BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Dalam era globalisasi manusia disuguhkan dengan berbagai bentuk kemajuan, mulai dalam sektor industri, telekomunikasi hingga dunia pendidikan. Seiring perkembangan tersebut, kebutuhan dalam mengelola data menjadi suatu tantangan tersendiri yang harus dipelajari, guna memecahkan masalah di kemudian hari. Oleh karena itu, muncullah istilah *data mining* (penambangan data) yang menjadi pencetus awal mula berbagai jenis metode pengolahan data.

Pada mulanya, *data mining* dikembangkan dalam dunia bisnis yang bertujuan untuk mengatasi permintaan informasi yang semakin kompleks yang tak mudah dilihat melalui analisis manual [1]. *Data mining* bekerja dengan cara mengenali pola tersembunyi serta mencari hubungan data dalam sebuah *database*, proses ini sering kali dikenal dengan istilah ekplorasi data maupun pembelajaran deduktif berbasis data (*deductive learning*) [2].

Seiring perkembangan zaman, *data mining* tidak hanya digunakan dalam dunia bisnis datanya berbasis angka atau numerik. Akan tetapi, *data mining* juga dikembangkan untuk mampu memahami data yang berbasis tekstual atau lebih dikenal dalam istilah *text mining* [3].

Dalam komputasi, teks merupakan suatu data yag berbentuk abstrak dan tidak terstruktur, berbeda dengan data berbasis numerik, data teks tak bisa diolah secara langsung karena setiap kata dianggap terpisah sebelum dilakukan pemrosesan, sehingga tiap kata belum memiliki makna pasti secara kontekstual [4].

Data berbasis teks memerlukan sebuah metode atau cara khusus untuk mempelajari strukturnya, data teks akan melewati serangkaian proses sebelum memasuki tujuan akhir pengolahannya. Salah satu metode yang tekenal adalah *natural language processing* (*NLP*).

Natural language processing (*NLP*) merupakan analisis data linguistik yang membangun representasi teks dari data linguistik tidak terstruktur. Dengan memanfaatkan kekayaan linguistik, *NLP* bekerja dengan menangkap hubungan gramatikal maupun mempelajari hubungan semantik, guna menangkap makna kata sesungguhnya [5].

Berbicara tentang kekayaan gramatikal, Al-Qur'an sebagai kitab suci umat Islam merupakan sebuah data tertulis yang sangat kaya dengan kosakata, selain itu juga, kata dalam Al-Qur'an sering kali tidak pada satu makna yang tetap, hubungan similar yang terdapat pada Al-Qur'an bisa saja memiliki banyak makna berbeda, sehingga perlu ada langkah khusus untuk mempelajari tentang makna tersebut.

Word2vec sebagai salah satu model terkemuka dalam pemrosesan *NLP* merupakan jawaban yang tepat dalam mengatasi hubungan semantik antar kata . *Word2vec* mampu merepresentasikan dengan baik makna semantik dalam kata, dengan metode *CBOW* maupun *SKIP-GRAM* setiap kata mampu mengenali makna tersendiri sesuai hubungan dengan kata disekitarnya [11] [15].

Akan tetapi, seperti yang dijelaskan sebelumnya bahwa didalam Al-Qur'an sering kali terjadi duplikasi makna pada sebuah kata. Misalnya, kata "neraka" bisa dimaknai sebagai sebuah siksaaan dan juga bisa dimaknai sebagai nama-nama, misalnya "neraka jahannam" atau "neraka hawiyyah". Oleh karena itu, dengan adanya kemiripan makna dalam sebuah kata, perlu adanya metode tambahan untuk mengkategorikan kata yang memiliki makna serupa.

Clustering atau metode pengelompokan adalah salah satu cabang keilmuan dari *data mining* yang bekerja dengan cara mengelompokan data ke dalam sebuah *cluster* dengan kategori serupa. Hal tersebut dicetuskan oleh Tyron dan Bailey pada tahun 1970 yang mengatakan "Untuk memahami dunia kita memerlukan konseptualisasi persamaan dan perbedaan antara entitas yang menyusunnya" [20].

Dengan variasi data yang terdapat dalam Al-Qur'an, kata yang terkandung tak selamanya memiliki makna tersurat (langsung), akan tetapi sering juga dijumpai makna tersirat dalam sebuah kalimat. Sehingga baru dapat dimengerti jika dibaca atau digabung dengan beberapa kata di sekitarnya, hal ini tentu saja tergambarkan

adanya variansi kepadatan data (kata) untuk memahaminya. *Clustering* sebagai salah satu metode penambangan data memiliki suatu metode analisis yang mampu mengenali pola kepadatan dalam data yang bernama *OPTICS clustering* [23].

Struktur kepadatan intrinsik pada data Al-Qur'an tidak dapat dicirikan dengan kepadatan global saja, kepadatan lokal yang sangat berbeda (panjang ayat) mungkin sangat dibutuhkan untuk mengungkap *cluster* di berbagai daerah. *OPTICS* akan menganalisa perubahan kepadatan dari sekumpulan *data point* sehingga menghasilkan *cluster* yang sangat kuat secara makna semantik maupun hubungan antar kata satu sama lain.

Dengan melakukan kombinasi metode komputasi di atas, tentunya akan menghasilkan sebuah hasil yang cukup memuaskan. Selain itu, penulis akan melakukan uji validasi hasil *clustering* dengan menghitung nilai *silhouette score* [23], *Calinski-Harabasz Index* [24] dan *Density-Based Clustering Validation* [28]. Kedua cara ini umum dilakukan untuk meninjau seberapa besar tingkat akurasi *cluster* setelah dilakukan pengelompokan. Langkah-langkah ini akan sangat bermanfaat dalam proses pembelajaran, khususnya dalam syiar agama islam, karena dengan cara tersebut seseorang mampu dengan mudah memahami dan menemukan kelompok ayat dalam Al-Qur'an sesuai maknanya.

1.2 Rumusan Masalah

Berdasarkan penjelasan pada bagian latar belakang, maka didapat beberapa rincian masalah yang dapat digaris bawahi, diantaranya :

1. Bagaimana peran *training word2vec* dalam mengatasi pemahaman *similarity* kata pada data Al-Qur'an terjemah bahasa Indonesia?
2. Bagaimana peran *OPTICS clustering* dalam pengelompokan kata pada data Al-Qur'an terjemah Bahasa Indonesia?
3. Bagaimana peran jangkauan kepadatan dalam pembentukan *cluster*, guna mempelajari pola dan makna kata pada data Al-Qur'an.
4. Bagaimana tingkat akurasi pengelompokan berdasarkan hubungan semantik kata dan hubungan *similarity* kata?

1.3 Batasan Masalah

Dalam penelitian ini, ada beberapa batasan masalah yang diterapkan, di antaranya :

1. Data yang digunakan adalah Al-Quran terjemah bahasa Indonesia.
2. Metode *training* yang digunakan dalam penelitian ini adalah *word2vec model*.
3. Arsitektur *word2vec* yang digunakan adalah *skip-gram*
4. Menggunakan *OPTICS clustering method* sebagai pengelompokan berbasis kepadatan dengan memanfaatkan jangkauan metrik *cosine*.
5. Dalam menguji kualitas *cluster* digunakan metode validasi *silhouette coefficient*, *Calinski Harabsz Index*, *cosine similarity* dan *Density-Based Clustering Validation*.

1.4 Tujuan Penelitian

Tujuan penelitian ini dilakukan, antara lain sebagai berikut :

1. Dapat melakukan proses *pre-training* dan menemukan parameter terbaik untuk *training word2vec*.
2. Mengimplementasikan algoritma *OPTICS* serta memvisualisasikan hasil *clustering* secara informatif dan terstruktur.
3. Dapat memahami pengaruh jangkauan kepadatan terhadap hasil *clustering*.
4. Dapat memahami hubungan *semantic similarity* tiap *cluster* dan menganalisa pola relasi *data points* dalam ayat Al-Qur'an.

1.5 Metode Penelitian

Metode penelitian yang digunakan dalam menyelesaikan tugas akhir ini adalah sebagai berikut :

1. Studi literatur

   Pada tahap ini, penulis mengumpulkan informasi dari berbagai sumber referensi sebagai rujukan untuk membangun pengetahuan mengenai penelitian yang akan dilakukan. Mulai dari buku, jurnal, paper atau karya tulis ilmiah lain yang berkaitan dengan pengolahan bahasa (*word2vec*) dan metode pengelompokan (*OPTICS clustering*).

2. Simulasi Percobaan

   Pada metode ini penulis melakukan eksperimen dengan membuat model *training word2vec*. Selanjutnya, model yang dihasilkan dikelompokkan menggunakan algoritma *OPTICS clustering* serta melakukan analisis dan visualisi hasil *cluster* nya didalam ruang vektor.

3. Kesimpulan

   Pada tahap ini, diperoleh kesimpulan dari evaluasi metode yang telah diterapkan, menunjukkan bahwa pendekatan yang digunakan menghasilkan representasi optimal sesuai dengan tujuan penelitian. Analisis terhadap parameter yang berpengaruh mengindikasikan bahwa model yang dikembangkan mampu menangkap hubungan semantik dengan baik, sehingga dapat digunakan sebagai dasar dalam proses clustering dan eksplorasi lebih lanjut.

1.6 Sistematika Penulisan

Pada laporan skripsi ini terdapat lima bab utama dan beberapa subbab yang menjadi sistematika penulisan. Selain itu juga memiliki daftar pustaka sebagai rujukan yang digunakan dalam penelitian . Antara lain sebagai berikut :

BAB I PENDAHULUAN

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

BAB II LANDASAN TEORI

Pada bab ini berisi tentang dasar-dasar teori yang digunakan dalam proses penelitian, seperti penambangan data (*data mining*), penambangan teks (*text mining*), pengolahan bahasa alami (*natural language processing*), pra-pemrosesan (*pre-processing*), *word embedding*, *clustering, clustering* berbasis kepadatan (*density-based clustering*), *Ordering Points to Identify the Clustering Structure* (*OPTICS*), serta evaluasi dan validasi *clustering*.

BAB III IDENTIFIKASI RELASI SEMANTIK KATA PADA DATA AL-QUR'AN TERJEMAH BAHASA INDONESIA MENGGUNAKAN *OPTICS CLUSTERING*

Pada bab ini berisi rangkaian metodologi penelitian yang lebih terperinci selama penelitian berlangsung, seperti pengambilan dataset, pra-pemrosesan data, *embedding* kata menggunakan *word2vec* dan metode pengelompokkan kata menggunakan *OPTICS clustering*.

BAB IV ANALISIS *CLUSTERING OPTICS* PADA DATA AL-QUR'AN TERJEMAH

Pada bab ini berisi pembahasan mengenai analisis studi kasus yang dilakukan pada penelitian kali ini. Studi kasus yang dilakukan berkaitan dengan hasil penelitian yang telah dilakukan, seperti : *embedding* kata yang dilakukan menggunakan *word2vec*, pengelompokkan kata menggunakan *OPTICS clustering*, serta analisis *data points* pada ayat Al-Qur'an terjemah.

BAB V PENUTUP

Bab ini berisi kesimpulan dari serangkaian proses penelitian yang telah dilakukan , selain itu bab ini memberikan saran dan masukan untuk pengembangan metode yang bisa dilakukan dimasa mendatang.

DAFTAR PUSTAKA