BAB I
PENDAHULUAN

1.1. Latar Belakang Masalah

Dalam beberapa tahun terakhir pertumbuhan data digital meningkat, sehingga penemuan pengetahuan dan penambangan data telah menarik perhatian, dengan kebutuhan yang muncul untuk mengubah data tersebut menjadi informasi dan pengetahuan yang berguna. Penggunaan informasi dan pengetahuan yang diambil dari sejumlah besar data bermanfaat bagi banyak aplikasi seperti analisis pasar dan manajemen bisnis [1]. Agama Islam memiliki beberapa sumber hukum dan pengetahuan yang mengatur perilaku umatnya (Muslim) ketika bertindak sebagai hamba dan khalifah di muka bumi. Sumber hukum Islam merupakan dasar utama dalam pengambilan Istinbat hukum. Oleh karena itu, apapun pokok persoalan yang dimaksud, harus didasarkan pada sumber hukum. Sumber hukum pertama adalah Al-Qur'an, Wahyu atau Kalamullah, yang dijamin keasliannya dan juga dilindungi dari campur tangan manusia. Sebagaimana firman Allah Subhanahu Wa Ta'ala dalam Qur'an Surat Al-Hijr ayat 9.

إِنَّا نَحْنُ نَزَّلْنَا الذِّكْرَ وَإِنَّا لَهُ لَحَافِظُونَ

Artinya : "*Sesungguhnya Kami-lah yang menurunkan Al-Qur'an dan sesungguhnya kami benar-benar memeliharanya.*" (Q.S. Al-Hijr : 9)

Dengan demikian, hal ini menegaskan posisi Al-Qur'an sebagai sumber hukum utama. Dalam keberadaannya, sumber hukum Islam tidak hanya Al-Qur'an, tetapi juga Hadits, Ijma', dan Qiyas. Ketiganya hanyalah sumber sekunder hukum Islam, dan sumber-sumber ini tidak berfungsi sebagai pelengkap Al-Qur'an tetapi untuk pemahaman manusia tentang Maqasid al-Syariah. Pemahaman manusia tidak sempurna, tetapi Al-Qur'an sempurna, sehingga penjelasan diperlukan sebagai tindakan menguraikan sesuatu yang tidak sepenuhnya dipahami [2]. Sehingga, Hadits juga merupakan salah satu sumber pengetahuan dalam Islam. Nabi Muhammad Shallallahu 'Alaihi Wa Sallam bersabda

كِتَابَ اللهِ وَ سُنَّةَ رَسُوْلِهِ :تَرَكْتُ فِيْكُمْ أَمْرَيْنِ لَنْ تَضِلُّوْا مَا تَمَسَّكْتُمْ بِهِمَا

Artinya : "*Aku telah tinggalkan kepada kamu dua perkara. Kamu tidak akan sesat selama berpegang kepada keduanya, (yaitu) Kitab Allah dan Sunnah Rasul-Nya*." (HR. Malik; Al-Hakim, Al-Baihaqi, Ibnu Nashr, Ibnu Hazm. Hadits ini disahihkan oleh Syaikh Salim Al-Hilali di dalam At-Ta'zhim wa Al-Minnah fi Al-Intishar As-Sunnah, hlm. 12-13).

Dalam banyak aplikasi, *database* menyimpan informasi dalam bentuk teks sehingga *text mining* (penambangan teks) adalah salah satu bidang penelitian baru. Namun, penambangan teks juga merupakan tugas yang jauh lebih kompleks (daripada penambangan data) karena harus berurusan dengan data tekstual yang tidak terstruktur dan ambigu [1]. Sekitar 80% data dunia berupa teks tidak terstruktur. Teks yang tidak terstruktur ini tidak dapat dengan mudah digunakan oleh komputer untuk diproses lebih lanjut. Sehingga diperlukan suatu teknik yang berguna untuk mengekstrak beberapa informasi berharga dari teks tersebut [3]. Penambangan teks mengekstrak informasi tersembunyi dari data tidak terstruktur tersebut ke data semi terstruktur. Penambangan teks merupakan langkah penting dalam proses penemuan pengetahuan. Penambangan teks adalah bidang interdisipliner yang mencakup *information retrieval* (pencarian informasi), *text analysis* (analisis teks), *information extraction* (ekstraksi informasi), *clustering* (pengelompokan), *classification* (klasifikasi), *visualization* (visualisasi), *database technology* (teknik database), machine learning (pembelajaran mesin), dan *data mining* (penambangan data) [4].

Text clustering (pengelompokan teks) dan *text classification* (klasifikasi teks) merupakan tugas mendasar dalam penambangan teks. Klasifikasi teks digunakan sebagian besar sebagai metode pembelajaran yang diawasi (*supervised learning)*, sedangkan pengelompokan teks digunakan untuk pembelajaran yang tidak diawasi (*unsupervised learning*). Tujuan pengelompokan teks bersifat deskriptif, sedangkan tujuan klasifikasi teks bersifat prediktif (Veyssieres dan Plant, 1998). Dikarenakan tujuan pengelompokan adalah untuk menemukan seperangkat kategori baru, sehingga kelompok-kelompok baru tersebut terbentuk berdasarkan keterkaitan satu sama lain, dan penilaian yang dihasilkan bersifat intrinsik [5].

Document Clustering (pengelompokan dokumen) adalah tugas mendasar dari penambangan teks yang secara efisien mengelola organisasi, navigasi, peringkasan, dan pengambilan dokumen. Pengelompokan Dokumen mencoba untuk secara otomatis membagi dokumen yang tidak berlabel ke dalam grup. Kelompok-kelompok tersebut sesuai dengan tema, topik, atau kategori korpus asli [6].

Secara umum, metode *clustering* utama dikelompokkan menjadi beberapa kelompok, diantaranya *partitioning methods*, *hirarchical methods*, *density-based methods*, *grid-based methods*, dan *model based methods*. *Partitioning methods* membangun dan mempartisi data, dimana setiap partisi mewakili sebuah *cluster* dengan syarat setiap *cluster* harus berisi setidaknya satu objek dan setiap objek harus memiliki tepat satu *cluster*. *Hirarchical metods* membuat dekomposisi hirarki dari kumpulan objek data yang diberikan. Sebuah metode hirarki dapat diklasifikasikan sebagai *agglomerative* atau *divisive*, berdasarkan bagaimana dekomposisi hirarki terbentuk. *Density-based methods* didasarkan pada pengertian densitas. Ide umumnya adalah untuk terus menumbuhkan *cluster* yang diberikan selama kepadatan (jumlah objek) melebihi beberapa ambang batas. *Grid-based methods* mengkuantisasi ruang objek menjadi sejumlah sel terbatas yang membentuk struktur *grid*, kemudian melakukan semua operasi *clustering* pada struktur *grid*. *Model-based methods* menghipotesiskan model untuk setiap *cluster* dan menemukan data yang paling cocok dengan distribusi spasial titik data model tersebut.

Partitioning methods yang paling dikenal dan umum digunakan adalah *k-means* yang diusulkan oleh MacQueen tahun 1967 dan *k-medoids* yang diusulkan oleh Kaufman dan Rosseeuw tahun 1987. Algoritma *k-means* mengambil parameter input *k* dan mempartisi sekumpulan *n* objek ke dalam *k cluster*, sehingga menghasilkan *intracluster similarity* tinggi, sedangkan *intercluster similarity* rendah. *Cluster similarity* diukur berdasarkan nilai rata-rata objek dalam *cluster* yang dapat dilihat sebagai pusat gravitasi *cluster*. Algoritma ini mencoba untuk menentukan *k* partisi yang meminimalkan *square-error function*. Berbeda dengan algoritma *k-means*, algoritma *k-medoids* mengambil objek representatif dalam sebuah *cluster*, yang disebut *medoid*, yang merupakan titik paling sentral

dalam sebuah *cluster*. *PAM* (*Partitioning Around Medoids*) adalah algoritma *clustering* tipe *k-medoids* yang pertama diperkenalkan [7]. Pada penelitian ini akan dilakukan analisis pada kedua algoritma diatas, yaitu algoritma *k-means* dan algoritma *k-medoids* (algoritma yang digunkan adalah algoritma *PAM*) untuk melihat efektifitas kedua metode tersebut dalam pengelompokkan teks.

Sebelum melakukan pengelompokkan, hal yang perlu dilakukan adalah menentukan ukuran kemiripan/ ukuran jarak. Ukuran tersebut mencerminkan tingkat kedekatan atau pemisahan objek dan harus konsisten dengan fitur yang diyakini dapat membedakan *cluster* yang tertanam dalam data. Hal ini sering bergantung pada data atau konteks masalah, dan tidak ada ukuran terbaik secara umum untuk semua jenis masalah pengelompokan [8]. Kemiripan antara dua objek (*similarity measure)* dihitung dengan menggunakan ukuran jarak.

Terdapat banyak ukuran jarak telah diusulkan dalam literatur untuk pengelompokan data. Ukuran jarak yang paling sering digunakan adalah fungsi metrik, seperti *euclidean distance*, *manhattan distance*, *minkowski distance*, dan *hamming distance*. *Jaccard index*, *cosine similarity* dan *dice coefficient* juga merupakan ukuran jarak yang populer digunakan [9]. *Euclidean distance* adalah metrik standar untuk masalah geometri. Ukuran ini adalah jarak normal antara dua titik dan dapat dengan mudah diukur menggunakan penggaris dalam ruang 2D atau 3D. *Euclidean distance* sering digunakan dalam masalah pengelompokan seperti pengelompokan teks. *Euclidean distance* juga merupakan ukuran jarak yang sering digunakan pada algoritma K-Means. Jika dokumen direpresentasikan sebagai vektor istilah, kesamaan dua dokumen sama dengan korelasi antara vektor. Ini dikuantifikasi sebagai kosinus sudut antara vektor, sehingga disebut dengan *cosine similarity*. *Cosine similarity* adalah salah satu ukuran kesamaan yang paling umum diterapkan pada dokumen teks, termasuk berbagai pencarian informasi dan pengelompokan aplikasi [8]. Oleh karena itu, pada Skripsi ini digunakan dua teknik pengelompokkan, yaitu *k-means* dan *k-medoids* dengan dua similarity measure, yaitu *euclidean distance* dan *cosine similarity* untuk mengelompokkan data terjemahan Hadits bahasa Indonesia pada kitab Shahih Bukhari dan Shahih Muslim.

Data dimensi tinggi menjadi salah satu permasalahan yang menyebabkan buruknya performa algoritma pengelompokkan. Dengan data dimensi tinggi, sulit untuk memahami struktur yang mendasarinya. Selain itu, penyimpanan, transmisi, dan pemrosesan data berdimensi tinggi memberikan tuntutan besar pada sistem. Oleh karena itu dilakukan pengurangan ukuran data tetapi sebanyak mungkin struktur asli dipertahankan [10]. Reduksi dimensi adalah proses pengurangan jumlah variabel acak yang sedang dipertimbangkan, dan dapat dibagi menjadi seleksi fitur dan ekstraksi fitur. Saat dimensi meningkat, kinerja kueri dalam struktur indeks menurun. Algoritma reduksi dimensi adalah satu-satunya solusi yang diketahui yang mendukung pengambilan objek yang terukur dan memenuhi presisi hasil kueri. Fitur mengubah data dalam ruang berdimensi tinggi ke ruang dengan dimensi yang lebih sedikit. Secara umum, teknik pengelompokkan pada data berdimensi tinggi merupakan tugas yang sulit karena jumlah variabel yang terlibat lebih banyak. Untuk meningkatkan efisiensi, data noise dan outlier dapat dihapus sehingga dapat meminimalkan runtime. Hal ini juga dapat mengurangi jumlah variabel dalam kumpulan data asli. Untuk melakukannya, kita dapat memilih metode reduksi dimensi seperti *principal componen analysis* (PCA) [11]. PCA adalah teknik untuk mengekstraksi faktor (komponen) signifikan dari sejumlah besar variabel yang dapat diakses dalam kumpulan data. Ini mengekstrak set item dimensi rendah dari set data dimensi tinggi dengan tujuan mendapatkan informasi sebanyak mungkin. Dengan elemen yang lebih sedikit, representasi menjadi jauh lebih penting [12].

Menentukan jumlah *cluster* yang optimal untuk suatu kumpulan data merupakan masalah penting dalam beberapa algoritma *clustering*. Tidak ada metode tunggal untuk menentukan nilai k, nilai optimal untuk kumpulan data tertentu mungkin bergantung pada metode yang digunakan untuk mengukur kesamaan (*similarity measure*) dan nilai awal yang digunakan [13]. Saat ini telah dikembangkan berbagai teknik untuk menentukan nilai *k* yang paling optimal sesuai kondisi dan situasi tertentu, salah satunya adalah *Silhouette Score*. *Silhouette Score* dihitung untuk melihat kualitas cluster dan kekuatannya, seberapa baik suatu objek untuk dimasukkan ke dalam cluster. Ini adalah kombinasi dari *cohesion method* dan *separation method*. *Silhouette Score* berkisar

antara -1 sampai 1, semakin kecil nilai yang diperoleh maka banyak objek dalam suatu *cluster* yang tidak sesuai. Dataset pada *cluster* yang sesuai jika memiliki nilai mendekati 1 [14].

Berdasarkan hasil rujukan penelitian diatas terbukti bahwa ukuran jarak yang digunakan menghasilkan hasil yang berbeda saat digunakan pada algoritma *clustering*. Selain itu, banyaknya fitur yang tidak relevan mempengaruhi hasil dari proses *clustering*. Pada penelitian ini diharapkan adanya pertimbangan pada parameter diatas agar dapat diketahui faktor apa saja yang dapat mempengaruhi hasil *clustering* data agar menjadi lebih baik.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah penulis sampaikan sebelumnya, maka rumusan masalah yang akan diteliti pada Skripsi ini adalah sebagai berikut:

1. Bagaimana pengaruh reduksi dimensi dengan metode reduksi *Principal Component Analysis* (PCA) terhadap proses *text clustering* dengan *K-Means* dan *K-Meodoids*?
2. Bagaimana perbedaan pengelompokkan data terjemahan hadis menggunakan algoritma *K-Means* dan *K-Meodoids* dengan kombinasi *proximity measure* ?
3. Bagaimana perbandingan hasil jumlah *cluster* terbaik pada masing-masing algoritma *clustering K-Means* dan *K-Medoids* dengan kombinasi *proximity measure* tanpa reduksi dan menggunakan reduksi?

1.3. Batasan Masalah

Untuk menjaga agar penelitian Skripsi ini dapat fokus pada rumusan masalah dan tidak menyimpang dari tujuan yang ingin diperoleh, maka penulis menentukan batasan masalah sebagai berikut:

1. *Dataset* yang digunakan yaitu berupa data teks terjemahan hadis Bahasa Indonesia, pada kitab Sahih Bukhari dan Sahih Muslin, berdasarkan 5 kategori hadis yang sama pada setiap kitab.
2. Digunakan dua jenis *similarity measure* yaitu *cosine distance* dan *euclidean distance.*

3. Algoritma *clustering* yang digunakan adalah algoritma *K-Means* dan algoritma *PAM.*
4. Metode yang digunakan untuk mereduksi fitur adalah metode *Principal Component Analysis* (PCA) pada *feature extraction*.
5. Metode yang digunakan untuk mengevaluasi hasil dan menentukan jumlah *cluster* terbaik *cluster* adalah metode *Silhouette Coefficient* (SC).

1.4. Tujuan Penelitian

Berdasarkan latar belakang masalah dan rumusan masalah yang telah dijelaskan, terdapat beberapa tujuan yang ingin dicapai dalam penelitian Skripsi ini, antara lain:

1. Sebagai implementasi konsep wahyu memandu ilmu, dimana objek dalam penelitian ini merupakan hadis Nabi Muhammad *Shallallahu 'alaihi Wa Sallam* yang diintegrasikan dengan perkembangan teknologi.
2. Dapat mengoptimalkan kinerja algoritma *clustering K-Means* dan *K-Medoids* dalam mengelompokkan data teks.
3. Mendapatkan perbandingan hasil jumlah *cluster* terbaik pada algoritma *clustering K-Means* dan *K-Medoids* tanpa reduksi dan menggunakan reduksi PCA pada setiap *proximity measure* yang digunakan.

Adapun manfaat dari penelitian ini adalah sebagai berikut:

1. Hasil penelitian ini diharapkan menjadi salah satu bentuk pengembangan dan pengetahuan dalam kajian *clustering* khususnya dalam *clustering* data teks terjemahan hadis.
2. Memberikan pemahaman mengenai cara menentukan jumlah *cluster* terbaik pada *clustering* hadits.

Hasil penelitian ini diharapkan menjadi tambahan informasi mengenai parameter yang mempengaruhi akurasi *cluster* pada performa teknik *clustering* dalam mengelompokkan data teks dan bermanfaat bagi umat islam mengelompokkan hadis.

1.5. Metode Penelitian

1. Studi Literatur

   Tahap studi literatur merupakan tahap untuk mengumpulkan data, materi, dan informasi mengenai *clustering* algoritma *partitioning method* terutama algoritma *K-Means* dan *K-Medoids*, *similarity measure,* dan reduksi dimensi dari berbagai sumber, diantaranya buku, jurnal, artikel, dan lain sebagainya.

2. Analisis

   Pada tahap ini, penulis mengkaji dan menganalisis hasil dari setiap tahap studi literatur sesuai dengan masalah yang dipilih dalam Skripsi ini. Kemudian di tahap ini juga dilakukan pengelompokan *dataset* hadis dengan variasi 2 sampai 10 *cluster*.

3. Simulasi

   Pada tahap ini penulis melakukan pengujian metode *clustering K-Means* dan *K-Medoids* menjadi empat skenario, dimana skenario pertama dilakukan untuk *dataset* yang tidak direduksi dan menggunakan *euclidean distance* sebagai *similarity measure*, skenario kedua dilakukan untuk *dataset* yang tidak direduksi dan menggunakan *cosine distance* sebagai *similarity measure,* skenario ketiga dilakukan untuk *dataset* yang telah direduksi oleh PCA dan menggunakan *euclidean distance* sebagai *similarity measure*, dan skenario keempat dilakukan untuk *dataset* yang telah direduksi oleh PCA dan menggunakan *cosine distance* sebagai *similarity measure*, menggunakan Bahasa pemrograman *python* yang dijalankan di *Pycharm*. Kemudian dari hasil pengujian tersebut akan dianalisis nilai akurasi *cluster* ketika diuji dengan teknik evaluasi *cluster* internal menggunakan metode *Silhouette Coefficient* (SC) yang juga berfungsi sebagai acuan untuk menentukan jumlah *cluster* terbaik.

1.6. Sistematika Penulisan

Sistematika penulisan pada Skripsi ini terdiri dari lima bab dan di dalam setiap bab terdiri dari beberapa subbab. Dengan sistematika penulisan sebagai berikut:

BAB I : PENDAHULUAN

Bab ini berisi tentang pemaparan latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, metode penelitian, serta sistematika penulisan Skripsi.

BAB II : LANDASAN TEORI

Bab ini berisi penjelasan mengenai teori-teori yang berkaitan dengan masalah yang akan dikaji.

BAB III : ALGORITMA K-MEANS DAN K-MEDOIDS DENGAN KOMBINASI PROXIMITY MEASURE DAN REDUKSI DIMENSI PCA

Bab ini berisi tentang penelitian yang dilakukan dari pengambilan *dataset*, lalu tahap *text preprocessing,* kemudian melakukan reduksi fitur menggunakan metode *Principal Component Analysis* (PCA), lalu dilakukan *clustering* menggunakan metode *clustering K-Means* dan *K-Medoids*, dengan dua metode *similarity measure* yaitu *cosine similarity* dan *euclidean distance*, dan terakhir melakukan evaluasi dengan metode *Silhouette Coefficient* (SC).

BAB IV : ANALISIS HASIL PENENTUAN JUMLAH CLUSTER TERBAIK PADA ALGORITMA K-MEANS DAN K-MEDOIDS

Bab ini berisi penjelasan mengenai hasil pengujian metode *clustering* pada data teks terjemahan hadis Bahasa Indonesia dengan beberapa variasi jumlah *cluster* (*k*), *similarity measure*, dan persentase reduksi.

BAB V : **PENUTUP**

Bab ini berisi penjelasan mengenai beberapa hal yang menjadi kesimpulan atas penelitian yang telah dilakukan serta beberapa saran yang berisi rekomendasi untuk pengembangan tulisan ini.