

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Al-Qur'an merupakan kitab suci umat Islam yang diturunkan oleh Allah SWT kepada Nabi Muhammad SAW sebagai mukjizat sekaligus pedoman hidup bagi umat manusia. Kandungannya mencakup berbagai aspek kehidupan, seperti keimanan, hukum, sejarah, serta kehidupan manusia di dunia dan akhirat. Salah satu tema penting dalam ajaran Islam adalah eskatologi, yaitu pembahasan mengenai kematian, hari kebangkitan, pembalasan amal perbuatan, dan kehidupan setelah kematian. Tema eskatologi memiliki kedudukan penting karena berkaitan dengan keyakinan, orientasi moral, dan pertanggungjawaban manusia dalam ajaran Islam [1]. Kepastian datangnya hari akhir dan kebangkitan ditegaskan dalam Al-Qur'an Surah Al-Hajj ayat 7 sebagai berikut:

وَأَنَّ السَّاعَةَ آتِيَةٌ لَا رَيْبَ فِيهَا وَأَنَّ اللَّهَ يَبْعَثُ مَنْ فِي الْقُبُورِ

“Dan sungguh, hari Kiamat itu pasti datang, tidak ada keraguan padanya; dan sungguh, Allah akan membangkitkan apa yang ada di dalam kubur.”

Ayat tersebut memuat dua konsep eskatologi, yaitu kepastian datangnya hari Kiamat dan kebangkitan manusia dari kubur. Kedua konsep ini berkaitan dengan kategori tematik kiamat dan akhirat serta kebangkitan yang digunakan dalam penelitian ini. Contoh tersebut juga menunjukkan bahwa satu ayat dapat memuat lebih dari satu konsep eskatologi sekaligus, sehingga pelabelan tema bersifat *multi-label*.

Pesan-pesan eskatologi dalam Al-Qur'an disampaikan melalui pilihan kata, susunan kalimat, dan konteks yang beragam. Untuk membantu pemahaman pembaca berbahasa Indonesia, tersedia berbagai sumber terjemahan dan tafsir, termasuk terjemahan Al-Qur'an Kementerian Agama Republik Indonesia (Kemenag RI) dan Tafsir Jalalayn. Perbedaan pilihan kata serta gaya penyampaian antarsumber dapat menghasilkan variasi bentuk bahasa dan representasi makna. Kondisi tersebut menyebabkan relasi semantik antarteks tidak selalu mudah dipetakan hanya melalui pembacaan manual, sehingga diperlukan pendekatan yang mampu menganalisis kemiripan makna secara sistematis dan terukur.

Perkembangan *Natural Language Processing* (NLP) memungkinkan teks bahasa alami diubah menjadi representasi numerik untuk dianalisis menggunakan

metode komputasional [2]. Dalam analisis teks Al-Qur'an, perbedaan unit analisis menjadi hal yang penting. Ayat berfungsi sebagai sumber data dan konteks asal teks, sedangkan kalimat digunakan sebagai unit analisis karena satu ayat dapat terdiri atas lebih dari satu kalimat. Pemecahan ayat menjadi unit-unit kalimat memungkinkan setiap bagian teks direpresentasikan dan dianalisis secara lebih terfokus. Dengan demikian, kata berperan sebagai fitur pembentuk representasi, kalimat menjadi objek analisis, dan ayat tetap dipertahankan sebagai informasi sumber serta konteks tekstual.

Salah satu pendekatan untuk membentuk representasi semantik kata adalah model *Word2Vec*. Model ini memetakan kata ke dalam ruang vektor berdasarkan pola kemunculan dan konteksnya di dalam korpus. Arsitektur *Skip-gram* mempelajari representasi kata dengan memprediksi kata-kata konteks berdasarkan kata pusat, sehingga kata-kata yang memiliki konteks penggunaan serupa cenderung memiliki posisi yang berdekatan dalam ruang vektor [3]. Meskipun demikian, representasi pada tingkat kata belum secara langsung menggambarkan makna sebuah kalimat secara utuh. Oleh sebab itu, sejumlah vektor kata perlu digabungkan menjadi satu representasi kalimat menggunakan metode *Smooth Inverse Frequency* (SIF). Metode SIF membentuk rata-rata berbobot dari vektor kata. Representasi kalimat hasil SIF kemudian diproses menggunakan *Common Component Removal* (CCR) untuk mengurangi komponen umum yang terdapat pada sebagian besar representasi kalimat [4].

Representasi kalimat yang telah dibentuk kemudian dapat dibandingkan berdasarkan arah vektornya menggunakan *Cosine Similarity*. Nilai tersebut selanjutnya dapat dikonversi menjadi *Cosine Distance* untuk digunakan dalam proses pengelompokan. Selain itu, PCA dapat digunakan sebagai metode reduksi dimensi untuk membantu memvisualisasikan sebaran vektor kalimat berdimensi tinggi. Dalam penelitian ini, PCA hanya digunakan untuk visualisasi dan tidak menggantikan vektor berdimensi asli yang digunakan dalam analisis kemiripan maupun proses klusterisasi [5].

Pola keterkaitan semantik antarkalimat yang kompleks masih sulit diidentifikasi secara manual meskipun setiap kalimat telah direpresentasikan ke dalam bentuk vektor. Teknik *clustering* dapat digunakan untuk mengelompokkan kalimat berdasarkan kedekatan representasi semantiknya. Salah satu metode yang sesuai adalah *Hierarchical Agglomerative Clustering* (HAC), yaitu metode pengelompokan hierarkis yang bekerja secara *bottom-up* dengan menggabungkan kluster secara bertahap. Berbeda dari metode *partitioning* seperti *K-Means* yang

memerlukan jumlah klaster sejak awal, HAC dapat membentuk struktur hierarki dan menghasilkan dendrogram. Pada penelitian ini, kriteria *Average Linkage* digunakan untuk menghitung jarak rata-rata seluruh pasangan objek antarklaster, sedangkan metrik *Cosine Distance* digunakan untuk mengukur perbedaan arah vektor.

Berdasarkan permasalahan tersebut, penelitian berjudul “Klasterisasi Tema Eskatologi pada Teks Al-Qur’an Berbahasa Indonesia Menggunakan *Hierarchical Agglomerative Clustering* (HAC) Berbasis *Word2Vec*” mengintegrasikan model *Word2Vec*, metode *Smooth Inverse Frequency* (SIF), *Common Component Removal* (CCR), normalisasi norma- L_2 , pengukuran *Cosine Distance*, serta HAC untuk memetakan struktur semantik kalimat-kalimat bertema eskatologi. Korpus Kemenag dan Tafsir Jalalayn dimanfaatkan sebagai sumber pelatihan model representasi kata, sedangkan kalimat Kemenag bertema eskatologi digunakan sebagai objek klasterisasi. Hasil pengelompokan kemudian diinterpretasikan berdasarkan kata kunci khas yang diperoleh melalui pembobotan TF-IDF. Kata dominan berdasarkan frekuensi digunakan sebagai informasi pendukung, sedangkan isi kalimat diperiksa untuk memastikan kesesuaian makna dengan kata kunci dan tema yang terbentuk. Dengan demikian, tema klaster dipahami sebagai hasil interpretasi semantik berbasis data dan bukan sebagai label otomatis yang dihasilkan secara langsung oleh algoritma HAC.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan, rumusan masalah dalam penelitian ini dijabarkan sebagai berikut:

1. Bagaimana membentuk representasi vektor kalimat bertema eskatologi Islam menggunakan vektor kata *Word2Vec*, metode *Smooth Inverse Frequency* (SIF), *Common Component Removal* (CCR), dan normalisasi norma- L_2 untuk menangkap relasi semantik antarkalimat?
2. Bagaimana kalimat-kalimat Kemenag bertema eskatologi dapat dikelompokkan berdasarkan kedekatan representasi semantiknya menggunakan algoritma *Hierarchical Agglomerative Clustering* (HAC) dengan metode *Average Linkage* dan metrik *Cosine Distance*?
3. Bagaimana kualitas struktur klaster yang terbentuk serta bagaimana interpretasi tema setiap klaster berdasarkan kata kunci khas yang diperoleh melalui skor TF-IDF, kata dominan berdasarkan frekuensi, dan pemeriksaan isi kalimat bertema eskatologi?

1.3 Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut:

1. Data yang digunakan berupa teks Al-Qur'an berbahasa Arab, terjemahan Kementerian Agama Republik Indonesia (Kemenag), dan Tafsir Jalalayn berbahasa Indonesia. Data Arab digunakan untuk menyaring kandidat tema eskatologi, sedangkan teks Kemenag dan Tafsir Jalalayn digunakan sebagai korpus bahasa Indonesia.
2. Data penelitian difokuskan pada teks yang teridentifikasi mengandung konsep eskatologi Islam berdasarkan pencocokan kata kunci pada teks Arab yang telah melalui tahap pembersihan.
3. Ayat digunakan sebagai sumber dan konteks teks, sedangkan unit analisis utama adalah kalimat hasil segmentasi ayat. Kata digunakan sebagai fitur pembentuk representasi dan bukan sebagai objek klasterisasi.
4. Representasi kata dibentuk menggunakan model *Word2Vec* arsitektur *Skip-gram* dengan dimensi *embedding* 300 dan ukuran jendela konteks 5.
5. Representasi kalimat dibentuk menggunakan metode *Smooth Inverse Frequency* (SIF), kemudian diproses menggunakan *Common Component Removal* (CCR) dan normalisasi norma- L_2 .
6. Pengelompokan kalimat dilakukan menggunakan *Hierarchical Agglomerative Clustering* (HAC) dengan metode *Average Linkage* dan metrik *Cosine Distance*. Jumlah klaster dibatasi pada rentang $10 \leq K \leq 40$ dan ukuran klaster minimum diuji pada tiga nilai kandidat (3, 2, dan 1). Partisi yang dipilih adalah kombinasi ambang dan ukuran minimum yang menghasilkan nilai *Silhouette Score* tertinggi secara keseluruhan.
7. Kualitas klaster dievaluasi menggunakan *Silhouette Score*, *Cophenetic Correlation Coefficient* (CCC), kohesi intraklaster, dan *inter-cluster similarity*. Kohesi intraklaster dihitung berdasarkan rata-rata *Cosine Similarity* antarkalimat dalam klaster yang sama, sedangkan *inter-cluster similarity* dihitung berdasarkan rata-rata *Cosine Similarity* antarcentroid klaster.
7. Tema klaster ditentukan menggunakan kata kunci hasil pembobotan *TF-IDF*. Kata dominan berdasarkan frekuensi digunakan sebagai informasi pendukung, sedangkan isi kalimat digunakan untuk membantu interpretasi tema secara semantik.

8. Perbandingan terhadap skenario cakupan data lain (D1 sampai D6) dilakukan sebagai analisis sensitivitas pendukung. Analisis tematik dan interpretasi utama hanya didasarkan pada skenario D0 yang memuat seluruh kategori tematik.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah disusun, tujuan penelitian ini adalah sebagai berikut:

1. Membentuk dan menganalisis representasi vektor kalimat bertema eskatologi Islam pada teks terjemahan Al-Qur'an bahasa Indonesia menggunakan model *Word2Vec*, metode *Smooth Inverse Frequency* (SIF), *Common Component Removal* (CCR), dan normalisasi norma- L_2 , serta mengkaji kemampuannya dalam menangkap relasi semantik antarkalimat.
2. Menganalisis proses pengelompokan kalimat-kalimat Kemenag bertema eskatologi berdasarkan kedekatan semantik menggunakan algoritma *Hierarchical Agglomerative Clustering* (HAC) dengan metode *Average Linkage* dan metrik *Cosine Distance*.
3. Mengevaluasi kualitas kluster yang terbentuk menggunakan *Silhouette Score*, *Cophenetic Correlation Coefficient* (CCC), rata-rata kohesi intrakluster, dan rata-rata *Cosine Similarity* antarcentroid kluster.
4. Menginterpretasikan tema-tema semantik pada kluster kalimat bertema eskatologi berdasarkan kata kunci khas hasil pembobotan TF-IDF, kata dominan berdasarkan frekuensi sebagai informasi pendukung, serta pemeriksaan isi kalimat secara manual.

1.5 Metodologi Penelitian

Metodologi penelitian yang digunakan pada penelitian ini terdiri atas beberapa tahapan sebagai berikut:

1. Studi Literatur

Tahap ini dilakukan untuk mengumpulkan teori dasar, fakta-fakta, serta referensi ilmiah yang berkaitan dengan model *Word2Vec*, *Smooth Inverse Frequency* (SIF), *Common Component Removal* (CCR), normalisasi norma- L_2 , pembobotan TF-IDF, teknik klasterisasi teks, khususnya

Hierarchical Agglomerative Clustering (HAC), serta metrik evaluasi kualitas kluster. Sumber literatur berupa buku, artikel jurnal, dan prosiding yang relevan dengan pemrosesan bahasa alami dan klusterisasi teks. Hasil studi literatur menjadi dasar penyusunan landasan teori serta perancangan tahapan penelitian.

2. Analisis

Pada tahap ini dilakukan perancangan alur sistem serta pengkajian terhadap model *Word2Vec* dalam menghasilkan representasi vektor dari teks terjemahan Al-Qur'an dan Tafsir Jalalayn berbahasa Indonesia. Teks Arab digunakan sebagai dasar penyaringan ayat bertema eskatologi menggunakan pendekatan berbasis aturan (*rule-based*). Teks Kemenag dan Tafsir Jalalayn yang memiliki nomor surat dan ayat yang sama kemudian digunakan sebagai korpus bahasa Indonesia.

Tahap analisis dilanjutkan dengan perancangan *sentence embedding* menggunakan metode *Smooth Inverse Frequency* (SIF), *Common Component Removal* (CCR), dan normalisasi norma- L_2 . Selain itu, dirancang skema visualisasi menggunakan *Principal Component Analysis* (PCA), klusterisasi kalimat menggunakan *Hierarchical Agglomerative Clustering* (HAC), evaluasi kualitas kluster, serta pelabelan tema menggunakan TF-IDF. Rancangan ini disusun agar setiap tahap dapat ditelusuri mulai dari data mentah hingga interpretasi tema.

3. Simulasi

Pada tahap ini dilakukan eksperimen dan simulasi pemodelan menggunakan korpus gabungan teks Kemenag dan Tafsir Jalalayn untuk melatih model *Word2Vec*. Kalimat-kalimat Kemenag yang termasuk dalam Dataset Ayat Eskatologi digunakan sebagai objek analisis dan klusterisasi. Selain skenario utama yang memuat seluruh tiga belas kategori tematik (D0), dilakukan pula simulasi pada enam skenario cakupan data lain (D1 sampai D6) sebagai analisis sensitivitas.

Proses komputasi diimplementasikan menggunakan bahasa pemrograman Python untuk melakukan prapemrosesan teks, melatih model *Word2Vec*, membentuk representasi kalimat menggunakan SIF, CCR, dan normalisasi norma- L_2 , melakukan visualisasi PCA, mengeksekusi algoritma *Hierarchical Agglomerative Clustering* (HAC), menghitung metrik evaluasi, serta menghasilkan kata kunci tema berbasis TF-IDF. Model *Word2Vec* dilatih

dengan nilai *seed* yang tetap agar hasil pelatihan dapat direproduksi. Hasil eksperimen disajikan dalam bentuk tabel, grafik, dan dendrogram untuk memudahkan analisis.

1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini disusun ke dalam lima bab yang saling berkaitan, yaitu sebagai berikut:

BAB I PENDAHULUAN

Bab ini memuat latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, metodologi penelitian, dan sistematika penulisan. Latar belakang menjelaskan alasan pemilihan tema eskatologi dan pendekatan yang digunakan. Rumusan masalah, batasan masalah, dan tujuan penelitian menentukan cakupan serta arah penelitian ini.

BAB II LANDASAN TEORI

Bab ini berisi penjelasan mengenai teori-teori yang berkaitan dengan penelitian, meliputi *data mining*, *Natural Language Processing*, prapemrosesan teks, *word embedding*, model *Word2Vec*, pembentukan *sentence embedding* menggunakan metode SIF dan CCR, normalisasi norma- L_2 , PCA, *Cosine Similarity*, *Cosine Distance*, pembobotan TF-IDF, metode klasterisasi, khususnya HAC, serta validasi kluster. Teori-teori tersebut menjadi dasar bagi tahapan penelitian pada bab berikutnya. Rumus dan istilah yang dijelaskan pada bab ini digunakan kembali pada pembahasan hasil.

BAB III KLASTERISASI TEMA ESKATOLOGI PADA TEKS AL-QUR'AN BERBAHASA INDONESIA MENGGUNAKAN HIERARCHICAL AGGLOMERATIVE CLUSTERING (HAC) BERBASIS WORD2VEC

Bab ini berisi penjelasan mengenai tahapan penelitian secara rinci, meliputi pengumpulan data, penyaringan ayat bertema eskatologi berbasis aturan, prapemrosesan teks, pelatihan model *Word2Vec*, pembentukan *sentence embedding* menggunakan SIF dan CCR, normalisasi norma- L_2 , visualisasi menggunakan PCA, pengelompokan kalimat menggunakan *Hierarchical Agglomerative Clustering* (HAC), evaluasi kualitas kluster, serta pelabelan tema menggunakan TF-IDF. Penjelasan disusun berurutan sesuai alur pengerjaan. Rancangan skenario perbandingan cakupan data juga dijelaskan pada bab ini.

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

Bab ini berisi hasil implementasi dan eksperimen terhadap data teks Al-Qur'an berbahasa Indonesia yang telah disaring berdasarkan tema eskatologi. Pembahasan meliputi hasil pembentukan dataset, pelatihan model *Word2Vec* menggunakan korpus gabungan Kemenag dan Tafsir Jalalayn, pembentukan representasi kalimat, hasil visualisasi PCA, pengukuran *Cosine Similarity* pada tingkat kata dan kalimat, serta pengelompokan kalimat Kemenag menggunakan HAC. Selain itu, bab ini menyajikan hasil evaluasi kualitas klaster, kata kunci hasil pembobotan TF-IDF, kata dominan berdasarkan frekuensi, interpretasi tema semantik pada klaster yang terbentuk, dan perbandingan hasil pada skenario cakupan data D0 sampai D6.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan penelitian berdasarkan rumusan masalah, tujuan penelitian, dan hasil eksperimen. Kesimpulan disusun untuk menjawab setiap rumusan masalah yang diajukan pada Bab I. Bab ini juga memuat saran untuk pengembangan penelitian selanjutnya.

