BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Kehilangan kepercayaan diri dan tidak dapat untuk hidup mandiri merupakan kesulitan yang dialami oleh penyandang disabilitas dalam kesehariannya. Penyandang disabilitas kerap dipandang sebagai sosok yang inferior oleh masyarakat dan tidak jarang mendapat panggilan seperti ‘aneh’ dan ‘idiot’ karena ketidakmampuannya dalam melakukan kegiatan sehari-hari. Media tentu memiliki kemampuan untuk merubah perlakuan negatif dan meningkatkan *awareness* masyarakat terhadap isu yang dihadapi oleh penyandang disabilitas dengan berbagai cara seperti memperbanyak informasi yang akurat mengenai penyandang disabilitas dan melakukan wawancara terhadap sosok penyandang disabilitas yang sukses (Jadallah and Khasawneh, 2020). Akan tetapi, simpati media massa di Indonesia terhadap para penyandang disabilitas masih rendah (Dewan Pers, 2020).

Di Indonesia, tidak banyak media massa yang dengan khusus memberikan ruang berita bagi para penyandang disabilitas. Menurut International Labour Organization (2014), sejumlah wartawan di Indonesia beralasan bahwa liputan mengenai penyandang disabilitas tidaklah menarik bagi khalayak media massa. Adapun juga survei yang dilakukan oleh Dewan Pers menunjukkan bahwa, kesetaraan akses informasi bagi kelompok rentan dan perlindungan hukum bagi penyandang disabilitas memiliki nilai konsistensi terendah pada kondisi lingkungan

fisik dan politik dan juga kondisi lingkungan hukum (Dewan Pers, 2020). Tidak jarang berita yang menyinggung penyandang disabilitas justru menampilkan penyandang disabilitas sebagai sebuah aib dan membuat seolah-olah kondisi dialami penyandang disabilitas saat ini merupakan sebuah kutukan ataupun kemalangan yang perlu dikasihani (ILO, 2014).

Minimnya liputan mengenai penyandang disabilitas, seolah menyembunyikan ketidaksetaraan yang penyandang disabilitas alami. Informasi mengenai ketidaksetaraan yang terjadi pada penyandang disabilitas tidak dapat terlihat oleh khalayak umum. Tentu hal tersebut dapat memberi gambaran yang salah dan memperluas kesalapahaman terhadap penyandang disabilitas (ILO, 2014). Sementara itu, sulitnya akses informasi mengenai penyandang disabilitas juga dapat mempersulit penyandang disabilitas itu sendiri dalam memperjuangkan hak untuk memperoleh pendidikan, kesempatan kerja, kesehatan, dan lainnya (Dewan Pers, 2020). Oleh karena itu, diperlukannya informasi yang akurat mengenai seluruh kelompok masyarakat dari pihak media massa (ILO, 2014).

Media massa sebagai pihak penyedia informasi pun merasa kesulitan dalam melakukan liputan mengenai penyandang disabilitas. Membuat liputan mengenai penyandang disabilitas dianggap kompleks dan membutuhkan kepekaan serta analisa yang mendalam (ILO, 2014). Jurnalis di seluruh dunia mengalami permasalahan yang sama yaitu *time pressure*. Jurnalis dituntut untuk bisa memproduksi berita secepat mungkin sehingga mempersulit para jurnalis dalam melakukan riset dan analisis yang mendalam (Fernandes and Jorge, 2017; Hawksworth *et al.*, 2017; Himma-Kadakas, 2017; Hovden and Kristensen, 2018;

Perreault and Stanfield, 2019). Oleh karena itu, diperlukan sebuah cara untuk membantu para jurnalis dalam melakukan riset dengan lebih cepat. Salah satu cara yang bisa dilakukan dalam membantu mempercepat riset jurnalis adalah dengan membuat ringkasan dari berita yang akan digunakan sebagai bahan riset.

Ringkasan berita dapat membantu jurnalis dan juga masyarakat dalam menerima informasi dengan lebih cepat karena tidak perlu membaca teks berita secara menyeluruh melainkan cukup membaca hasil ringkasan yang telah dibuat (Ashari and Riasetiawan, 2017; Musyaffanto, Budi Herwanto and Riasetiawan, 2019; Ullah and Al Islam, 2019; Upasani *et al.*, 2020; El-Kassas *et al.*, 2021). Membuat ringkasan dari sebuah berita secara manual membutuhkan banyak sekali waktu dan usaha (Jain, Borah and Biswas, 2020a; El-Kassas *et al.*, 2021). Hal ini diakibatkan karena seseorang perlu membaca teks tersebut secara menyeluruh dan memahaminya dan setelah itu baru dapat menulis kesimpulan dari teks tersebut (Allahyari *et al.*, 2017). Oleh karena itu, diperlukan sebuah peringkas teks yang dapat bekerja secara otomatis (*automatic text summarization*).

TextRank merupakan salah satu metode yang paling populer dan sering digunakan menjadi standar dalam peringkasan teks otomatis (Ramón-Hernández *et al.*, 2020). Metode TextRank menggunakan pendekatan secara abstraktif, bekerja secara *unsupervised* dan menggunakan *graph-based model* dalam melakukan peringkasan teks (Barrios *et al.*, 2016; Ashari and Riasetiawan, 2017; Gunawan, Harahap and Fadillah Rahmat, 2019; Jain, Borah and Biswas, 2020a; Marekšuppa and Adamec, 2020; Shah *et al.*, 2020). Permasalahan pada metode TextRank dan algoritma berbasis *graph-based model* lainnya yaitu terdapat kalimat *redundant*

pada hasil akhir teks ringkasan (Khan *et al.*, 2018; Gunawan, Harahap and Fadillah Rahmat, 2019; Ullah and Al Islam, 2019; El-Kassas *et al.*, 2021). Pengimplementasian metode Maximum Marginal Relevance (MMR) dapat menyelesaikan permasalahan tersebut (Sahba *et al.*, 2018; Gunawan, Harahap and Fadillah Rahmat, 2019; Musyaffanto, Budi Herwanto and Riasetiawan, 2019; Ullah and Al Islam, 2019; Vijaya and Reddy, 2019; Lebanoff, Song and Liu, 2020).

Maximum Marginal Relevance (MMR) merupakan salah satu metode peringkas teks dengan pendekatan ekstraktif yang paling sukses dan *straightforward* (Lebanoff, Song and Liu, 2020). Tujuan utama penggunaan metode Maximum Marginal Relevance (MMR) adalah untuk mengurangi penggunaan kalimat yang memiliki makna serupa atau dianggap *redundant* pada calon hasil teks ringkasan (Khan *et al.*, 2018; Gunawan, Harahap and Fadillah Rahmat, 2019; Musyaffanto, Budi Herwanto and Riasetiawan, 2019; Ullah and Al Islam, 2019; Mao *et al.*, 2020). Hasil ringkasan dari metode Maximum Marginal Relevance (MMR) juga dianggap mampu mendekati hasil ringkasan manusia (Musyaffanto, Budi Herwanto and Riasetiawan, 2019).

Metode Textrank maupun metode Maximum Marginal Relevance (MMR) membutuhkan bentuk *vector* dari berita yang ingin diringkas (Steinberger and Ježek, 2009; Patil *et al.*, 2014; Barrios *et al.*, 2016; Gunawan *et al.*, 2016; Ashari and Riasetiawan, 2017; Neha *et al.*, 2018; Xiong *et al.*, 2018; Kurniawan and Louvan, 2018; Akhtar, Beg and Javed, 2019; Li *et al.*, 2019; Musyaffanto, Budi Herwanto and Riasetiawan, 2019; Ullah and Al Islam, 2019; Vijaya and Reddy, 2019; Gunawan, Harahap and Fadillah Rahmat, 2019; Mao *et al.*, 2020;

Marekšuppa and Adamec, 2020; Ramón-Hernández *et al.*, 2020; Shah *et al.*, 2020; Upasani *et al.*, 2020; Bordoloi *et al.*, 2020; Jain, Borah and Biswas, 2020b; Lebanoff, Song and Liu, 2020; El-Kassas *et al.*, 2021). Terdapat berbagai macam cara dalam merepresentasikan dokumen ke dalam bentuk vektor dan penelitian ini menguji 3 metode berbeda yaitu One-hot Encoding, Term Frequency – Inverse Document Frequency (TF-IDF) dan juga FastText.

One-hot Encoding merupakan salah satu cara merepresentasikan kata dalam bentuk *vector* dan bahkan merupakan metode *encoding* yang terbilang simpel dan banyak digunakan (Potdar, S. and D., 2017; Cerda, Varoquaux and Kégl, 2018). Jain (2020b) melakukan penelitian yang berjudul Fine-Tuning Textrank for Legal Document Summarization: A Bayesian Optimization Based Approach juga menggunakan *One-hot vector based TextRank* sebagai *baseline model*nya. Shengli (2019) melakukan penelitian dengan judul Abstractive text summarization using LSTM-CNN based deep learning dan mengimplementasikan metode one-hot encoding dalam penelitiannya.

TF-IDF merupakan salah satu metode yang digunakan dalam merepresentasikan dokumen ke dalam *multi-dimensional vector* dan termasuk metode *term-weighting* yang paling banyak digunakan (Beel, Joeran and Langer, Stefan and Gipp, 2017; Irfan *et al.*, 2018; Khan, Qian and Naeem, 2019; Marwah and Beel, 2020; Marcińczuk Michałand Gniewkowski, Walkowiak and Kedkowski, 2021). Khan (2019) melakukan penelitian dengan judul Extractive based Text Summarization Using K- Means and TF-IDF dalam mengimplementasikan metode TF-IDF dalam merepresentasikan dokumennya ke dalam bentuk vektor. Irfan

(2018) dalam penelitiannya yang berjudul Implementation of Fuzzy C-Means Algorithm and TF-IDF on English Journal Summary juga mengimplementasikan metode TF-IDF dan bahkan mengatakan bahwa metode TF-IDF merupakan inti dari *preprocessing text* yang dilakukan juga metode yang simpel dan efisien untuk diimplementasikan.

FastText mampu merepresentasikan kata yang tidak diketahui ke dalam bentuk *vector* karena FastText mempertimbangkan suku kata (*subwords*) dari sebuah kata sedangkan model lain biasanya mengabaikan atau memberikan *random vector* (Bojanowski *et al.*, 2017; Ruseti, 2019). Terdapat 2 model FastText dalam komputasi *word representation* yaitu skipgram dan CBOW (continuous-bag-of-words). Skipgram mampu belajar dari *training* data yang terbilang kecil dan mampu merepresentasikan *rare words and phrases* secara akurat. CBOW memiliki waktu *training* yang lebih singkat dan dapat merepresentasikan kata yang sering muncul secara akurat (Giatsoglou *et al.*, 2017; Naili, Chaibi and Ben Ghezala, 2017).

Bansal (2018) melakukan penelitian dengan judul Sentiment classification of online consumer reviews using word vector representations dan mengkombinasikan model skipgram dan CBOW pada algoritma SVM, Naïve Bayes, Logistic Regression dan Random Forest dan menyimpulkan bahwa kombinasi model CBOW dan Random Forest mampu mendapatkan akurasi tertinggi. Zhang (2020) dalam penelitiannya yang berjudul Neural Latent Extractive Document Summarization mengimplementasikan *pre-trained* FastText *vectors* pada seluruh model yang diuji.

Aayush Shah (2020) melakukan penelitian dengan judul Hindi History Note Generation with Unsupervised Extractive Summarization dan membandingkan beberapa metode *unsupervised* yang memiliki pendekatan ekstraktif. Metode yang dibandingkan adalah metode TextRank, LexRank, Luhn, dan KLSum. Penelitian tersebut menunjukkan performa metode TextRank merupakan yang terbaik di antara metode lainnya.

Dani Gunawan (2019) melakukan penelitian dengan judul Multi-document Summarization by using TextRank and Maximal Marginal Relevance for Text in Bahasa Indonesia dan menyatakan bahwa sebenarnya hasil teks ringkasan dari metode TextRank sudah baik tetapi masih dapat ditemukan beberapa kalimat dengan makna yang serupa atau *redundant*. Oleh karena itu, dilakukan pemrosesan kembali menggunakan metode Maximal Marginal Relevance (MMR) dan hasil akhir teks ringkasan tidak lagi menunjukkan kalimat dengan makna serupa atau *redundant*. Penelitian yang dilakukan oleh Dani Gunawan (2019), melakukan perhitungan *similarity* antar kalimat berdasarkan perbandingan *unit overlap* dengan jumlah kalimat pada kedua kalimat sedangkan penelitian ini menggunakan perhitungan *cosine similarity* terhadap bentuk *vector* kalimat dan juga menguji 5 (lima) buah metode dalam merepresentasikan kata ke dalam bentuk *vector*.

Marekšuppa (2020) melakukan penelitian dengan judul A Summarization Dataset of Slovak News Articles dan menyatakan bahwa metrik evaluasi otomatis yang paling banyak digunakan adalah Recall-Oriented Understudy for Gisting Evaluation (ROUGE) Score namun kekurangan utama dari ROUGE Score adalah hanya dapat digunakan pada Bahasa Inggris. Rike Adelia (2019) melakukan

penelitian dengan judul Indonesian Abstractive Text Summarization Using Bidirectional Gated Recurrent Unit dan menyadari bahwa perbedaan kaidah Bahasa Indonesia dan Bahasa inggris dapat menyebabkan rendahnya nilai ROUGE Score dimana penggunaan kata yang berbeda dalam sumber dan ringkasan berita tetapi memiliki makna yang sama. Marekšuppa memberi contoh dimana dalam artikel berita sebelum diringkas menggunakan kata "perkembangan" untuk mendeskripsikan "*growth*" sedangkan pada ringkasan menggunakan kata "pertumbuhan". Dani Gunawan (2016), Rendra Budi (2017), Periantu Marhendri (2018), Reinert Yoshua (2019), dan Dessyanto Boedi (2020) melakukan penelitian mengenai peringkas teks otomatis dengan menggunakan teks Bahasa Indonesia dan menggunakan nilai F1-score, precision, dan recall sebagai alternatif ROGUE Score pada tahap evaluasi.

Steinberg (2009) dalam artikelnya yang berjudul Evaluation measures for text summarization membahas evaluasi dengan *content-based measures*. Evaluasi menggunakan *content-based measures* memiliki kelebihan dibandingkan dengan *Co-selection measures* (*F1-score, precision, recall*) yaitu *content-based measures* tetap dapat mengevaluasi hasil ringkasan teks meskipun kalimat dalam ringkasan manusia merupakan hasil parafrase atau tidak sama persis dengan kalimat dalam artikel berita yang asli. Steinberg memperkenalkan perhitungan nilai *cosine similarity* antara hasil ringkasan mesin dan ringkasan manusia sebagai bagian dari evaluasi dengan *content-based measures.*

Berdasarkan latar belakang di atas, dilakukan penelitian terhadap implementasi metode TextRank dan Maximum Marginal Relevance (MMR) pada peringkas berita difabel otomatis.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang sudah diuraikan di atas, rumusan masalah yang dipilih adalah sebagai berikut.

1. Bagaimana cara mengimplementasikan metode *TextRank* dan *Maximum Marginal Relevance* (MMR) dalam meringkas artikel berita mengenai penyandang disabilitas?
2. Berapa nilai *F1-score, precision, recall,* dan *cosine similarity* dari implementasi metode *TextRank* dan *Maximum Marginal Relevance* (MMR) dalam meringkas artikel berita mengenai penyandang disabilitas?

1.3 Batasan Masalah

Agar pembahasan yang dilakukan tidak menyimpang dari latar belakang dan rumusan masalah, maka batasan masalah yang diambil dalam penelitian ini adalah sebagai berikut.

1. Terdapat 2 (dua) *dataset* yang digunakan yaitu *dataset* IndoSum dan *dataset* yang berisi berita *online* dari *website* difabel.tempo.co, liputan6.com, newsdifabel.com, kompas.com, cnnindonesia.com, kumparan.com, merdeka.com, juara.bolasport.com dan antaranews.com.
2. Pada *dataset* hasil *scraping* dilakukan peringkasan berita oleh seorang pakar.

3. Pakar membuat teks ringkasan dengan cara memilih paling sedikit 5 (lima) kalimat terpenting untuk diikutsertakan dalam ringkasan dan kalimat yang dipilih tidak diparafrase.
4. Peringkasan berita dilakukan hingga jumlah kalimat dalam ringkasan sama dengan jumlah kalimat dalam ringkasan pakar.
5. Data yang digunakan adalah berita yang melakukan pembahasan mengenai kaum disabilitas.
6. Data yang digunakan merupakan berita dengan Bahasa Indonesia.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut.

1. Mengimplementasikan metode *TextRank* dan *Maximum Marginal Relevance* (MMR) dalam meringkas artikel berita mengenai penyandang disabilitas.
2. Mengetahui nilai F1-*score*, *precision*, *recall,* dan *cosine similarity* dari hasil implementasi metode *TextRank* dan *Maximum Marginal Relevance* dalam meringkas artikel berita mengenai penyandang disabilitas.

1.5 Manfaat Penelitian

Manfaat dari dilakukannya penelitian ini adalah untuk menghemat waktu para pembaca berita dalam memahami isi berita melalui teks ringkasan sehingga informasi mengenai penyandang disabilitas dapat diketahui oleh masyarakat luas dan juga dilakukannya penelitian lain yang mengangkat topik mengenai penyandang disabilitas.

1.6 Sistematika Penulisan

Sistematika penulisan laporan skripsi ini terdiri dari 5 bab utama, yang dijabarkan sebagai berikut.

1. BAB 1 PENDAHULUAN

   Bab satu terdiri dari latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan laporan.

2. BAB 2 LANDASAN TEORI

   Bab dua terdiri dari definisi penyandang disabilitas, definisi dan konsep dasar peringkas teks otomatis, konsep dasar *one-hot encoding*, konsep dasar *term frequency - inverse document frequency* (TF-IDF), konsep dasar FastText, konsep dasar metode TextRank, konsep dasar meode maximal marginal relevance, konsep dasar *confusion matrix*.

3. BAB 3 METODOLOGI PENELITIAN

   Bab tiga membahas tahapan-tahapan pada metode yang dilakukan dalam penilitan. Bab ketiga juga berisi diagram alur dari setiap tahapan yang dilalui.

4. BAB 4 HASIL DAN DISKUSI

   Bab empat terdiri dari implementasi kode, skenario pengujian yang dilakukan, hasil dari pengujian yang dilakukan, serta evaluasi terhadap hasil yang diperoleh.

5. BAB 5 SIMPULAN DAN SARAN

   Bab lima terdiri dari simpulan yang dapat diambil dari hasil penelitian dan saran untuk penelitian yang dapat dilakukan selanjutnya.