BAB I
PENDAHULUAN

1.1. Latar Belakang

Al-Qur'an merupakan sumber pengetahuan yang paling penting dan dianggap sebagai acuan teks standar yang memuat berbagai cerita berharga, dan hikmah warisan umat Islam. Al-Qur'an sebagai sumber hukum Islam, memiliki peran vital dalam membentuk pemahaman dan praktik agama Islam. Al-Quran adalah kitab suci yang diturunkan oleh Allah *Subhanahu wa ta'ala*, kepada hambanya yang mulia, nabi Muhammad *Shallallahu' alaihi wasallam* melalui malaikat jibril alaihis salam, sebagai pedoman hidup dan pemberi peringatan kepada umat manusia[1]. Seperti yang difirmankan Allah subhanahu wa ta'ala dalam Quran Surat Asy-Syu'ara ayat 192-194:

وَاِنَّهٗ لَتَنْزِيْلُ رَبِّ الْعٰلَمِيْنَ ١٩٢ نَزَلَ بِهِ الرُّوْحُ الْاَمِيْنُۙ ١٩٣ عَلٰى قَلْبِكَ لِتَكُوْنَ مِنَ الْمُنْذِرِيْنَۙ

"Sesungguhnya ia (Al-Qur'an) benar-benar diturunkan Tuhan semesta alam (192), Ia (Al-Qur'an) dibawa turun oleh Ruhulamin (Jibril) (193), (Diturunkan) ke dalam hatimu (Nabi Muhammad) agar engkau menjadi salah seorang pemberi peringatan(194)."

Adapun hadis yang berkaitan dengan turunnya Al-Quran diriwayatkan oleh Al-Hakim no.787 yang berbunyi

أنزلَ ٱلْقُرْأنَ جُمْلَةً وَاحِدَةً اِلىَ سَمَآءِ الدُّنْيَا وَكَانَ بِمَوَاقِعِ النُّجُوْمِ وَكَانَ اللهُ يَنْزِيْلُهُ عَلىَ رَسُوْلِهِ بَعْضَهُ فِيْ بَعْضٍ

"Al-Qur'an diturunkan secara sekaligus ke langit dunia, dan hal itu adalah seperti perpindahan bintang-bintang. Allah menurunkannya kepada Nabi Muhammad SAW sedikit demi sedikit ". (HR. Al-Hakim) [No. 787].

Al-Qur'an sebagai kitab suci umat Islam memiliki pesan-pesan yang dalam dan kompleks, yang sering kali diungkapkan melalui susunan kata-kata yang kaya akan makna. Oleh karena itu, berbagai versi terjemahan dan tafsir Al-Qur'an dalam Bahasa Indonesia telah dikembangkan untuk memperluas pemahaman terhadap makna ayat-ayat suci. Masing-masing versi baik terjemahan literal maupun tafsir

kontekstual menghadirkan nuansa berbeda dalam penyampaian pesan dan penggunaan kata. Perbedaan ini dapat menyebabkan variasi dalam pola kemunculan kata, sehingga model pemahaman semantik seperti *Word2Vec* yang mengandalkan konteks sekitar kata berpotensi menghasilkan representasi yang berbeda untuk kata yang sama tergantung pada sumber datanya. Namun, proses penerjemahan tidak selalu mampu menangkap seluruh makna semantik dari kata-kata asli dalam bahasa Arab, terutama ketika satu kata dapat memiliki beragam makna tergantung pada konteksnya. Misalnya di antara istilah penting yang digunakan untuk menyebut Tuhan adalah *Rabb* dan *Ilah*. Meskipun keduanya sering kali diterjemahkan serupa dalam bahasa Indonesia, keduanya digunakan dalam konteks yang berbeda dan memiliki nuansa makna yang khas. Untuk memahami relasi semantik antara kata *Rabb* dan *Ilah*, diperlukan pendekatan yang tidak hanya mengandalkan terjemahan literal, tetapi juga mempertimbangkan konteks kemunculannya dalam ayat-ayat Al-Qur'an. *Kesamaan semantik* merupakan suatu pengukuran untuk mencari nilai yang menyatakan tingkat kesamaan atau kedekatan secara semantik, baik antar kata, kalimat, maupun teks[2]. Sepasang kata dikatakan memiliki kesamaan semantik apabila memiliki kesamaan dari sisi makna atau konsep[3]. Perhitungan kesamaan semantik dilatarbelakangi oleh tantangan dalam pemrosesan bahasa alami, di mana mesin belum dapat menyamakan persepsi manusia dengan baik. Oleh karena itu, perhitungan kesamaan semantik digunakan untuk membantu mesin memahami bahasa manusia secara lebih akurat.

Untuk memahami hubungan semantik antar kata secara lebih mendalam, pendekatan berbasis teknologi pengolahan bahasa alami (*Natural Language Processing*) dapat dimanfaatkan[3]. Konsep kesamaan semantik telah diterapkan secara luas di berbagai bidang seperti *information retrieval, query suggestion, automatic summarization*, dan pencarian berbasis gambar[4]. Selain itu, kajian ini juga menjadi topik penting dalam disiplin ilmu seperti psikologi, linguistik, ilmu kognitif, biomedis, dan kecerdasan buatan (*Artificial Intellegent*)[5]. Tugas utama dalam kesamaan semantik adalah menentukan tingkat kedekatan antara dua kata. Karena komputer hanya dapat memproses data dalam bentuk numerik, maka kata-kata harus terlebih dahulu dikonversi menjadi representasi angka. Pendekatan yang

dapat digunakan untuk mengukur kedekatan antar kata adalah dengan menganalisis vektor kata yang mewakili maknanya. Salah satu metode yang umum digunakan dalam pendekatan tersebut adalah kesamaan semantik berbasis vektor dengan menghitung sudut antara dua vektor menggunakan *cosine similarity[6]*. Teknik ini memungkinkan untuk menilai seberapa mirip dua kata berdasarkan orientasi vektornya dalam ruang multidimensi. Dalam konteks NLP, salah satu tantangan utamanya adalah bagaimana mengubah representasi kata dari bentuk simbolik menjadi bentuk numerik agar dapat diproses oleh komputer. Representasi numerik inilah yang memungkinkan sistem untuk melakukan perhitungan kesamaan makna secara matematis.

Salah satu pendekatan populer dalam merepresentasikan kata dalam bentuk vektor adalah dengan menggunakan model *Word2Vec*, yang diperkenalkan oleh Mikolov et al. pada tahun 2013[7]. Model ini bekerja dengan cara memetakan kata ke dalam ruang vektor berdimensi tinggi berdasarkan distribusi kata-kata lain yang muncul di sekitarnya. Dengan demikian, kata-kata yang memiliki konteks penggunaan serupa akan memiliki representasi vektor yang saling berdekatan. Model ini didasarkan pada *hipotesis distribusional* yang menyatakan bahwa kata-kata yang muncul dalam konteks serupa cenderung memiliki makna yang mirip[8]. Dengan demikian, representasi semantik dari suatu kata sangat bergantung pada konteks kata-kata yang mengelilinginya dalam sumber data yang digunakan. Model *Skip-gram*, yang merupakan salah satu arsitektur dalam *Word2Vec*, terbukti efisien dalam melatih representasi vektor kata pada dataset berukuran besar.

Meskipun pendekatan berbasis vektor seperti *Word2Vec* dan perhitungan cosine similarity cukup efektif dalam mengukur kedekatan makna antar kata, metode ini memiliki keterbatasan. Analisis kesamaan berbasis pasangan kata (*pairwise similarity*) belum cukup mampu mengungkap pola keterkaitan semantik yang lebih kompleks dan tersembunyi dalam keseluruhan teks, seperti hubungan antar kata yang tersebar dalam berbagai ayat dalam Al-Qur'an. Oleh karena itu, dibutuhkan pendekatan yang lebih mendalam dan struktural untuk menganalisis relasi semantik antar kata secara lebih komprehensif.

Salah satu pendekatan yang dapat digunakan untuk mendalami relasi semantik adalah teknik *clustering*, yang merupakan bagian dari metode *data mining[9]*. *Clustering* bertujuan untuk mengelompokkan data ke dalam kelompok atau klaster yang memiliki karakteristik serupa[10]. Dalam konteks ini, clustering menjadi teknik yang sangat penting untuk mengelompokkan kata-kata berdasarkan representasi vektornya. *Clustering* memungkinkan untuk menemukan pola semantik tersembunyi yang tidak bisa langsung diamati melalui perhitungan kesamaan pasangan kata (*pairwise similarity*) saja. Pendekatan ini dapat membantu mengungkap struktur semantik dalam teks Al-Qur'an, seperti kelompok kata yang terkait dengan tema Aqidah dan Iman, Akhirat, atau Kisah Nabi, tanpa perlu mendefinisikan hubungan tersebut secara eksplisit.

Metode clustering tradisional seperti *K-Means* merupakan algoritma yang banyak digunakan dalam pengelompokan data karena kesederhanaan dan efisiensinya. Namun, salah satu kelemahan mendasar dari *K-Means* adalah asumsi bentuk klaster yang linier dan cenderung bulat (isotropis). Akibatnya, algoritma ini kurang efektif dalam mengelompokkan data yang memiliki struktur non-linier atau distribusi kompleks, seperti yang sering ditemui pada representasi vektor kata hasil pembelajaran model *Word2Vec*. Dalam pengelompokan data semantik yang kompleks seperti ayat Al-Qur'an, struktur cluster tidak selalu berbentuk sferis atau terpisah secara linear. Hal ini menyebabkan *K-Means* memiliki keterbatasan dalam mengenali pola cluster yang bersifat non-konveks atau saling bertumpuk, sehingga hasil clustering bisa kurang optimal dan kurang merefleksikan hubungan semantik yang sesungguhnya.

Sebagai solusi atas keterbatasan ini*, Spectral Clustering* diusulkan sebagai metode alternatif yang mampu mengatasi permasalahan tersebut. *Spectral Clustering* bekerja dengan membangun graf yang merepresentasikan hubungan kemiripan antar data melalui matriks afinitas, lalu menggunakan informasi spektral (eigenvector dari matriks Laplacian) untuk mereduksi dimensi dan menemukan struktur cluster yang lebih kompleks dan fleksibel

Spectral Clustering merupakan algoritma berbasis graf yang bekerja dengan memanfaatkan struktur konektivitas data dalam bentuk graf melalui matriks

afinitas[11]. Berbeda dari algoritma tradisional seperti *K-Means* yang hanya bekerja optimal untuk bentuk klaster *sferis*, *Spectral Clustering* mampu mendeteksi bentuk klaster yang tidak teratur dan kompleks, serta menyatu ke solusi global yang optimal. Proses utamanya melibatkan transformasi data menjadi graf, menghitung vektor eigen dari matriks Laplacian, dan memanfaatkan hasilnya untuk proses pengelompokan akhir[12]. Evaluasi hasil clustering menjadi hal penting untuk memastikan bahwa klaster yang terbentuk memiliki kualitas yang baik dan bermakna secara semantik. *Silhouette coefficient dan Conductance* menjadi acuan untuk menilai hasil *spectral clustering*. *Silhouette coefficient* mengukur seberapa baik suatu kata cocok dengan klaster tempat ia berada dibandingkan dengan klaster lain. Nilainya berkisar antara -1 hingga 1, dengan nilai mendekati 1 menunjukkan bahwa kata tersebut berada pada klaster yang sesuai. Sementara itu, *Conductance* merupakan metrik berbasis graf yang menghitung seberapa kuat kohesi internal suatu klaster dibandingkan dengan koneksi keluarannya. Nilai *conductance* yang rendah menunjukkan bahwa klaster tersebut memiliki keterikatan yang tinggi antar anggotanya dan terpisah dengan baik dari klaster lainnya.

Penelitian-penelitian sebelumnya yang melakukan pengelompokan menggunakan Spectral Clustering, diantaranya oleh Li Zhang dkk pada publikasinya yang berjudul "*Automatic Synonym Extraction using Word2vec and Spectral clustering*"[13]. Dalam penelitian tersebut diusulkan sebuah metode untuk mengekstraksi sinonim dari teks dengan menggabungkan *Word2Vec* dan *Spectral Clustering*. Metode ini diawali dengan pelatihan model *Word2Vec* menggunakan korpus besar seperti *Wikipedia* bahasa Inggris, berukuran 1.1 GB. Untuk mengevaluasi efektivitas metode, dilakukan perbandingan antara *Spectral Clustering* dan *K-Means*. Sebanyak 30 teks dipilih untuk pengujian, dan sinonim dari setiap teks dikumpulkan serta divalidasi oleh 20 orang. Sinonim yang diperoleh manusia digunakan sebagai *ground truth*. Untuk mengevaluasi kinerja metode ekstraksi sinonim, digunakan tiga metrik utama, yaitu *Precision*, *Recall*, dan *F1-score*. Hasil pengujian pada 30 teks menunjukkan bahwa *Spectral Clustering* secara signifikan lebih unggul dibandingkan *K-Means*. *Spectral Clustering* mencapai *precision* 0,808, *recall* 0,744, dan *F1-score* 0,775, sedangkan *K-Means* hanya memperoleh *precision* 0,279, *recall* 0,473, dan *F1-score* 0,351. Temuan ini

membuktikan bahwa *Spectral Clustering* lebih akurat dan seimbang dalam mengekstraksi sinonim.

Penelitian lainnya terkait implemantasi algoritma *Spectral Clustering* dilakukan oleh Cahyaningrum dkk dalam publikasinya yang berjudul "*Implementation of spectral clustering with partitioning around medoids (PAM) algorithm on microarray data of carcinoma*" pendekatan *Spectral Clustering* telah diterapkan dalam bidang biomedis, penelitian tersebut bertujuan untuk mengelompokkan data microarray adenoma[14]. Dalam penelitian tersebut, digunakan algoritma *Spectral Clustering* yang dipadukan dengan *Partitioning Around Medoid* (PAM) untuk mengelompokkan 7.087 data microarray menjadi tiga klaster berdasarkan nilai kesamaan ekspresi gen, dengan hasil nilai average *silhouette coefficient* sebesar 0,528. Evaluasi clustering dalam penelitian tersebut hanya mengandalkan satu metrik, yaitu *silhouette coefficient*, untuk mengukur kualitas pengelompokan.

Sebagai bentuk pengembangan dari penelitian sebelumnya, penelitian ini menggunakan dataset mengenai ayat-ayat yang berkaitan dengan tema ketuhanan yang bersumber dari terjemahan Al-Quran bahasa Indonesia versi Kemenag. Dalam penelitian ini tidak hanya hanya menggunakan *silhouette coefficient* sebagai metrik evaluasi, tetapi juga menambahkan metrik *conductance*. Penggunaan *conductance* memberikan sudut pandang evaluasi tambahan yang relevan dalam konteks graf, karena metrik ini mengukur sejauh mana suatu klaster memiliki keterpisahan yang baik dari klaster lainnya dalam struktur graf afinitas. Dengan kombinasi dua metrik ini, hasil pengelompokan diharapkan dapat dievaluasi secara lebih menyeluruh, baik dari segi kedekatan intra-klaster maupun keterpisahan antar-klaster.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menangkap relasi semantik kata dengan mengelompokan ayat-ayat Al-Qur'an menggunakan representasi vektor *word2vec* dan algoritma *Spectral Clustering*. Evaluasi terhadap hasil clustering dilakukan menggunakan metrik *silhouette coefficient* dan *Conductance* guna memastikan kualitas dan validitas hasil pengelompokan. Diharapkan, pendekatan ini dapat memberikan kontribusi dalam pengembangan pemahaman semantik berbasis teknologi terhadap Al-Qur'an, serta

memperkaya kajian tafsir modern melalui perspektif analisis data dan kecerdasan buatan.

1.2. Rumusan Masalah

Berdasarkan latar belakang masalah yang telah disampaikan sebelumnya, rumusan masalah yang akan menjadi konsentrasi pada penelitian skripsi ini adalah:

1. Perbedaan pola kemunculan kata dalam dataset yang berasal dari sumber yang berbeda seperti terjemahan literal dan tafsir Al-Qur'an memengaruhi keakuratan model *Word2Vec* dalam merepresentasikan hubungan semantik antar kata.
2. Bagaimana transformasi spektral dapat memperbaiki hasil klasterisasi *K-Means* pada data dengan struktur non-linier seperti vektor kata *Word2Vec*?
3. Bagaimana persebaran semantik kata *ilāh* dan *rabb* dalam ayat-ayat Al-Qur'an berdasarkan model Word2Vec dan algoritma clustering

1.3. Batasan Masalah

Batasan masalah pada skripsi ini di antaranya sebagai berikut:

1. Data yang digunakan adalah beberapa sumber terjemahan Al-Qur'an bahasa Indonesia yaitu versi terjemahan literal Kemenag dan versi tafsir Jalalayn.
2. Analisis semantik dilakukan pada sejumlah pasangan kata terpilih yang dianggap memiliki hubungan semantik, berdasarkan kemiripan makna secara linguistik.
3. Representasi kata dibangun menggunakan model *Word2Vec* menggunakan pendekatan *Skip-gram* dengan dimensi embedding yang digunakan adalah 300 dan ukuran *window* sebesar 5.
4. Teknik *clustering* yang digunakan berbasis graf yaitu spectral clustering dan *K-Means* sebagai algoritma partisi.
5. Pembentukan matriks afinitas yang digunakan pada *spectral clustering* yaitu *K-Nearest neighbor* dengan *k* tetangga terdekatnya yaitu 5.
6. Validasi *clustering* berupa pengukuran kualitas internal *cluster* dengan menggunakan metrik *silhouette coefficient* dan *conductance score*.

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

1.4. Tujuan dan Manfaat Penelitian

Berdasarkan latar belakang dan rumusan masalah yang telah diuraikan, terdapat beberapa tujuan yang diharapkan dapat tercapai pada penelitian skripsi ini, antara lain:

1. Menganalisis pengaruh perbedaan pola kemunculan kata dalam dataset yang berasal dari sumber yang berbeda, seperti terjemahan literal dan tafsir Al-Qur'an, terhadap keakuratan model *Word2Vec* dalam merepresentasikan hubungan semantik antar kata.
2. Mengevaluasi efektivitas transformasi spektral dalam memperbaiki hasil klasterisasi k-means menggunakan algoritma *spectral clustering*, khususnya pada data vektor kata *Word2Vec* yang memiliki struktur non-linier, dengan membandingkan kualitas hasil klasterisasi sebelum dan sesudah dilakukan transformasi.
3. Menganalisis persebaran semantik kata *ilāh* dan *rabb* dalam ayat-ayat Al-Qur'an dengan menggunakan model Word2Vec dan algoritma clustering.

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

1. Memberikan pemahaman tentang pengaruh variasi sumber terjemahan dan tafsir Al-Qur'an terhadap representasi semantik kata dalam model *Word2Vec*.
2. Membantu dalam mempelajari isi ayat-ayat Al-Qur'an dengan mengelompokkan ayat-ayat berdasarkan keterkaitan antar ayatnya, sehingga dapat mengungkap relasi semantik dan pengetahuan tersembunyi dalam teks.

1.5. Metode Penelitian

Metode penelitian pada tugas akhir ini yaitu sebagai berikut:

1. Studi Literatur

Tahap studi literatur dilakukan pengumpulan fakta-fakta dan referensi literatur yang berhubungan dengan model word2vec dan teknik pengelompokan data teks

2. Analisis

   Pada tahap ini dilakukan pengkajian dan analisis terhadap hasil model word2vec dari beberapa sumber terjemahan Al-Qur'an bahasa Indonesia dalam menghasilkan representasi vektor, kemudian melakukan clustering terhadap ayat-ayat al-quran.

3. Simulasi

   Pada tahap ini dilakukan simulasi percobaan pada terjemah Al-Quran literal bahasa Indonesia dari Kemenag dan versi tafsir Jalalayn Al-Quran bahasa Indonesia omenggunakan model word2vec dan algoritma *spectral clustering* menggunakan Bahasa pemrograman *python* yang dijalankan di *visual studio code*.

1.6. Sistematika Penulisan

Sistematika penulisan pada skripsi ini terdiri dari lima bab dan di dalam setiap bab terdiri dari 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*), pra-pemrosesan (*pre-processing*), *natural language processing*, *word embedding*, metode *clustering*, validasi *clustering*, dan *Python*

BAB III : CLUSTERING AYAT AL-QURAN MENGGUNAKAN ALGORITMA SPECTRAL CLUSTERING

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 *spectral clustering*.

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

Pada bab ini berisi pembahasan mengenai proses, hasil, dan analisis studi kasus yang dilakukan dalam penelitian ini. Dimulai dari penjelasan mengenai langkah-langkah yang dilakukan dalam pemilihan dan pengambilan dataset yang digunakan, kemudian *pre-processing* dataset, *training dataset* menggunakan *word2vec*, pengempokkan kata menggunakan *spectral clusering*, 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.