BAB I PENDAHULUAN

I.1 Latar Belakang

Terciptanya internet, jejaring sosial, forum, dan teknologi informasi yang tersebar secara cepat, menyebabkan interaksi terhadap informasi semakin sulit untuk dipahami, dibuat, dikembangkan, dan disimpan. Dengan besarnya akan kebutuhan informasi dan pesatnya perkembangan internet menyebabkan dorongan pertumbuhan situs media *online* di Indonesia. Dengan luasnya informasi sehingga hampir tidak mungkin untuk seorang pun untuk memproses dan meringkas semua data informasi yang tersedia. Konsumen tidak tertarik membaca teks yang Panjang dan oleh karena itu, konsumen sering melewatkan bagian penting dari informasi tersebut (Bhargava & Sharma, 2020; Mutlu et al., 2020).

Hutama et al. (2017) menyatakan bahwa membaca sebuah artikel berita merupakan salah satu cara termudah untuk mendapatkan informasi yang di mana setiap harinya media berita *online*, seperti Kompas, Detik, Kumparan, CNN Indonesia, IDN Times, dan berbagai media berita *online* lain mampu mempublikasikan berita hingga lebih dari 2000 artikel di Indonesia. Sebagian besar berita berbentuk cerita pendek dengan teks yang tidak terlalu panjang. Namun ada juga berita yang memiliki pembahasan yang panjang dan berbelit-belit.

Membaca merupakan salah satu kegiatan yang tidak dapat terpisahkan bagi manusia, baik membaca buku, majalah, maupun artikel berita. Tetapi, permasalahan akan muncul ketika sebuah teks atau artikel yang akan dibaca memiliki isi yang banyak dan panjang karena membutuhkan waktu yang cukup lama untuk dapat memahami isi dari teks tersebut (Savanti Widya Gotami et al., 2018). Hal ini menjadi sebuah tantangan di tengah rendahnya tingkat literasi di Indonesia.

Indonesia memiliki literasi yang sangat rendah dari negara lain. Berdasarkan hasil *survey* yang dilakukan oleh *Program for International Student Assessment* (PISA) yang di rilis oleh *Organization for Economic Co-operation and Development* (OECD) pada tahun 2019, Indonesia berada di peringkat 62 dari 70 negara, yang berarti Indonesia berada di 10 peringkat terbawah (Ilham, 2022). *United Nations*

Educational, Scientific and Cultural Organization (UNESCO) menyebutkan bahwa Indonesia memiliki indeks minat membaca yang sangat rendah, yaitu hanya 0,001% yang berarti dari 1000 orang Indonesia, hanya 1 orang saja yang rajin membaca (Devega, 2017).

Beberapa faktor yang menyebabkan rendahnya literasi di Indonesia adalah belum membiasakan membaca buku dari rumah, perkembangan teknologi yang semakin pesat, sarana membaca yang minim, kurangnya motivasi untuk membaca, dan malas untuk mengembangkan gagasan (Jessica, 2017). Mengingat jadwal yang padat dan banyaknya informasi yang tersedia, terdapat peningkatan kebutuhan akan ringkasan dari artikel berita (Sethi et al., 2018).

Automatic text summarization adalah salah satu alternatif teknologi yang bisa digunakan untuk menyelesaikan masalah di atas (Mubarok, 2021). *Automatic text summarization* merupakan cabang dari ilmu *Natural Language Processing* (NLP) yang bertujuan untuk merepresentasikan dokumen teks panjang yang dikompresi sehingga informasi yang didapatkan dengan cepat dimengerti dan dapat dibaca oleh pengguna (Joshi et al., 2019). Menurut Saziyabegum & Sajja (2016), *text summarization* bertujuan untuk mengurangi teks sebuah sumber menjadi versi lebih ringkas yang akan mempertahankan konten dan makna umum. Salah satu keuntungan dari *text summarization* adalah meminimalkan waktu dan upaya dalam membaca.

Text summarization dapat dikategorikan menjadi 2 jenis yaitu *single document summarization* yang dimana meringkas satu dokumen secara keseluruhan dan *multi document summarization* yang dimana menggunakan beberapa dokumen dengan subjek yang sama secara bersamaan dan digunakan untuk proses tinjauan literatur karya ilmiah untuk menerima informasi ringkas tentang suatu subjek yang mengurangi redundansi (Akhmetov et al., 2021). Lalu *text summarization* juga dikelompokkan menjadi *query-based text summarization,* yaitu peringkasan yang menyediakan informasi penting dan relevan berdasarkan permintaan dari *user* dan *generic text summarization*, peringkasan yang berisi semua informasi penting dari sumber dokumen (Annisa & Khodra, 2017).

Text Summarization terdiri dari dua metode, yaitu *abstractive text summarization* dan *extractive text summarization*. *Abstractive text summarization* memparafrasakan teks secara keseluruhan sehingga ringkasan tersebut memiliki kosa kata yang bervariasi. Sedangkan, *extractive text summarization* melibatkan pemilihan kalimat penting dari naskah asli dan menggabungkannya menjadi kalimat yang lebih pendek tanpa kehilangan informasi penting (Adelia et al., 2019; Moratanch & Chitrakala, 2017a; Singh et al., 2019). Kategori *text summarization* bisa dilihat pada Tabel I.1.

Tabel I.1 Tabel *Text Summarization*

Text Summarization		
Input	*Output*	*Purpose*
Single Document	*Abstractive*	*Generic*
Multi Document	*Extractive*	*Query Based*

Metode *extractive text summarization* diklasifikasikan menjadi dua, yaitu *Unsupervised Learning* dan *Supervised Learning*. *Unsupervised Learning* memiliki metode pendekatan seperti *graph-based approach*, *fuzzy logic-based approach*, *concept-based approach*, dan *Latent Semantic Analysis Method*, Sedangkan *Supervised Learning* memiliki pendekatan seperti *Machine Learning Approach based on Bayes Rule*, *Neural Network based Approach*, dan *Conditional Random Fields*. (Moratanch & Chitrakala, 2017a)

Sudah ada beberapa penelitian tentang penerapan *text summarization* dengan berbagai algoritma untuk Bahasa Indonesia. Sebagai contoh, Savanti Widya Gotami et al. (2018) menerapkan *text summarization* terhadap artikel berita kesehatan Indonesia dengan menggunakan metode *Latent Semantic Analysis* (LSA). Lalu, ada Musyaffanto et al. (2019) menerapkan *text summarization* dengan menggunakan metode *maximal marginal relevance and non-negative matrix factorization* terhadap sebuah artikel berita Indonesia.

Setyawan et al. (2021) melakukan penelitian dengan menerapkan *text summarization* berdasarkan pendekatan statistika pada dokumen berbahasa Indonesia dan Saputra (2021) menerapkan *text summarization* terhadap artikel berita berbahasa Indonesia secara abstraktif dengan menggunakan metode *Long Short-Term Memory* (LSTM). Ada juga Mubarok (2021) menerapkan *text*

summarization secara abstraktif terhadap artikel berita indonesia berbasis *deep learning* dengan menggunakan metode *Sequence to Sequence Long Short-Term Memory*.

Sudah banyak penelitian tentang *automatic text summarization* berbasis bahasa Indonesia terhadap sebuah artikel berita dan dokumen secara ekstraktif maupun abstraktif, tetapi beberapa peneliti menggunakan *dataset* yang dikumpulkan sendiri dan tidak dipublikasikan sehingga tidak mudah untuk digunakan sebagai *benchmark text summarization* berbahasa Indonesia. Dari permasalahan di atas, peneliti ingin menyelidiki *extractive text summarization* berbahasa Indonesia berbasis *machine learning*. *Dataset* yang akan digunakan dalam penelitian ini adalah Indosum yang dibuat oleh Kurniawan & Louvan (2019) yang tersedia secara publik. Dengan ini, peneliti mengambil judul: "*Extractive Text Summarization* Terhadap Artikel Berita Indonesia Berbasis *Machine Learning*".

I.2 Perumusan Masalah

Rumusan masalah yang mendasari penelitian ini adalah:

1. Bagaimana akurasi *extractive text summarization* terhadap sebuah artikel berita Indonesia dengan algoritma *machine learning*?
2. Bagaimana hasil *extractive text summarization* terhadap artikel berita Indonesia dengan algoritma *machine learning*?

I.3 Tujuan Penelitian

Penelitian ini bertujuan untuk:

1. Mengetahui cara kerja *extractive text summarization* dengan algoritma *machine learning*.
2. Mengetahui hasil evaluasi kualitas *extractive text summarization* dengan algoritma *machine learning* terhadap artikel berita Indonesia.

I.4 Batasan Penelitian

1. Data yang digunakan dalam penelitian ini adalah Indosum yang dibuat oleh Kurniawan & Louvan (2019) yang tersedia secara publik.
2. Metode yang digunakan dalam penelitian ini adalah *Word2Vec*.

I.5 Manfaat Penelitian

Manfaat penelitian ini:

1. Mengetahui hasil kualitas ringkasan yang dihasilkan.

2. Memahami penerapan *extractive text summarization* berbasis *machine learning*.