BAB I
PENDAHULUAN

1.1 Latar Belakang

Analisis sentiment merupakan salah satu bidang dari *Natural Language Processing* (NLP) yang digunakan untuk melakukan klasifikasi pada data dalam bentuk teks. Data dalam bentuk teks ini bisa diperoleh dari berbagai macam sumber di internet, seperti twitter, instagram, facebook, google play review, dll. Tujuan dilakukan sentimen analisis yaitu untuk mendapatkan opini dari pengguna platform untuk topik tertentu, atau digunakan sebuah perusahaan untuk melihat bagaimana *feedback* konsumen terhadap produk mereka. Pada proses analisis sentimen terdapat suatu proses untuk mengubah teks menjadi vektor, metode yang paling sering digunakan untuk mengubah teks menjadi vektor adalah *Bag of Words* (BoW). BoW merupakan metode vektorisasi yang paling sederhana, karena metode ini hanya menghitung frekuensi kemunculan kata-kata unik yang terdapat pada suatu kalimat. Metode BoW ini sendiri memiliki kelemahan seperti hilangnya konteks dari kalimat, urutan kata, dan hubungan semantik antar kata[1]. Kelemahan lainnya yaitu ukuran penyimpanan, karena semakin banyak kata untuk yang tersimpan dalam *corpus* artinya semakin besar pula ukuran penyimpanan yang dibutuhkan.

Tahun 2003, muncul metode vektorisasi baru yaitu *word embedding*[2]. Word embedding bekerja dengan cara menandai seluruh kata yang terdapat pada sebuah dokumen kedalam *dense vector*, dimana sebuah vektor merepresentasikan proyeksi kata di dalam ruang vektor[3]. Lalu pada tahun 2013, Tomas Mikolov memperkenalkan *word2vec*. *Word2vec* memiliki dua model arsitektur yang dapat digunakan yaitu *Continuous Bag of Words* (CBOW) dan *skip-gram*[4]. Berikutnya tahun 2017, *facebook* memperkenalkan *fasttext*. *Fasttext* sendiri merupakan pengembangan dari metode *word2vec*. Sama seperti *word2vec*, *fasttext* juga memiliki fitur cbow dan skip-gram. *Fasttext* sendiri mempunyai keunggulan untuk menangani kata-kata yang belum pernah muncul atau juga sering disebut sebagai

out of vocabulary (OOV),yang mana apabila terjadi pada *word2vec* akan menimbulkan *error*[5].

1.2 Rumusan Masalah

Berdasarkan penjelasan dari latar belakang,maka dapat dirumuskan masalah yang nantinya akan menjadi fokus masalah dalam penelitian ini yaitu bagaimana pengaruh *CBOW* dan *Skip-gram* pada model LSTM terhadap tingkat akurasi untuk sentimen analisis berdasarkan ukuran *Word Dimension Vektor*.

1.3 Batasan Masalah

Berdasarkan rumusan masalah yang ada, penelitian ini diberikan batasan-batasan masalah agar penelitian ini tidak keluar dari pokok permasalahan. Adapun batasan masalah tersebut yaitu :

1. Dataset yang digunakan merupakan data open source dari penelitian *Analisis Sentimen Tentang Opini Pilkada Dki 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan NaïveBayes dan Pembobotan Emoji*[6].
2. Data yang digunakan hanya data yang berbahasa indonesia.
3. Data yang digunakan hanya berupa text dan tidak menggandung gambar.
4. Klasifikasi dibagi menjadi menjadi dua kelas sentimen, yaitu positif dan negatif.
5. Metode yang digunakan adalah *Long Short Term Memory* (LSTM).
6. Penelitian ini tidak berfokus pada optimasi (*Hyper Tunning Parameter*) algortima *machine learning*.
7. Perbandingan hanya menggunakan tingkat akurasi sebagai alat ukur.
8. *Word dimension* yang digunakan adalah 50, 100, 150, 200, 300.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk membandingkan *Continuous Bag of Words Model* (CBOW) dan *Skip-gram* pada proses *word embedding fasttext*, dan melihat perbedaan tingkat akurasi yang dihasilkan.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah untuk mengetahui nilai akurasi dari masing-masing model arsitektur *CBOW* dan *skip-gram* pada *word embedding fasttext* dengan model LSTM berdasarkan ukuran *word dimension* vektor.

1.6 Metode Penelitian

Metode yang dilakukan pada penelitian ini agar mendapat hasil yang sesuai dengan tujuan, maka menggunakan beberapa metode untuk memperoleh data dan informasi yang dibutuhkan. Metode yang dilakukan antara lain:

1. Studi Pustaka

   Melakukan studi pustaka guna untuk mencari bahan-bahan referensi baik berupa jurnal, buku, artikel yang dapat mendukung penelitian ini mengenai metode *word embedding*, dan sentimen analisis.

2. Metode Pengolahan Data

   Data yang sudah dikumpulkan akan melalui proses preprocessing sebagai berikut :

   A. *Data cleanning*
   B. *Tokenization*
   C. *Stopword Removal*
   D. *Stemming*

3. Metode Analisis

   Analisa data dilakukan dengan membandingkan tingkat akurasi dari dua model arsitektur *fasttext* yaitu, *CBOW* dan *Skip-gram* pada model LSTM berdasarkan ukuran *word dimension* vektor.

1.7 Sistematika Penulisan

Sistematika penulisan merupakan suatu metode atau urutan dalam mengerjakan penelitian,karya tulis, dsb. Hal ini penting untuk diperhatikan supaya penulisan lebih terstruktur dan memudahkan pembaca untuk memahami apa yang dijabarkan dalam laporan skripsi. Sistematika penulisan skripsi ini adalah sebagai berikut :

BAB I **PENDAHULUAN**

Pada bab ini menjabarkan tentang latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

BAB II **LANDASAN TEORI**

Pada bab dua ini dijelaskan secara rinci teori – teori yang digunakan dalam penulisan skripsi, seperti pemahaman tentang *Natural Language Processing*, sentimen analisis,*preprocessing* dsb.

BAB III **METODOLOGI PENELITIAN**

Pada bab ini menjelaskan metode apa saja yang digunakan selama penelitian, dan menjelaskan proses apa saja yang dilakukan selama penelitian.

BAB IV **HASIL PENELITIAN DAN PEMBAHASAN**

Pada bab ini menjabarkan hasil dari penelitian yang telah diimplementasikan dan pembahasan hasil penelitian.

BAB V **KESIMPULAN DAN SARAN**

Bab ini berisi hasil kesimpulan yang didapat dari hasil penelitian yang telah dilakukan, serta saran yang dapat dijadikan bahan evaluasi.