BAB I
PENDAHULUAN

1.1. Latar Belakang

Kemajuan pesat dalam bidang teknologi pada era globalisasi saat ini telah mendorong pertumbuhan data yang semakin meningkat dan beragam. Hal ini menimbulkan tantangan dalam mengolah dan menggali informasi penting dari sekumpulan data besar tersebut. Salah satu bidang yang menarik perhatian adalah penambangan data (*data mining*) yang bertujuan untuk menggali informasi berharga yang tersembunyi dari suatu data dalam jumlah besar [1]. Penggalian informasi dari sejumlah besar data bermanfaat untuk banyak tugas seperti analisis bisnis, analisis perbankan, dan analisis teks.

Dalam beberapa tugas, *data mining* dapat diaplikasikan pada berbagai jenis data termasuk data berbentuk teks di mana objeknya dapat berupa dokumen, paragraf, atau kata. Penambangan teks (*text mining*) merupakan tugas yang memiliki tingkat kompleksitas yang tinggi karena data tekstual memiliki sifat ambigu (suatu kata dapat memiliki lebih dari satu makna). Oleh karena itu, dibutuhkan suatu teknik untuk menggali informasi penting yang tersembunyi dari teks tersebut. Salah satu teknik yang banyak digunakan dalam *text mining* yaitu *clustering*, suatu metode pembelajaran tanpa pengawasan (*unsupervised learning*) untuk melakukan pengelompokkan data yang tidak berlabel [2],[3]. Tujuan utama dari algoritma *clustering* adalah untuk memaksimalkan homogenitas dalam setiap *cluster* dan heterogenitas antar *cluster* yang berbeda [4]. Dengan kata lain, objek yang termasuk dalam *cluster* yang sama harus memiliki banyak fitur yang sama, tetapi sangat berbeda dengan objek data yang tidak termasuk dalam *cluster* tersebut. Objek-objek data yang dimaksud bisa berupa angka, kata, kalimat, dokumen, gambar dan lain sebagainya.

Secara umum, terdapat beberapa metode *clustering* yang populer yang banyak digunakan di antaranya yaitu *partitioning methods*, *hierarchical methods*, dan *density-base methods.* Proses *clustering* dengan *partitioning methods* yaitu

mengelompokkan setiap objek data ke dalam *k* partisi yang ditentukan oleh pengguna ketika proses *clustering* dimulai. Suatu objek dianggap mirip dengan objek-objek yang lain di dalam partisinya dan dianggap berbeda dengan objek-objek data yang berada di luar partisi tersebut. Setiap partisi mewakili sebuah *cluster* dengan syarat harus berisi setidaknya satu objek untuk setiap *cluster*nya dan setiap objek harus memiliki tepat satu *cluster*. Proses *clustering* dengan *hierarchical methods* terdiri dari dua teknik yaitu *divisive* dan *agglomerative*. Teknik *divisive* memulai dengan mengelompokkan seluruh objek data menjadi beberapa *cluster* besar, selanjutnya membaginya menjadi lebih kecil agar menghasilkan *cluster* yang lebih spesifik. Sementara teknik *agglomerative* memulai dengan mengelompokkan data menjadi *cluster-cluster* kecil kemudian menggabungkannya menjadi *cluster* yang lebih besar. Proses *clustering* dengan *density-base methods* melibatkan pengelompokkan objek data berdasarkan kepadatan dan jarak antara objek data. Metode ini memulai dengan sejumlah objek acak, kemudian memperluas *cluster* dengan memasukkan semua objek dalam jarak dan kepadatan yang ditentukan [5].

Salah satu algoritma *clustering* yang menggunakan *partitioning methods* adalah algoritma *k-means*, yang diusulkan oleh Mac Queen pada tahun 1967. Algoritma ini dapat digunakan di banyak bidang karena kesederhanaannya. Cara kerja algoritma *k-means* adalah dengan mempartisi data menjadi *k cluster* berdasarkan jarak rata-rata setiap objek data dengan *centroid* terdekat sesuai dengan nilai parameter *k* yang ditentukan oleh pengguna ketika proses *clustering* dimulai. Dengan partisi secara iteratif ini, *k-means* meminimalkan jarak antara setiap data dan kelompoknya. Karena kesederhanaan dan kemampuan untuk mengelompokkan data yang relatif cepat dan efisien, algoritma *k-means* menjadi sangat populer.

Di samping kepopulerannya itu, algoritma *k-means* memiliki kelemahan yaitu hasil akhir *k-means* sangat tergantung pada penentuan nilai parameter *k* untuk menentukan banyaknya *centroid* sebagai titik pusat dari setiap *cluster*, karena posisi *centroid* yang dipilih secara acak saat proses *clustering* dimulai dapat menghasilkan hasil *clustering* yang tidak optimal atau tidak seimbang secara langsung menggunakan seluruh data dalam satu langkah untuk dataset yang ukurannya

sangat besar dan non-globular sehingga algoritma *k-means* tidak menjamin untuk menghasilkan *cluster* yang unik [6]. Algoritma *k-means* tidak memiliki mekanisme bawaan untuk menentukan jumlah *cluster* yang optimal, sehingga pengguna harus mencoba beberapa variasi nilai *k* yang berbeda saat proses *clustering* dimulai untuk mendapatkan jumlah *cluster* yang optimal. Kemudian mengevaluasi *cluster* yang terbentuk dari beberapa variasi nilai *k* yang digunakan dalam percobaan untuk menentukan *cluster* terbaik berdasarkan evaluasi tertentu, misalnya dengan menggunakan metrik *silhouette coefficient, Davies-Bouldin Index* atau *Dunn Index*.

Penentuan nilai parameter *k* dalam algoritma *k-means* merupakan keputusan yang sangat penting karena penentuan nilai *k* yang tidak tepat dapat mempengaruhi secara signifikan kualitas dan keseimbangan hasil *clustering* yang diperoleh. Jika nilai *k* terlalu kecil akan menyebabkan hasil *clustering* menjadi *underfitting* yang ditandai dengan jumlah *cluster* yang terbentuk terlalu sedikit dibandingkan dengan tingkat heterogenitas yang ada dalam data. Ketidakseimbangan dalam ukuran cluster ini menyebabkan banyak poin data yang jauh dari *centroid cluster* mereka, di mana beberapa titik data yang memiliki tingkat kemiripan rendah atau bahkan tidak sama sekali akan dipaksa untuk dikelompokkan dalam satu *cluster* dengan titik-titik data yang memiliki tingkat kemiripan tinggi sehingga hasil *clustering* menjadi kurang representatif dan tidak mampu menangkap pola tersembunyi dalam data secara efektif. Sebaliknya, jika nilai *k* terlalu besar akan menyebabkan hasil *clustering* menjadi *overfitting* yang ditandai dengan jumlah *cluster* yang terbentuk terlalu banyak dibandingkan dengan tingkat heterogenitas yang ada dalam data, di mana beberapa titik data yang memiliki tingkat kemiripan tinggi akan dipaksa untuk dikelompokkan dalam beberapa *cluster* yang berbeda. Sebagian besar *cluster* yang terbentuk berukuran sangat kecil, seringkali hanya terdiri dari beberapa poin data atau bahkan satu poin data saja sehingga *cluster* yang terbentuk tidak memiliki makna yang jelas sehingga akan sulit untuk diinterpretasikan, hanya akan menambah kompleksitas komputasi tanpa memberikan informasi penting tentang pola yang tersembunyi dari data.

Beberapa pengembangan yang dilakukan untuk menyelesaikan kekurangan-kekurangan dari algoritma k-means standar di antaranya, yaitu pada menggunakan

pendekatan dengan mengombinasikan *k-means clustering* dan *hierarchical clustering* [6]. Prosesnya dimulai dengan menjalankan k-means beberapa iterasi, setiap iterasi menghasilkan *centroid* akhir yang berbeda karena inisialisasi *centroid* secara acak. Hasil dari iterasi-iterasi ini kemudian digabungkan menjadi satu set data baru yang terdiri dari semua *centroid* akhir yang dihasilkan. Algoritma hierarki kemudian diterapkan pada set data baru ini untuk menghasilkan *centroid* awal yang lebih baik untuk iterasi *k-means* berikutnya. Sementara pada menggunakan algoritma bisecting *k-means* [7], [8], pada saat inisialisasi, dimulai dengan menjadikan seluruh objek data menjadi satu *cluster* tunggal. Setelah *cluster* tunggal terbentuk, *cluster* tersebut selanjutnya dikelompokkan lagi menjadi dua *cluster* yang lebih kecil dengan menggunakan algoritma *k-means*. Sekali lagi *cluster* berikutnya dipecah secara rekursif menjadi dua *cluster* lagi dengan algoritma *k-means* dan begitu pula seterusnya hingga mencapai jumlah *cluster* yang diinginkan terpenuhi atau tidak ada lagi *cluster* yang dapat dipecah kembali (angota dari suatu *cluster* kurang dari dua anggota.

Sebelum melakukan *text mining*, kita harus merepresentasikan kata ke dalam bentuk vektor numerik. Karena sebagian besar algoritma *machine learning* hanya menerima input berupa vektor numerik. Metode representasi vektor kata tradisional di antaranya, yaitu *bag-of-words* (BOW) dan *term frequency-inverse document frequency* (TF-IDF) mengubah dokumen menjadi vektor yang sangat besar karena ukurannya sama dengan jumlah kata dalam kosakata [9]. Namun, kedua metode ini tidak memperhitungkan konteks di mana kata muncul, sehingga menghasilkan vektor yang jarang dan boros memori. Pada metode *bag-of-words* (BOW) ini memiliki kekurangan yaitu mengabaikan urutan kata, sehingga jika ada dua kalimat yang berbeda dapat mempunyai representasi vektor yang sama jika memiliki himpunan kata yang identik. Selain itu, *bag of words* juga kurang sensitif terhadap makna antar kata. Misalnya, kata “kucing” dapat mempunyai jarak yang lebih dekat dengan kata “Depok” dibandingkan dengan kata “kelinci”, meskipun secara semantik, kata “kucing” lebih dekat dengan kata “kelinci” karena keduanya adalah nama-nama binatang bukan nama-nama daerah. Sementara pada *term frequency-inverse document frequency* (TF-IDF) memiliki beberapa kelemahan yang signifikan, meskipun efektif dalam banyak situasi. Ketidakmampuan TF-IDF untuk

menangkap makna semantik dari kata-kata adalah salah satu kelemahan utamanya. Meskipun TF-IDF menghitung frekuensi kata yang muncul dalam dokumen dan mengurangi bobot kata yang sering muncul, TF-IDF tidak memperhitungkan hubungan semantik antar kata. Misalnya, kata "kucing" dan "binatang" mungkin memiliki makna yang mirip, namun dalam konteks TF-IDF, mereka dianggap sebagai entitas yang berbeda yang tidak mempunyai keterikatan satu sama lain. BOW dan TF-IDF hanya menghitung berapa kali kata muncul tanpa mempertimbangkan makna kata dalam kalimat. TF-IDF bahkan lebih mahal secara komputasi dalam kasus di mana korpus datanya besar. Karena keduanya tidak mempertahankan informasi kontekstual dari kata-kata, sehingga tidak dapat memahami makna yang lebih dalam dari suatu kata.

Pada tahun 2013 Tomas Mikolov dan kawan-kawannya di Google memperkenalkan *word2vec* sebagai model representasi vektor kata yang lebih canggih dibandingkan BOW dan TF-IDF [9], [10]. Model ini mengubah kata-kata menjadi vektor angka yang lebih efisien dan tidak menghilangkan makna kontekstual mereka. Misalnya, kata-kata yang sering muncul bersama dalam konteks yang sama akan memiliki vektor yang dekat satu sama lain di dalam ruang vektor. *Word2vec* menghasilkan vektor yang padat, yang menghemat lebih banyak memori dan mempercepat pemrosesan. Selain itu, *word2vec* lebih efektif dalam menangani tugas-tugas seperti pengelompokkan teks dibandingkan dengan BOW dan TF-IDF karena desainnya yang sederhana dan berfungsi baik untuk dataset kecil maupun besar. Hal ini karena *word2vec* mempertahankan hubungan semantik antar kata-kata dalam kalimat atau dokumen, sehingga makna kontekstual tidak hilang.

Relasi semantik merupakan topik kajian dalam bidang linguistik, terutama pada pengolahan bahasa alami (NLP) yang menarik perhatian untuk penelitian saat ini. Relasi semantik sangat penting dalam salah satu tugas pengolahan bahasa alami, yaitu pengelompokkan kata. Pengelompokkan kata merujuk pada proses mengelompokkan kata menjadi beberapa *cluster* berdasarkan kemiripannya sehingga kata-kata yang berada dalam *cluster* yang sama memiliki tingkat kemiripan yang lebih tinggi dari pada kata-kata yang berada dalam *cluster* yang

lainnya. Kemiripan *cluster-cluster* tersebut dapat berdasarkan relasi semantik, sinonim, atau kategori yang lainnya.

Penelitian-penelitian sebelumnya yang terkait mengenai pengelompokan kata menggunakan algoritma *k-means* juga telah dilakukan, di antaranya oleh Habib M et al., pada [11] dan oleh Soliman A et al., pada [12]. Kedua penelitian tersebut juga menggunakan algoritma *k-means* dalam mengelompokkan kata menjadi beberapa *cluster* berdasarkan kemiripannya. Namun kedua penelitian tersebut tidak menggunakan seluruh data dalam proses *clustering*nya, melainkan hanya mengambil beberapa sampel secara acak untuk digunakan dalam proses *clustering*. Sehingga kedua penelitian tersebut tidak melakukan *clustering* kata secara menyeluruh dari dataset yang digunakan. Karena *k-means* kesulitan untuk melakukan *clustering* secara langsung menggunakan seluruh data dalam satu langkah pada dataset yang jumlahnya sangat besar dan kompleks.

Penelitian pada tugas akhir ini mengusulkan metode algoritma *k-means* bertingkat yang mirip seperti *bisecting k-means*. Metode *k-means* bertingkat bertujuan untuk menemukan struktur kelompok kata berdasarkan kemiripannya dengan membaginya secara berulang (rekursif) ke dalam *cluster* yang lebih kecil dan spesifik. Pada *k-means* bertingkat dimulai dengan menentukan jumlah *cluster* menggunakan *k-means* untuk membagi seluruh dataset menjadi *k cluster* awal (*cluster* induk). Setelah *k cluster* awal terbentuk, setiap *cluster* tersebut selanjutnya dikelompokkan lagi menjadi *k cluster* yang lebih kecil dengan menggunakan kembali algoritma *k-means.* Sekali lagi *cluster* berikutnya dipecah secara rekursif menjadi *k cluster* lagi dengan algoritma *k-means* dan begitu pula seterusnya. Proses *clustering* ini terus dilakukan berulang kali pada setiap tingkatnya dan akan berhenti ketika mencapai jumlah anggota dalam suatu *cluster* tidak dapat dipecah lagi atau hingga mencapai jumlah iterasi yang telah ditentukan terpenuhi. Dengan pendekatan ini memungkinkan kita untuk mengungkap struktur hierarkis dalam data, di mana setiap tingkat hierarki menunjukkan pengelompokan data yang lebih spesifik dan terfokus. Ini sangat berguna untuk dataset yang jumlahnya sangat besar dan kompleks yang tidak dapat dicapai dengan menggunakan *k-means* standar yang melakukan *clustering* secara langsung menggunakan seluruh data dalam satu

langkah. Sehingga diharapkan *k-means* bahwa metode bertingkat yang diusulkan pada penelitian ini untuk menemukan kelompok-kelompok kata berdasarkan keterkaitan dan kesamaan semantiknya dapat memperbaiki kekurangan *k-means* standar dan memberikan wawasan baru yang lebih komprehensif pada dataset yang jumlahnya sangat besar dan kompleks.

Dataset yang dipilih pada penelitian ini adalah data teks Al-Qur'an. Al-Qur'an adalah kitab suci yang diturunkan kepada Nabi Muhammad saw. sebagai sumber utama hukum, ilmu pengetahuan, hikmah, dan petunjuk bagi umat manusia [13]. Al-Qur'an terdiri dari 114 surah dan 30 juz. Namun, terdapat perbedaan pendapat para ulama tentang banyaknya ayat Al-Qur'an [14], Terdapat tujuh mazhab yang terkenal menghitung jumlah ayat al-Qur'an. Yang pertama, Al-Madanî al-Awwal menghitung 6.217 atau 6.214 ayat, yang kedua, Al-Madanî al-Akhîr menghitung 6.214 ayat, yang ketiga, Ahl Mekkah menghitung 6.210 ayat, yang keempat, Ahl Bashrah menghitung 6.204 ayat, yang kelima, Ahl Damaskus menghitung 6.227 atau 6.226 ayat, dan yang keenam, Al-Humushi menghitung 6.232 ayat, dan yang ketujuh, ahl Kufah menghitung sebanyak 6.236 ayat. Mushaf al-Qur'an yang diterbitkan di Indonesia total ayatnya sebanyak 6.236 ayat. Al-Qur'an memiliki banyak informasi yang tersebar di dalamnya mengenai kata-kata yang memiliki makna yang terkait atau mirip. Salah satu cara untuk memahami Al-Qur'an adalah dengan mencoba memahami keterkaitan makna antar kata yang terkandung di dalam ayat-ayat Al-Qur'an. Pada penelitian ini bertujuan untuk menemukan kelompok-kelompok kata berdasarkan kemiripan dan keterkaitan maknanya dengan menggunakan metode *clustering*.

Penelitian-penelitian sebelumnya mengenai keterkaitan semantik antar kata pada Al-Qur'an juga telah dilakukan, di antaranya yaitu pada [15] menganalisis pencarian keterkaitan kata-kata yang termasuk kata benda saja. Penelitian pada tugas akhir ini merupakan pengembangan dari penelitian pada [16] yang membahas pembangunan daftar kata terkait pada kosa kata Al-Qur'an berdasarkan kesamaan distribusional dari vektor *word2vec* dan *cosine-similarity*, terdapat beberapa kekurangan yang perlu diperhatikan. Salah satu yang paling signifikan adalah

bahwa penelitian tersebut tidak menganalisis pola relasi semantik antar kata pada ayat-ayat Al-Qur'an.

Metode kesamaan distribusional dari vektor *word2vec* dan *cosine-similarity* cenderung berkonsentrasi pada menghitung kesamaan antar kata secara langsung berdasarkan representasi vektor kata, tetapi tidak menggambarkan berdasarkan pola keterikatan antar kata tertentu yang lebih kompleks atau tersembunyi dalam ayat-ayat Al-Qur'an. Oleh karena itu, dengan menggunakan teknik *clustering* pada penelitian ini untuk mengelompokkan kata-kata dalam Al-Qur'an berdasarkan kemiripannya, dapat memungkinkan untuk menemukan pola hubungan antar kata yang lebih kompleks dan mendalam. Sehingga diharapkan dapat memperbaiki kekurangan dalam analisis sebelumnya dan memberikan wawasan baru yang lebih komprehensif mengenai keterikatan antar kata dalam ayat-ayat Al-Qur'an.

Motivasi pemilihan data teks Al-Qur'an didasari oleh kewajiban penulis sebagai seorang muslim untuk mempelajari Al-Qur'an, termasuk melalui penelitian tugas akhir ini. Penulis berharap dengan mengelompokkan kata-kata yang terkandung dalam Al-Qur'an berdasarkan kemiripan dan keterkaitan maknanya untuk menemukan dan mempelajari bagaimana kata-kata dalam Al-Qur'an berhubungan satu sama lain sehingga dapat membantu mempelajari isi ayat-ayat Al-Qur'an dan menggali pengetahuan yang tersembunyi di dalamnya. Terdapat banyak ayat-ayat Al-Qur'an yang menjelaskan perintah dan keutamaan-keutamaan dalam mempelajari Al-Qur'an, sebagai contoh, Allah swt. berfirman:

كِتٰبٌ اَنْزَلْنٰهُ اِلَيْكَ مُبٰرَكٌ لِّيَدَّبَّرُوْٓا اٰيٰتِهٖ وَلِيَتَذَكَّرَ اُولُوا الْاَلْبَابِ

Artinya: *"(Al-Qur'an ini adalah) kitab yang Kami turunkan kepadamu (Nabi Muhammad) yang penuh berkah supaya mereka menghayati ayat-ayatnya dan orang-orang yang berakal sehat mendapat pelajaran."* (Q.S. Sad [38]: 29).

Rasulullah saw. juga bersabda mengenai keutamaan mempelajari Al-Qur'an, salah satunya terdapat dalam hadis berikut:

وَعَنْ عُثْمَانَ بْنِ عَفَّانَ – رَضِيَ اللهُ عَنْهُ – ، قَالَ : قَالَ رَسُوْلُ اللهِ – صَلَّى اللهُ عَلَيْهِ

وَسَلَّمَ – : ((خَيْرُكُمْ مَنْ تَعَلَّمَ الْقُرْآنَ وَعَلَّمَهُ)) رَوَاهُ الْبُخَارِيُّ .

Utsman bin 'Affan radhiyallahu 'anhu berkata bahwa Rasulullah shallallahu 'alaihi wa sallam bersabda, "Sebaik-baik orang di antara kalian adalah yang belajar Al-Qur'an dan mengajarkannya." (HR. Bukhari, no. 5.027).

Oleh karena itu, berdasarkan hal-hal tersebut di atas, penulis akan melakukan penelitian yang berjudul *"Relasi Semantik Kata pada Data Terjemahan Al-Qur'an Bahasa Indonesia Menggunakan Algoritma Clustering K-Means Bertingkat"*. Teknik *clustering k-means* bertingkat diusulkan mengingat terdapat 6.236 ayat dalam Al-Qur'an yang diharapkan mampu menekan biaya komputasi yang akan sangat besar dan mengoptimalkan kinerja algoritma *k-means* pada data bervolume besar. Selanjutnya, untuk mengetahui seberapa baik *cluster* yang dihasilkan yaitu menggunakan metode pengevaluasian internal *clustering* seperti *silhouette coefficient*. Sementara untuk mengevaluasi keterkaitan semantik antar kata dari kelompok-kelompok kata yang dihasilkan yaitu dengan menghitung metrik *cosine similarity*. Langkah terakhir dari penelitian ini yaitu melakukan analisis dengan mengamati pola relasi semantik kata dari *cluster-cluster* yang dihasilkan pada ayat-ayat Al-Qur'an. Proses ini bertujuan untuk melihat bagaimana kata-kata tersebut muncul dalam konteks ayat-ayat Al-Qur'an.

1.2. Rumusan Masalah

Rumusan masalah pada skripsi ini di antaranya sebagai berikut:

1. Penentuan nilai parameter *k* untuk inisialisasi *centroid* pada algoritma *k-means* mengalami kesulitan atau gagal dalam membentuk *cluster* dengan ukuran yang seimbang secara langsung dalam satu langkah ketika data yang dikelompokkan memiliki volume yang besar dan bentuk non-globular.
2. Bagaimana pola karakteristik relasi semantik kata dari *cluster* yang diperoleh pada ayat-ayat Al-Qur'an.

1.3. Batasan Masalah

Batasan masalah pada skripsi ini di antaranya sebagai berikut:

1. Data yang digunakan adalah data terjemahan Al-Qur'an bahasa Indonesia.

2. Teknik representasi vektor kata yang digunakan yaitu *word2vec* dengan dimensi *embedding* yang digunakan adalah 300 dan ukuran *window* sebesar 10.

3. Teknik *clustering* yang digunakan berbasis partisi yaitu *k-means* bertingkat.

4. Validasi *clustering* berupa pengukuran kualitas internal *cluster* dengan menggunakan metrik *silhouette coefficient*.

5. Validasi relasi semantik kata menggunakan metrik *cosine similarity.*

1.4. Tujuan Penelitian

Tujuan pada penelitian skripsi ini adalah sebagai berikut:

1. Menjelaskan algoritma *k-means* bertingkat untuk mengatasi masalah penentuan nilai parameter *k* untuk inisialisasi *centroid* pada algoritma *k-means* mengalami kesulitan atau gagal dalam membentuk *cluster* dengan ukuran yang seimbang secara langsung dalam satu langkah ketika data yang dikelompokkan memiliki volume yang besar dan bentuk non-globular.

2. Menganalisis pola karakteristik relasi semantik kata dari *cluster* yang diperoleh pada ayat-ayat Al-Qur’an.

Adapun manfaat yang dapat diperoleh dari penelitian ini di antaranya sebagai berikut:

1. Hasil penelitian ini diharapkan menjadi salah satu bentuk pengembangan dari algoritma *k-means* untuk menghasilkan *cluster* yang optimal dan seimbang.

2. Dengan mengelompokkan kata-kata yang terkandung dalam Al-Qur’an berdasarkan keterkaitan maknanya untuk menemukan dan mempelajari bagaimana kata-kata dalam Al-Qur'an berhubungan satu sama lain sehingga dapat membantu mempelajari isi ayat-ayat Al-Qur’an dan menggali pengetahuan yang tersembunyi di dalamnya.

1.5. Metode Penelitian

Metode penelitian pada tugas akhir ini yaitu sebagai berikut

1. Studi Literatur

Tahap studi literatur dilakukan pengumpulan data dan referensi literatur berupa buku, jurnal, karya ilmiah, dan artikel yang berkaitan dengan metode-metode yang digunakan yaitu, teknik representasi vektor kata untuk data teks (*word embeddings*), dan teknik pengelompokan data teks.

2. Simulasi

Pada tahap ini dilakukan simulasi percobaan pada data Al-Qur'an menggunakan algoritma *clustering k-means* bertingkat serta melakukan analisis pada hasil *cluster* yang diperoleh.

1.6. Sistematika Penulisan

Berdasarkan sistematika penulisannya, tugas akhir ini terdiri atas lima bab serta daftar pustaka, di mana dalam setiap bab terdapat beberapa subbab.

BAB I : PENDAHULUAN

Bab ini berisi tentang pembahasan latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, metode penelitian, dan sistematika penulisan.

BAB II : LANDASAN TEORI

Bab ini berisi penjelasan mengenai teori-teori yang berkaitan dengan penelitian ini, di antaranya yaitu penambangan data (*data mining*), penambangan teks (*text mining*), pra-pemrosesan (*pre-processing*), natural language processing, jaringan saraf tiruan (*artificial neural network*), *word embedding*, metode *clustering*, validasi *clustering*, dan *literature review*

BAB III : RELASI SEMANTIK KATA PADA DATA TERJEMAHAN AL-QUR'AN BAHASA INDONESIA MENGGUNAKAN ALGORITMA *CLUSTERING K-MEANS* BERTINGKAT

Pada bab ini berisi penjelasan tentang inti metodologi penelitian yang dilakukan, berupa pembahasan rinci tentang pengumpulan data, *pre-processing*, metode *word embeddings word2vec* dan algoritma *clustering k-means* bertingkat.

BAB IV : ANALISIS STUDI KASUS PADA DATA TEKS AL-QUR'AN BAHASA INDONESIA

Pada bab ini berisi pembahasan mengenai analisis hasil studi kasus yang dilakukan dalam penelitian ini. Dimulai dari penjelasan dataset yang digunakan, *embedding* kata menggunakan *word2vec*, pengelompokkan kata dengan menggunakan *k-means* bertingkat, dan analisis keterkaitan semantik antar kata pada kelompok-kelompok kata yang terbentuk yang diperoleh.

BAB V : PENUTUP

Pada bab ini berisi kesimpulan dari penelitian yang dilakukan dan saran untuk pengembangan penelitian ini selanjutnya.