

BAB I

PENDAHULUAN

1.1.Latar Belakang

Pemrosesan bahasa alami atau *Natural Language Processing* (NLP) merupakan topik penelitian dan pengaplikasian yang membahas cara komputer dapat digunakan sebagai sarana memanipulasi serta memahami teks atau ucapan bahasa alami untuk melakukan hal-hal yang bermanfaat. Bahasa alami adalah bahasa yang secara umum digunakan oleh manusia untuk saling berkomunikasi dan berinteraksi dalam kehidupan sehari-hari[1]. Komputer membutuhkan sebuah proses dalam memahami bahasa yang diterima dapat diolah dengan baik oleh komputer. Banyak terapan sistem, salah satu area utama penelitian pada *field* pemrosesan bahasa alami adalah peringkasan teks.

Peringkasan teks adalah cara menghasilkan ringkasan yang berisi kalimat penting dan mencakup semua informasi penting yang relevan dari dokumen asli. Hasil dari peringkasan ini memiliki poin utama dari teks yang sebenarnya. Pada peringkasan teks memiliki dua pokok substansi penelitian, yaitu ekstraksi dan abstraksi[2]. Salah satu topik pembahasan dalam penelitian peringkasan teks adalah artikel berita.

Artikel berita merupakan laporan tercepat dari terjadinya suatu peristiwa atau kejadian yang faktual, penting, serta menarik bagi sebagian besar pembaca, serta menyangkut kepentingan mereka. Bagi surat kabar, berita adalah informasi tentang peristiwa dan pernyataan yang penting bagi pembaca untuk mewujudkan falsafah hidup mereka. Berita adalah fakta, ide, dan opini aktual yang menarik, akurat, dan dianggap penting oleh banyak pembaca, pendengar, dan pemirsa[3]. Pada media publik Indonesia terdapat banyak situs berita online, seperti *detik.com*, *kompas.com*, *cnnindonesia.com* dan lain sebagainya. Berita yang disajikan dalam media elektronik, biasanya dikelompokkan dalam beberapa golongan berdasarkan isi artikel berita tersebut, misalnya ekonomi, olahraga, politik, teknologi, Kesehatan, serta lainnya.

Permasalahan yang dapat ditimbulkan adalah penggunaan media informasi elektronik dalam menyampaikan informasi mengakibatkan jumlah dari artikel berita yang diterbitkan dari portal berita setiap harinya meningkat. Hal ini mengakibatkan jumlah dari artikel berita yang tersedia semakin melimpah. Dalam setiap artikel berita tentu memiliki teks yang sangat panjang dan harus menemukan topik utama dalam artikel yang diberitakan oleh penulis berita[4]. Hal ini dapat merepotkan bagi pengguna yang memiliki sedikit waktu untuk membaca berita.

Berdasarkan dari permasalahan yang timbul tersebut diperlukannya sebuah aplikasi peringkasan dokumen dari artikel berita. Salah satu hal yang dapat dilakukan dengan akurat serta tidak mengubah inti dari isi artikel adalah dengan melakukan peringkasan teks otomatis (*Automatic Text Summarization*). Ringkasan teks otomatis (*Automatic Text Summarization*) adalah teks dari satu buku atau lebih, di mana teks tersebut memberikan informasi penting dari sumber buku aslinya, secara otomatis dan teks tersebut tidak melebihi setengah dari sumber aslinya.[5]. Peringkasan teks otomatis adalah proses pembuatan teks yang sebelumnya telah direduksi dengan menggunakan program komputer untuk menghasilkan ringkasan teks dengan frasa-frasa penting di mana ringkasan tersebut tidak melebihi setengah dari teks aslinya.

Penelitian oleh [6] dengan metode *Cross Latent Semantic Analysis* (CLSA) untuk peringkasan modul teks pembelajaran. Uji validitas dan modul pembelajaran pengelompokan memberikan nilai *f-score*, *precision* dan *recall* pada kompresi 20% dengan nilai kontinu 0.3853, 0.432, dan 0.3715.

Penelitian yang dilakukan oleh [7] hasil dari pengujian didapatkan algoritma *textrank* mampu memberikan hasil ringkasan teks dengan konten yang informatif hingga 82,48% untuk teks ringkasan sebesar 50% dan konten informatif 93,76% untuk teks ringkasan yang dirangkum 75%.

Pada penelitian yang sudah dilakukan oleh [8] hasil penelitian ini dari proses *Maximal Marginal Relevance* (MMR) adalah ringkasan teks akhir. Evaluasi menggunakan ROUGE-1 dan ROUGE-2 dengan rata-rata *F-score* 0,5103 dan 0,4257.

Dengan kesempatan ini peneliti hendak membangun mesin peringkasan teks otomatis. Salah satu algoritma dalam peringkasan teks yaitu menggunakan algoritma *TextRank* yang termasuk ke dalam kategori peringkasan *extractive summarization*. Algoritma *TextRank* merupakan algoritma sederhana dengan menilai setiap kalimat dan melakukan pemeringkatan tanpa mengubah isi teks asli yang berbasis grafik untuk memproses teks.

1.2. Perumusan Masalah

Perumusan masalah pada penelitian ini sebagai berikut :

1.2.1. Permasalahan

Dalam meringkas sebuah teks artikel berita Bahasa Indonesia terdapat banyak metode yang dapat dipergunakan, salah satu metode tersebut adalah *graph-based*. Metode ini mengambil sebuah kesimpulan dari seberapa banyak *edge* yang terhubung pada sebuah *node*. Bagaimana membuat sistem yang dapat menghasilkan peringkasan teks berita Bahasa Indonesia dengan menggunakan metode *graph-based* pada arsitektur *TextRank*,

1.2.2. Metode Usulan

Berdasarkan permasalahan yang telah diuraikan, maka peneliti menggunakan metode algoritma *TextRank* untuk meringkas sebuah artikel berita.

1.2.3. Pertanyaan Penelitian

Berapakah nilai skor kesamaan dari peringkasan artikel berita Bahasa Indonesia dengan menggunakan algoritma *textrank* dengan peringkasan manual jika peringkasan tersebut diukur menggunakan metode *ROUGE*?

1.3. Tujuan dan Manfaat

Tujuan dan manfaat dari penelitian ini adalah sebagai berikut :

1.3.1. Tujuan Penelitian

Tujuan dari penelitian ini untuk meringkas sebuah artikel berita Bahasa Indonesia menjadi sebuah teks artikel yang lebih ringkas dan mengetahui relevansi

dari hasil peringkasan mesin menggunakan algoritma *TextRank* yang diukur dengan menggunakan metode *ROUGE*.

1.3.2. Manfaat Penelitian

Manfaat yang diharapkan dapat diperoleh dari skripsi ini adalah :

Pembuatan mesin peringkasan otomatis untuk menjadikan sebuah teks artikel berita Bahasa Indonesia yang relative panjang ke dalam bentuk yang lebih ringkas dengan menggunakan algoritma *TextRank*.

1.4. Batasan-Batasan

Berdasarkan pada perumusan masalah yang telah disebutkan diatas, maka batasan masalah penelitian ini adalah sebagai berikut:

1. Data artikel berita yang digunakan dalam penelitian ini bersumber dari <https://github.com/kata-ai/indosume>. Dokumen yang diambil untuk penelitian ini sebanyak 10 teks artikel berita dari file dev.01.jsonl pada kategori “Tajuk Utama” dari portal berita CNN Indonesia.
2. Link artikel berita yang digunakan dalam penelitian ini hanya terbatas pada portal berita CNN Indonesia.

1.5. Sistematika Laporan

Rencana susunan sistematika penulisan laporan tugas akhir yang akan dibuat. Sistematika penulisan laporan tugas akhir ini meliputi :

BAB I PENDAHULUAN

Bab ini berisi latar belakang, perumusan masalah, tujuan & manfaat, batasan-batasan yang akan dilakukan dan sistematika penulisan.

BAB II KAJIAN PUSTAKA

Bab ini berisi penjelasan tentang teori-teori dan konsep yang digunakan dalam penelitian.

BAB III METODE USULAN

Bab ini berisi penjelasan tentang metode yang diusulkan, jalannya penelitian, dan proses rekayasa sistem penelitian yang akan berlangsung.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi penjelasan hasil dari implementasi metode, penjelasan *source code*, dan *user interview website*.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi tentang kesimpulan dari penelitian dan saran untuk penelitian selanjutnya.