

BAB I

PENDAHULUAN

1.1. Latar Belakang

Berita menjadi salah satu kebutuhan penting bagi masyarakat. Dengan adanya berita, masyarakat dapat mengetahui peristiwa-peristiwa atau informasi berupa fakta terkini yang berkaitan baik dengan lingkungan sekitarnya maupun dunia luar. Adanya internet membuat berita menjadi semakin mudah didapatkan karena tersebar di berbagai media, contohnya media sosial dan koran elektronik. Koran elektronik ini dapat diakses melalui portal-portal berita seperti Kompas, Liputan6, Detik, dan lain-lain. Bahkan pada tahun 2018, berdasarkan pernyataan dari Menteri Komunikasi dan Informatika saat itu, Rudiantara, jumlah portal berita online yang ada di Indonesia mencapai angka 43.000. Namun, hanya kurang dari 100 portal berita saja yang telah diverifikasi oleh Dewan Pers (Kementerian Komunikasi dan Informatika, 2018). Banyaknya portal berita membuat banyaknya informasi yang tersebar di masyarakat semakin banyak.

Sebagai salah satu sumber informasi, berita menjadi kebutuhan bagi masyarakat. Namun, bagi orang-orang dengan waktu luang yang terbatas, membaca berita mungkin saja tidak sempat dilakukan. Oleh karena itu, menyajikan berita yang ringkas akan membantu mereka dalam menghemat waktu dan mengetahui informasi inti dari berita dengan cepat. Hal ini dapat dilakukan dengan meringkas teks berita tersebut menjadi lebih singkat, padat, dan jelas.

Peringkasan teks secara manual memang sudah biasa dilakukan sejak dulu. Walaupun ringkasan yang dihasilkan sangat akurat, tetapi hal ini sangat memakan waktu karena kita harus membaca dan memahami keseluruhan isi teks terlebih dahulu. Perkembangan teknologi yang cepat membuat orang-orang memunculkan inovasi baru, salah satunya dalam metode peringkasan teks. Teknologi dapat membuat peringkasan teks dilakukan secara otomatis. Tentu saja hal ini akan menghemat waktu baik dalam proses peringkasan teks, maupun saat membaca teks.

Peringkasan teks secara otomatis ini dilakukan dengan menggunakan pemrosesan bahasa alami atau *Natural Language Processing* (NLP). NLP ini merupakan salah satu cabang dari Kecerdasan Buatan atau *Artificial Intelligence*

yang merupakan bagian dari Ilmu Komputer. NLP bekerja dengan memproses dan menghasilkan bahasa manusia dengan menggunakan teknik komputasi (Hirschberg & Manning, 2015). Penerapan NLP ini sudah semakin banyak digunakan, contohnya untuk *chatbot*, *speech recognition* seperti *Google Assistant* dan *Siri*, pendeteksi plagiarisme, dan juga peringkasan teks. Sistem peringkasan teks ini dapat merangkum berbagai jenis teks, seperti teks pada buku, novel, karya tulis ilmiah, *e-mail*, berita, dan lain-lain (Mridha dkk., 2021).

Salah satu cara yang dapat digunakan untuk meringkas teks berita secara otomatis adalah dengan menggunakan *Deep Learning*. Munculnya *Artificial Neural Network* (ANN) yang mengadopsi struktur dan cara kerja sel neuron pada sistem saraf manusia menjadi konsep awal dari terbentuknya *Deep Learning*. *Deep Learning* merupakan salah satu teknik dari *Machine Learning* yang digunakan untuk menyelesaikan data dengan abstraksi yang lebih kompleks, masalah ini biasanya tidak dapat diselesaikan oleh algoritma *Machine Learning* biasa. *Deep Learning* membantu manusia dalam mengefisienkan waktu dan tenaga mereka untuk melakukan pekerjaan, seperti pada kegiatan meringkas teks.

Perkembangan *Deep Learning* terbilang cukup lambat pada awal kemunculannya di tahun 1960-an dibandingkan dengan perkembangan algoritma *Machine Learning* baru lainnya yang mulai berkembang pesat di tahun 1990-an (Géron, 2020). Konsep *backpropagation* pertama dikembangkan pada tahun 1985 dan mengawali perkembangan *Recurrent Neural Network* (RNN) yang digunakan untuk menangani data sekuensial, seperti teks dan data *time series* yang tidak dapat ditangani oleh algoritma *feed-forward* seperti *Convolutional Neural Network* (CNN). Lalu pada tahun 1997, *Long Short-Term Memory* mulai dikembangkan sebagai variasi dari RNN yang mengatasi masalah saat melatih sekuens yang terlalu panjang (Di dkk, 2018).

Namun, dalam bidang NLP, bahasa Indonesia masih belum banyak dieksplorasi seperti halnya bahasa Inggris. Hal ini disebabkan oleh belum adanya dataset berskala besar, begitu juga pada peringkasan teks otomatis (Koto dkk, 2020). Dataset peringkasan teks pertama berbahasa Indonesia mulai tersedia di tahun 2018, dataset tersebut adalah dataset Indosum (Kurniawan & Louvan, 2018). Dataset tersebut berisi sekitar 19.000 data artikel berita beserta ringkasan aslinya.

atau ketidaklengkapan kalimat. Selain itu, juga terdapat perulangan kata-kata yang tidak benar. Kesalahan dalam peringkasan teks secara otomatis ini ditunjukkan melalui contoh pada Gambar 1.2.

Oleh karena itu, penelitian ini dilakukan untuk membuat sistem peringkasan teks berbahasa Indonesia secara otomatis dengan menggunakan algoritma LSTM yang juga diaplikasikan mekanisme atensi Bahdanau (Bahdanau dkk., 2016). Mekanisme atensi berhasil menjadi *state-of-the-art* pada permasalahan *sequence-to-sequence* seperti peringkasan teks dan mesin translasi. Selain itu, perkembangan Transformer (Vaswani dkk., 2017) yang juga dapat digunakan pada permasalahan *sequence-to-sequence* menghasilkan beberapa variasi lain seperti BART dan T5. Penelitian ini juga akan menggunakan *pre-trained model* BART dan T5 untuk membuat sistem peringkasan teks berbahasa Indonesia. Masih banyak peluang untuk mengurangi kesalahan tata bahasa pada kata atau kalimat dalam peringkasan teks secara otomatis, sehingga penelitian ini akan mencoba untuk memaksimalkan keakuratan ringkasan teks.

1.2. Rumusan Masalah

Sesuai latar belakang masalah yang telah diuraikan pada sub bab sebelumnya, rumusan masalah yang muncul adalah sebagai berikut:

1. Bagaimana cara mengembangkan model pembelajaran mesin untuk sistem peringkasan teks berita berbahasa Indonesia secara otomatis dengan menggunakan *Long Short-Term Memory* dan *Transformer*?
2. Bagaimana analisis hasil komparasi penggunaan *Long Short-Term Memory* dan *Transformer* terhadap kualitas hasil prediksi ringkasan model pembelajaran mesin untuk sistem peringkasan teks berita berbahasa Indonesia secara otomatis?

1.3. Tujuan Penelitian

Setelah diketahui rumusan masalahnya, maka tujuan dari penelitian ini adalah sebagai berikut:

1. Mengembangkan model sistem pembelajaran mesin untuk sistem peringkasan teks berita berbahasa Indonesia secara otomatis dengan menggunakan *Long Short-Term Memory* dan *Transformer*.

2. Menganalisis hasil komparasi penggunaan *Long Short-Term Memory* dan *Transformer* terhadap kualitas hasil prediksi ringkasan model pembelajaran mesin untuk sistem peringkasan teks berita berbahasa Indonesia secara otomatis.

1.4. Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah sebagai berikut :

1. Memberikan kemudahan bagi pembaca dalam mengetahui informasi dalam suatu teks berbahasa Indonesia secara ringkas, sehingga pembaca lebih mengerti isi dari berita dan menghemat waktu mereka.
2. Memberikan penelitian terkait peringkasan teks berbahasa Indonesia secara otomatis menggunakan algoritma *Long Short-Term Memory* dan *Transformer*.

1.5. Batasan Penelitian

Adapun batasan masalah dari penelitian ini adalah sebagai berikut:

1. Data yang digunakan merupakan teks berita berbahasa Indonesia.
2. Model dibangun dengan menggunakan bahasa pemrograman Python.
3. Panjang artikel berita dibatasi hanya sebanyak 500 kata.

1.6. Sistematika Penulisan

Sistematika penulisan pada penelitian ini terdiri dari lima bab dengan struktur sebagai berikut.

BAB I PENDAHULUAN

Pada bab ini akan dipaparkan latar belakang dari rumusan masalah dari penelitian peringkasan teks secara otomatis, tujuan yang dicapai, manfaat dan batasan-batasan pada penelitian, serta sistematika penulisan.

BAB II KAJIAN PUSTAKA

Pada bab ini membahas hasil studi literatur yang dilakukan untuk melakukan penelitian peringkasan teks berita menggunakan *Deep Learning*. Penjelasan dimulai dari definisi umum terkait berita, peringkasan teks, dan NLP. Lalu terkait perkembangan *Deep Learning* hingga algoritma *recurrence* seperti RNN dan LSTM. Bab ini juga membahas arsitektur model *encoder-decoder* yang akan diaplikasikan menggunakan LSTM dan mekanisme atensi. Terdapat

pembahasan mengenai *fine-tuning* dan *Transformer* beserta variasinya (T5 dan BART). Lalu dijelaskan juga materi *word embedding*, ROUGE yang akan digunakan sebagai metrik evaluasi, dataset Indosum, dan penelitian-penelitian peringkasan teks terdahulu.

BAB III METODOLOGI PENELITIAN

Pada bab ini terdapat pembahasan mengenai desain dari penelitian peringkasan teks mulai dari perumusan masalah hingga analisis dan evaluasi hasil dari eksperimen penelitian. Terdapat juga pembahasan mengenai alat serta bahan yang digunakan pada penelitian dan desain sistem.

BAB IV HASIL DAN PEMBAHASAN

Pada bab ini memuat kegiatan dilakukan selama penelitian seperti tahap praproses data dan detail perancangan skenario eksperimen. Hasil dan analisis dari eksperimen juga dijabarkan pada bab ini.

BAB V KESIMPULAN

Pada bab ini terdapat kesimpulan serta saran yang berkaitan dengan hasil eksperimen.