BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Banyaknya jumlah penyandang disabilitas atau difabel telah menjadi perhatian bagi beberapa orang, organisasi, bahkan negara. Tercatat jumlah penyandang disabilitas di dunia adalah sebanyak satu miliar orang, atau 15% dari populasi semua manusia di dunia (WHO, 2020), sedangkan Indonesia mempunyai jumlah penyandang disabilitas sebanyak sebelas juta orang (ILO, 2017). Dengan angka sebesar sebelas juta, masih banyak penyandang disabilitas di Indonesia yang masih kesulitan untuk mengakses fasilitas dari layanan pendidikan, kesehatan, transportasi, ketenagakerjaan dan pekerjaan, sehingga menyulitkan penyandang disabilitas untuk berpartisipasi secara maksimal (Cameron & Suarez., 2017; Kusumastuti *et al*, 2013; ILO, 2014).

Berbeda dengan keadaan di negara lain, seperti di negara Inggris yang telah mendukung penyandang disabilitas agar dapat kesetaraan akses ke peluang pasar pekerjaan dan akses kesehatan dengan cara mengakses yang telah disesuaikan dengan keadaan penyandang disabilitas (*Great Britain*, 2016). Negara Tiongkok pun juga telah mendukung beberapa akses fasilitas untuk penyandang disabilitas seperti memperpanjang wajib belajar sembilan tahun dari sekolah dasar hingga sekolah menengah pertama serta membebaskan biaya untuk jenjang sekolah menengah atas, pemerintah memberikan pelatihan kerja, rujukan perkerjaan, dan konseling pekerjaan (Tong, 2017). Negara Irlandia juga melakukan tindakan untuk

mengidentifikasi narapidana yang mempunyai keterbatasan intelektual agar dapat diberikan rehabilitas, dan bimbingan agar dapat menjalani hidup yang lebih baik (McNamee & Staunton Dr, 2017).

Menurut ILO (2014) penyebab kurangnya perhatian terhadap penyandang disabilitas di Indonesia adalah kurangnya informasi yang tidak terdokumentasi di media massa. Bentuk perhatian dari beberapa organisasi atau pemerintah, salah satunya berupa pemberitaan tentang penyandang disabilitas, seperti di *United States* dan China sudah mulai membuat penyandang disabilitas menjadi perhatian media (Tang & Bie, 2016). Seperti pada penelitian yang dilakukan oleh McAndrew, *et al* (2020) tentang representasi dari anak-anak dan remaja disabilitas di Irlandia dengan menggunakan data sebanyak 89 surat kabar dalam rentang waktu satu tahun yang berkaitan dengan disabilitas. Hal ini membuktikan sudah banyak negara yang sudah memperhatikan penyandang disabilitas salah satunya melalui media berita.

Berita berfungsi sebagai platform dalam menciptakan kesadaran publik tentang topik-topik disabilitas (Bendukurthi dan Raman, 2016). Menurut ILO (2014), Berita yang mempunyai konten positif dari penyandang disabilitas akan meningkatkan pemahaman masyarakat, serta mengubah pola pikir negatif mengenai persepsi keterbatasan dari penyandang disabilitas. Mendukung banyaknya pemberitaan penyandang disabilitas, Gurmeet dan Karan (2016) mengatakan bahwa dengan banyaknya jumlah berita membuat pembaca menjadi kesulitan untuk mengakses berita yang diminatinya sehingga menjadi sebuah keharusan untuk mengkategorikan berita agar lebih menarik dan mudah dibaca.

Mengkategorikan berita dengan jumlah banyak akan membutuhkan waktu yang cukup lama jika dilakukan oleh tenaga manusia, maka penting untuk memiliki

sistem yang efisien dalam memisahkan berita menjadi beberapa kategori yang berbeda dengan cepat, teknologi yang dapat digunakan adalah pembelajaran mesin (Deb, Jha, Panjiyar, & Gupta, 2020). Salah satu metode pembelajaran mesin adalah Logistic Regression, menurut Al-Tahrawi (2015) Logistic Regression adalah metode yang terkenal memiliki keunggulan menghasilkan model probabilitas untuk kategorisasi proba*list*ik. Logistic Regression sudah banyak diimplementasikan pada bidang sentimen analisis (Pranckevicius & Marcinkevicius, 2016; UmniySalamah, 2018), klasifikasi berita (Li, et al., 2016; Shah, et al., 2020), klasifikasi pengucapan (Pranckevicius & Marcinkevicius, 2017), dan Kategori teks arab (Al-Tahrawi, 2015). Keunggulan *Logistic Regression* dibuktikan dalam penelitian yang dilakukan oleh (Shah, Patel, Sanghvi, & Shah, 2020), mengenai perbandingan akurasi algoritma pembelajaran mesin yaitu Logistic Regression, Random Forest, dan K-Nearest Neighbors (KNN) untuk klasifikasi kategori berita BBC yang memiliki lima kategori berita dan Logistic Regression menjadi metode yang mempunyai hasil akurasi tertinggi sebesar 97%, sedangkan Random Forest sebesar 93%, dan KNN sebesar 92%.

Sebelum metode dari pembelajaran mesin menerima data yang dipelajari, harus masuk tahap pra pemrosesan agar data menjadi optimal, salah satu teknik yang dapat digunakan untuk optimalisasi data adalah metode *word embedding*. Dengan menggunakan metode *word embedding* berguna untuk merepresentasikan angka numerik dari sebuah kata dengan bentuk *vector*, mempelajari kata-kata yang mempunyai kemiripan semantik (Mandelbaum & Shalev, 2016). Salah satu *library* yang dapat menerapkan metode *word embedding* adalah FastText. Keunggulan FastText dapat mempelajari kata yang tidak terdapat di dalam data yang digunakan,

karena FastText juga mempelajari *subword* dari sebuah kata (Bojanowski *et al,* 2017; Faiza *et al*, 2019). Dalam penelitian perbandingan metode *word embedding* yang dilakukan oleh Ibrahim *et al* (2019), membandingkan metode *Word2vec*, FastText, dan Glove untuk sentimen analisis yang mempunyai tiga label yaitu positif, negatif, dan netral yang menyimpulkan bahwa FastText dapat megoptimalisasi sebuah metode pembelajaran mesin sehingga menghasilkan akurasi tertinggi. Pada penelitan yang dilakukan oleh Major, *et al* (2018) yang bertujuan mendemonstrasikan kegunaan dari *pre-trained* model dari *word embedding* untuk klasifikasi artikel berita dan wikipedia, salah satunya membandingkan FastText CBOW dan skip-gram yang menyimpulkan skip-gram mempunyai performa yang lebih tinggi.

Data merupakan faktor yang penting dalam proses pembelajaran mesin, salah satu cara untuk mengoptimalkan data adalah dengan cara data *augmentation* yang merupakan teknik untuk membuat data baru berdasarkan data *training* yang telah ada (Prins, 2019). Salah satu teknik data *augmentation* adalah *back-translation* yang dapat membuat berbagai hasil parafrase dari kalimat aslinya dan meningkatkan hasil performa model pembelajaran mesin (Xie, Dai, Hovy, Luong, & Le, 2019), serta baik dalam memproses teks yang panjang (Gao, 2020). Pada penelitian Ma dan Li (2020) back-translation mampu meningkatkan performa model klasifikasi teks yang menggunakan data *chinese text* yang telah di-*augmentation* menggunakan back-translation.

Berdasarkan hasil penelitian yang telah dijabarkan di atas, maka penelitian ini menggunakan metode Logistic Regression yang dikombinasikan dengan

FastText sebagai model *pre-trained word embedding* untuk mengklasifikasi berita penyandang disabilitas.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan sebelumnya, maka rumusan masalah dalam penelitian ini adalah sebagai berikut.

1. Bagaimana mengimplementasi *word embedding* untuk klasifikasi berita penyandang disabilitas menggunakan Logistic Regression?
2. Bagaimana nilai dari *accuracy, precision, recall,* dan *f1-score* yang dihasilkan oleh metode Logistic Regression dengan *word embedding* untuk klasifikasi berita penyandang disabilitas?

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini dapat dijabarkan menjadi beberapa poin sebagai berikut.

1. *Dataset* yang digunakan adalah berita tentang penyandang disabilitas berbahasa Indonesia.
2. *Dataset* yang digunakan penelitian ini adalah Berita *online* dari *website* difabel.tempo.co, liputan6.com, newsdifabel.com, dan kompas.com.
3. *Dataset* yang digunakan terdiri dari sembilan kategori berita yaitu Hukum, Internasional, *Lifestyle*, Nasional, Olahraga, Pendidikan, Regional, Tekno, dan Tokoh.
4. *Word embedding* yang digunakan adalah FastText.

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut.

1. Mengimplementasikan *word embedding* untuk klasifikasi berita penyandang disabilitas menggunakan Logistic Regression.
2. Mengukur *accuracy, precision, recall* dan F1-Score dari Logistic Regression yang telah menggunakan *word embedding* untuk klasifikasi berita penyandang disabilitas.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah membantu editor berita dalam mengklasifikasi berita difabel dan membantu dewan pers Indonesia mengetahui kategori apa saja yang telah diberitakan tentang penyandang disabilitas di Indonesia.

1.6 Sistematika Penulisan

Sistematika dalam penulisan laporan tentang penelitian Implementasi *Word Embedding* untuk Klasifikasi Berita Penyandang Disabilitas Menggunakan Logistic Regression akan dibagi menjadi 5 bab, yang dijabarkan sebagai berikut.

BAB 1 PENDAHULUAN

Bab satu menjelaskan latar belakang masalah dari penelitian, rumusan masalah, batasan masalah, tujuan penelitian, manfaaat penelitan dan sistematikan penulisan laporan.

BAB 2 LANDASAN TEORI

Bab dua menjelaskan semua landasan teori yang digunakan sebagai ilmu dasar dari penelitian ini, antara lain pengertian dari difabel & disabilitas, *text preprocessing*, *word embedding* dengan FastText,

Logistic Regression, dan cara evaluasi model menggunakan *Confusion Matrix*.

BAB 3 METODOLOGI PENELITIAN

Bab tiga menjelaskan langkah-langkah dari proses penelitian ini, dimulai dari gambaran umum metodologi penelitian, studi literatur, pengumpulan data, data *preprocessing*, pembuatan model pembelajaran mesin, pembuatan aplikasi, membangun aplikasi, penulisan laporan, dan spesifikasi perangkat yang digunakan.

BAB 4 HASIL DAN DISKUSI

Bab empat merupakan bab yang akan menjabarkan cara mengimplementasi kode pemrograman, hasil dan pengujian dari model yang diteliti.

BAB 5 SIMPULAN DAN SARAN

Bab lima merupakan bab dari kesimpulan dari penelitian dan saran untuk penelitian selanjutnya.