BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan teknologi yang terus meningkat mengakibatkan peningkatan informasi dalam jumlah yang besar. Peningkatan aliran informasi ini menyebabkan banyaknya penumpukan data. Ketersediaan data yang semakin meningkat setiap harinya salah satunya yang dihasilkan dari penggunaan teknologi informasi di berbagai bidang kehidupan menimbulkan kebutuhan untuk memanfaatkan informasi dan pengetahuan yang terkandung di dalam data tersebut. Selain data yang bersumber dari teknologi dan informasi, sebagai umat beragama islam terdapat data yang merupakan sumber hukum syariat utama dalam agama islam yang bersumber dari Al-Qur'an yang merupakan firman Allah dan juga hadis yang merupakan segala sesuatu yang disandarkan kepada Nabi Muhammad *Shallallahu 'alaihi Wa Sallam*.

Sebagai umat beragama islam terdapat data yang merupakan sumber hukum syariat utama dalam agama islam yang bersumber dari Al-Qur'an yang merupakan firman Allah untuk menjadi pedoman hidup umat manusia agar bisa membedakan antara yang hak atau yang batil, ketaatan atau kemaksiatan, berpahala atau berdosa dan juga hadis yang merupakan segala sesuatu yang disandarkan kepada Nabi Muhammad *Shallallahu 'alaihi Wa Sallam*, baik ucapannya, perilakunya, hingga ketetapannya untuk menjawab permasalahan komunikasi umat beragama yang semakin mengkhawatirkan seiring dengan berkembangnya teknologi dan perubahan zaman. Seorang Muslim tentunya harus sangat memperhatikan dan terus mempelajari kedua landasan hukum dalam agamanya ini untuk memperoleh kebahagiaan di Dunia dan di Akhirat. Sebagaimana firman Allah *Subhanahu Wa Ta'ala* dalam Qur'an Surat An-Nuur ayat 56.

وَأَطِيعُوا الرَّسُولَ لَعَلَّكُمْ تُرْحَمُونَ

Artinya: *"Dan taatlah kepada Rasul supaya kamu diberi rahmat."* (QS. An-Nuur:56)

Maka berdasarkan firman Allah tersebut hendaklah wajib untuk mentaati Nabi *Shallallahu 'alaihi Wa Sallam* Sebagaimana firman Allah *Subhanahu Wa Ta'ala* dalam Surat lain yaitu Qur'an Surat An-Nisaa ayat 59.

يَا أَيُّهَا الَّذِينَ آمَنُوا أَطِيعُوا اللَّهَ وَأَطِيعُوا الرَّسُولَ

Artinya: *"Hai orang-orang yang beriman, taatilah Allah dan taatilah Rasul-nya"* (QS. An-Nisaa:59).

Disamping itu, terkadang perintah tersebut disampaikan dalam bentuk tunggal, tidak dibarengi kepada perintah yang lain, sebagaimana firman Allah *Subhanahu Wa Ta'ala* dalam Qur'an Surat An-Nisaa ayat 80.

مَنْ يُّطِعِ الرَّسُوْلَ فَقَدْ اَطَاعَ اللّٰهَ ۚ وَمَنْ تَوَلّٰى فَمَآ اَرْسَلْنٰكَ عَلَيْهِمْ حَفِيْظًا

Artinya: "Barang siapa mentaati Rasul (Muhammad), maka sesungguhnya dia telah menaati Allah. Dan barangsiapa berpaling (dari ketaatan itu), maka (ketahuilah) Kami tidak mengutusmu (Muhammad) untuk menjadi pemelihara mereka." (QS. An-Nisa:80).

Oleh karena itu, sangatlah penting bagi umat beragama Islam untuk mengenal lebih dalam dan mempelajari lebih luas akan suri tauladan Nabi Muhammad *Shallallahu 'alaihi Wa Sallam* dalam melakukan suatu tindakan apapun, sebagaimana sabda Rasulullah *Shallallahu 'alaihi Wa Sallam.*

مَن عَمِلَ عَمَلاً لَيْسَ عَلَيهِ أَمرُنَا فَهُوَ رَدٌّ

Artinya: *"Barang siapa melakukan suatu amalan yang tidak berdasarkan perintah kami, maka amalan itu tertolak."* (HR. Al-Bukhari, no. 2697 dan Muslim no. 1719)

Dan dalam Riwayat lain Rasulullah *Shallallahu 'alaihi Wa Sallam* bersabda

مَن رَغِبَ عَنْ سُنَّتِي فَلَيْسَ مِنِّي

Artinya: *"Barang siapa yang membenci sunahku, maka ia bukan termasuk golonganku."* (HR. Al-Bukhari, no. 5063 dan Muslim, no. 1401)

Menaati Rasul tidak dapat dikatakan perbuatan syirik, karena Rasul penyampai perintah Allah. Dengan demikian menaati Rasul adalah menaati Allah, bukan mempersekutukan nya dengan Allah. Oleh karenanya, penelitian mengenai hadis ini didasari pada firman Allah dan atas dasar rasa cinta kepada Rasulullah *Shallallahu 'alaihi Wa Sallam*.

Karena data di dunia nyata semakin hari semakin berkembang sehingga kumpulan data yang sangat besar dapat diidentifikasi menjadi pola yang menarik dengan pengelompokan [1]. Oleh karenanya dibutuhkan solusi matematika, karena matematika merupakan ilmu dasar yang memegang peranan penting dalam perkembangan ilmu pengetahuan dan teknologi di era modern. Salah satu dari solusi matematika yaitu *data mining*. *Data mining* muncul sebagai bidang yang berkaitan dengan ekstraksi informasi yang berguna dari data tersebut yang kemudian metode *data mining* diterapkan untuk memecahkan berbagai masalah di dunia nyata [2].

Metode *text mining* merupakan metode pengembangan dari *data mining*. *Text mining* adalah proses ekstraksi pola berupa pengetahuan dari sebagian besar jumlah data teks, data teks dapat berupa *paper*, berita, Al-Qur'an, dan hadis [3]. *Text mining* merupakan teknik yang digunakan untuk menangani masalah *klasifikasi*, *clustering*, *information extraction*, dan *information retrieval* [3]. *Text mining* merupakan konsep dan teknik *data mining* untuk mencari pola dalam teks, yaitu proses analisis teks yang dilakukan secara otomatis oleh komputer untuk menggali informasi yang berkualitas dari suatu rangkaian teks yang terangkum dalam sebuah dokumen. Prosedur utama metode ini terkait dengan menemukan kata-kata yang dapat mewakili isi dari dokumen untuk selanjutnya dilakukan analisis keterhubungan antar dokumen dengan menggunakan metode statistik, seperti teknik *clustering*.

Clustering teks merupakan bagian yang penting dalam metode *text mining* yang merupakan pengembangan *data mining*. *Clustering* teks adalah klasifikasi dokumen tanpa pengawasan, yang membagi koleksi teks menjadi beberapa *subset*

yang disebut *cluster*, masing-masing *cluster* memiliki kesamaan yang lebih besar daripada yang berada dalam *cluster* yang berbeda [4]. *Clustering text* mengacu pada proses mengelompokkan dokumen teks serupa bersama-sama. Masalah dalam pengelompokan dokumen dapat diformulasikan sebagai berikut: Dengan diberikan satu set dokumen, dokumen tersebut harus dibagi menjadi beberapa kelompok, sehingga dokumen dalam kelompok yang sama lebih mirip satu sama lain daripada dokumen dalam kelompok yang lainnya. Ada banyak penerapan dari teks *clustering* diantaranya: kategorisasi dokumen, peringkasan sebuah korpus, dan klasifikasi dokumen [5].

Teknik atau metode *clustering* bekerja dengan membagi atau mengelompokkan data-data berdasarkan kesamaan karakteristik suatu data dengan data lainnya. Dalam teknik *clustering*, terdapat beberapa algoritma yang dapat diterapkan, salah satu algoritma yang banyak digunakan dalam teknik *clustering* adalah algoritma berbasis partisi yaitu algoritma *K-Means* [6]. Kesederhanaan metode *K-Means* banyak digunakan di berbagai bidang karena memiliki beberapa keunggulan yaitu mudah diimplementasikan dan memiliki tingkat ketelitian yang cukup tinggi terhadap ukuran objek sehingga metode ini relatif lebih terukur dan efisien. Selain itu metode ini juga mudah dijalankan, relatif cepat dan mudah beradaptasi. Akan tetapi algoritma *K-Means* sangat sensitif terhadap penempatan awal pusat *cluster* yang disebabkan karena pemilihan titik awal pusat *cluster* dilakukan secara acak [7]. Beberapa penelitian telah dilakukan pengembangan untuk memperbaiki kekurangan algoritma *K-Means*, salah satu hasil dari pengembangan tersebut adalah algoritma *K-Means++*. Algoritma *K-Means++* digunakan untuk mengatasi permasalahan dalam kecepatan dan akurasi pada algoritma *K-Means* dengan memilih pusat *cluster* awal menggunakan perhitungan matematis [8]. Pada penelitian ini penulis hanya akan menganalisis metode algoritma *K-Means++* untuk melihat seberapa efektif metode tersebut dalam pengelompokan teks.

Rujukan penelitian sebelumnya [4], menggunakan algoritma *K-Means* untuk mengelompokkan terjemahan ayat Al-Qur’an dalam Bahasa Indonesia. Adapun percobaan yang dilakukannya menghasilkan nilai *Silhouette Coefficient* sebesar

0.3777 pada surat Al-Baqarah. Akurasi *cluster* yang tidak maksimal dikarenakan salah satunya pada pendefinisian awal pusat *cluster*nya yang secara acak.

Penelitian sebelumnya [9], menjelaskan bahwa algoritma *clustering K-Means* sangat sensitif terhadap pemilihan *cluster* awal dimana *cluster* awal sangat berpengaruh terhadap hasil pengelompokan sehingga penelitian ini mengoptimalkan *cluster* awal pusat dengan hasil pengelompokan yang menghasilkan akurasi rata-rata 80%. Pada penelitian sebelumnya [10], menggunakan algoritma *K-Means* mengelompokan surat-surat yang ada di dalam Al-Qur'an dan menghasilkan kelompok sebanyak 4 *cluster.* Akurasi *cluster* yang tidak maksimal dikarenakan salah satunya pada pendefinisian awal pusat *cluster*nya yang secara acak.

Adapun pada penelitian [11], menggunakan algoritma *Fuzzy C-means* dan algoritma *K-Means++* untuk mengelompokkan *dataset* Iris, *Soybean-small* dan *wine*. Adapun percobaan yang dilakukan menghasilkan nilai *Silhouette Coefficient* yang hampir sama pada kedua algoritma. Akan tetapi secara keseluruhan algoritma *K-Means++* memiliki tingkat akurasi yang lebih unggul dibandingkan algoritma *Fuzzy C-means* pada ketiga *dataset* tersebut.

Dalam mengimplementasikan *clustering,* pengukuran kedekatan (*proximity*) objek satu dengan lainnya menjadi proses yang sangat penting. *Proximity* bisa mengacu pada pengukuran seberapa mirip objek satu dengan lainnya (*similarity*) atau tidak seberapa mirip kedua objek tersebut (*dissimilarity*) [12]. Dalam hal ini istilah *dissimilarity* (*distance*) lebih sering digunakan daripada istilah *dissimilarity*.

Mengingat banyaknya metode *clustering* dan *proximity measure* yang dapat digunakan untuk mengelompokkan dokumen teks, dimana pemilihan kombinasi dari keduanya kemudian menjadi sangat penting guna mengoptimalkan hasil yang diperoleh karena *cluster* yang terbentuk dari masing-masing kombinasi pun tentunya akan beragam. Oleh karena itu pada Skripsi ini, teknik *clustering* yang digunakan yaitu algoritma *clustering K-Means++* dengan dua *proximity measure* berbasis jarak (*dissimilarity*) yaitu *cosine distance* dan *euclidean distance* [13] untuk mengelompokkan terjemahan hadis pada kitab Sahih Bukhari, Sahih Muslim dan Sunan Turmuzi dalam Bahasa Indonesia.

Salah satu yang menyebabkan performa buruk dari algoritma *clustering* adalah karena terdapat banyak fitur yang tidak relevan. Karenanya dimensi ruang fitur yang tinggi merupakan salah satu masalah utama yang harus dipertimbangkan dalam proses pengelompokan teks. Setelah dilakukan *preprocessing* setiap dokumen direpresentasikan menjadi vektor menggunakan *Vector Space Model* (VSM). Pada VSM setiap term yang terdapat di dalam dokumen merupakan representasi dari fitur yang berbeda. Semakin besar dokumen maka akan menghasilkan fitur yang semakin banyak, ratusan bahkan ribuan fitur [14]. Banyaknya fitur yang tidak relevan ini dapat mengakibatkan kinerja yang buruk untuk algoritma yang digunakan.

Teknik reduksi dimensi telah banyak dikembangkan oleh para peneliti untuk meningkatkan kinerja algoritma *clustering*. Dalam penelitian mengenai reduksi dimensi menggunakan PCA menunjukkan bahwa PCA memberikan solusi efektif untuk *clustering K-Means* [15], oleh karenanya dalam penelitian ini digunakan teknik reduksi dimensi *Principal Component Analysis* (PCA) pada *feature extraction*. *Principal Component Analysis* (PCA) adalah metode reduksi dimensi yang umum dan banyak digunakan [16]. *Principal Component Analysis* (PCA) merupakan salah satu bentuk teknik yang digunakan untuk mengumpulkan data berdimensi tinggi dan selanjutnya menggunakan dependensi antar variabel untuk merepresentasikan data secara lebih sistematis untuk membentuk dimensi rendah tanpa kehilangan banyak informasi yang ada dalam *dataset* [17].

Dari hasil rujukan penelitian diatas terbukti bahwa terdapatnya hasil perbedaan akurasi penelitian ketika pengelompokan data dengan tipe yang berbeda. Penentuan awal *cluster* sangat berpengaruh agar dapat menghasilkan akurasi yang tepat pada proses pengelompokan, dan banyaknya fitur yang tidak relevan dalam dokumen pada *clustering* teks menyebabkan performa yang kurang baik pada algoritma *clustering* [18]. 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 cara mengelompokkan dan mereduksi fitur data terjemahan hadis menggunakan algoritma *clustering K-Means++* dengan kombinasi *proximity measure* berbasis *dissimilarity* dan metode reduksi *Principal Component Analysis* (PCA)?
2. Bagaimana kinerja algoritma *K-Means++* dengan kombinasi *proximity measure* berbasis *dissimilarity* dan teknik reduksi fitur dengan menggunakan metode *Principal Component Analysis* (PCA)?
3. Bagaimana perbandingan hasil *clustering* data teks terjemahan hadis pada masing-masing algoritma *clustering K-Means++* dengan kombinasi *dissimilarity 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, Sahih Muslin, dan Sunan Turmuzi, berdasarkan 12 kategori hadis yang sama pada setiap kitab.
2. Digunakan dua jenis *proximity measure* berbasis *dissimilarity* yaitu *cosine distance* dan *euclidean distance.*
3. Metode *clustering* yang digunakan adalah metode algoritma *K-Means++*.
4. Metode yang digunakan untuk mereduksi fitur adalah metode *Principal Component Analysis* (PCA) pada *feature extraction.*
5. Parameter yang dijadikan tolak ukur untuk membandingkan performansi dari algoritma *clustering K-Means++* adalah nilai evaluasi *cluster* yang dihasilkan.
6. Parameter yang dijadikan tolak ukur untuk membandingkan variasi reduksi dimensi menggunakan PCA terhadap performansi dari algoritma *clustering*

K-Means++ adalah nilai evaluasi *cluster* yang dihasilkan dan *runtime* program.

7. Metode yang digunakan untuk mengevaluasi hasil *cluster* adalah metode *Davies-Bouldin Index* (DBI) dan *Silhouette Coefficient* (SC).

1.4 Tujuan dan Manfaat 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-Mean++* dalam mengelompokkan data teks.
3. Mendapatkan perbandingan hasil *clustering* menggunakan algoritma *K-Means++* 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 *clustering* algoritma *K-Means++* dengan kombinasi *proximity measure* berbasis *dissimilarity* dan teknik reduksi fitur dengan menggunakan metode *Principal Component Analysis* (PCA)
3. 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 dalam mencari dan mengelompokkan hadis.

1.5 Metode Penelitian

1. Studi Literatur

   Tahap studi literatur merupakan tahap untuk mengumpulkan data, materi, dan informasi mengenai *clustering* algoritma *K-Means++*, *proximity measure*, dan reduksi fitur 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 berdasarkan 12 macam kategori hadis.

3. Simulasi

   Pada tahap ini penulis melakukan pengujian metode *clustering K-Means++* menggunakan *proximity measure* berbasis *dissimilarity* menjadi dua skenario, dimana skenario pertama dilakukan untuk *dataset* yang tidak direduksi, dan skenario kedua dilakukan untuk *dataset* yang telah direduksi oleh PCA, 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 *Davies-Bouldin Index* (DBI) dan *Silhouette Coefficient* (SC).

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 *CLUSTERING K-MEANS++* PADA DATA TERJEMAHAN HADIS

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++*, dengan dua metode *proximity measure* berbasis *dissimilarity* yaitu *cosine distance* dan *euclidean distance*, dan terakhir melakukan evaluasi dengan metode *Davies-Bouldin Index* (DBI) dan *Silhouette Coefficient* (SC).

BAB IV : ANALISIS HASIL *CLUSTERING* ALGORITMA *K-MEANS++* PADA DATA TERJEMAHAN HADIS

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