1. PENDAHULUAN

1.1 Latar Belakang Permasalahan

Di era modern ini, informasi merupakan bagian yang penting dalam kehidupan sehari-hari. Dalam mendapatkan informasi terdapat beberapa hal yang dapat dilakukan dimana salah satunya adalah dengan membaca. Suatu informasi dapat diperoleh dengan berbagai cara seperti mengakses *web*. Akan tetapi, semakin banyaknya informasi yang ada di internet membuat manusia kerepotan untuk terus mengikuti perkembangannya (El-Kassas et al., 2020). Maka dari itu, diperlukan suatu pembuatan ringkasan terhadap informasi-informasi yang tersedia secara *online*.

Ringkasan merupakan suatu teks singkat yang dihasilkan dari kumpulan teks panjang namun tetap menyimpan informasi penting dari teks asalnya (Joshi et al., 2019). Dengan bantuan ringkasan akan membantu dalam membaca dengan waktu yang lebih cepat dibandingkan dengan harus membaca informasi secara utuh. Selain itu, ringkasan juga membantu dalam mengabaikan informasi yang tidak relevan tanpa harus kehilangan makna dari informasi yang dibaca. Dalam melakukan pembuatan ringkasan juga terkadang masih dilakukan secara manual oleh manusia yang memakan waktu lama karena jumlah informasi yang ada sangat banyak. Sehingga pembuatan ringkasan secara otomatis diperlukan untuk mengatasi hal ini.

Berita *online* merupakan salah satu sumber informasi yang ada di internet dengan jumlah yang sangat banyak serta topik yang beragam (Schmitt et al., 2017). Topik berita dapat meliputi mengenai politik, ekonomi, olahraga, teknologi, dan masih banyak lagi. Adanya variasi informasi yang luas dan sering digunakan sehingga membuat berita menjadi objek pada penelitian ini. Kemudian, pada penelitian ini dilakukan pada berita berbahasa Indonesia karena untuk pembuatan ringkasan secara otomatis pada berita berbahasa Indonesia belum terlalu berkembang seperti berita berbahasa Inggris. Selain itu, bahasa Indonesia juga masih tergolong ke dalam *low-resource language* bila dibandingkan dengan *high-resource language* seperti bahasa Inggris sehingga perlu dilakukan penelitian untuk pengembangan *natural language processing* berbahasa Indonesia (Hirschberg & Manning, 2015).

Pada model *Bidirectional Encoder Representations from Transformers* (BERT), dilakukan penerapan model secara dua arah dengan menggunakan *Masked LM* (MLM) dengan melihat konteks dari kalimat untuk prediksi terhadap kalimat yang sudah di-*mask*. Lalu, BERT bergantung

juga pada *transformer* yang merupakan mekanisme *attention* yang mempelajari relasi kontekstual antar kata dalam teks. Pada BERT menerima input berupa kalimat yang diubah terlebih dahulu menjadi *sequence tokens* dan diproses dalam *transformer* (Devlin et al., 2019).

Terdapat beberapa penelitian yang telah melakukan peringkasan secara otomatis pada berita *online* berbahasa Inggris dengan menggunakan *pre-trained encoder* BERT. Pada penelitian yang dilakukan dengan BERT digunakan tambahan susunan *transformer encoder* setelah dihasilkan *embedding* dari BERT. Evaluasi terhadap model yang diusulkan mampu menghasilkan ringkasan ekstraktif berita dengan skor yang paling tinggi bila dibandingkan dengan metode lainnya (Liu & Lapata, 2019).

Pada penelitian sebelumnya yang dilakukan oleh Kristian Halim (Halim et al., 2020), ringkasan ekstraktif yang dihasilkan dengan menggunakan *recurrent neural network* mampu mencapai skor ROUGE terbaik sekitar 80% dengan referensi ekstraktif dan 50% dengan referensi abstraktif. Berdasarkan pada penelitian yang telah dilakukan oleh Liu dan Lapata, untuk mengetahui penerapan dan perbandingan metode BERT dengan metode lain seperti *recurrent neural network* diperoleh bahwa metode BERT menghasilkan skor yang paling tinggi (Liu & Lapata, 2019). Selain itu, BERT juga dapat mempelajari kata yang tidak dikenali dengan mengubahnya menjadi sub-kata. Maka dari itu, pada penelitian yang akan dilakukan ini akan menggunakan BERT dalam menghasilkan ringkasan ekstraktif pada berita berbahasa Indonesia.

Dalam pembuatan ringkasan otomatis berdasarkan pada hasilnya dapat dibedakan menjadi dua yaitu abstraktif dan ekstraktif (El-Kassas et al., 2020). Pada penelitian ini akan digunakan metode ekstraktif dalam pembuatan ringkasannya karena pengguna internet lebih memilih teks ringkasan yang mirip dengan teks asli yang dibuat oleh penulis (Schmitt et al., 2017). Metode yang akan digunakan dalam penelitian ini untuk menghasilkan ringkasan ekstraktif adalah *Bidirectional Encoder Representations from Transformers* (BERT).

1.2 Perumusan Masalah

Berdasarkan dari latar belakang yang ada, dapat dirumuskan permasalahan sebagai berikut:

1. Bagaimana performa dari *Bidirectional Encoder Representations from Transformers* dalam melakukan peringkasan ekstraktif otomatis pada berita Berbahasa Indonesia ?
2. Seberapa besar perbedaan skor hasil ringkasan ekstraktif bila dibandingkan dengan referensi ekstraktif dan abstraktif dengan menggunakan *Bidirectional Encoder Representations from Transformers* ?

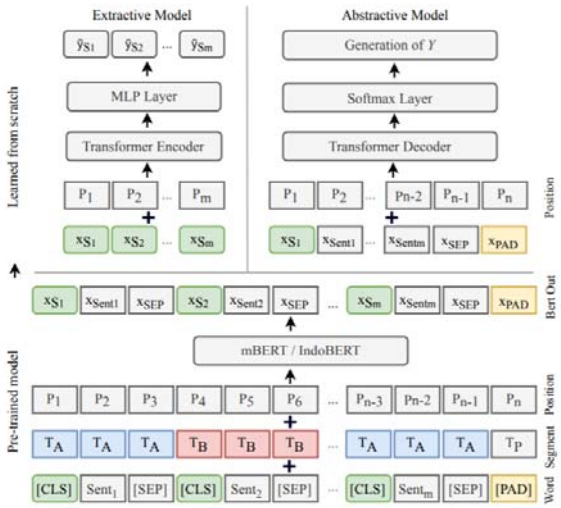
1.3 Tujuan Skripsi

Tujuan dari skripsi ini adalah membuat peringkasan otomatis secara ekstraktif pada berita berbahasa Indonesia dengan menggunakan *Bidirectional Encoder Representations from Transformers* sehingga dapat mengurangi waktu baca dan mendapatkan informasi yang relevan.

1.4 Ruang Lingkup

Ruang lingkup dibatasi pada:

1. Input yang akan digunakan berupa dataset berita berbahasa Indonesia dengan format json yang diambil dari Indosum yang merupakan hasil penelitian yang dilakukan oleh Kurniawan dan Louvan (Kurniawan & Louvan, 2018).
2. Output berupa kalimat-kalimat yang dinilai penting oleh metode dan diambil dari bacaan berita aslinya.
3. Hasil ringkasan dapat dimodifikasi panjangnya sesuai dengan persentase maksimum kalimat yang dapat diinput oleh pengguna.
4. *Pre-trained model* yang digunakan adalah indobert-base-uncased (Koto et al., 2020).
5. Evaluasi dari ringkasan yang dihasilkan akan menggunakan *ROUGE* dimana akan dibandingkan metode BERT dengan metode-metode lain pada dataset yang sama.
6. Menggunakan bahasa pemrograman php dan python.
7. Alur program yang akan dibuat mengikuti bagan berikut (mengikuti model ekstraktif):



Gambar 1.1 Arsitektur model BERT dalam menghasilkan ringkasan

Sumber: Koto, F., Lau, J. H., & Baldwin, T. (2020). *Liputan6: A Large-scale Indonesian Dataset for Text Summarization*. http://arxiv.org/abs/2011.00679

1.5 Metodologi Penelitian

Dalam melakukan penelitian, terdapat beberapa langkah yang dilakukan yaitu:

1. Studi Literatur
    1.1. Teori *News Summarization*
    1.2. Metode *Bidirectional Encoder Representations from Transformers*
    1.3. Arsitektur *Transformer*
    1.4. Evaluasi *ROUGE*
2. Pengambilan dataset
    2.1. Pengambilan dataset dari Indosum
    2.2. Analisa dataset untuk identifikasi pola dataset
3. Perencanaan dan Pembuatan Program
    3.1. Melakukan pembacaan dataset ke dalam sistem
    3.2. Melakukan *preprocessing* sederhana terhadap data
    3.3. Mengimplementasikan *BERT Summarization* ke dalam sistem
4. Pengujian dan Analisis Program
    4.1. Melakukan *testing* evaluasi ringkasan yang dihasilkan sistem dengan *ROUGE*
    4.2. Melakukan analisa metode *BERT* terhadap metode-metode lainnya
5. Pengambilan Kesimpulan
    5.1. Membuat kesimpulan mengenai hasil penelitian dari yang sudah dilakukan
    5.2. Membuat saran untuk penelitian serupa kedepannya

1.6 Sistematika Penulisan

Sistematika penulisan untuk penyusunan skripsi ini dibagi menjadi beberapa bab, yaitu:

BAB I: PENDAHULUAN

Bab 1, berisi penjelasan tentang latar belakang, perumusan masalah, tujuan skripsi, ruang lingkup, metodologi penelitian, sistematika penulisan mengenai penelitian yang dilakukan dalam skripsi yang dikerjakan.

BAB II: LANDASAN TEORI

Bab 2, berisi teori-teori yang menjelaskan mengenai dataset yang digunakan, *Bidirectional Encoder Representations from Transformers*, *Transformer*, dan *ROUGE score*.

BAB III: ANALISIS DAN DESAIN SISTEM

Bab 3, berisikan analisis dan perencanaan alur pembuatan keseluruhan sistem dari aplikasi yang dibuat.

BAB IV: IMPLEMENTASI SISTEM

Bab 4 berisikan tentang pembuatan aplikasi dan segmen-segmen program yang dibuat sesuai dengan desain sistem pada Bab 3.

BAB V: PENGUJIAN SISTEM

Bab 5 berisikan hasil dari pengujian implementasi aplikasi dan hasil pengujian yang telah dilakukan pada aplikasi.

BAB VI: KESIMPULAN DAN SARAN

Bab 6 berisikan kesimpulan dari penjelasan hasil pengujian sistem yang dicapai dan menjelaskan saran untuk pengembangan lebih lanjut.