BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Data merupakan fakta dan statistik yang telah dikumpulkan secara bersama-sama untuk digunakan dalam berbagai macam analisis atau dijadikan sebagai referensi-referensi dalam mendukung berbagai macam penelitian atau pendapat-pendapat[1]. Perkembangan teknologi dan informasi mengakibatkan jumlah data yang disimpan dalam *file* komputer dan *database* tumbuh pada tingkat yang fenomenal. Seiring waktu, para pengguna data ini mengharapkan sistem pengolahan informasi yang lebih canggih dari zaman sebelumnya[2]. Salah satu metode yang dapat diimplementasikan adalah *data mining*.

Data mining atau penambangan data merupakan proses penggalian infromasi dan pola yang bermanfaat dari data yang sangat besar, *data mining* juga merupakan proses logis untuk menemukan informasi yang berguna. Setelah ditemukan informasi dan pola dapat digunakan untuk alat pendukung dalam pengambilan keputusan[2]. Data dalam berbagai jenis dari berbagai sumber menjadi suatu informasi yang penting apabila ditemukan suatu makna atau informasi didalamnya. Beberapa sumber data yang ada terkadang diabaikan dan jarang menjadi bahan penelitian. Salah satunya data yang bersumber dari dokumen hadis. Data dokumen hadis ini berisi setiap hadis baik berbahasa arab atau terjemahannya namun terkadang ada yang keduanya.

Hadis menurut bahasa ialah sesuatu yang baru. Secara istilah sama dengan *As-Sunnah* menurut Jumhur Ulama. *As-Sunnah* menurut istilah *syari'at* ialah segala sesuatu yang bersumber dari Nabi *Shallallahu 'alaihi Wa Sallam* dalam bentuk *qaul* (ucapan), *fi'il* (perbuatan), *taqrir* (pensyari'atan) bagi ummat islam. As-Sunnah menurut istilah ulama ushul fiqih ialah segala sesuatu yang bersumber dari Nabi *Shallallahu 'alaihi Wa Sallam* selain dari Al-Qur'an, baik perbuatan, perkataan, *taqrir* (penetapan) yang baik untuk menjadi dalil bagi hukum *syar'i*[3].

As-Sunnah menurut istilah ahli fiqih (*fuqaha'*) ialah segala sesuatu yang sudah tetap dari Nabi Muhammad *Shallallahu 'alaihi Wa Sallam* dan hukumnya tidak *fardhu* dan tidak wajib, yakni hukumnya *sunnah*. *As-Sunnah* menurut ulama salaf adalah petunjuk yang dan dilaksanakan oleh Rasulullah *Shallallahu 'alaihi Wa Sallam* dan para sahabatnya, baik tentang ilmu, *I'tiqaad* (keyakinan), perkataan maupun perbuatannya. Diantara makna *sunnah* Rasulullah *Shallallahu 'alaihi Wa Sallam*, maksudnya adalah *sunnah* sebagai sumber nilai *tasyri'*. Al-Qur'an menyifatkan *As-Sunnah* dengan makna hikmah[3].

Allah *Subhanahu Wa Ta'ala* berfirman:

رَبَّنَا وَابْعَثْ فِيهِمْ رَسُولًا مِّنْهُمْ يَتْلُو عَلَيْهِمْ آيَاتِكَ وَيُعَلِّمُهُمُ الْكِتَابَ وَالْحِكْمَةَ وَيُزَكِّيهِمْ ۚ إِنَّكَ أَنتَ الْعَزِيزُ الْحَكِيمُ

Artinya: *"Ya Rabb kami, utuslah kepada mereka seorang Rasul di antara mereka yang akan membacakan ayat-ayat-Mu kepada mereka dan mengajarkan Al-Kitab dan Al-Hikmah kepada mereka dan mensucikan mereka (dari kelakuan-kelakuan yang keji), sesungguhnya Engkau Mahamulia lagi Mahabijaksana."* (QS. Al-Baqarah : 129)

Dalam surat lain Allah *Subhanahu Wa Ta'ala* berfirman:

لَقَدْ مَنَّ اللَّهُ عَلَى الْمُؤْمِنِينَ إِذْ بَعَثَ فِيهِمْ رَسُولًا مِّنْ أَنفُسِهِمْ يَتْلُو عَلَيْهِمْ آيَاتِهِ وَيُزَكِّيهِمْ وَيُعَلِّمُهُمُ الْكِتَابَ وَالْحِكْمَةَ وَإِن كَانُوا مِن قَبْلُ لَفِي ضَلَالٍ مُّبِينٍ

Artinya: *"Sesungguhnya Allah telah memberi karunia bagi orang-orang yang beriman, Ketika Dia mengutus di antara mereka seorang Rasul dari golongan mereka sendiri, yang membacakan ayat-ayat-Nya dan membersihkan mereka (dari sifat-sifat jahat), dan mengajarkan Al-Kitab (Al-Qur'an) dan Al-Hikmah (As-Sunnah). Sesungguhnya mereka sebelum itu dalam kesesatan yang nyata"* (QS. Ali 'Imran : 164)

Dalam hadis memiliki bab yang membahas tentang berbagai perkara diantaranya, sholat, akhlak terpuji, haji dan lain-lain. Pada sumbernya data hadis

ditulis dengan bahasa arab namun seiring berjalannya waktu, data hadis banyak ditulis disertai terjemahannya. Banyaknya bab yang dibahas dalam hadis membuat data hadis menarik untuk diteliti. Berbagai metode digunakan untuk mengolah data hadis, kasus tersebut termasuk kedalam topik *text mining*.

Text mining adalah bidang interdisipliner yang mengacu pada *information retrieval*, *data mining*, *machine learning*, *statistic*, dan *computational linguistic*. Sebagian besar informasi disimpan sebagai teks seperti artikel berita, makalah, buku, perpustakaan *digital*, *email*, *blog*, dan halaman *web*. Tujuan penting dari *text mining* adalah untuk memperoleh informasi berkualitas tinggi dari teks[2]. Namun karena program komputer membutuhkan data numerik untuk diolah sehingga data teks yang di *input* ke program tidak langsung diolah melainkan dilakukan proses *preprocessing* terlebih dahulu.

Preprocessing merupakan proses pengolahan data mentah sebelum masuk ke proses *data mining*, di mana hasilnya berupa *dataset* yang siap diolah sesuai metode yang akan dipakai[4]. Menurut (Danubianu. M, 2012), data *preprocessing* difokuskan terutama pada dua masalah, pertama data harus diatur dalam bentuk yang tepat untuk algoritma *data mining*, kedua set data yang digunakan harus mengarah pada kinerja dan kualitas terbaik untuk model yang diperoleh dari oprasi *data mining*[1]. Tujuannya dari *preprocessing* adalah agar model atau algoritma yang digunakan bisa disesuaikan dengan data[2]. Ada berbagai macam algoritma atau model yang digunakan untuk mendapatkan suatu informasi dari data teks atau bisa juga disebut dengan *text mining task*. Beberapa jenis *text mining task* yang umum adalah *classification text*, *clustering text*, *concept/entity extraction*, *production of granular taxonomies*, *sentiment analysis*, *document summarization*, dan *entity-relation modeling*[5].

Clustering text atau pengelompokkan teks adalah sebuah cara untuk mengelompokkan dokumen-dokumen digital secara otomatis kedalam bentuk klaster berdasarkan karakteristik *intrinsic* dokumen tersebut. Tujuan metode ini adalah untuk meminimalkan variasi antar data yang berada dalam satu *cluster* dan memaksimalkan variasi dengan data yang berada dalam *cluster* lain Metode ini dikelompokkan kedalam empat kategori yaitu, berbasis partisi (*partitioning*

methods), berbasis hirarki (*hierarchical methods*), berbasis kepadatan (*density based*), berbasis kisi (*grid based*)[6]. *Partitioning method* adalah bentuk paling dasar dalam metode *clustering* yang mengatur objek-objek dari suatu himpunan menjadi beberapa kelompok atau *cluster* eksklusif.

K-Means merupakan salah satu algoritma *partitioning method* yang dimulai dengan memilih titik perwakilan *k* sebagai *centroid* awal. Setiap titik kemudian ditetapkan ke *centroid* terdekat berdasarkan ukuran kedekatan tertentu yang dipilih. Setelah *cluster* terbentuk, *centroid* untuk setiap *cluster* diperbarui. Kemudian kedua langkah tersebut dilakukan berulang hingga *centroid* tidak berubah atau konvergen[7]. Salah satu kelemahan dari algoritma ini adalah terdapat data pencilan atau *outlier* yang dihasilkan dari *cluster* yang diperoleh.

Outlier adalah objek data yang menyimpang secara signifikan dari objek lainnya, seolah-olah dihasilkan oleh mekanisme yang berbeda[5]. *Outlier* juga biasa disebut sebagai sumbang, kelainan, menyimpang, atau anomali dalam *data mining* dan literatur statistik[8]. Menurut Weissberg (1985), jika terdapat masalah yang berkaitan dengan *outlier*, maka diperlukan alat diagnosis yang dapat mengidentifikasi masalah *outlier*, salah satunya dengan menyisihkan *outlier* dari kelompok data kemudian menganalisa data tanpa *outlier*[9]. Data *outlier* sangat mengganggu pada proses pengelompokkan *k-means* karena menyebabkan pusat *cluster* terdistorsi sehingga tidak merepresentasikan kelompok sebenarnya. Lalu pengaruh terhadap perhitungan jarak dan mempengaruhi terhadap penentuan jumlah *cluster* yang optimal.

Deteksi *outlier* adalah suatu teknik untuk mengidentifikasi *outlier* di mana objek tersebut mempunyai perilaku berbeda dibandingkan objek-objek lain pada umumnya. Beberapa algoritma yang dapat digunakan untuk mendeteksi adanya *outlier* pada suatu *dataset* adalah *cluster based, distance based,* dan *density based*[10]. Algoritma *distance based* digunakan untuk menghitung nilai jarak maksimum untuk setiap *cluster*. Jika jarak maksimum ini lebih besar dari nilai ambang batas (*threshold*) maka dinyatakan sebagai *outlier* sebaliknya sebagai objek nyata atau *inliers*. Sedangkan algoritma *cluster based* bekerja dengan mengasumsikan *outlier* adalah objek yang tidak termasuk kedalam *cluster*

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG

manapun, atau termasuk dalam *cluster* yang sangat kecil, atau dipaksa menjadi bagian dari *cluster* yang sangat berbeda dari anggota lain[11]. Dalam mendefinisikan *cluster* kecil, mengikuti Loureiro (2004) yaitu *cluster* dengan jumlah titik lebih kecil dari setengah rata-rata titik dalam *k cluster*[12].

Penelitian ini merujuk ke penelitian sebelumnya dari[4]. Dengan mengelompokkan data terjemahan hadis menggunakan algoritma *k-means* dan *fuzzy c-means* yang menyimpulkan algoritma *fuzzy c-means* lebih baik daripada algoritma *k-means* setelah hasil *cluster* dievaluasi. Adapun penelitian lain seperti[13]. Dengan mengelompokkan data terjemahan hadis menggunakan algoritma *k-means++* yang menyimpulkan pengelompokkan data hadis menggunakan algoritma *k-means++* lebih efektif menggunakan teknik pengukuran *dissimilarity measure* dengan *cosine similarity* dan dilakukan reduksi fitur menggunakan *principal component analysis* (PCA). Lalu penelitian[8]. Dengan menganalisis *cluster* yang mengandung data pencilan dengan menggunakan metode *smoothed self organizing maps* pada data covid-19 di Jawa Barat yang menyimpulkan metode *self organizing maps* memiliki hasil yang kurang memuaskan, sementara itu metode *smoothed self organizing maps* memberikan hasil yang jauh lebih cepat menurut dibandingkan dengan *self organizing maps*.

Selain itu penelitian ini melanjutkan dari studi literatur yaitu[14]. Di mana penelitian tersebut melakukan percobaan dalam mendeteksi *outlier* menggunakan algoritma *k-means distance based* dan *k-means cluster based* pada data Data_User_Modeling_*Dataset* yang memiliki 259 *raw data* dan 5 atribut/kolom dengan diantaranya 4 atribut berjenis data numerik dan 1 atribut berjenis data *class* yang menyimpulkan, nilai *silhouette index* dan akurasi meningkat ketika *outlier* dihapus untuk kedua algoritma.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah disampaikan sebelumnya, maka rumusan masalah yang disimpulkan sebagai berikut:

1. Apa perbedaan hasil pengelompokkan *k-means* menggunakan data hasil reduksi fitur menggunakan metode *principal component analysis* (PCA) dan

data hasil pembobotan *term frequency – inverse document frequency* untuk proses deteksi *outlier*?

2. Apa pengaruh pengelompokkan *k-means* berdasarkan kedekatan jarak menggunakan *euclidean distance* dan *cosine similarity* terhadap deteksi *outlier*?
3. Bagaimana perbandingan hasil antara data hasil pengelompokkan *k-means* sebelum *outlier* dihapus dan setelah *outlier* dihapus?

1.3 Batasan Masalah

Agar penelitian Skripsi ini tidak ambigu dan tidak menyimpang dari topik pembahasan, maka penulis menentukan batasan masalah sebagai berikut:

1. *Dataset* yang digunakan berupa data teks terjemahan hadis Bahasa Indonesia pada kitab Sahih Bukhari, berdasarkan 5 kategori dengan data terbanyak dalam hadis tersebut yaitu Tafsir Al-Qur'an, Peperangan, Perilaku yang terpuji, Haji dan Jum'at.
2. Metode reduksi fitur yang digunakan adalah *principal component analysis* (PCA).
3. *Proximity measure* yang digunakan untuk pengukuran jarak yaitu *euclidean distance* dan *cosine similarity*.
4. Algoritma deteksi *outlier* yang digunakan adalah *k-means distance based* dan *k-means cluster based*.
5. Metode evaluasi hasil *cluster* yang digunakan adalah *silhouette coefficient* (SC).

1.4 Tujuan dan Manfaat Penelitian

Berdasarkan latar belakang masalah dan rumusan masalah yang telah dijelaskan, beberapa tujuan yang ingin dicapai dalam penelitian sebagai berikut:

1. Sebagai implementasi konsep wahyu memandu ilmu, di mana objek dalam penelitian ini merupakan hadis Nabi Muhammad *Shallallahu 'alaihi Wa Sallam* yang diintegrasikan dengan perkembangan teknologi.
2. Mengoptimalkan kinerja algoritma *k-means* dengan data terjemahan hadis.

UIN SUNAN GUNUNG DJATI BANDUNG

3. Mendapatkan perbandingan hasil sebelum dan setelah *outlier* dihapus

Adapun manfaat dari penelitian ini adalah sebagai berikut:

1. Hasil penelitian ini diharapkan menjadi salah satu acuan dan pengetahuan untuk penelitian selanjutnya khusunya pada topik *clustering k-means* dan deteksi *outlier*.
2. Memberikan pemahaman mengenai proses mendeteksi *outlier* pada algoritma *k-means*.
3. Memberikan gambaran mengenai hasil yang diperoleh saat sebelum dan setelah *outlier* dihapus.

1.5 Metode Penelitian

1. Studi Literatur

   Tahap studi literatur merupakan tahap untuk mengumpulkan data, referensi, dan informasi mengenai topik *clustering k-means* dan deteksi *outlier*.
2. Analisis

   Pada tahap ini, topik yang diteliti akan dikaji dan dianalisis berdasarkan dari hasil studi literatur. Kemudian kekurangan dari penelitian sebelumnya dijadikan bahan kajian untuk penelitian selanjutnya
3. Simulasi

   Pada tahapan ini dilakukan pengujian metode *clustering k-means* menggunakan *euclidean distance* dan *cosine similarity* sebagai perhitungan kedekatan jarak kemudian dikombinasikan dengan menggunakan data yang direduksi fitur menggunakan *principal component analysis* (PCA) dan data yang tidak direduksi lalu hasilnya dibandingkan antara sebelum dan sesudah *outlier* dihapus, yang menjadi tolak ukur perbandingannya adalah hasil uji validitas *cluster* yang menggunakan *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:

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG

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 : PERBANDINGAN PERFORMA METODE DETEKSI *OUTLIER DISTANCE-BASED* DAN *CLUSTER-BASED* PADA ALGORITMA *K-MEANS* DENGAN DATA TEKS HADIS

Bab ini berisi penelitian yang dilakukan dari pengambilan *dataset*, lalu tahap *text preprocessing*, kemudian dilakukan reduksi fitur menggunakan metode *principal component analysis* (PCA), lalu dilakukan pengelompokkan menggunakan metode *clustering k-means* berdasarkan *euclidean distance* dan *cosine similarity* sebagai *proximity measure*, kemudian *outlier* dideteksi menggunakan algoritma *k-means distance based* dan *k-means cluster based*, kemudian hasil dari uji validitas *cluster* dibandingkan menggunakan *silhouette coefficient*.

BAB IV : ANALISIS HASIL IMPLEMENTASI DAN PERCOBAAN PERBANDINGAN PERFORMA METODE DETEKSI *OUTLIER DISTANCE-BASED* DAN *CLUSTER-BASED* PADA ALGORITMA *K-MEANS* DENGAN DATA TEKS HADIS

Bab ini berisi penjelasan mengenai hasil pengujian algoritma *k-means distance based dan cluster based* pada data teks terjemahan hadis Bahasa Indonesia dengan menggunakan perhitungan jarak *euclidean distance dan cosine similarity* dikombinasikan menggunakan data hasil reduksi fitur dan tidak reduksi.

BAB V : **PENUTUP**

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG

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.

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG