BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Al-Qur'an merupakan pedoman hidup umat Islam yang diturunkan sebagai petunjuk dan rahmat bagi seluruh manusia. Dalam surat Al-Baqarah ayat 2 disebutkan :

ذٰلِكَ الْكِتٰبُ لَا رَيْبَ ۛ فِيْهِ ۛ هُدًى لِّلْمُتَّقِيْنَ ۙ ﴿٢﴾

Kitab (Al-Qur'an) ini tidak ada keraguan di dalamnya; (ia merupakan) petunjuk bagi orang-orang yang bertakwa,

Frasa "petunjuk bagi mereka yang bertakwa" mengisyaratkan bahwa bimbingan Al-Qur'an tidak bersifat pasif. Perintah ilahi pertama yang diturunkan kepada Nabi Muhammad Saw yaitu "Iqra" (bacalah!) sebagaimana termaktub dalam surat Al-Alaq ayat 1-5 yaitu :

اِقْرَأْ بِاسْمِ رَبِّكَ الَّذِيْ خَلَقَ ۚ ﴿١﴾ خَلَقَ الْاِنْسَانَ مِنْ عَلَقٍ ۚ ﴿٢﴾ اِقْرَأْ وَرَبُّكَ الْاَكْرَمُ ۙ ﴿٣﴾ الَّذِيْ

عَلَّمَ بِالْقَلَمِ ۙ ﴿٤﴾ عَلَّمَ الْاِنْسَانَ مَا لَمْ يَعْلَمْ ؕ ﴿٥﴾

Artinya : (1) Bacalah dengan (menyebut) nama Tuhanmu yang menciptakan, (2) Dia menciptakan manusia dari segumpal darah, (3) Bacalah! Tuhanmu Yang Maha Mulia, (4) Yang mengajar (manusia) dengan pena. (5) Sekali-kali tidak! Sesungguhnya manusia itu benar-benar melampaui batas.

Ayat ini menegaskan bahwa Al-Qur'an merupakan sumber petunjuk yang otentik dan bebas dari keraguan. Oleh karena itu, memahami isi kandungan Al-Qur'an termasuk melalui pendekatan teknologi dan ilmu komputer, menjadi salah satu bentuk kontribusi keilmuan dalam menjadikan Al-Qur'an lebih mudah dipahami oleh seluruh kalangan.

Teknologi kecerdasan buatan (AI) adalah salah satu inovasi teknologi yang berkembang paling cepat dalam era digital yang semakin maju. Ini memiliki potensi besar untuk diterapkan dalam banyak bidang seperti pendidikan khususnya pendidikan agama. AI dapat membantu umat Islam dalam belajar membaca dan

memahami Al-Qur'an dengan cara yang lebih interaktif dan adaptif [1]. Dengan memberikan akses yang lebih luas dan mendalam terhadap pengetahuan Al-Qur'an, kecerdasan buatan dapat digunakan untuk mempelajari Al-Qur'an [2]. Dengan memanfaatkan AI, proses tadabbur dapat diperkaya melalui pendekatan komputasional yang mampu memproses dan menganalisis data tekstual dalam skala besar secara objektif dan efisien. Ini sejalan dengan anjuran Rasulullah Saw :
"Barang siapa menempuh jalan untuk mencari ilmu, maka Allah akan memudahkan baginya jalan menuju surga" (HR.Tirmidzi).

Salah satu tantangan utama dalam kajian Al-Qur'an adalah memahami hubungan antar konsep *similarity* kata yang tersebar dalam berbagai surat dan ayat. Tradisionalnya, pendekatan manual dalam tafsir Al-Qur'an memerlukan waktu yang lama dan rentan terhadap subjektivitas. Oleh karena itu, diperlukan metode yang dapat mengotomatisasi proses ini dengan tetap mempertahankan akurasi dan analisis. Salah satu pendekatan yang semakin penting dalam menangani masalah ini adalah *text mining. Text mining* adalah proses menambang data yang berupa teks, biasanya berasal dari dokumen. Tujuan dari proses ini adalah untuk menemukan kata-kata yang dapat mewakili isi dokumen sehingga dapat dilakukan analisis hubungan antara dokumen.

Penerapan kecerdasan buatan (AI) dan pemrosesan bahasa alami (NLP) dalam analisis teks keagamaan adalah solusi baru. Penelitian telah menunjukkan bahwa model *FastText* yang dikembangkan oleh *facebook AI research* dapat mengatasi kelemahan model sebelumnya dalam menangkap makna kata dengan mempertimbangkan subkata (*subword)*. Ini memungkinkan *FastText* untuk memahami konteks dan makna kata lebih jauh, terutama dalam bahasa yang kompleks seperti bahasa Arab dan terjemahan [3].

Pemanfaatan *word embedding* untuk analisis Al-Qur'an telah dilakukan dalam beberapa penelitian sebelumnya, namun dengan fokus yang berbeda. Sebagai contoh, Qorina (2020) dalam tesisnya melakukan analisis perbandingan antara metode *FastText* dan *Word2Vec* untuk sistem temu kembali informasi pada teks Sirah Nabawiyah. Hasilnya menunjukkan keunggulan *FastText* dalam menangani variasi kata, namun penelitian tersebut berfokus pada perbandingan performa model secara umum dan tidak secara spesifik menyelidiki bagaimana pengaruh

parameter internal *FastText* dalam menangani variasi kata, namun penelitian tersebut berfokus pada perbandingan performa model secara umum dan tidak secara spesifik menyelidiki bagaimana pengaruh parameter internal *FastText,* terutama n-gram, terhadap kualitas hasil similaritas kata [3]. Sementara itu Guntara (2019) melakukan penelitian untuk membangun daftar kata terkait dalam kosakata Al-Qur'an berdasarkan konsep kesamaan distribusional. Penelitian ini memberikan kontribusi penting dalam pemetaan semantic, namun metode yang digunakan tidak berbasis *sub-word* seperti *FastText* sehingga kemampuannya untuk menangani variasi morfologis dan kata-kata yang jarang muncul lebih terbatas [2]. Dari tinjauan di atas, terlihat adanya sebuah celah penelitian yang jelas, yaitu yang menjadi pembeda utama dari penelitian ini adalah fokusnya yang lebih mendalam dan spesifik. Berbeda dengan penelitian Qorina (2020) yang membandingkan performa antar model secara umum, penelitian ini melakukan analisis pada level parameter internal *FastText* yang secara khusus menguji pengaruh pengaturan parameter n-gram terhadap kualitas representasi kata. Selain itu, penelitian Guntara (2019) menggunakan metode kesamaan distribusional yang tidak berbasis *subword*, Penelitian ini secara fundamental memanfaatkan arsitektur berbasis sub-word pada *FastText.* Pendekatan ini secara teoritis lebih andal dalam menangani variasi morfologis dan kata-kata langka yang merupakan keterbatasan pada metode sebelumnya. Meskipun *FastText* diakui unggul karena mekanisme *sub-word* nya, belum ada penelitian yang secara sistematis dan mendalam menganalisis sejauh mana pengaruh pengaturan parameter n-gram (panjang minimum dan maksimum dari *sub-word*) saat proses pelatihan model terhadap kualitas similaritas kata yang dihasilkan, khususnya pada korpus terjemahan Al-Qur'an bahasa Inggris.

Penggabungan metode data, NLP, *word embedding, cosine similarity*, memberikan pendekatan yang komprehensif dan efektif dalam analisis data teks yang kompleks seperti data terjemahan Al-Qur'an Bahasa Inggris. Tujuan penelitian ini adalah untuk membuat model *FastText* yang dapat menunjukkan hubungan antar *similarity* kata dalam terjemahan Al-Qur'an. Dengan menggunakan teknik penerapan kata, diharapkan model ini dapat memberikan visualisasi yang memudahkan pemahaman dan mengidentifikasi keterkaitan antar ayat. Metode ini tidak hanya membuat analisis teks keagamaan lebih mudah, tetapi juga

menawarkan peluang untuk membangun aplikasi eduktif berbasis kecerdasan buatan yang dapat digunakan oleh berbagai kelompok masyarakat.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, didapat rumusan masalah sebagai berikut

1. Bagaimana pengaruh pengaturan parameter n-gram (panjang minimum dan maksimum) pada training model *FastText* terhadap kemampuannya dalam mengukur similaritas kata, khususnya dalam menangani variasi morfologis pada data terjemahan Al-Qur'an bahasa Inggris?
2. Sejauh mana mode *FastText* yang telah dilatih dengan parameter n-gram yang optimal mampu mengidentifikasi dan memetakan klaster-klaster tematik yang relevan secara teologis dalam korpus terjemahan Al-Qur'an?
3. Bagaimana visualisasi menggunakan *Principal Component Analysis* (PCA) dapat membantu dalam menganalisis dan menginterpretasi hubungan semantic antar kata yang dihasilkan oleh model *FastText?*

1.3 Batasan Masalah

Permasalahan yang dibahas dalam tugas akhir ini memiliki batasan sebagai berikut

1. Penelitian ini hanya menggunakan teks terjemahan Al-Qur'an bahasa Inggris. Bukan teks asli dalam bahasa Arab, analisis hubungan tematik hanya terbatas pada struktur semantik dalam bahasa Inggris.
2. Pemodelan *word embedding* secara spesifik hanya menggunakan algoritma *FastText* dengan arsitektur *skip-gram.* Penelitian ini tidak membandingkan performa *FastText* dengan algoritma lain seperti *Word2Vec, GloVe* atau model yang lebih modern.
3. Analisis pengaruh parameter secara mendalam difokuskan pada parameter n-gram. Meskipun parameter lain digunakan, analisis pengaruhnya tidak menjadi fokus utama penelitian ini.
4. Evaluasi hasil kemiripan kata dilakukan secara kuantitatif dengan metrik *cosine similarity* dan visualisasi PCA. Penelitian ini tidak melibatkan validasi kualitatif dari ahli tafsir atau pakar di bidang studi ke-Islaman.

5. Untuk analisis tematik yang mendalam pada korpus utama, pengujian similaritas kata di fokuskan pada satu kata target utama, yaitu kata *"lord".* Kata ini di pilih sebagai studi kasus karena frekuensi kemunculan yang tinggi dalam teks.

1.4 Tujuan Penelitian

Adapun tujuan yang ingin di capai dari penelitian ini adalah sebagai berikut :

1. Menganalisis secara sistematis pengaruh variasi parameter n-gram terhadap kualitas representasi vektor kata yang dihasilkan oleh model *FastText.*
2. Membangun model *FastText* yang optimal untuk mengidentifikasi hubungan similaritas kata dan mengungkap klaster-klaster tematik yang bermakna dalam data terjemahan Al-Qur'an bahasa Inggris.
3. Memvalidasi hasil similaritas kata secara kuantitatif menggunakan metrik *cosine similarity* dan secara kualitatif melalui visualisasi PCA untuk mempermudah interpretasi.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut :

1. Memberikan kontribusi pada bidang Pemrosesan Bahasa Alami (NLP) dan linguistic komputasi, khususnya mengenai pemahaman mendalam tentang peran mekanisme sub-word pada model *FastText* dalam menangani teks keagamaan yang kompleks dan kaya secara morfologis.
2. Hasil dari penelitian ini dapat menjadi dasar untuk pengembangan aplikasi berbasis kecerdasan buatan yang lebih canggih, seperti sistem pencarian tematik Al-Qur'an, alat bantu tadabbur, atau platform edukasi ke-Islaman yang interaktif.
3. Menyajikan alur kerja yang komprehensif, mulai dari pra-pemrosesan data hingga analisis dan visualisasi yang dapat di replikasi atau diadaptasi oleh peneliti lain untuk menganalisi korpus teks yang serupa.

1.6 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 *FastText* dan *cosine similarity* melalui studi literatur yang mencakup buku, jurnal, dan skripsi.

2. Penelitian

   Pada tahap penelitian penulis melakukan pengujian menggunakan korpus dummy yang bervariasi dan melakukan *pre-processisng* data terjemahan Al-Qur'an bahasa Inggris dengan menggunakan algoritma *FastText* dan data disimpan dalam dokumen untuk digunakan dalam evaluasi *cosine similarity.*

1.7 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, manfaat 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, text mining, Natural Language Processing* (NLP), *FastText* serta metode evaluasi meliputi *cosine similarity.*

BAB III **PENGARUH PARAMETER *N-GRAM* PADA TRAINING *FASTTEXT* TERHADAP *SIMILARITY* KATA MENGGUNAKAN DATA TERJEMAHAN AL-QUR'AN BAHASA INGGRIS**

Tujuan dari bab ini adalah mengetahui bagaimana variasi parameter pada model *FastText* dipengaruhi oleh variasi parameter *n-gram.* Tahapan penelitian ini mencakup proses pengumpulan data, tahapan *preprocessing teks,*

	pelatihan model *FastText* serta metode yang digunakan untuk mengukur kemiripan kata dengan *cosine similarity*
BAB IV	**ANALISIS HASIL PENGUJIAN PARAMETER *N-GRAM* PADA TRAINING *FASTTEXT***
	Bagian ini membahas dan menganalisis hasil eksperimen. Untuk tujuan analisis ini, skor *similarity* tertinggi di kumpulkan dari berbagai kombinasi parameter pelatihan. Bab ini juga membahas visualisasi penerapan menggunakan metode PCA dan penilaian representasi kata yang dihasilkan.
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 *FastText* dan *cosine similarity.*
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.