

BAB I PENDAHULUAN

1.1 Latar Belakang

Sejumlah informasi meningkat secara eksponensial dan manusia harus berurusan dengan ratusan bahkan ribuan informasi di internet [1]. Dengan banyaknya informasi yang tersedia di internet, membuat manusia kesulitan untuk mengikuti informasi tersebut. Maka dari itu, seinformasi dari dokumen teks perlu diringkas secara efektif agar bermanfaat. Ketersediaan dokumen yang meningkat ini menuntut penelitian mendalam di bidang NLP untuk peringkasan teks otomatis. Maka dari itu diperlukan adanya sistem peringkasan dari informasi yang tersedia untuk membantu manusia dengan cepat memahami informasi. Dengan adanya ATS (*Automatic Text Summarization*) untuk memenuhi jumlah indeks dokumen yang berupa ringkasan otomatis dapat membuat kalimat menjadi pendek dan dapat dipahami manusia.

ATS telah menjadi perhatian khusus dalam beberapa tahun terakhir karena semakin banyaknya informasi yang tersedia di internet dan media sosial, yang membuat sulit bagi pengguna untuk membaca semua informasi yang tersedia secara keseluruhan [2]. Oleh karena itu, peringkasan teks otomatis menjadi sangat penting untuk membantu pengguna dalam mencari informasi yang relevan dan menghemat waktu dengan menyediakan ringkasan yang tepat dari teks yang panjang dan kompleks. Selain itu, peringkasan teks otomatis juga bisa bermanfaat dalam berbagai bidang, seperti penelitian, jurnalistik, dan bisnis, untuk membantu mengelompokkan dan mengolah informasi yang tersedia secara efisien. Automatic text summarization juga bisa digunakan dalam aplikasi-aplikasi lain, seperti pembuatan dokumen otomatis, pembuatan laporan otomatis, dan pembuatan presentasi otomatis. Walaupun *automatic text summarization* telah menjadi bidang yang penting dalam pemrosesan teks otomatis, masih terdapat beberapa tantangan yang harus dihadapi dalam pengembangan sistem peringkasan teks otomatis, seperti memahami konteks dan makna kata secara akurat, mengelompokkan informasi yang relevan, dan menghasilkan ringkasan yang tepat sesuai dengan tujuan dan

konteks penggunaan teks. Oleh karena itu, masih terdapat banyak ruang untuk mengembangkan teknik-teknik baru dalam peringkasan teks otomatis untuk meningkatkan kinerja sistem dan membantu pengguna mencari informasi yang relevan dengan lebih efisien. Dalam peringkasan teks otomatis, terdapat dua jenis utama yaitu peringkasan teks otomatis yang terstruktur (*extractive text summarization*) dan peringkasan teks otomatis yang tidak terstruktur (*abstractive text summarization*) [3].

Peringkasan teks otomatis yang terstruktur (*extractive text summarization*) merupakan teknik yang mengelompokkan informasi yang ada dalam teks asli ke dalam ringkasan dengan mengambil kalimat-kalimat yang dianggap penting atau relevan dari teks asli [4]. Teknik ini cenderung lebih mudah untuk dilakukan dibandingkan dengan peringkasan teks otomatis yang tidak terstruktur, namun hasil yang dihasilkan cenderung kurang terstruktur dan tidak selalu menyampaikan makna yang sama dengan teks asli. Sedangkan peringkasan teks otomatis yang tidak terstruktur (*abstractive text summarization*) merupakan teknik yang menghasilkan ringkasan dengan cara menggabungkan dan menyusun kembali informasi yang ada dalam teks asli menjadi kalimat-kalimat yang lebih pendek dan ringkas [5]. Teknik ini cenderung lebih sulit untuk dilakukan dibandingkan dengan peringkasan teks otomatis yang terstruktur, namun hasil yang dihasilkan cenderung lebih terstruktur dan lebih mampu menyampaikan makna yang sama dengan teks asli. Kedua jenis peringkasan teks otomatis tersebut memiliki kelebihan dan kekurangan masing-masing, dan pemilihan teknik yang tepat tergantung pada tujuan dan konteks penggunaan teks yang akan diperingkaskan. Baik peringkasan teks otomatis yang terstruktur maupun yang tidak terstruktur dapat menggunakan teknik-teknik seperti stemming dan stopwords untuk meningkatkan kinerja sistem, tergantung pada tujuan dan konteks penggunaan teks tersebut.

Preprocessing merupakan tahap yang sangat penting untuk ATS karena dengan adanya tahap pengolahan di awal data karena data mentah sangat rentan terhadap masalah seperti noise, missing value, dan inkonsistensi yang dapat menurunkan akurasi hasil akhir. Proses ini meliputi pemilihan dokumen teks agar

lebih terstruktur dan mudah dibaca oleh mesin. Oleh karena itu, penelitian ini akan fokus pada penerapan teknik preprocessing seperti *data cleaning*, *stemming*, *case folding*, dan *stopwords*. Sehingga hasil dari pengaruh preprocessing terhadap kinerja *automatic text summarization* dapat diidentifikasi.

1.2 Rumusan Masalah

Berdasar uraian latar belakang diatas, maka dapat dirumuskan beberapa masalah berikut ini:

1. Apakah tahap *preprocessing* berpengaruh terhadap hasil akhir ringkasan?
2. Teknik kombinasi *preprocessing* apa yang sangat berpengaruh terhadap hasil akhir ringkasan?

1.3 Batasan Masalah

Berdasarkan rumusan masalah, maka dibutuhkan sebuah batasan masalah untuk membatasi ruang lingkup pembahasan yang akan dilakukan. Ruang lingkup yang akan dibahas yaitu:

1. Dataset yang digunakan pada penelitian yaitu INDOSUM.
2. Pada dataset yang digunakan, terdapat 7 kolom yaitu *category*, *gold_label*, *paragraps*, *source_url*, dan *summary*. Kolom yang akan diproses yaitu *paragraps*, dan *summary*.
3. *Pre-trained* model yang digunakan yaitu IndoBERT.
4. Tahap *preprocessing* *stemming* dan *stopwords* menggunakan Sastrawi.
5. Metode yang digunakan untuk mengevaluasi hasil adalah ROUGE.

1.4 Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini:

1. Mengetahui pengaruh tahap *preprocessing* terhadap peringkasan teks otomatis.
2. Mengetahui teknik *preprocessing* yang berpengaruh terhadap kualitas hasil ringkasan.

1.5 Manfaat Penelitian

Adapun manfaat dari penelitian ini antara lain:

1. Mengetahui penerapan, kinerja dan tingkat akurasi dari *pre-trained*

IndoBERT dalam peringkasan teks otomatis.

2. Menjadi sumber referensi dalam melakukan penelitian lanjutan khususnya pada peringkasan teks otomatis dan penerapan model BERT.

1.6 Sistematika Penulisan

BAB I PENDAHULUAN, Bab ini membahas uraian mengenai latar belakang, rumusan masalah, batasan masalah, tujuan, manfaat dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA, Bab ini menguraikan penjelasan tentang landasan teori yang berkaitan dengan permasalahan yang diangkat dalam penelitian ini yang memuat berbagai konsep teoritis yang digunakan sebagai kerangka atau landasan untuk mendukung pemahaman dalam penelitian yang dilakukan.

BAB III METODE PENELITIAN, Bab ini menjelaskan metode penelitian, rancangan system, dan metode pengujian yang dilakukan dalam penelitian.

BAB IV HASIL DAN PEMBAHASAN, Bab ini menjelaskan hasil dan pembahasan dari pengujian yang telah dilakukan.

BAB V PENUTUP, berisi kesimpulan dan saran yang dapat peneliti rangkum selama proses penelitian.