlah kluster yang optimal dengan mengidentifikasi titik di mana penurunan dalam kualitas klasterisasi mulai melambat. Penelitian ini menunjukkan bahwa kombinasi kedua metode ini tidak hanya meningkatkan kualitas klasterisasi tetapi juga memberikan fleksibilitas dalam penyesuaian jumlah kluster yang dibutuhkan. Hasil penelitian ini menegaskan bahwa metode *K-Means* yang dipadukan dengan metode *Elbow* adalah solusi yang efektif untuk mengelompokkan berita, terutama dalam konteks data yang besar dan beragam, sehingga menghasilkan pola yang jelas dan relevan.

Penelitian untuk menerapkan metode *K-Means* dalam klasterisasi judul berita yang dipadukan dengan *TF-IDF* dan metode *Elbow* penting karena menawarkan fleksibilitas dalam inisialisasi kluster diawal, efisiensi dalam menangani dataset besar karena metode *K-Means* hanya perlu menyimpan informasi tentang *centroid* dan jarak terhadap *centroid*, sedangkan *AHC* membutuhkan lebih banyak memori karena harus menyimpan seluruh matriks jarak antar semua elemen, dan menawarkan efektivitas dalam waktu klasterisasi karena pendekatan metode *K-Means* yang berbasis iterasi, sedangkan pada *AHC*, semua elemen data awalnya dianggap sebagai kluster individu dan proses penggabungan berlangsung hingga semua kluster bergabung. Teknik pra-pemrosesan teks yang canggih dapat mengatasi tantangan semantik dan sintaktik. Dengan fokus pada pengelompokan judul berita, diharapkan

penelitian ini mampu menghasilkan klaster dengan pola yang jelas dapat memberikan kontribusi signifikan dalam pengembangan sistem rekomendasi berita dan aplikasi pengelolaan informasi.

1.2. Rumusan Masalah

Rumusan masalah pada penelitian ini adalah bagaimana klasterisasi judul berita menggunakan metode *K-Means*?

1.3. Tujuan dan Manfaat

Tujuan pada penelitian ini adalah mengetahui kualitas metode *K-Means* terhadap data judul berita dalam menampilkan suatu pola pada setiap klaster.

Manfaat pada penelitian ini adalah sebagai berikut :

1. Dapat menjadi acuan dalam pemilihan berita kesukaan atau menyesuaikan kategori berita itu sendiri.
2. Dapat menjadi referensi untuk penelitian selanjutnya terkait bidang klasterisasi dengan data teks.

1.4. Ruang Lingkup

Ruang lingkup dalam penelitian ini adalah sebagai berikut.

1. Melakukan klasterisasi data judul berita menggunakan metode *K-Means*
2. Penentuan jumlah klaster paling optimal menggunakan metode *Elbow*.
3. Vektorisasi kata dari data judul berita menggunakan metode *TF-IDF*.
4. Data yang digunakan dalam penelitian ini merupakan data judul berita pada situs *detik.com* pada tahun 2020. Pengumpulan data dilakukan dengan menggunakan dataset yang diunduh dari GitHub, yang terdiri dari 12.180 data judul berita dari tahun 2020 pada situs *detik.com*.
5. Batasan dalam penelitian adalah keterbatasan jumlah data judul berita, kompleksitas linguistik dalam judul berita, dan tantangan dalam menentukan parameter yang optimal untuk metode *K-Means*.

1.5. Sistematika Penulisan

Sistematika penulisan skripsi yang disusun untuk memberikan gambaran yang runtut dan jelas mengenai penyusunan skripsi ini meliputi beberapa bagian utama.

BAB I PENDAHULUAN

Bab ini menyajikan latar belakang masalah, rumusan masalah, batasan masalah, tujuan, manfaat, serta sistematika penulisan laporan penelitian ini.

BAB II TINJAUAN PUSTAKA

Bab ini menyajikan dasar teori dan referensi yang digunakan sebagai dasar untuk melakukan klasterisasi *K-Means* menggunakan metode *Elbow* untuk klasterisasi judul berita.

BAB III METODE PENELITIAN

Bab ini menguraikan tentang metode serta langkah kerja yang dilakukan dalam melakukan klasterisasi judul berita menggunakan metode *K-Means* dengan vektorisasi kata menggunakan *TF-IDF* dan penentuan jumlah klaster menggunakan metode *Elbow*, mulai dari garis besar penyelesaian masalah, pengumpulan data, pra-pemrosesan data, proses klasterisasi, hingga proses evaluasi.

BAB IV HASIL DAN ANALISIS

Bab ini menyajikan tentang implementasi metode *K-Means* dengan penentuan jumlah klaster menggunakan metode *Elbow* dalam melakukan klasterisasi judul berita, disertai dengan proses, hasil, dan pengujian dari implementasi klasterisasi.

BAB V PENUTUP

Bab ini menyajikan kesimpulan dari bab-bab yang telah dibahas sebelumnya dalam skripsi ini, serta saran-saran untuk pengembangan penelitian di masa mendatang.