BAB I

PENDAHULUAN

1.1 Latar Belakang

Ilmu pengetahuan dan teknologi telah berkembang dengan pesat sehingga berdampak besar pada pola hidup masyarakat. Perkembangan teknologi yang terus menerus terjadi telah menyebabkan perubahan yang sangat signifikan pada kehidupan manusia (Wahyudi and Sukmasari, 2014). Tidak dapat dipungkiri bahwa berbagai aktivitas dalam kehidupan sehari-hari menjadi lebih mudah disebabkan oleh teknologi yang semakin berkembang. Tanpa kemajuan teknologi, berbagai aktivitas tidak akan berjalan dengan lebih mudah dan efisien. Perkembangan zaman yang disaksikan dunia saat ini semuanya dimungkinkan karena perkembangan teknologi di berbagai bidang. Sebagai contohnya adalah seseorang dimudahkan oleh *google translate* yang merupakan salah satu media yang dapat menerjamahkan teks dari suatu bahasa ke bahasa lainnya (Alam, 2020). Terdapat *chat bot* yang merupakan program komputer yang diprogram untuk dapat membalas pesan pengguna secara otomatis dan natural (Pratama and Al Irsyadi, 2021). Pada *email* terdapat fitur *email filtering* yang dapat membedakan antara *spam email* dan bukan dengan proses klasifikasi teks yang mengelompokkan teks ke dalam beberapa kategori tertentu (Wibisono, Rizkiono and Wantoro, 2020). Contoh-contoh tersebut merupakan penerapan dari salah satu bidang pada *Artificial Intelligence* (AI) yaitu *Natural Language Processing* (NLP).

Natural language processing adalah salah satu cabang ilmu dari AI yang menggunakan bahasa alami untuk memahami komunikasi antara manusia dan komputer (Nila and Afrianto, 2015). Penerapan NLP adalah salah satu bidang di dalam AI yang membantu mesin untuk memahami bahasa manusia lebih akurat. Dengan menggunakan ilmu linguistik dan ilmu komputer, NLP dapat mengajarkan mesin untuk menganalisis makna dari rangkaian kata-kata. Penerapan NLP memiliki manfaat dalam memfasilitasi komunikasi antara manusia dan komputer untuk mencari informasi, sehingga adanya interaksi antara keduanya menggunakan bahasa alami. Terdapat dua metode yang digunakan dalam NLP untuk menganalisis bahasa manusia, yaitu analisis semantik dan analisis sintaktik. Analisis semantik berfokus menganalisis pada makna per kata, sementara analisis sintaktik berfokus menganalisis pada susunan antara kata-kata dalam suatu kalimat (Bahri, Saputra and Wajhillah, 2017). Dengan NLP, manusia dapat berinteraksi dengan komputer menggunakan bahasa yang dimengerti oleh komputer.

Mesin tidak memahami bahasa yang digunakan oleh manusia, sehingga diperlukan penggunaan bahasa yang dimengerti oleh mesin ketika ingin menggunakannya untuk mempermudah berbagai aktivitas yang ingin dilakukan. Bahasa alami yang digunakan oleh NLP memungkinkan mesin melakukan interaksi dengan manusia, atau bahkan berinteraksi dengan mesin lainnya (Chandra, Kurniawan and Musa, 2020). Salah satu bahasa alami yang dapat dimengerti oleh mesin adalah bahasa pemograman yang di mana salah satu dari bahasa pemgoraman tersebut dapat berupa angka-angka yang termuat dalam sebuah vektor. Teknik NLP

mempelajari bagaimana sebuah kata dapat dikonversikan ke dalam sebuah vektor dengan menggunakan metode *Word Embedding*.

Word Embedding merupakan metode yang ampuh yang banyak digunakan dalam berbagai permasalahan yang ada pada NLP (Wang dkk., 2019), seperti analisis semantik, pencarian informasi, penerjamahan mesin, dan lain-lain. *Word Embedding* adalah salah satu cabang dari NLP yang digunakan untuk mengubah sebuah kata menjadi sebuah *word vector* yang di mana *word vector* ini merepresentasikan sebuah kata di dalam ruang vektor (Nurdin dkk., 2020). Dimensi dari *word vector* tersebut dapat ditentukan secara acak, ketika dimensi dari *word vector* tersebut semakin besar maka proses training dari *word vector* tersebut menjadi lebih baik sehingga dapat merepresentasikan sebuah kata dengan lebih akurat. Dengan vektor inilah suatu mesin dapat memahami bahasa yang dimaksud oleh pengguna. *Word Embedding* dapat menangkap informasi secara semantik dan sintaktik kata dari sebuah korpus (Lai dkk., 2015).

Dalam *Word Embedding* terdapat beberapa metode yang dapat digunakan untuk mendapatkan representasi kata tersebut dalam *word vector*. Secara umum *Word Embedding* dibagi menjadi dua metode (Almeida dan Xexéo, 2019). Metode yang memperkirakan kata yang diinginkan (*focus word*) dari konteks yang ada disebut dengan *Prediction-based Model*. Di sisi lain, metode yang menggunakan frekuensi atau jumlah dari kata yang muncul dalam satu konteks disebut dengan *Count-based Model* (Bollegala, Hayashi and Kawarabayashi, 2017).

Mikolov dkk. (2013) memperkenalkan salah satu cabang dari *Prediction-based Model* yang disebut dengan Word2Vec, yang merupakan salah satu metode

yang populer dalam pembentukan *Word Embedding*. Word2Vec adalah model yang merepresentasikan kata ke dalam vektor yang dapat membawa makna semantik dari kata tersebut (Meyer, 2016). Metode Word2Vec mempelajari *Word Embedding* untuk memperoleh *word vector* dengan cara menghitung probabilitas suatu kata dari kata-kata disekitarnya. Secara umum Word2Vec memiliki 2 model di dalamnya, yaitu Continous Bag-Of-Word (CBOW) dan Skip-Gram. CBOW bekerja dengan menjadikan sebuah kata utama sebagai input untuk menghitung probabilitas dari kata-kata yang ada di sekitar kata utama tersebut. Sementara Skip-gram bekerja dengan menghitung probabilitas kata utama dengan menjadikan kata-kata yang ada di sekitar kata utama tersebut menjadi input.

Namun, terdapat ketidakkonsistenan dalam pengukuran jarak antara kata-kata konteksnya dalam model Skip-gram (Xing dkk., 2015). Telah ditemukan oleh Xing dkk. bahwa adanya ketidakkonsistenan antara fungsi objektif untuk proses *learning word vector*, pengukuran jarak antara *word vector*, serta fungsi objektif untuk transformasi linear dalam menerjemahkan dari satu bahasa ke bahasa yang lainnya. Masalah tersebut dapat diatasi dengan menormalisasikan sekumpulan *word vector* dan mengubah matriks *embedding* menjadi matriks yang bersifat ortogonal. Normalisasi *word vector* bertujuan untuk menempatkan *word vector* tersebut terletak pada *hypersphere* di mana sekumpulan *word vector* tersebut menjadi sebuah *unit vector* sehingga jarak antara *word vector* tersebut menjadi lebih dekat. Xing dkk. (2015) menyebutkan bahwa dengan jarak *word vector* yang lebih berdekatan telah mengatasi ketidakkonsistenan pada pengukuran jarak antara kata-kata konteks tersebut. Oleh karena itu pada penelitian ini bertujuan untuk mengkaji

proses *Word Embedding* dengan model Skip-gram dan normalisasi *word vector* yang mengakibatkan jarak antara *word vector* tersebut menjadi lebih dekat.

1.2 Identifikasi Masalah

Berdasarkan latar belakang yang telah dipaparkan, permasalahan yang dibahas dalam skripsi ini dapat diidentifikasikan sebagai berikut:

1. Bagaimana proses *Word Embedding* dengan model Skip-gram?
2. Bagaimana hasil menormalisasikan *word vector*?
3. Bagaimana jarak dan letak antara *word vector* setelah melalui proses normalisasi?

1.3 Batasan Masalah

Permasalahan yang dikaji hanya terbatas pada beberapa hal:

1. Model Skip-gram yang dikaji pada penelitian ini dibatasi dengan mengkaji *word vector* pada dimensi 3.
2. Tidak mengkaji pembaharuan *weight matrix* pada model Skip-gram.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

1. Mengetahui proses *Word Embedding* dengan model Skip-gram dalam sudut pandang aljabar.
2. Mengetahui hasil dari normalisasi *word vector* dalam model Skip-gram.

3. Mengetahui jarak dan letak antara *word vector* setelah melalui proses normalisasi.

1.5 Kegunaan Penelitian

Dari penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Menjadi referensi terkait kajian tentang *Word Embedding* dalam sudut aljabar.
2. Menjadi referensi terkait solusi dari ketidakkonsistenan yang berada pada model Skip-gram.
3. Memberikan arahan lain untuk penelitian lainnya yang terkait dengan proses *Word Embedding*.

1.6 Metodologi Penelitian

Metodologi yang digunakan dalam penelitian ini dibagi menjadi beberapa bagian :

1. Studi Literatur

   Studi literatur yang dilakukan bersumber pada beberapa jurnal dan artikel dari penelitian-penelitian sebelumnya yang berkaitan dengan proses *Word Embedding* dalam model Skip-gram. Penelitian ini secara khusus mengacu pada paper yang dipublikasikan oleh Xing dkk. (2015), pada penelitian tersebut ditemukan ketidakkonsistenan antara fungsi objektif untuk proses *learning word vector*, pengukuran jarak antara *word vector*, serta fungsi objektif untuk transformasi linear dalam menerjemahkan dari satu bahasa

ke bahasa yang lainnya. Masalah ini dapat diperbaiki dengan menormalisasikan sekumpulan *word vector* tersebut dan membuat matriks *embedding* menjadi matriks yang bersifat ortogonal.

2. Studi Eksperimental

Dalam skripsi ini melakukan proses *Word Embedding* dengan model Skip-gram menormalisasi sekumpulan *word vector*. *Word vector* dinormalisasikan dengan membagi setiap entri pada vektor tersebut dengan panjang dari *word* vector tersebut. Bentuk dari *word vector* yang telah dinormalisasi menjadi *unit vector* yang di mana *word vector* tersebut memiliki panjang 1 satuan.

1.7 Sistematika Penulisan

Sistematika penulisan skrips ini adalah sebagai berikut:

BAB I – PENDAHULUAN

Bab ini menjelaskan latar belakang permasalahan, identifikasi masalah, tujuan penelitian, batasan penelitian, kegunaan penelitian, metodologi penelitian, dan sistematika penulisan.

BAB II – LANDASAN TEORI

Bab ini mencantumkan teori-teori yang relevan untuk digunakan untuk penelitian dalam skripsi ini.

BAB III – OBJEK DAN METODE PENELITIAN

Bab ini berisi tentang objek yang diteliti, metode penelitian yang digunakan, dan alur penelitian yang digunakan dalam skripsi ini.

BAB IV – HASIL DAN PEMBAHASAN

Bab ini membahas hasil penelitian terhadap objek yang diteliti yaitu metode *Word Embedding* dengan model Skip-gram dan normalisasi *word vector* yang diperoleh dari model Skip-gram.

BAB V – SIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari hasil penelitian yang dilakukan dan saran bagi peneliti lain yang akan melakukan penelitian selanjutnya yang berkaitan dengan skripsi ini.