BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Dalam era digital saat ini, volume data yang dihasilkan setiap hari telah mencapai jumlah yang sangat besar dan terus meningkat. Data-data ini tidak hanya berupa angka dan teks sederhana, tetapi juga mencakup berbagai bentuk dan format yang kompleks. Oleh karena itu, diperlukan metode dan teknik yang canggih untuk mengelola, menganalisis, dan mengekstraksi informasi yang bermanfaat dari data-data tersebut. Salah satu pendekatan yang semakin penting dalam menangani masalah ini adalah *data mining*. *Data mining* adalah proses penemuan pola, hubungan, dan wawasan dari sejumlah besar data melalui berbagai teknik analisis [1] . Dalam konteks data teks, *Natural Language Processing* (NLP) menjadi salah satu alat utama untuk menganalisis dan memahami data bahasa alami. NLP mencakup berbagai teknik dan metode yang memungkinkan komputer untuk memproses dan menganalisis data teks secara efektif [2].

Salah satu perkembangan signifikan dalam bidang NLP adalah pengenalan *word embeddings*. *Word embeddings* adalah representasi kata-kata dalam bentuk *vector* berdimensi tinggi yang menangkap hubungan semantik antar kata [3]. Teknik ini memungkinkan komputer untuk memahami konteks dan makna kata-kata dalam suatu teks, yang sangat penting dalam berbagai aplikasi NLP seperti analisis sentimen, pengenalan entitas, dan terjemahan mesin. Contohnya, penggunaan *word2vec* pada terjemahan Al-Qur'an dalam bahasa Inggris dapat memberikan wawasan yang lebih dalam mengenai hubungan semantik antar kata-kata yang terdapat dalam teks suci tersebut. Dengan *word2vec*, kita dapat menangkap makna tersembunyi dan korelasi yang ada dalam teks, yang seringkali sulit ditangkap oleh metode tradisional. Misalnya, kata "*mercy*" dan "*compassion*" mungkin memiliki *vector* yang sangat dekat satu sama lain, menunjukkan kesamaan semantik yang kuat dalam konteks Al-Qur'an.

Dalam *data mining*, *clustering* adalah teknik penting yang digunakan untuk mengelompokkan data ke dalam *cluster* yang memiliki karakteristik serupa [4]. Salah satu algoritma *clustering* yang efektif untuk menangani dataset besar adalah

CLARANS (*Clustering Large Applications based on Randomized Search*). Algoritma ini menggunakan pendekatan berbasis *medoid* dan pencarian acak untuk mengurangi kompleksitas komputasi, sehingga mampu menghasilkan *cluster* yang lebih baik dalam waktu yang lebih singkat. Evaluasi hasil *clustering* ini sangat penting untuk memastikan kualitasnya. *Best cost* menjadi acuan utama dalam menilai hasil *clustering* CLARANS, *best cost* adalah nilai biaya (*cost*) terendah yang dicapai selama proses pencarian pusat *cluster* (*medoid*) terbaik. Biaya ini dihitung berdasarkan jarak antara data dan pusat *cluster* yang diusulkan (*medoid*). *Best medoid* adalah titik data dalam *cluster* yang dipilih sebagai pusat *cluster* yang optimal atau terbaik. *Medoid* ini adalah representasi dari *cluster*, yang dipilih sehingga total jarak antara semua titik dalam *cluster* ke *medoid* adalah minimum. Metrik yang umum digunakan adalah *silhouette coefficient* (SC). *Silhouette coefficient* mengukur seberapa mirip objek dengan *cluster* mereka sendiri dibandingkan dengan *cluster* lainnya[5].

Semantic similarity adalah konsep yang digunakan dalam pemrosesan bahasa alami (NLP) untuk mengukur seberapa mirip potongan kata dalam hal makna. *Wordnet* adalah salah satu alat utama yang digunakan dalam pendekatan awal untuk menghitung *semantic similarity*, penggunaan *ontologi* dalam *wordnet* juga menjadi penting dalam NLP dan *data mining*. *Wordnet* adalah database leksikal bahasa Inggris yang mengelompokkan kata-kata ke dalam set sinonim yang disebut *synset*, menyediakan definisi singkat, dan merekam berbagai hubungan antara set-set sinonim ini [6]. Penggunaan *wordnet* dalam NLP memungkinkan peningkatan pemahaman konteks dan makna kata-kata, yang pada gilirannya meningkatkan akurasi dan efektivitas analisis teks. Dalam konteks *semantic similarity*, *wordnet* memungkinkan kita untuk mengukur kesamaan semantik antara dua kata berdasarkan hubungan sinonim dan hiponim yang ada dalam database. Hal ini sangat berguna dalam aplikasi seperti pengenalan entitas, pencarian informasi, dan analisis teks, di mana pemahaman yang mendalam tentang makna kata sangat diperlukan.

Penggabungan metode data, NLP, *word embeddings*, *clustering* CLARANS, dan evaluasi menggunakan *silhouette coefficient*, bersama dengan pemanfaatan *wordnet*, memberikan pendekatan yang komprehensif dan efektif dalam analisis

data teks yang kompleks seperti data terjemahan Al-Qur'an bahasa inggris. Pendekatan ini tidak hanya meningkatkan kemampuan untuk mengekstraksi informasi yang bermanfaat dari data yang besar, tetapi juga membuka peluang baru dalam berbagai aplikasi praktis, mulai dari analisis sentimen hingga pengelompokan dokumen.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, didapat rumusan masalah sebagai berikut

1. Bagaimana *pre-processing* data terjemahan Al-Qur'an bahasa inggris dan *training word2vec*?
2. Bagaimana hasil *best cost* dan *best medoid* algoritma *clustering* CLARANS (*Clustering Algorithm Bassed on Randomized Search*) dan visualisasinya dengan data terjemahan Al-Qur'an bahasa inggris?
3. Bagaimana *semantic similarity* kata yang dihasilkan dari *clustering* CLARANS jika dilihat dari data terjemahan Al-Qur'an bahasa inggris?

1.3 Batasan Masalah

Permasalahan yang dibahas dalam tugas akhir ini memiliki batasan sebagai berikut

1. Metode *clustering* yang digunakan adalah metode CLARANS.
2. Metode yang digunakan untuk mengevaluasi hasil *cluster* adalah *best cost* dengan nilai *Silhouette Coefficient* (SC).
3. Focus penelitian ini akan dibatasi pada penggunaan *Wordnet* sebagai *ontology* untuk *semantic similarity* kata dalam hasil *clustering* CLARANS.
4. Bahasa pemrograman menggunakan Python 3.7.
5. Dataset yang digunakan adalah data terjemahan Al-Qur'an bahasa inggris versi Maududi.
6. Algoritma *training word2vec* digunakan dalam *pre-processing* data.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut

1. Dapat melakukan *pre-processing* dengan data terjemahan Al-Qur'an bahasa inggris dan *training word2vec*.

2. Dapat melakukan *clustering* CLARANS dan mendapatkan hasil *best cost* dan *best medoid* serta menampilkan hasil visualisasinya.
3. Dapat menampilkan *semantic similarity* dari hasil *clustering* CLARANS dan menganalisis similaritasnya dari terjemahan Al-Qur'an bahasa inggris.

1.5 Metode Penelitian

Metode yang digunakan oleh penulis dalam menyelesaikan skripsi ini adalah sebagai berikut

1. Studi Literatur

   Pada tahap ini, penulis memperoleh pemahaman mengenai konsep metode *word embeddings* dan CLARANS melalui studi literatur yang mencakup buku, jurnal, skripsi, dan tesis.
2. Penelitian

   Pada tahap penelitian penulis melakukan *pre-processing* data terjemahan Al-Qur'an bahasa inggris dengan menggunakan algoritma *word2vec* dan data disimpan dalam dokumen untuk digunakan dalam *clustering* menggunakan algoritma CLARANS, kemudian akan mendapatkan hasil *best cost* dan *best medoid* serta hasil pengelompokan kata sesuai *cluster* dan *semantic similarity* dinilai menggunakan *wordnet*.

1.6 Sistematika Penulisan

Pada skripsi ini terdapat lima bab sistematika penulisan yang diantaranya.

BAB I PENDAHULUAN

Bab pendahuluan ini berisi latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, metode penelitian, dan sistematika penulisan dari masalah yang dikaji.

BAB II LANDASAN TEORI

Bab landasan teori ini menjelaskan tentang teori-teori yang melandasi pembahasan inti yang saling berkaitan dan sebagai penunjang dalam penulisan skripsi, seperti *data mining*, *Natural Language Processing* (NLP), *clustering*, *Medoid*, *Partitioning Around Medoids* (PAM), *Clustering Large Applications* (CLARA), *Clustering Large*

Applications Based on Randomized Search (CLARANS) serta metode evaluasi meliputi *best cost, silhouette coefficient*, dan *wordnet* .

BAB III *SEMANTIC SIMILARITY WORD EMBEDDINGS* DENGAN *CLUSTERING* CLARANS MENGGUNAKAN DATA TERJEMAHAN AL-QUR'AN BAHASA INGGRIS

Pada bab ini berisi pembahasan tentang penelitian yang di lakukan dari pengambilan dataset, lalu tahap *text pre-processing* dengan metode *word embeddings* lalu dilakukan *clustering* menggunakan metode *clustering* CLARANS, beberapa hasil *cluster* akan didapatkan pengelompokan kata sesuai jumlah *cluster* yang diatur dalam parameter CLARANS, dan terakhir melakukan evaluasi *best cost* dengan *Silhoutte Coefficient* (SC), kemudian *wordnet* digunakan untuk *semantic similarity* pada pengelompokan kata dalam sebuah *cluster*.

BAB IV ANALISIS HASIL *CLUSTERING* CLARANS

Bab ini berisi pemaparan mengenai analisis hasil *clustering* yang sudah dilakukan pada bab sebelumnya. Analisis hasil *clustering* yang dilakukan berupa perbandingan nilai *best cost, best medoid* dari setiap parameter *cluster*, kemudian hasil pengelompokan kata dianalisis pada dataset yang digunakan.

BAB V PENUTUP

Bab penutup berisi hasil simpulan dari rumusan masalah yang telah dijelaskan dan berisi saran yang diperuntukan untuk penelitian berikutnya sebagai pengembangan dari *word embeddings* dan CLARANS.

DAFTAR PUSTAKA