BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi dan informasi yang pesat mengakibatkan informasi yang tersedia baik online maupun offline semakin bertambah salah satunya adalah artikel berita. Dari banyaknya kalimat pada artikel berita, hanya beberapa informasi penting yang merupakan representasi pokok pikiran dari penulis [1]. Artikel berita bersifat *real time* dan *up to date* yang membuat pembaca harus meluangkan waktu dalam membaca dan mengikuti perkembangan berita. Untuk mempermudah pembaca dalam menangkap isi artikel berita, dibutuhkan sebuah peringkasan teks otomatis (*Automatic Text Summarization*). Dengan cara ini, pembaca hanya membutuhkan waktu lebih sedikit dibandingkan membaca keseluruhan isi berita [2].

Peringkasan teks otomatis (*Automatic Text Summarization*) merupakan peringkasan teks yang dilakukan oleh komputer untuk membuat sebuah artikel menjadi lebih sederhana dengan tidak menghilangkan makna dan mempermudah pembaca dalam mengambil sebuah kesimpulan tanpa harus membaca keseluruhan artikel [2]. Terdapat dua pendekatan peringkasan teks otomatis, yaitu peringkasan teks secara abstraktif (*Abstractive Summarization*) dengan membuat dan menyusun kalimat baru yang merupakan intisari dari artikel yang diringkas. Peringkasan teks secara ekstraktif (*Extractive Summarization*) yaitu mengambil dan menyusun kalimat dari artikel teks asli yang mengandung informasi penting [3].

Peringkasan ekstraktif memiliki keunggulan yaitu lebih mudah memberikan hasil ringkasan yang diharapkan dan lebih baik daripada peringkasan teks abstraktif yang kompleks [4].

Penelitian tentang peringkasan teks ekstraktif sudah banyak dilakukan, salah satunya adalah peringkasan bahasa Indonesia [5]. Meskipun sudah ada penelitian tentang peringkasan teks bahasa Indonesia, masih jarang yang dilatih atau dievaluasi menggunakan dataset besar yang tersedia untuk publik [6].

Dalam peringkasan ekstraktif terdapat beberapa metode peringkasan, seperti berbasis statistik, semantik, graph, dll [7]. Salah satu metode yang tidak terlalu sering digunakan dalam peringkasan teks selama 10 tahun terakhir adalah *TextRank* [4]. *TextRank* merupakan metode perangkingan kalimat berbasis *graph* dengan menentukan nilai tertinggi dari sebuah kalimat. Setiap kalimat dianggap sebagai sebuah *vertex*, semakin tinggi skor sebuah *vertex*, maka semakin penting *vertex* tersebut [3].

Rani dan Bidhan (2021) telah melakukan penelitian dengan membandingkan peringkasan ekstraktif antara *TextRank*, TF-IDF dan LDA menggunakan dataset review, *news*, dan legal berbahasa Inggris dievaluasi menggunakan ROUGE. Dalam penelitian tersebut pendekatan *TextRank* mendominasi nilai tertinggi pada setiap dataset. Pada dataset review nilai *F-measure* ROUGE-L *TextRank* adalah 0.179, TF-IDF 0.103, dan LDA 0.112. Pada dataset *news* nilai *F-measure* ROUGE-L *TextRank* adalah 0.652, TF-IDF 0.486, dan LDA 0.602. Pada dataset legal nilai *F-measure* ROUGE-L *TextRank* adalah 0.234, TF-IDF 0.233, dan LDA 0.210. Penelitian tersebut menunjukkan bahwa *TextRank* lebih baik daripada TF-IDF dan LDA yang dievaluasi menggunakan

ROUGE [8]. ROUGE digunakan untuk menentukan kualitas hasil ringkasan dengan membandingkannya menggunakan ringkasan yang dibuat oleh manusia [9].

Maka dari itu, penelitian ini menggunakan algoritma *TextRank* untuk melakukan peringkasan pada artikel berita bahasa Indonesia dengan pendekatan ekstraktif. Hasil yang didapat dari peringkasan teks akan dievaluasi menggunakan ROUGE.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah disampaikan maka dirumuskan masalah yaitu, berapa nilai kualitas hasil rangkuman ekstraktif menggunakan algoritma *TextRank* pada artikel berita Bahasa Indonesia yang diukur dengan ROUGE?

1.3 Batasan Masalah

Adapun batasan masalah yang ada dalam penelitian ini adalah sebagai berikut:

1. Artikel berita yang digunakan untuk peringkasan adalah artikel berita berbahasa Indonesia.
2. Artikel berita yang digunakan adalah dataset Liputan6 sebanyak 193.883 data.
3. Versi *Library* yang digunakan adalah pandas 1.3.0, numpy 1.21.0, nltk 3.6.5, regex 2021.10.23, networx 2.6.3, matplotlib 3.4.2, python 3.9.5, rouge_score 0.0.4.

4. *Preprocessing* yang dilakukan hanya *segmentation, case folding,* dan *stopword*.
5. Ekstraksi fitur yang digunakan berupa *pre-trained word embedding fastText* berdimendi 300.

1.4 Maksud dan Tujuan Penelitian

1.4.1 Maksud

Maksud dari penelitian ini adalah untuk mendapatkan ringkasan dari artikel berita dengan menggunakan algoritma *TextRank* serta mengevaluasi hasil ringkasan menggunakan ROUGE.

1.4.2 Tujuan

Tujuan dari penelitian ini adalah untuk mendapatkan nilai ROUGE dari hasil ringkasan menggunakan algoritma *TextRank* pada teks berbahasa Indonesia.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah untuk membantu memberikan referensi bagi para peneliti dalam pengembangan sistem atau aplikasi peringkasan teks khususnya artikel berita bahasa Indonesia.

1.6 Metode Penelitian

Metode awal yang digunakan adalah melakukan pengumpulan data. Data yang digunakan berupa dataset sekunder publik Liputan6 dari penelitian yang dilakukan oleh Koto, Lau dan Baldwin [10] yang tersedia pada github [11]. Langkah selanjutnya adalah *preprocessing* berupa *segmentation, case folding,* dan *stopword*. Tahap setelah *preprocessing* adalah ekstraksi fitur (*feature extraction*)

menggunakan fitur *word embedding fastText* dimensi 300 [12]. Setelah itu adalah tahap *processing* yang terdiri dari *similarity*, penerapan algoritma *TextRank* dengan perhitungan bobot *vertex* menggunakan *PageRank*, dan *sentence ranking* dan langkah terakhir adalah evaluasi hasil ringkasan menggunakan ROUGE.

1.7 Sistematika Penulisan

Sistematika penulisan penelitian terdiri atas beberapa bagian yang disusun untuk menggambarkan penelitian yang akan dijalankan. Sistematika penulisan ini adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini berisi latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metode penelitian, dan sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini berisi kajian pustaka tentang penelitian lain yang telah dilakukan, dasar teori yang menjadi referensi dalam mendukung pengerjaan dan metode yang diterapkan.

BAB III METODE PENELITIAN

Bab ini berisi langkah-langkah pengerjaan dalam penelitian dan komponen-komponen yang digunakan untuk mencapai tujuan penelitian.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi hasil pengujian algoritma yang telah dilakukan dalam penelitian, serta pembahasan terhadap hasil pengujian algoritma.

BAB V PENUTUP

Berisi kesimpulan dan saran yang dapat peneliti rangkum selama proses penelitian sebagai rekomendasi untuk penelitian selanjutnya.