BAB I
PENDAHULUAN

1.1 Latar Belakang

Pemrosesan bahasa alami (*Natural Language Processing* / NLP) adalah salah satu cabang dari kecerdasan buatan yang berfokus pada pengolahan teks dan ucapan manusia, dengan tujuan agar komputer dapat memahami dan memanipulasi bahasa. Salah satu tantangan utama dalam NLP adalah merepresentasikan kata-kata dalam bentuk yang dapat dipahami oleh model komputer. *Word embedding* menjadi solusi untuk tantangan ini, di mana kata-kata direpresentasikan sebagai vektor dalam ruang dimensi tinggi [1]. Metode Word2Vec, yang diperkenalkan Mikolov et al. (2013), mempelopori pendekatan ini dengan dua arsitektur utama, yaitu *Continous Bag of Words* (CBoW) dan Skip-gram [2].

Namun, Word2Vec memiliki keterbatasan dalam menangani bahasa yang morfologinya kompleks seperti bahasa Arab, karena tidak memperhatikan struktur internal kata. Bahasa Arab, khususnya dalam teks Al-Qur'an, memiliki karakteristik unik berupa sistem akar kata, imbuhan yang rumit, dan bentuk kata yang bervariasi sehingga menuntut representasi kata yang fleksibel. Menjawab kekurangan tersebut, Bojanowski et al. (2017) mengembangkan FastText, model lanjutan dari Word2Vec yang memperhitungkan n-gram karakter sebagai sub-kata dalam pelatihan model, sehingga lebih sensitif terhadap morfologi kata dan sangat cocok untuk bahasa Arab.

Beberapa penelitian sebelumnya telah menunjukkan efektivitas penggunaan FastText dalam NLP berbahasa Arab. Belinkov dan Glass (2015) dalam studinya mengenai "*Arabic Diacritization with Reccurent Neural Networks*" mengaplikasikan *embedding* sebagai input pada model RNN untuk memulihkan harakat dalam teks Arab, menekankan pentingnya kualitas representasi kata dalam pemrosesan morfologi bahasa Arab. Zalmout dan Habash (2020) mengembangkan model *multitask adverensial learning* untuk menangani dialek dan fitur morfologis Arab secara bersamaan, penelitian ini menunjukkan bahwa pendekatan berbasis embedding memainkan peran krusial alam generalisasi antar variasi bahasa Arab.

Selain itu, beberapa penelitian terdahulu telah menunjukkan potensi FastText dalam memahami bahasa yang kompleks secara morfologis. Misalnya, penelitian oleh Grave et al. (2018) memperluas FastText dengan *supervised learning* untuk meningkatkan representasi kata pada berbagai bahasa dan menunjukkan bahwa FastText unggul dalam menangani kata yang jarang muncul dan kata baru (*out-of-vocabulary*). Dalam konteks bahasa Arab, Zahran et al. (2021) menggunakan FastText untuk membangun model representasi semantik dalam pengelompokan topik pada teks keislaman dan menunjukkan hasil yang unggul dibandingkan Word2Vec dan GloVe. Sementara itu, El Mahdaouy et al. (2022) mengevaluasi performa FastText dalam klasifikasi sentimen teks Arab dan menunjukkan bahwa model ini menghasilkan akurasi yang lebih baik pada data dengan variasi bentuk kata. Penelitian-penelitian ini memperkuat bahwa FastText memiliki keunggulan dalam menangani bahasa seperti Arab yang sangat dipengaruhi oleh akar kata, imbuhan, dan struktur fleksibel.

Lebih lanjut, studi oleh Adewumi et al. (2022) berjudul "*Word2Ve: Optimal Hyperparameters and Their Impact on NLP Task*" mengevaluasi pengaruh konfigurasi hyperparameter terhadap performa model dalam berbagai tugas NLP. Hasilnya menegaskan bahwa parameter seperti dimensi vektor, *window size*, dan epoch sangat mempengaruhi kualitas *semantic Similarity*, terutama dalam konteks bahasa non-Inggris seperti Arab. Selain itu, penelitian oleh Darwish et al. (2020) melalui "*A Panoramic Survey of NLP in the Arab World*" menyoroti bahwa meskipun penelitian NLP dalam bahasa Arab meningkat, terdapat kesenjangan besar antara kebutuhan dan teknologi yang tersedia.

Dengan latar belakang tersebut, penelitian ini menggunakan FastText sebagai model *word embedding* untuk merepresentasikan kata dalam teks Al-Qur'an berbahasa Arab, dan berfokus pada analisis pengaruh variasi hyperparameter terhadap kualitas *semantic similarity*. Evaluasi dilakukan secara kuantitatif menggunakan *cosine Similarity*, untuk mengetahui sejauh mana model dapat memberikan kontribusi dalam pengembangan sistem NLP berbasis bahasa Arab, khususnya untuk kebutuhan aplikasi seperti sistem pencarian Al-Qur'an, klasifikasi tematik, hingga penerjemah kontekstual berbasis makna.

Dalam penelitian NLP berbahasa Arab, teks Al-Qur'an sering menjadi sumber data utama. Al-Qur'an memiliki struktur bahasa yang kompleks, kaya makna, dan bervariasi bentuk katanya mengikuti aturan morfologi khas. Bahasa Arab Al-Qur'an berbeda dari bahasa Arab modern, sehingga memerlukan pendekatan khusus dalam pemrosesan dan representasi katanya. Tantangan utamanya meliputi kompleksitas morfologi, pola akar kata, sistem afiks, dan makna kata yang sangat bergantung pada konteks ayat [3]. Hal ini sejalan dengan firman Allah SWT dalam Surat Yusuf ayat 2 :

اِنَّآ اَنْزَلْنٰهُ قُرْاٰنًا عَرَبِيًّا لَّعَلَّكُمْ تَعْقِلُوْنَ ٢

Artinya: "Sesungguhnya Kami menurunkan Al-Qur'an berbahasa Arab, agar kamu memahaminya." (Q.S Yusuf: 2)

Ayat ini menekankan pentingnya pemahaman bahasa Arab sebagai kunci dalam memahami kandungan Al-Qur'an. Pemahaman tersebut menjadi semakin penting ketika dikaitkan dengan perkembangan teknologi pemrosesan bahasa yang dapat memudahkan akses terhadap makna dan konteks ayat. Pentingnya bahasa Arab dalam memahami ajaran Islam juga dipertegas dalam berbagai riwayat, sebagaimana sabda Rasulullah Shallallahu 'alaihi wa sallam:

"تَعَلَّمُوا الْعَرَبِيَّةَ ، فَإِنَّهَا جُزْءٌ مِنْ دِينِكُمْ"

Artinya: "Belajarlah bahasa Arab, karena ia adalah bagian dari agamamu." (Hadits ini diriwayatkan oleh Imam Al-Baihaqi dalam Syu'abul Iman, dan dishahihkan oleh Al-Albani dalam Silsilah Al-Ahadits Ash-Shahihah no. 1297).

Hadits ini menekankan pentingnya mempelajari bahasa Arab bagi umat Muslim karena keterkaitannya yang erat dengan pemahaman agama, termasuk Al-Qur'an. Hal ini semakin menguatkan relevansi penelitian ini dalam upaya meningkatkan pemahaman terhadap teks Al-Qur'an melalui teknologi NLP. Lebih dari itu, semangat menuntut ilmu yang mendasari penelitian ini juga didasari firman Allah dalam Surah Thaha ayat 114:

رَّبِّ زِدْنِيْ عِلْمًا ١١٤

Artinya: "Ya Tuhanku, tambahkanlah kepadaku ilmu pengetahuan." (Q.S Thaha: 114)

Melalui teknologi NLP dan model seperti FastText, pemeliharaan makna dan konteks Al-Qur'an dapat diupayakan secara ilmiah dan sistematis, demi memperluas akses terhadap pemahaman yang lebih mendalam. Rasulullah SAW juga bersabda:

"خَيْرُكُمْ مَنْ تَعَلَّمَ الْقُرْآنَ وَعَلَّمَهُ"

"Sebaik-baik kalian adalah yang mempelajari Al-Qur'an dan mengajarkannya." (H.R Tirmidzi No. 2907)

Dengan demikian, penelitian ini tidak hanya memiliki kontribusi akademik dalam pengembangan NLP bahasa Arab, tetapi juga bernilai ibadah dan dakwah dalam memperluas pemahaman terhadap Al-Qur'an dengan pendekatan teknologi yang ilmiah dan terstruktur.

Seiring dengan kompleksitas bahasa Arab dalam Al-Qur'an yang kaya makna, sistem akar kata, dan morfologi yang bervariasi, maka dibutuhkan pendekatan teknologi yang mampu merepresentasikan kata secara kontekstual. Di sinilah peran penting model representasi kata seperti Word2Vec dan pengembangannya, yaitu FastText, menjadi sangat relevan.

Namun, penerapan Word2Vec pada bahasa arab, khususnya dalam teks Al-Qur'an memiliki keterbatasan yang disebabkan oleh kompleksitas morfologi, struktur sintaksis yang berbeda, serta keterbatasan dataset yang tersedia, sehingga diperlukan metode representasi kata yang lebih akurat [4]. Bahasa Arab adalah bahasa flektif yang sangat kaya, di mana satu akar kata dapat menghasilkan banyak bentuk kata berbeda.

Oleh karena itu, penelitian ini menggunakan FastText, pengembangan dari Word2Vec yang tidak hanya merepresentasikan makna secara utuh tetapi juga mempertimbangkan karakter sub kata (n-gram). Pendekatan ini diharapkan lebih mampu menangkap makna kata berdasarkan bentuk morfologisnya, yang sangat relevan untuk bahasa Arab [5].

Penggunaan FastText diharapkan meningkatkan pemahaman hubungan semantik antar kata dalam teks Al-Qur'an. Kualitas representasi kata sangat dipengaruhi oleh konfigurasi hyperprameter seperti *vektor size, window size, minimum count,* epoch, dan metode pelatihan [6]. Pengaturan hyperparameter yang

tepat dapat secara signifikan meningkatkan kinerja model dalam memahami teks Al-Qur'an dan pada akhirnya meningkatkan akurasi tugas-tugas NLP lainnya.

Selain itu, penelitian terkait NLP dalam bahasa Arab terutama dalam pemrosesan teks Al-Qur'an, masih tergolong kurang dibandingkan dengan bahasa lain. Dengan meningkatnya permintaan terhadap aplikasi berbasis NLP yang mendukung bahasa arab, seperti mesin pencari berbasis Al-Qur'an, dan sistem terjemahan otomatis, penelitian ini diharapkan dapat berkontribusi dalam mengembangkan model representasi kata yang lebih baik untuk bahasa Arab. Temuan ini dapat menjadi acuan bagi pengembang sistem NLP dan peneliti, dalam mengoptimalkan pemrosesan bahasa Arab [6].

Dengan tetap merujuk pada kerangka Word2Vec, penelitian ini bertujuan untuk mengevaluasi bagaimana penerapan dan evaluasi FastText dengan berbagai konfigurasi hyperparameter dapat mempengaruhi hasil pemrosesan bahasa alami untuk dataset bahasa Arab, khususnya dalam teks Al-Qur'an.

1.2 Rumusan Masalah

Adapun rumusan masalah yang akan dikaji pada penelitian ini sebagai berikut:

1. Bagaimana pengaruh konfigurasi hyperparameter dalam FastText terhadap kualitas representasi kata pada pemrosesan bahasa alami menggunakan dataset Al-Qur'an bahasa Arab?
2. Bagaimana efektivitas FastText dalam merepresentasikan relasi semantik antar kata dalam teks Al-Qur'an, serta dampaknya terhadap akurasi evaluasi *semantic similarity*?

1.3 Batasan Masalah

Batasan masalah pada penelitian ini, sebagai berikut:

1. Penelitian ini menggunakan FastText sebagai model word embedding, yang dikembangkan dari Word2Vec.
2. Penggunaan Hyperparameter berfokus pada utama *vektor size*, *window size*, *learning rate*, *minimum count,* dan jumlah iterasi (epoch).
3. Penelitian ini menggunakan dataset Al-Qur'an bahasa Arab.

4. Evaluasi yang dilakukan berbentuk representasi kata dalam ruang vektor dan hubungan *semantic similarity*.
5. Penelitian ini hanya memfokuskan analisis *semantic similarity* pada kata target خير.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Menganalisis pengaruh hyperparameter FastText (*vector size, windows size,* jumlah iterasi (epoch), *minimum count,* dan metode pelatihan) terhadap kualitas representasi kata dalam pemrosesan teks Al-Qur'an bahasa Arab, serta menentukan konfigurasi yang optimal untuk meningkatkan akurasi *semantic similarity.*
2. Mengevaluasi efektivitas FastText dalam merepresentasikan relasi semantik antar kata dalam teks Al-Qur'an dan dampaknya terhadap hasil pengukuran *semantic similarity*.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Memberikan kontribusi ilmiah dalam bidang pemrosesan bahasa alami (*Natural Language Processing*), khususnya dalam pengembangan representasi kata berbasis FastText untuk bahasa Arab Al-Qur'an. Penelitian ini dapat menjadi referensi bagi studi lanjutan terkait pemilihan hyperparameter optimal dalam *word embedding*.
2. Memberikan dasar implementasi model FastText yang lebih efektif untuk digunakan dalam berbagai aplikasi berbasis bahasa Arab, seperti sistem pencarian makna Al-Qur'an, klasifikasi tematik ayat, dan penerjemahan kontekstual.
3. Mendukung upaya pemahaman Al-Qur'an secara lebih mendalam melalui teknologi, sebagai bentuk kontribusi keilmuan dalam memahami kandungan wahyu dengan pendekatan semantik dan kontekstual.

1.6 Metode Penelitian

Metodologi dalam penelitian ini di antaranya:

1. Studi Literatur

   Pada tahap studi literatur ini bertujuan untuk mengumpulkan informasi dari berbagai sumber seperti jurnal, buku, dan penelitian sebelumnya yang berkaitan dengan Word2Vec dan FastText, *word embedding*, hyperparameter, karakteristik bahasa Arab, dan evaluasi model.

2. Penelitian

   Pada tahap penelitian ini penulis melakukan tahap *pre-processing* data Al-Qur'an bahasa Arab menggunakan algoritma FastText dan dilatih menggunakan berbagai kombinasi hyperparameter. Setelah pelatihan, model dievaluasi. Hasil evaluasi dianalisis untuk menentukan konfigurasi hyperparameter yang optimal dalam meningkatkan kualitas representasi kata.

1.7 Sistematika Penulisan

Pada skripsi ini terdapat lima bab sistematika penulisan di antaranya:

BAB I PENDAHULUAN

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

BAB II LANDASAN TEORI

Bab landasan teori menjelaskan tentang teori-teori yang melandasi pembahasan inti yang saling berkaitan dan sebagai penunjang dalam penulisan skripsi, seperti *Natural Language Processing* (NLP), *Word Embedding, Word2Vec, FastText,* hyperparameter, serta metode evaluasi.

BAB III PENGARUH HYPERPARAMETER DALAM WORD2VEC TERHADAP *SEMANTIC SIMILARITY* KATA MENGGUNAKAN DATASET Al-QUR'AN BAHASA ARAB

Pada bab ini berisi pembahasan tentang penelitian yang dilakukan dari pengambilan dataset Al-Qur'an, lalu tahap text *pre-processing*

dengan metode algoritma FastText dan dilatih menggunakan berbagai kombinasi hyperparameter. Setelah pelatihan, model dievaluasi menggunakan *cosine similarity*. Hasil evaluasi dianalisis untuk menentukan konfigurasi hyperparameter yang optimal dalam meningkatkan kualitas representasi kata dalam teks Al-Qur'an.

BAB IV **EKSPERIMEN DAN ANALISIS**

Bab ini berisi pemaparan mengenai analisis hasil yang telah dilakukan pada bab sebelumnya. Hasil evaluasi dari *cosine similarity* digunakan sebagai dasar dalam menilai efektivitas model yang dilatih. Dari hasil analisis, konfigurasi optimal dapat diidentifikasi untuk meningkatkan akurasi dalam pemrosesan bahasa alami berbasis bahasa Arab. Hasil ini diharapkan dapat menjadi acuan dalam penelitian lebih lanjut dan pengembangan aplikasi NLP yang lebih efektif.

BAB V **PENUTUP**

Bab penutup berisi hasil simpulan dari rumusan masalah yang telah dijelaskan dan berisi saran yang diperuntukkan untuk penelitian berikutnya sebagai pengembangan dari Word2Vec dan FastText.

DAFTAR PUSTAKA Bagian ini berisi daftar seluruh sumber referensi yang menjadi acuan, seperti jurnal ilmiah, buku dan penelitian sebelumnya, yang digunakan sebagai landasan teori dan penguat dalam penyusunan skripsi ini