BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Saat ini, dengan kemajuan pesat dalam bidang teknologi, kita dapat mengumpulkan sejumlah besar data yang bermacam-macam. Penggalian data (*data mining*) muncul sebagai bidang yang berkaitan dengan ekstraksi informasi yang berguna dari data tersebut [1]. Salah satu alat *data mining* yang sering digunakan adalah *clustering*. *Clustering* merupakan teknik *data mining* tanpa pengawasan ketika label objek data tidak diketahui (*unsupervised*). Adalah tugas *clustering* untuk mengidentifikasi kategorisasi objek data yang sedang diperiksa. *Clustering* dapat diterapkan pada berbagai jenis data, termasuk teks. Saat berhadapan dengan data teks, objek bisa berupa dokumen, paragraf, atau kata [3].

Text clustering mengacu pada proses mengelompokkan dokumen teks serupa bersama-sama. Masalah dalam pengelompokan dokumen secara umum dapat diformulasikan sebagai berikut: Dengan diberikan suatu set dokumen, dokumen tersebut akan dibagi menjadi beberapa kelompok, sehingga dokumen dalam kelompok yang sama akan lebih mirip satu sama lain dibandingkan dokumen dalam kelompok lain. Ada banyak penerapan dari pengelompokan data teks, di antaranya: pengkategorisasian dokumen, peringkasan sebuah korpus, dan klasifikasi dokumen [16].

Text clustering dimulai dengan merepresentasikan dokumen teks ke dalam vektor dengan panjang tertentu. Representasi vektor yang paling umum untuk teks adalah *bag-of-words* dan *bag-of-n-gram* [8], karena kesederhanaan dan keefisienannya serta terkadang memiliki akurasi yang menakjubkan. Akan tetapi, tentu saja keduanya masih memiliki kelemahan. Sebagai contoh, pada *bag-of-words*, susunan kata dalam sebuah kalimat diacuhkan, sehingga dua kalimat yang memiliki susunan kata dan konteks yang berbeda bisa saja memiliki representasi vektor identik jika himpunan kata yang terkandung dalam kedua kalimat itu persis. Adapun pada *bag-of-n-grams* urutan kata

tetap diperhatikan meskipun dalam konteks yang pendek. Akan tetapi tingginya dimensi data menyebabkan mahalnya biaya komputasi. Selain itu, *bag-of-words* dan *bag-of-n-gram* juga memiliki kesensitifan yang kecil terhadap semantik sebuah kata, atau secara formalnya adalah jarak antar kata. Misal, kata "jeruk" bisa saja memiliki jarak yang lebih dekat dengan kata "Bandung" dibandingkan dengan kata "anggur" meskipun faktanya secara semantik, kata "jeruk" harusnya lebih dekat dengan "anggur" di mana keduanya merupakan nama buah-buahan dibandingkan "Bandung" yang merupakan nama sebuah kota.

Tomas Mikolov et al., pada [8] mengemukakan sebuah model bernama *paragraph vector* atau yang umum dikenal dengan nama *doc2vec*. Pada model tersebut, representasi vektor sebuah dokumen dilatih untuk memprediksi kata-kata dalam sebuah dokumen. Lebih tepatnya, model tersebut menggabungkan vektor dokumen dengan vektor kata yang dikandungnya, untuk kemudian memprediksi kata lainnya dalam konteks yang diberikan. Pelatihan dilakukan secara terus menerus untuk setiap potongan teks. Teks yang digunakan bervariasi, dapat berupa sebuah kalimat, paragraf, hingga sebuah dokumen. Hasil empiris menunjukkan bahwa *paragraph vector* mengungguli model *bag-of-words* serta teknik lainnya untuk representasi teks [9].

Penelitian pada tugas akhir ini merupakan pengembangan dari penelitian penulis sebelumnya untuk teks Al-Qur'an pada [16]. Penelitian tersebut lebih menitikberatkan penganalisisan pengaruh *proximity measure* pada macam-macam algoritma *clustering* berbasis partisi. Sebuah ayat pada Surat *Al-Baqarah* direpresentasikan sebagai sebuah dokumen menggunakan pendekatan *bag-of-words*. Vektor dokumen yang dihasilkan kemudian dikelompokkan dengan harapan ayat-ayat yang memiliki konteks yang sama dapat membentuk sebuah tema utuh. Penelitian-penelitian mengenai pengelompokan (*clustering*) ayat-ayat Al-Qur'an sebelumnya juga telah banyak dilakukan, di antaranya oleh Cepy Slamet et al., pada [17] dan oleh Miftachur Robani et al., pada [18]. Kedua penelitian tersebut juga

menggunakan pendekatan *bag-of-words* untuk merepresentasikan teks dengan asumsi yang sama: satu ayat, satu dokumen.

Pemilihan representasi dokumen sebagai sebuah ayat pada penelitian-penelitian sebelumnya menyebabkan penghirauan konteks yang dikandung, mengingat untuk mengetahui konteks sebuah ayat pada Al-Qur'an perlu meneliti ayat sebelum dan/atau sesudahnya [6]. Oleh karena itu, pada penelitian ini sebuah dokumen direpresentasikan oleh satu atau lebih ayat yang berurutan berdasarkan pokok bahasannya. Dengan demikian, dokumen yang diperoleh akan menempatkan ayat sesuai pada konteksnya. Analisis dan pembahasan juga lebih difokuskan pada representasi teks yang diperoleh menggunakan *paragraph vector*. Selain itu, data teks yang digunakan tidak terbatas pada terjemahan ayat-ayat Al-Qur'an saja, tetapi juga berbagai macam terjemahan tafsirnya.

Setiap muslim tentu meyakini bahwa Al-Qur'an adalah kitab suci yang berfungsi sebagai pedoman dan landasan utama untuk menjalani setiap aspek kehidupan. Dari ketentuan yang mengatur bagaimana manusia harus bersikap, hingga penciptaan alam semesta, semuanya tertuang dalam Al-Qur'an meski bahasannya tidak dijelaskan secara rinci. Suatu kenyataan yang dijumpai, bahwa seseorang membaca Al-Qur'an secara berurutan, ayat demi ayat, surat demi surat. Padahal belum tentu sebuah tema secara utuh dikemukakan dan tertuang dalam blok ayat yang berurutan [6]. Dengan kondisi yang demikian, pencarian tema dalam Al-Qur'an menjadi sangat sulit dilakukan. Oleh karena itu dengan mengelompokkan ayat berdasarkan kemiripan konteks yang dikandungnya diharapkan terbentuk kelompok ayat yang bersesuaian yang mampu merepresentasikan masing-masing tema yang terdapat dalam Al-Qur'an.

Motivasi pemilihan data teks Al-Qur'an juga didasari rasa kewajiban penulis sebagai seorang muslim untuk mempelajari Al-Qur'an, salah satunya dengan cara melakukan penelitian terhadapnya. Sehingga dengan mengelompokkan ayat-ayat Al-Qur'an berdasarkan topik yang dikandungnya diharapkan dapat memudahkan penulis khususnya, umumnya pembaca penelitian tugas akhir ini dalam menemukan sebuah topik/tema utuh dalam

Al-Qur’an. Banyak sekali perintah dan keutamaan-keutamaan dalam mempelajari Al-Qur’an, sebagai contoh, Allah *Subhanahu Wata’ala* berfirman:

كِتَابٌ أَنْزَلْنَاهُ إِلَيْكَ مُبَارَكٌ لِيَدَّبَّرُوا آيَاتِهِ وَلِيَتَذَكَّرَ أُولُو الْأَلْبَابِ

Artinya: “*Ini adalah sebuah kitab yang kami turunkan kepadamu penuh dengan berkah supaya mereka memperhatikan ayat-ayatnya dan supaya orang yang mempunyai pikiran mendapat pelajaran*” (QS. Shaad [38]: 29). Rasulullah *Shalallahu ‘Alaihi Wassalam* juga bersabda mengenai pentingnya mempelajari Al-Qur’an dalam hadis:

عَنْ عُثْمَانَ – رضى الله عنه- عَنِ النَّبِيِّ -صلى الله عليه وسلم- قَالَ «خَيْرُكُمْ مَنْ تَعَلَّمَ الْقُرْآنَ وَعَلَّمَهُ» رواه البخاري

Artinya: “*Ustman bin Affan radhiyallahu ‘anhu berkata: “Bahwa Rasulullah shallallahu ‘alaihi wasallam bersabda: “Sebaik-baik kalian adalah yang belajar Al-Qur’an dan mengajarkannya*.” (HR. Bukhari).

Setiap penelitian tentu tidak terlepas dari peran-peran metode sains dan teknologi yang digunakan. Agama Islam sendiri tidak pernah mengekang umatnya untuk maju dan modern. Sebaliknya, Islam sangat mendukung penganutnya untuk melakukan penelitian dalam bidang apapun. Akan tetapi di sisi lain, keberhasilan masyarakat modern dalam mengembangkan sains dan teknologi canggih untuk mengatasi berbagai masalah kehidupan dunia nyata tidak serta-merta menjamin penumbuhan akhlak yang mulia. Banyak produk-produk sains dan teknologi yang justru bertentangan dan jauh dari syariat Islam. Padahal seorang muslim yang taat akan meyakini bahwa Al-Qur’an adalah sumber segala pengetahuan dan kebenaran dengan ketentuan-ketentuan yang berlaku tidak lain adalah untuk kemaslahatan umat manusia.

Untuk itu, muncullah gagasan mengenai Islamisasi sains dan teknologi, di mana syariat Islam digunakan sebagai landasan utama untuk "menyaring" ide-ide pokok yang menyimpang sehingga produk-produk yang dihasilkan sesuai dengan ketentuan dan syariat-syariat-Nya. Hal inilah yang kemudian mendasari konsep "wahyu memandu ilmu" yang tertuang dalam visi Universitas Islam Negeri Sunan Gunung Djati Bandung: "*Menjadi Universitas Islam Negeri yang unggul dan kompetitif berbasis wahyu memandu ilmu dalam bingkai akhlak karimah di ASEAN tahun 2025.*" dengan misinya pada poin ke-2: "*Menyelenggarakan proses pembelajaran, penelitian, dan kajian ilmiah dengan bingkai akhlak karimah berbasis wahyu memandu ilmu untuk mengembangkan pengetahuan dan teknologi*". Sehingga jika syariat Islam sudah dijadikan sebagai landasan dan batasan aturan dalam penelitian, produk yang dihasilkan tentu akan sangat bermanfaat bagi kesejahteraan dan mampu menumbuhkan nilai moralitas (akhlak) penggunanya. Adapun konsep "wahyu memandu ilmu" ini merujuk pada ayat Al-Qur'an berikut.

إِنَّ فِي خَلْقِ السَّمَاوَاتِ وَالْأَرْضِ وَاخْتِلَافِ اللَّيْلِ وَالنَّهَارِ لَآيَاتٍ لِأُولِي الْأَلْبَابِ (١٩٠)
الَّذِينَ يَذْكُرُونَ اللَّهَ قِيَامًا وَقُعُودًا وَعَلَىٰ جُنُوبِهِمْ وَيَتَفَكَّرُونَ فِي خَلْقِ السَّمَاوَاتِ وَالْأَرْضِ
رَبَّنَا مَا خَلَقْتَ هَٰذَا بَاطِلًا سُبْحَانَكَ فَقِنَا عَذَابَ النَّارِ(١٩١)

Artinya: "*Sesungguhnya dalam penciptaan langit dan bumi, dan pergantian malam dan siang terdapat tanda-tanda (kebesaran Allah) bagi orang yang berakal, (yaitu) orang-orang yang mengingat Allah sambil berdiri, duduk atau dalam keadaan berbaring, dan mereka memikirkan tentang penciptaan langit dan bumi (seraya berkata), "Ya Tuhan kami, tidaklah Engkau menciptakan semua ini sia-sia; Mahasuci Engkau, lindungilah kami dari azab neraka*." (QS. Ali Imran [3]: 190-191).

Oleh karena itu, berdasarkan hal-hal tersebut di atas, penulis merasa tertarik mengangkat judul "*Document Embedding Menggunakan Paragraph*

Vector untuk Clustering Terjemahan Ayat-ayat Al-Qur'an". Teknik *clustering* CLARANS (*Clustering LArge Applications based on RANdomized Search*) kumudian dipilih mengingat terdapat 6.236 ayat dalam Al-Qur'an yang diharapkan mampu menekan biaya komputasi yang akan sangat besar. Terakhir, *cluster-cluster* yang terbentuk dievaluasi menggunakan metode pengevaluasian internal *cluster*, yakni *Davies-Bouldin Index* dan *Silhouette Coefficient* untuk mengetahui seberapa baik *cluster* yang diperoleh.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, maka permasalahan yang akan dibahas adalah sebagai berikut.

1. Bagaimana data teks direpresentasikan menggunakan *paragraph vector*?
2. Bagaimana karakteristik vektor yang dihasilkan menggunakan *paragraph vector* pada data teks terjemahan dan tafsir Al-Qur'an?
3. Bagaimana pengaruh *window* dan dimensi vektor terhadap nilai validasi yang diperoleh?
4. Bagaimana *cluster* yang terbentuk yang diperoleh CLARANS pada data terjemahan Al-Qur'an?
5. Bagaimana karakteristik *cluster* terbaik berdasarkan *Davies-Bouldin Index* dan *Silhouette Coefficient*?

1.3 Batasan Masalah

Adapun batasan masalah yang terdapat dalam penelitian ini dijelaskan sebagai berikut.

1. Data teks yang digunakan berupa terjemahan Al-Qur'an dan 5 buah tafsir Al-Qur'an dalam bahasa Indonesia.
2. Memfokuskan bahasan mengenai proses dan hasil representasi data teks menggunakan *paragraph vector*.
3. Validasi vektor berupa tes analogi pada vektor kata yang dikemukakan pada [9].

4. Teknik *clustering* yang digunakan berupa teknik *clustering* berbasis partisi: CLARANS.
5. Validasi *clustering* berupa pengukuran kualitas internal *cluster*: *Davies-Bouldin Index* dan *Silhouette Coefficient.*

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang dikemukakan di atas, maka tugas akhir ini memiliki tujuan sebagai berikut.

1. Merepresentasikan data teks ke dalam bentuk vektor menggunakan *paragraph vector*.
2. Menganalisis karakteristik vektor yang diperoleh menggunakan *paragraph vector* pada data terjemahan dan tafsir Al-Qur'an.
3. Menganalisis pengaruh *window* dan dimensi vektor terhadap nilai validasi yang diperoleh.
4. Menganalisis *cluster* yang terbentuk yang diperoleh CLARANS pada data terjemahan Al-Qur'an.
5. Menganalisis karakteristik *cluster* terbaik berdasarkan nilai *Davies-Bouldin Index* dan *Silhouette Coefficient*.

1.5 Metode Penelitian

Metode penelitian pada tugas akhir ini menggunakan pendekatan studi literatur. Pengkajian dilakukan dengan mencari referensi literatur berupa buku, jurnal, karya ilmiah, dan artikel yang berkaitan dengan metode-metode yang digunakan: struktur dan kandungan Al-Qur'an, representasi numerik untuk data teks (*word* dan *document embedding*), dan pengelompokan data teks (*text clustering*).

1.6 Sistematika Penulisan

BAB I PENDAHULUAN

Dalam bab ini dipaparkan latar belakang masalah berupa motivasi pemilihan topik penelitian-penelitian yang sudah

dilakukan sebelumnya, rumusan masalah, batasan masalah, tujuan penelitian, metode penelitian, serta sistematika penulisan dari penelitian yang akan dikaji.

BAB II LANDASAN TEORI

Bab ini mencakup hal-hal yang berkaitan dengan konsep dasar penelitian ini, di antaranya: penggalian data (*data mining*), penggalian teks (*text mining*), *clustering* teks (*text clustering*), pra-pemrosesan (*preprocessing*), jaringan saraf tiruan (*artificial neural network*), *word* dan *document embedding*, evaluasi *embedding*, algoritma *clustering* (*clustering algorithm*), validasi *clustering* (*clustering validation*), dan Pyhton.

BAB III *TEXT EMBEDDING* DAN ALGORITMA *CLUSTERING* CLARANS

Bab ini memaparkan inti penelitian matematika yang dilakukan, berupa penjelasan rinci (langkah demi langkah) model-model pada *paragraph vector* bekerja serta latar belakang, parameter, dan algoritma dari metode *clustering* yang digunakan, yakni CLARANS.

BAB IV *DOCUMENT EMBEDDING* MENGGUNAKAN *PARAGRAPH VECTOR* UNTUK *CLUSTERING* TERJEMAHAN AYAT-AYAT AL-QUR'AN.

Bab ini menjelaskan proses, hasil, dan analisis studi kasus yang dilakukan dalam penelitian ini. Dimulai dengan memberikan contoh proses *paragraph vector* bekerja untuk data teks sederhana, penjelasan langkah-langkah yang dilakukan pada data teks terjemahan dan tafsir Al-Qur'an (pengumpulan, pra-pemrosesan, dan pelatihan data), analisis *embedding* vektor, analisis *cluster-cluster* yang terbentuk yang diperoleh

CLARANS, hingga analisis *cluster* terbaik berdasarkan nilai terbaik *Davies-Bouldin Index* dan *Silhouette Coefficient*.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari penelitian yang telah dikaji sebelumnya, serta pemberian saran untuk pengembangan penelitian-penelitian selanjutnya.