BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Natural Language Processing (NLP) merupakan salah satu bidang dalam kecerdasan buatan yang berkembang pesat dalam pengolahan dan analisis bahasa manusia menggunakan komputer. Agar dapat diproses secara komputasional, bahasa perlu direpresentasikan ke dalam bentuk numerik. Salah satu bentuk representasi yang banyak digunakan adalah *word embedding*, yaitu representasi kata dalam bentuk vektor pada ruang berdimensi tertentu. Representasi tersebut dibangun berdasarkan pola kemunculan kata dalam suatu korpus, sehingga kata-kata yang memiliki pola konteks yang serupa cenderung memiliki representasi yang berdekatan dalam ruang vektor [1, 2]. Melalui representasi tersebut, hubungan antarkata dapat dinyatakan secara matematis dan digunakan dalam berbagai tugas NLP, seperti klasifikasi teks, pencarian informasi, penerjemahan mesin, dan pengukuran kemiripan kata.

Perkembangan *word embedding* pada awalnya banyak dilakukan pada bahasa Inggris. Berbagai model representasi kata dikembangkan untuk mempelajari hubungan semantik dan sintaksis dari teks berbahasa Inggris. Mikolov dkk. memperkenalkan Word2Vec melalui model *Continuous Bag-of-Words* (CBOW) dan *Skip-Gram* yang mempelajari representasi kata berdasarkan konteks kemunculannya dalam korpus [1]. Pennington dkk. kemudian memperkenalkan GloVe yang membentuk representasi kata berdasarkan informasi statistik kemunculan bersama antarkata [2]. Perkembangan tersebut menunjukkan bahwa hubungan antarkata dapat direpresentasikan dalam ruang vektor dan digunakan secara efektif untuk berbagai tugas pengolahan bahasa.

Pengembangan *word embedding* kemudian meluas ke bahasa-bahasa lain, termasuk bahasa Arab. Penerapan NLP pada bahasa Arab memiliki karakteristik tersendiri karena bahasa Arab memiliki sistem morfologi yang kaya, struktur sintaksis yang kompleks, serta variasi bentuk kata yang berkaitan dengan penggunaan dan konteksnya. Karakteristik tersebut memengaruhi pola kemunculan kata dalam korpus dan dapat berpengaruh terhadap pembentukan representasi kata. Oleh karena itu, pengembangan *word embedding* bahasa Arab perlu memperhatikan karakteristik bahasa serta representasi yang dihasilkan.

Bahasa Arab juga memiliki kedudukan penting dalam kajian keislaman karena merupakan bahasa yang digunakan dalam Al-Qur'an. Allah Swt. berfirman dalam QS. Yusuf [12]: 2:

إِنَّآ أَنزَلْنَاهُ قُرْءَٰنًا عَرَبِيًّا لَّعَلَّكُمْ تَعْقِلُونَ

"Sesungguhnya Kami menurunkannya berupa Al-Qur'an berbahasa Arab, agar kamu mengerti."

Ayat tersebut menunjukkan kedudukan bahasa Arab sebagai bahasa Al-Qur'an dan sumber ilmu pengetahuan. Seiring dengan perkembangan teknologi informasi, teks berbahasa Arab tersedia dalam berbagai sumber digital yang dapat dimanfaatkan dalam penelitian NLP. Salah satu sumber tersebut adalah Wikipedia bahasa Arab yang memuat artikel dari berbagai bidang pengetahuan dengan karakteristik yang beragam. Perbedaan ukuran korpus serta topik domain seperti geografi dan sejarah di dalam Wikipedia akan menghasilkan struktur spektral matriks yang berbeda. Hal ini menjadi penting karena perbedaan karakteristik data tersebut berimplikasi langsung pada tingkat *noise* statistik ($\hat{\sigma}$) dan jumlah dimensi efektif ($\hat{r}$) yang dihasilkan melalui metode penyaringan seperti *Universal Singular Value Thresholding* (USVT). Oleh karena itu, penelitian ini menggunakan variasi korpus untuk mengkaji pengaruh karakteristik data terhadap parameter-parameter tersebut.

Seiring dengan berkembangnya penelitian *word embedding* bahasa Arab, perhatian tidak hanya diberikan pada metode pembentukan representasi, tetapi juga pada konfigurasi yang digunakan. Allahim dan Cherif mengkaji beberapa konfigurasi *word embedding* bahasa Arab dengan mempertimbangkan ukuran vektor, ukuran *window*, dan algoritma pelatihan. Hasil penelitian menunjukkan bahwa konfigurasi yang berbeda dapat menghasilkan kinerja yang berbeda pada tugas NLP yang digunakan [3]. Salah satu konfigurasi yang berkaitan langsung dengan bentuk representasi adalah ukuran vektor atau jumlah dimensi. Jumlah dimensi menentukan banyaknya komponen yang digunakan untuk merepresentasikan setiap kata sehingga pemilihannya menjadi bagian penting dalam pembentukan *word embedding*.

Permasalahan mengenai dimensi merupakan tantangan fundamental dalam berbagai bidang analisis data karena adanya konsep dimensi intrinsik (*intrinsic dimensionality*). Kraemer dkk. menunjukkan bahwa sekumpulan data kompleks dengan ribuan variabel sering kali dapat direpresentasikan secara efisien hanya dengan sedikit dimensi independen tanpa kehilangan informasi signifikan [4]. Fenomena ini juga relevan dalam pengolahan teks, di mana representasi berdimensi tinggi membutuhkan ruang penyimpanan dan waktu komputasi yang besar. Akritidis dan Bozanis menunjukkan bahwa reduksi dimensi dapat mengurangi waktu pemrosesan dengan dampak minimal terhadap akurasi [5]. Namun, pengurangan dimensi yang terlalu ekstrem dapat menimbulkan "kutukan dimensi rendah" (*curse of low dimensionality*) sebagaimana dikaji oleh Ohsaka dan Togashi, di mana keterbatasan kapasitas model menyebabkan hilangnya keberagaman informasi dan munculnya bias popularitas [6]. Hal ini memperkuat penelitian Bensalah dkk. pada *word embedding* bahasa Arab bahwa pemilihan teknik dan jumlah dimensi yang tepat sangat menentukan kinerja akhir representasi [7].

Perhatian terhadap ukuran dimensi juga terdapat dalam penelitian Nacar dkk. melalui model *General Arabic Text Embedding* (GATE). Penelitian tersebut menggunakan pendekatan *Matryoshka Representation Learning* sehingga satu model dapat digunakan pada beberapa ukuran representasi, yaitu 768, 512, 256, 128, dan 64 dimensi [8]. Penelitian tersebut menunjukkan bahwa ukuran dimensi merupakan aspek penting dalam pengembangan representasi bahasa Arab. Namun, ukuran dimensi yang digunakan dalam berbagai penelitian terdahulu umumnya ditentukan secara heuristik atau empiris berdasarkan uji coba coba-coba (*trial and error*) dari perancang model, dan belum memberikan dasar matematis yang kokoh untuk menentukan jumlah dimensi optimal berdasarkan struktur intrinsik data dan tingkat *noise* statistik korpus.

Pemilihan dimensi yang tepat menjadi penting karena jumlah dimensi berkaitan erat dengan keseimbangan informasi yang dipertahankan dalam suatu representasi. Jika jumlah dimensi terlalu kecil, sebagian besar informasi semantik penting yang terdapat dalam representasi dapat hilang. Sebaliknya, peningkatan jumlah dimensi yang tidak terkontrol dapat menyebabkan komponen yang kurang informatif dan dipengaruhi oleh *noise* ikut dipertahankan, sehingga meningkatkan kompleksitas representasi dan kesalahan estimasi. Kondisi tersebut menunjukkan adanya *trade-off* antara informasi yang hilang akibat dimensi yang terlalu rendah (*bias*) dan pengaruh komponen *noise* pada dimensi yang lebih tinggi (*variance*). Persoalan dimensionalitas *word embedding* tersebut dikaji secara teoritis oleh Yin dan Shen melalui kerangka *Pairwise Inner Product* (PIP) *Loss* [9].

PIP *Loss* digunakan untuk mengukur perbedaan struktur hubungan antarkata antara *embedding* pada dimensi tertentu dan *oracle embedding*. Dengan ukuran tersebut, kualitas representasi dapat dibandingkan secara sistematis pada berbagai dimensi. Yin dan Shen menunjukkan bahwa pemilihan dimensi berkaitan erat dengan *bias-variance trade-off* [9]. Oleh karena itu, penentuan dimensi perlu mempertimbangkan keseimbangan antara kedua sumber kesalahan tersebut menggunakan pendekatan analitis yang melibatkan Teori Perturbasi Matriks (*Matrix Perturbation Theory*) [10].

Dalam penelitian ini, hubungan antara kata dan konteks direpresentasikan menggunakan *Pointwise Mutual Information* (PMI), kemudian didekomposisi menggunakan *Singular Value Decomposition* (SVD). Untuk mengatasi adanya perturbasi statistik akibat keterbatasan data pengamatan pada matriks korpus, digunakan *Universal Singular Value Thresholding* (USVT) yang diperkenalkan oleh Chatterjee guna mengestimasi tingkat *noise* dan menyaring komponen spektral yang tidak relevan [11]. Berdasarkan karakteristik spektral dan estimasi *noise* tersebut, dimensi optimal ditentukan melalui pendekatan teoretis menggunakan batas atas Teorema 3. Namun, untuk memastikan keandalan prediksi teoretis tersebut, diperlukan validasi numerik melalui Simulasi Monte Carlo guna melihat kedekatan antara hasil perhitungan matematis dengan perilaku data saat diberikan perturbasi acak. Kemudian dimensi optimal yang diperoleh dilakukan evaluasi intrinsik pada tugas *word similarity* untuk memastikan bahwa interval dimensi yang dianggap optimal secara teoretis memang selaras dengan performa representasi makna kata dalam penggunaan dunia nyata.

Dengan demikian, penelitian ini fokus untuk mengkaji dan mengoptimalkan penentuan dimensi *word embedding* bahasa Arab pada korpus Wikipedia menggunakan kerangka PIP *Loss* berbasis Teori Perturbasi Matriks.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan, rumusan masalah dalam penelitian ini dijabarkan sebagai berikut:

1. Bagaimana perbedaan ukuran dan domain korpus memengaruhi kebutuhan representasi berdimensi rendah pada *word embedding* bahasa Arab?

2. Bagaimana menentukan jumlah dimensi yang dapat mempertahankan informasi penting dalam representasi sekaligus mengendalikan pengaruh *noise* statistik pada korpus bahasa Arab?

3. Sejauh mana dimensi yang ditentukan berdasarkan karakteristik intrinsik korpus mampu menghasilkan performa yang sesuai pada evaluasi intrinsik *word similarity*?

1.3 Batasan Masalah

1. Korpus data yang digunakan dibatasi pada tiga jenis korpus teks bahasa Arab baku (*Modern Standard Arabic*), yaitu Korpus Utama (skala umum), Korpus Geografi (spesifik domain), dan Korpus Sejarah (spesifik domain).
2. Ukuran vokabulari ($|V|$) dibatasi hanya untuk kata dengan frekuensi kemunculan tertinggi (*Top K*), yaitu sebesar 20.000 kata untuk Korpus Utama, serta masing-masing 10.000 kata untuk Korpus Geografi dan Korpus Sejarah.
3. Metode penentuan dimensi optimal dibatasi pada kerangka *Pairwise Inner Product* (PIP) *Loss* melalui pendekatan teoretis batas atas (*upper bound*) Teorema 3 dan pendekatan numerik menggunakan Simulasi Monte Carlo, dengan tahap penyaringan awal menggunakan *Universal Singular Value Thresholding* (USVT).
4. Pengujian kualitas vektor kata dibatasi pada evaluasi intrinsik melalui uji kemiripan kata (*word similarity*) menggunakan metrik *Cosine Similarity*, serta tidak mencakup pengujian pada tugas NLP hilir (*downstream tasks*) seperti analisis sentimen atau penerjemahan mesin.

1.4 Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Menganalisis pengaruh ukuran dan domain korpus terhadap karakteristik *noise* statistik dan struktur spektral matriks PMI.
2. Menentukan dimensi optimal *word embedding* bahasa Arab dengan mempertimbangkan keseimbangan antara informasi yang dipertahankan dan pengaruh *noise* menggunakan kerangka PIP *Loss*.
3. Memvalidasi dimensi optimal yang diperoleh melalui pendekatan teoretis dan numerik menggunakan evaluasi intrinsik *word similarity*.

1.5 Metodologi Penelitian

Metodologi penelitian yang digunakan pada penelitian ini terdiri atas beberapa tahapan sebagai berikut.

1. **Studi Literatur**

   Pada tahap ini dilakukan kajian terhadap berbagai literatur yang berkaitan dengan *Natural Language Processing* (NLP), *word embedding*, algoritma *Word2Vec*, matriks *Pointwise Mutual Information* (PMI), *Singular Value Decomposition* (SVD), *Pairwise Inner Product* (PIP) *Loss*, Teori Perturbasi Matriks, *Universal Singular Value Thresholding* (USVT), serta simulasi Monte Carlo. Literatur diperoleh dari buku, jurnal ilmiah, prosiding konferensi, dan penelitian terdahulu yang relevan sebagai landasan teoritis penelitian.

2. **Pengumpulan dan Prapemrosesan Korpus**

   Tahap ini diawali dengan pengambilan data Wikipedia bahasa Arab sebagai sumber korpus penelitian. Selanjutnya dilakukan proses *pre-processing* yang meliputi normalisasi karakter Arab, tokenisasi, penghapusan karakter yang tidak diperlukan, serta penyaringan kosakata berdasarkan frekuensi kemunculan. Hasil dari tahap ini berupa korpus akhir yang telah bersih dan siap digunakan pada proses analisis.

3. **Pembentukan Korpus Berdasarkan Domain**

   Korpus akhir hasil prapemrosesan selanjutnya dibagi menjadi tiga kelompok sesuai domain penelitian, yaitu Korpus Utama yang mewakili domain umum, Korpus Geografi, dan Korpus Sejarah. Pembagian ini dilakukan untuk menganalisis pengaruh ukuran korpus dan karakteristik domain terhadap proses penentuan dimensi *word embedding*.

4. **Pembentukan Matriks PMI**

   Setiap korpus digunakan untuk membangun matriks ko-okurensi kata berdasarkan jendela konteks yang telah ditentukan. Selanjutnya matriks ko-okurensi ditransformasikan menjadi matriks *Pointwise Mutual Information* (PMI) sebagai representasi statistik hubungan antar kata yang akan digunakan pada proses dekomposisi matriks.

5. **Estimasi *Noise* dan Penyaringan Nilai Singular**

Matriks PMI didekomposisi menggunakan *Singular Value Decomposition* (SVD). Selanjutnya dilakukan estimasi tingkat *noise* menggunakan metode *Count-Twice Trick*. Nilai estimasi tersebut digunakan sebagai dasar dalam metode *Universal Singular Value Thresholding* (USVT) untuk memperoleh nilai singular yang merepresentasikan komponen sinyal utama beserta jumlah dimensi efektif ($\hat{r}$).

6. **Penentuan Dimensi Terbaik Menggunakan Simulasi Monte Carlo**

   Nilai singular hasil penyaringan USVT digunakan untuk membentuk matriks *oracle*, kemudian diberi *noise* Gaussian secara berulang dan diuraikan menggunakan SVD untuk menghitung rata-rata PIP *Loss* pada setiap kandidat dimensi. Dimensi terbaik (k) ditentukan sebagai dimensi dengan rata-rata PIP *Loss* paling rendah, karena nilai PIP *Loss* yang sebenarnya tidak dapat dihitung secara analitis dalam bentuk tertutup.

7. **Perhitungan Batas Atas PIP *Loss* Berdasarkan Teorema 3**

   Nilai singular yang sama digunakan sebagai input pada Teorema 3 dari [9] untuk menghitung komponen *bias*, *variance magnitude*, dan *variance rotation*, sehingga diperoleh batas atas (*upper bound*) PIP *Loss* secara analitis. Hasil ini berfungsi sebagai landasan teoretis yang memperkuat struktur bias-varians pada simulasi Monte Carlo, bukan sebagai pembanding atau validasi terhadap hasilnya.

8. **Evaluasi *Word Embedding***

   Representasi kata yang diperoleh pada dimensi terbaik kemudian dievaluasi menggunakan pengujian intrinsik *word similarity*. Tingkat kemiripan antar pasangan kata dihitung menggunakan *Cosine Similarity*, kemudian dibandingkan dengan penilaian manusia menggunakan koefisien korelasi Spearman. Hasil evaluasi digunakan untuk mengetahui pengaruh dimensi yang dipilih terhadap kualitas representasi semantik pada masing-masing korpus.

1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini disusun ke dalam lima bab yang saling berkaitan, yaitu sebagai berikut.

BAB I PENDAHULUAN

Pada bab I dibahas mengenai pendahuluan yang merupakan garis besar penulisan skripsi. Bab ini menguraikan latar belakang penelitian, rumusan masalah, batasan masalah, tujuan penelitian, metodologi penelitian, dan sistematika penulisan.

BAB II LANDASAN TEORI

Pada bab II diuraikan landasan teori yang digunakan dalam pembahasan skripsi. Secara umum, bab ini membahas konsep-konsep yang berkaitan dengan pemrosesan bahasa alami dan representasi kata, serta dasar matematis yang digunakan untuk menganalisis dan mengoptimalkan dimensi *word embedding*.

BAB III METODE PENELITIAN

Pada bab III dibahas prosedur penelitian yang meliputi uraian mengenai tahapan, metode, dan model matematis yang digunakan untuk mengestimasi serta menentukan dimensi optimal *word embedding* pada korpus teks bahasa Arab.

BAB IV IMPLEMENTASI DAN SIMULASI

Pada bab IV disajikan hasil uji coba dan analisis dari implementasi metode optimisasi dimensi yang diusulkan. Selain itu, bab ini juga menyajikan evaluasi performa model melalui pengujian kesamaan semantik (*word similarity*) untuk menentukan kualitas representasi pada berbagai domain korpus.

BAB V KESIMPULAN DAN SARAN

Pada bab V diuraikan penutup yang menyajikan kesimpulan berdasarkan hasil analisis dan pengujian metode optimisasi dimensi *word embedding*. Selain itu, bab ini juga memuat saran-saran yang relevan sebagai acuan bagi pengembangan penelitian selanjutnya di bidang pemrosesan bahasa alami.