BAB I
PEDAHULUAN

1.1 Latar Belakang

Internet telah menjadi kebutuhan penting dalam kehidupan sehari – hari, karena memberikan efektivitas dan efisiensi dalam berbagai aktivitas manusia. Dengan pesatnya penggunaan internet, jumlah informasi dan dokumen *online* pun semakin meningkat. Menurut (helmy, 2019) terdapat lebih dari 1.6 miliar situs web dan 500 juta blog di seluruh dunia. Dalam konteks ini, representasi yang akurat dari informasi teks (rangkuman) menjadi penting untuk membantu memahami isi teks secara cepat (Yuliska & Syaliman, 2020). Rangkuman memiliki peran penting dalam menganalisis informasi secara singkat dan membantu pembaca memahami inti informasi. Rangkuman yang baik harus objektif dan efektif dalam menyampaikan ide atau pesan secara jelas dan singkat. Rangkuman juga berguna dalam presentasi publik di mana waktu terbatas. Namun, menulis rangkuman yang panjang bisa rumit, oleh karena itu sistem otomatis diperlukan untuk melakukan rangkuman dengan lebih objektif dan efisien.

Text summarization atau ringkasan teks adalah pembahasan penting dalam NLP *(Natural Language Processing*). Tujuannya adalah menghasilkan ringkasan yang padat dari teks yang panjang dan kompleks. Terdapat berbagai metode maupun algoritma yang dapat digunakan dalam teknik ini. Peringkasan teks otomatis (*Automatic Text Summarization*) dilakukan menggunakan aplikasi komputer yang mengekstrak informasi penting dari dokumen aslinya (Setyadi et al., 2018). *Automatic Text Summarization* (ATS) juga dikenal sebagai ringkasan teks otomatis yang merupakan proses komputasional untuk menghasilkan ringkasan atau abstraksi dari teks yang panjang dan kompleks.

Dalam melakukan peringkasan, terdapat dua metode yang digunakan, yaitu abstraktif dan ekstraktif. Metode abstraktif adalah pendekatan yang melibatkan pembentukan kalimat – kalimat baru yang tidak ada dalam teks asli. Metode ini menciptakan ringkasan dengan menyusun kata – kata baru yang

dikombinasikan dengan kata – kata asli. Dengan menggunakan teknik ini, kalimat – kalimat baru yang terbentuk dapat memberikan inti atau esensi dari teks asli secara lebih bebas dan kreatif. Sedangkan metode ekstraktif adalah pendekatan yang mengambil kata – kata atau kalimat – kalimat dari teks asli dan menggunakan mereka untuk membuat ringkasan. Pada metode ini, tidak ada perubahan yang dilakukan pada kata – kata atau kalimat – kalimat yang diambil dari teks asli. Dengan kata lain, ringkasan dibentuk dengan menggabungkan potongan – potongan teks yang relevan dari teks asli (Setyawan et al., 2021).

Dalam mengambil suatu data dari *website* dapat menggunakan teknik *web scraping*. *Web Scraping* adalah teknik ekstraksi data dari *website* secara otomatis menggunakan *software* atau *script* khusus. Tujuannya adalah mengambil dan mengumpulkan data informasi dari *website* secara cepat dan efisien, termasuk teks, gambar, video, dan elemen lainnya. Dalam penelitian ini data yang akan di ambil hanya berupa teks. *Web scraping* melibatkan penggunaan teknologi seperti HTML dan menggunakan metode seperti parsing HTML, *regular expression*, dan lainnya. Meskipun tidak ada teknik *web scraping* yang 100% efektif, penerapannya membutuhkan pengetahuan tentang struktur halaman *website* yang dituju.

Penulis melakukan observasi pada similarweb.com ditemukan bahwa detik.com merupakan *website* paling populer pada kategori *news & publishers* di Indonesia pada bulan Maret 2023. Oleh karena itu penulis berencana menggunakan teknik *web scraping* untuk mengambil informasi dari postingan berita di situs tersebut. Situs tersebut memiliki beberapa kategori, salah satunya adalah *news* dengan *sub domain* news.detik.com. Berita dengan *sub domain* inilah yang nantinya akan dilakukan proses *scraping*.

Dari penjelasan sebelumnya penulis berencana membangun aplikasi sistem peringkas teks (*Automatic Text Summarization*) dan menerapkan teknik *web scraping* dengan menggunakan algoritma TF – IDF (*Term Frequency Inverse Document Frequency*) untuk menghasilkan *extractive summarization*. Proses *Automatic Text Summarization* (ATS) akan menggunakan data teks dari postingan *website* news.detik.com yang diambil menggunakan teknik *web scraping*.

1.2 Rumusan Masalah

Dari latar belakang yang telah dipaparkan oleh penulis sebelumnya maka terdapat beberapa beberapa rumusan masalah diantaranya :

1. Bagaimana melakukan *summarization* terhadap isi berita pada *website* berita dengan menggunakan algoritma TF – IDF?
2. Bagaimana cara mengekstraksi teks dari *website* berita?
3. Bagaimana melakukan evaluasi hasil ringkasan teks berita yang dihasilkan oleh sistem?

1.3 Batasan Masalah

Adapun ruang lingkup serta batasan masalah pada penelitian ini adalah :

1. Berita berasal dari *website* news.detik.com yang merupakan *sub domain* dari detik.com, digunakan sebagai target untuk melakukan *web scraping* dengan teknik *HTML Parsing* dan implemenntasi *Automatic Text Summarization* (ATS) dengan algoritma TF – IDF pada *content* yang telah di *scraping*.
2. Fokus pembahasan lebih mengutamakan pada penerapan algoritma TF-IDF dalam melakukan peringkasan teks daripada pembahasan mekanisme melakukan *scraping*.
3. Bahasa dalam berita merupakan bahasa Indonesia.
4. Jenis hasil ringkasan dalam penelitian ini adalah *extractive summarization*.
5. Hasil dari pembobotan setiap *term* akan di jumlahkan sebagai nilai bobot pada kalimat. Kalimat dengan bobot tertinggi dan terpilih akan menjadi kandidat ringkasan. Kandidat tersebut akan digabungkan sesuai dengan urutan kalimat aslinya untuk membentuk sebuah paragraph yang bagus.