BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Pada era perkembangan teknologi yang semakin maju, manusia disuguhkan dengan berbagai bentuk kemajuan. Perkembangan tersebut dimulai dari bidang industri, telekomunikasi, hingga dunia pendidikan. Seiring dengan perkembangan teknologi, kebutuhan dalam bidang teknologi informasi meningkat, karena kemudahan dalam pengambilan informasi. Karena akses yang sangat mudah, jumlah data teks juga meningkat. Perkembangan jumlah data teks yang cepat di internet mendorong kebutuhan metode representasi kata tidak hanya efisien secara komputasi, tetapi juga kaya akan informasi semantik[1].

Data dalam sebuah teks memerlukan metode khusus untuk mempelajari strukturnya. Sebelum mencapai tujuan akhir pengolahan, data harus melalui beberapa tahapan pemrosesan. Salah satu metode yang terkenal adalah *Natural Language Processing (NLP)*

Natural Language Processing (NLP) adalah bidang yang mempelajari teknik untuk mengolah dan menganalisis teks tidak terstruktur. *NLP* membangun representasi numerik (vektor) yang membuat komputer dapat memahami pola dan makna dalam teks. *NLP* memanfaatkan kekayaan linguistik untuk menangkap hubungan gramatikal dan semantik sehingga dapat merepresentasikan makna kata secara lebih akurat [2].

Berbicara tentang kekayaan gramatikal, Al-Qur’an sebagai kitab suci umat Islam merupakan sebuah data tertulis yang sangat kaya akan kosakata, Al-Qur’an mengandung nilai-nilai spiritual, moral, dan hukum yang kaya. Teks Al-Qur’an sering memiliki makna kata yang mendalam dan konteks beragam, sehingga pemahaman akurat menjadi sangat penting. Dalam Al-Qur’an, sebuah kata sering kali sering kali tidak hanya memiliki satu makna yang pasti, hubungan *similar* yang terdapat pada Al-Qur’an bisa saja memiliki banyak makna berbeda, sehingga perlu langkah khusus untuk mempelajari tentang makna tersebut.

Word2Vec, yang diperkenalkan oleh Mikolov et al. (2013), merupakan teknik representasi kata yang populer yang mampu menangkap hubungan semantik antarkata secara efektif. Model ini memiliki dua arsitektur utama yaitu *CBOW* dan *Skip-Gram*. Melalui pelatihan yang sesuai, *Word2Vec* dapat memahami konteks dan makna suatu kata berdasarkan kata-kata di sekitarnya [1]. Meskipun *Word2Vec* dilatih dengan pendekatan pembelajaran tidak terawasi, vektor representasi kata yang dihasilkan dapat dimanfaatkan dalam berbagai aplikasi *NLP*, seperti klasifikasi teks, analisis sentimen, pemodelan topik dan pemrosesan teks lainnya[3]. Hal ini sangat relevan untuk menganalisis teks Al-Qur'an yang memiliki struktur dan konteks yang unik. Dengan menggunakan arsitektur *Continuous Bag of Words (CBOW)* maupun *Skip-Gram*, setiap kata mampu mengenali makna tersendiri sesuai hubungan dengan kata di sekitarnya. Pendekatan berbasis *Continuous Bag of Words (CBOW)* yang digunakan dalam penelitian ini terbukti efektif dalam menangkap makna kata melalui konteks lokal dalam kalimat, sehingga cocok digunakan untuk mempelajari teks religius seperti Al-Qur'an.

Penggunaan *Word2Vec* dalam analisis teks Al-Qur'an memungkinkan peneliti menangkap nuansa dan variasi makna yang mungkin tidak terdeteksi dengan metode tradisional. Misalnya, dalam ayat-ayat yang memiliki kata-kata bermakna ganda atau bersinonim, *Word2Vec* dapat membantu mengidentifikasi konteks spesifik yang mempengaruhi makna kata tersebut. Penelitian Goldberg dan Levy (2014) menunjukkan bahwa teknik ini mampu meningkatkan akurasi dalam memahami hubungan semantik antarkata dalam berbagai jenis teks[4].

Beberapa penelitian sebelumnya telah menerapkan model *Word2Vec* pada teks berbahasa Indonesia, tetapi masih belum banyak yang secara khusus meneliti bagaimana model ini bekerja dalam menganalisis relasi makna kata dalam korpus keislaman, seperti terjemahan Al-Qur'an Bahasa Indonesia. Padahal, kata-kata seperti iman, azab, dan surga memiliki keterkaitan makna yang kompleks dan menarik untuk dikaji secara komputasional.

Dengan memperhatikan celah penelitian tersebut, penelitian ini bertujuan untuk melatih model *Word2Vec* dengan arsitektur *CBOW* pada korpus terjemahan Al-Qur'an dalam Bahasa Indonesia, menghitung dan mengevaluasi nilai kesamaan makna antar kata-kata kunci keislaman menggunakan metode *cosine similarity*, dan

menganalisis pengaruh parameter pelatihan seperti ukuran vektor (*vector size*) serta jendela (*window size*) terhadap kualitas representasi kata yang dihasilkan.

Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan model representasi kata berbahasa Indonesia, khususnya pada domain keislaman, serta menjadi dasar bagi pengembangan sistem cerdas berbasis makna seperti pencarian ayat tematik atau klasifikasi konsep dalam Al-Qur'an.

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang yang telah disampaikan sebelumnya maka, terdapat beberapa permasalahan utama yang harus diselesaikan.
Permasalahan tersebut adalah sebagai berikut :

1. Bagaimana cara membangun model *Word2Vec* dengan arsitektur *Continuous Bag of Words (CBOW)* dan fungsi aktivasi *Hierarchical Softmax* menggunakan korpus terjemahan Al-Qur'an dalam Bahasa Indonesia?
2. Bagaimana menghitung dan mengevaluasi tingkat kemiripan makna antar kata kunci keislaman dalam Al-Qur'an dengan menggunakan metode *cosine similarity* terhadap hasil representasi vektor dari model *Word2Vec*?
3. Bagaimana pengaruh variasi parameter pelatihan, seperti ukuran vektor *(vector size)* dan lebar jendela konteks *(window size)*, terhadap kualitas representasi semantik kata dalam model *CBOW* yang dioptimasi dengan *Hierarchical Softmax*?

1.3 Batasan Masalah

Untuk mengarahkan pembahasan pada penelitian ini, maka diperlukan adanya batasan masalah sehingga pembahasan tidak menyimpang dari sasaran yang dituju pada hal-hal berikut:

1. Model representasi kata yang digunakan adalah *Word2Vec* dengan arsitektur *Continuous Bag of Words* (*CBOW*) dan menggunakan *Hierarchical Softmax* sebagai pendekatan perhitungan fungsi probabilitas output. *Skip-Gram* atau metode lain seperti *FastText* dan *Glove* tidak dibahas dalam penelitian ini.

2. Korpus yang digunakan terbatas pada terjemahan Al-Qur'an Bahasa Indonesia, yang diperoleh dari sumber Tanzil.net. Versi terjemahan lain atau tafsir tidak digunakan.
3. Evaluasi kesamaan semantik antar kata dilakukan dengan menggunakan *cosine similarity*, tanpa membandingkan metode pengukuran lain seperti *Euclidean distance*, Jaccard, atau *Pearson correlation*.
4. Kata-kata yang dianalisis merupakan kata kunci keislaman terpilih seperti iman, azab, surga, dan kata lainnya yang memiliki konteks semantik penting dalam Al-Qur'an. Pemilihan kata dilakukan secara purposif.
5. Eksperimen dilakukan dengan variasi parameter pelatihan terbatas pada ukuran dimensi vektor *(vector size)* dan lebar jendela konteks *(window size)*. Parameter lain seperti *negative sampling*, epoch, dan *learning rate* tidak dijadikan variabel utama dalam analisis.
6. Hasil penelitian tidak diarahkan untuk membangun aplikasi, melainkan bersifat eksperimental untuk menilai sejauh mana model *Word2Vec CBOW* mampu memahami hubungan makna antar kata (representasi semantik) dari korpus Al-Qur'an.

1.4 Tujuan Penelitian

Tujuan dari skripsi ini adalah :

1. Membangun model *Word2Vec* dengan arsitektur *Continuous Bag of Words (CBOW)* dengan metode *Hierarchical Softmax* sebagai pendekatan untuk mengurangi kompleksitas perhitungan probabilitas output yang dilatih menggunakan korpus terjemahan Al-Qur'an dalam Bahasa Indonesia.
2. Mengukur dan mengevaluasi kesamaan semantik antar kata kunci keislaman dalam korpus Al-Qur'an menggunakan teknik *cosine similarity* terhadap hasil vektor kata dari model yang dibangun.
3. Menganalisis pengaruh variasi parameter pelatihan seperti ukuran dimensi vektor (*vector size*) dan lebar jendela konteks *(window size)* terhadap performa representasi semantik yang dihasilkan oleh model *Word2Vec* arsitektur *CBOW*.

1.5 Metode Penelitian

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

1. Studi literatur

   Pada tahap ini, penulis mengumpulkan data dan informasi 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 metode-metode yang digunakan yaitu *Word2Vec.*

2. Analisis

   Pada tahap ini dilakukan pengkajian dan analisis terhadap hasil referensi literatur yang diperoleh yang sesuai dengan topik masalah yang akan diteliti dalam skripsi ini.

3. Simulasi

   pada tahap ini penulis melakukan simulasi model *Word2Vec* dengan menilai bagaimana perubahan parameter *vector size* dan *window* memengaruhi hasil dengan menggunakan data *training* dan *testing* dengan konteks Al-Qur'an menggunakan bahasa pemrograman *Python.* Kemudian dari hasil simulasi tersebut akan dihitung nilai akurasi ketika diuji dengan data *testing* sebagai evaluasi.

1.6 Sistematika Penulisan

Dalam skripsi ini terdapat lima bab utama dan beberapa sub bab, serta daftar pustaka dan lampiran. Berikut adalah penjabarannya :

BAB I PENDAHULUAN

Bab ini berisi Latar Belakang, Rumusan Masalah, Batasan Masalah, Tujuan Penelitian, Metode Penelitian, dan Sistematika Penulisan.

BAB II LANDASAN TEORI

Landasan teori berisi tentang penjelasan teori-teori yang berkaitan dengan pembahasan dari penelitian ini. Didapatkan dari sumber yang terdapat pada buku, artikel, dan penelitian terdahulu.

BAB III ANALISIS *SEMANTIC SIMILARITY* ANTAR KATA MENGGUNAKAN MODEL *WORD2VEC*

Bab ini berisi tentang pembahasan utama mengenai metode yang digunakan, berupa pembahasan rinci tentang pengumpulan data, *pre-processing, Train Word2vec* Model, dan pengukuran *similarity* dengan *cosine similarity*.

BAB IV STUDI KASUS DAN ANALISIS

Pada bab ini akan dijelaskan mengenai analisis hasil pengaplikasian dan percobaan kombinasi dari *Word2Vec* untuk mencari S*emantic Similarity* pada terjemahan Al-Qur'an dalam Bahasa Indonesia. Pada penelitian ini, percobaan dilakukan dengan menggunakan bahasa pemrograman *Python* dan dijalankan di Jupyter Notebook.

BAB V KESIMPULAN DAN SARAN

Pada bab ini, berisi tentang kesimpulan yang dapat diambil dari hasil analisis yang telah dilakukan di bab sebelumnya. Dan juga terdapat saran untuk pengembangan penelitian lanjutan yang dapat dilakukan dari penelitian ini