BAB I
PENDAHULUAN

Bagian ini berisi tentang latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, ruang lingkup penelitian, dan sistematika penulisan.

1.1 Latar Belakang Masalah

Seiring dengan meningkatnya penggunaan internet di era digital ini, jumlah dokumen digital meningkat secara eksponensial sehingga proses pengolahan dokumen menjadi kebutuhan yang tidak dapat dipisahkan lagi. *Text clustering* merupakan sebuah cara pengelompokkan teks digital berdasarkan karakteristiknya. *Text clustering* diterapkan pada beberapa ranah aplikasi seperti pengorganisasian hasil yang dikembalikan oleh mesin pencarian sebagai tanggapan atas *query* pengguna [1], menelusuri dokumen dalam koleksi yang besar [2], pendeteksian topik [3], menghasilkan hierarki dokumen-dokumen yang ada pada web [4].

Banyak metode klasterisasi yang telah diusulkan oleh para ahli. Akan tetapi, pada tugas akhir ini penulis hanya akan membandingkan dua metode klasterisasi yatu metode *k-means* dengan metode DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) untuk melihat metode mana yang lebih baik dalam pengelompokkan teks.

Selain pemilihan metode klasterisasi, dimensi ruang fitur yang tinggi juga merupakan salah satu masalah utama yang harus diperhatikan dalam proses pengelompokkan teks. Semakin besar dokumen, maka akan menghasilkan fitur yang semakin banyak, ratusan bahkan ribuan fitur. Jumlah fitur yang banyak dapat membuat tingginya komputasi, selain itu jika terdapat banyak fitur yang tidak relevan, dapat menyebabkan performa yang buruk dari algoritma *clustering*. Salah satu cara untuk mengatasi dimensi fitur yang tinggi adalah dengan reduksi dimensi [5].

Berbagai metode reduksi dimensi telah diperkenalkan pada berbagai literatur untuk memilih atau menyeleksi subfitur yang informatif. Masing-masing metode reduksi menggunakan strategi yang berbeda-beda untuk memilih subfitur, hasil yang didapat pun akan berbeda meskipun *dataset*-nya sama. Oleh karena itu,

pendekatan *hybrid* yang meliputi strategi yang berbeda dalam memilih subfitur mendapat banyak perhatian. Biasanya, metode *union* dan metode *intersection* digunakan untuk menggabungkan subfitur yang dipilih dengan metode reduksi yang berbeda. *Union* memilih (menggabungkan) semua fitur dan *intersection* hanya memilih fitur umum (irisan) dari fitur-fitur yang dipertimbangkan, sehingga pendekatan *union* menyebabkan terjadinya peningkatan dimensi fitur dan pendekatan *intersection* menyebabkan hilangnya beberapa fitur penting. Maka dari itu, untuk mengambil kelebihan dari suatu metode dan mengurangi kelemahannya, diajukan pendekatan baru yakni *modified union*. Pendekatan ini menerapkan metode *union* untuk memilih fitur-fitur peringkat atas dan menerapkan metode *intersection* pada sisa fitur lainnya [6].

Pada kasus ini, *feature selection* menggunakan metode *term variance* (tv) dan metode *document frequency* (df) untuk menghitung nilai relevansi tiap fitur. Selanjutnya, *feature extraction* menggunakan metode *principal component analysis* (PCA) untuk mengurangi dimensi ruang fitur tanpa kehilangan banyak informasi.

1.2 Rumusan Masalah

Berdasarkan latar belakang diatas, maka dalam skripsi ini dapat dibuat rumusan masalah sebagai berikut:

1. Bagaimana cara mereduksi fitur-fitur dengan mengintegrasikan metode *feature selection* dan *feature extraction*?
2. Bagaimana cara melakukan klasterisasi dengan menggunakan metode *k-means* dan DBSCAN?
3. Bagaimana perbandingan hasil klasterisasi menggunakan metode *k-means* dengan metode DBSCAN?

1.3 Batasan Masalah

Adapun batasan masalah dalam skripsi ini sebagai berikut:

1. *Dataset* yang digunakan yaitu berupa terjemahan Hadits Shahih Bukhari, Sahih Muslim, Abu-Dawud, dan Malik's Muwatta sebanyak 892 hadits dalam bahasa Inggris. Terdapat lima kategori hadits yaitu: *Adzan, Wudlu, Zakat, Knowledge*, dan *Tawheed*.

2. Metode yang digunakan untuk mereduksi fitur adalah metode *term variance* (tv) dan metode *document frequency* (df) pada *feature selection*, kemudian metode *principal component analysis* (PCA) pada *feature extraction*.
3. Untuk menggabungkan fitur yang telah direduksi, digunakan metode *modified union.*
4. Metode klasterisasi yang digunakan adalah metode *k-means* dan metode DBSCAN.
5. Metode yang digunakan untuk mengevaluasi hasil klaster adalah metode Davies-Bouldin Index (DBI) dan *Silhoutte Coefficient* (SC).

1.4 Tujuan Penelitian

Tujuan dari skripsi ini adalah sebagai berikut:

1. Memahami bagaimana cara mereduksi fitur-fitur dengan mengintegrasikan metode *feature selection* dan *feature extraction.*
2. Memahami bagaimana cara melakukan klasterisasi dengan menggunakan metode *k-means* dan DBSCAN.
3. Mengetahui bagaimana perbandingan hasil klasterisasi menggunakan metode *k-means* dengan metode DBSCAN.

1.5 Metode Penelitian

Metode yang penulis tempuh dalam menyelesaikan skripsi ini adalah menggunakan pendekatan teoritis atau studi literatur. Dengan cara mencari dan mengumpulkan data berupa informasi yang mendukung pengerjaan skripsi ini, yaitu mengenai *text clustering*, *feature selection*, *feature extraction*, *term variance*, *document frequency*, *principal component analysis*, *k-means*, dan DBSCAN. Sumber-sumber tersebut dapat diperoleh dari sumber yang berbentuk artikel, jurnal, skripsi, buku, dan yang lainnya. Kemudian melakukan pengkajian dan analisis terhadap sumber-sumber yang berkaitan dengan skripsi ini. Adapun dalam penerapan model terhadap data studi kasus dilakukan menggunakan data sekunder yang diperoleh peneliti dari sumber yang sudah ada.

1.6 Sistematika Penulisan

Berdasarkan sistematika penulisan, skripsi ini terdiri atas lima bab serta daftar pustaka di mana dalam setiap bab terdapat beberapa subbab.

BAB I PENDAHULUAN

Bagian ini mengemukakan beberapa hal mengenai pendahuluan yang mendukung dalam penulisan skripsi ini. Pendahuluan tersebut berupa latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, ruang lingkup penelitian, dan sistematika penulisan dari masalah yang dikaji.

BAB II LANDASAN TEORI

Bagian ini mengemukakan landasan teori yang menunjang dalam penulisan skripsi, seperti *text clustering*, *feature selection*, *feature extraction*, *term variance*, *document frequency*, *principal component analysis*, *k-means*, dan DBSCAN.

BAB III ANALISIS PERBANDINGAN ALGORITMA *CLUSTERING K-MEANS* DAN DBSCAN UNTUK DATA TEKS TERJEMAH HADITS MENGGUNAKAN EKSTRAKSI CIRI *HYBRID*

Bab ini berisi pembahasan tentang penelitian yang dilakukan. Mulai dari *text preprocessing,* lalu melakukan reduksi fitur menggunakan metode *term variance* (tv) dan metode *document frequency* (df), kemudian digabungkan menggunakan metode *union, intersection,* dan *modified union*, setelah itu direduksi lagi menggunakan metode *principal component analysis* (PCA), lalu di klastering menggunakan metode *k-means* dan DBSCAN, dan terakhir melakukan evaluasi dengan metode Davies-Bouldin Index (DBI) dan *Silhoutte Coefficient*.

BAB IV ANALISIS HASIL KLASTERISASI

Bagian ini akan dipaparkan mengenai analisis hasil klasterisasi yang sudah dilakukan di bab III. Kemudian dilakukan juga analisis *dataset* yang digunakan.

BAB V PENUTUP

Bagian ini akan menjelaskan beberapa hal mengenai kesimpulan untuk jawaban dari rumusan masalah yang diajukan serta beberapa saran untuk pengembangan tulisan dan analisis dari masalah yang dikaji dalam skripsi ini.

DAFTAR PUSTAKA