BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan teknologi informasi telah membawa dunia ke dalam era digital yang penuh dengan data dan informasi. Data adalah pikiran dan imajinasi yang tidak terwujud (Koupaee & Wang, 2018) . Informasi kini tersedia secara masif dalam bentuk teks, baik melalui media sosial, artikel ilmiah, hingga portal berita daring. Namun, dengan volume informasi yang besar, pengguna menghadapi tantangan dalam menyerap isi konten secara efisien. Terlebih lagi dalam konteks berita *online*, dimana pembaca membutuhkan akses cepat terhadap inti informasi tanpa harus membaca keseluruhan isi artikel.

Salah satu solusi untuk permasalahan tersebut adalah melalui penerapan sistem peringkasan teks otomatis. Dengan menggunakan teknologi pemrosesan bahasa alami (*Natural Language Processing* / NLP), sistem ini dapat menghasilkan ringkasan yang lebih ringkas namun tetap mempertahankan makna utama dari konten aslinya. Menurut penelitian Budaya et al. (2022) teknik vektorisasi seperti Word2Vec mampu meningkatkan performa model *Neural Machine Tranlation* dari bahasa Kawi ke bahasa Indonesia. Hal ini menegaskan bahwa pentingnya tahap pra pemrosesan dan representasi kata dalam tugas NLP. Dalam konteks NLP, peringkasan teks terbagi menjadi dua pendekatan utama yaitu ekstraktif dan abstraktif. Pendekatan ekstraktif dibentuk dengan menggunakan sebagian dari dokumen asli, seperti kalimat dan menggabungkan nya menjadi sebuah ringkasan,

sedangkan pendekatan abstraktif dibentuk dengan menginterpretasikan atau melakukan parafrase pada dokumen asli dan menghasilkan kalimat-kalimat baru yang bermakna dalam versi yang lebih pendek (Abdel-Salam & Rafea, 2022). Penelitian sebelumnya tentang NLP lebih banyak diarahkan pada analisis sentimen berbasis teks, dibuktikan dengan adanya penelitian oleh Ariyani et al. (2025) yang membandingkan metode *Naïve Bayes* dan *K-Nearest Neighbor* dalam mengklasifikasikan opini publik di Twitter (sekarang X) terkait COVID-19, dengan hasil bahwa *Naïve Bayes* lebih unggul dalam akurasi. Sejalan dengan hal tersebut, Sasmita et al. (2022) mengembangkan sistem analisis sentimen untuk evaluasi kinerja dosen menggunakan algoritma *Naïve Bayes* yang terbukti efektif dengan *precision* rata-rata 90.84% dan *recall* 95.73%. Kedua penelitian tersebut memperlihatkan bahwa pendekatan NLP berbasis pembelajaran mesin telah dimanfaatkan secara luas untuk *domain* kesehatan dan pendidikan sehingga memperkuat landasan penelitian ini dalam mengembangkan model NLP untuk tugas peringkasan teks.

Model IndoBERT dan PEGASUS merupakan dua model yang saat ini banyak digunakan dalam tugas peringkasan teks. IndoBERT adalah sebuah model bahasa *pre-trained* yang dikembangkan khusus untuk bahasa Indonesia. Model ini merupakan salah satu dari sedikit model BERT monolingual yang ada untuk bahasa Indonesia (Koto et al., 2020). Seperti BERT, IndoBERT juga menggunakan arsitektur Transformer yang terdiri dari sejumlah lapisan *encoder*. Dalam hal ini, *encoder* bertanggung jawab untuk memahami konteks dan hubungan antara kata-kata dalam teks. Model IndoBERT dilatih dari *dataset* besar dan bersih yang disebut

Indo4B, yang dikumpulkan dari berbagai sumber publik seperti teks, media sosial, blog berita dan situs web (Wilie et al., 2020).

PEGASUS merupakan salah satu model bahasa *pre-trained* seperti hal nya IndoBERT yang dikembangkan khusus untuk peringkasan teks abstraktif oleh Google Research. Dalam penelitian yang berjudul "*Faster (Multi) Document Summarization Using PRIMERA & PEGASUS*" dijelaskan bahwa PEGASUS memiliki pendekatan *pre-training* yang memungkinkan model memahami dan merepresentasikan nuansa bahasa Inggris dengan baik (Patil et al., 2024). Menggunakan teknik yang disebut dengan *Gap-Sentence Generation* membantu model memahami esensi dokumen dengan mengisi kalimat yang hilang, sehingga mampu menghasilkan ringkasan yang abstraktif dan relevan secara kontekstual. PEGASUS juga menggunakan arsitektur standar *encoder-decoder* dari Transformer, dimana GSG dan MLM diretapkan secara bersamaan (J. Zhang et al., 2020). Pemanfaatan model Transformer untuk teks berbahasa Indonesia juga telah dilakukan sebelumnya oleh Wijaya et al. (2025), yang melakukan *fine-tuning* terhadap model IndoBERT pada *domain* berita pariwisata dan memperoleh akurasi 77%, meskipun masih menghadapi tantangan dalam klasifikasi sentimen netral. Sementara itu, (Dewi et al., 2024) mengembangkan model klasifikasi judul berita pariwisata berbasis *Support Vector Machine* (SVM) dengan pendekatan *Binary Term Presence* yang berhasil mencapai akurasi 87.80%. Kedua penelitian ini menunjukkan bahwa model IndoBERT dan metode *machine learning* tradisional sama-sama relevan untuk teks Indonesia dan menjadi pijakan penting bagi penelitian ini yang memperluas cakupan kearah peringkasan teks abstraktif.

Penelitian ini berupaya untuk membandingkan performa kedua model tersebut dalam tugas peringkasan teks otomatis pada dataset berita berbahasa Indonesia. Penelitian ini penting mengingat belum banyak studi yang secara langsung membandingkan performa IndoBERT dan PEGASUS dalam konteks bahasa Indonesia, terutama untuk tugas peringkasan teks abstraktif.

1.2 Identifikasi Masalah

Dari uraian latar belakang diatas, dapat dikenali masalah yang akan menjadi fokus penelitian adalah sebagai berikut:

a. Informasi yang tersedia di internet memiliki volume yang sangat besar dan sangat kompleks.
b. Manusia memiliki keterbatasan waktu dalam membaca dan memahami informasi yang masif.
c. Diperlukan suatu model atau sistem yang mampu merangkum teks dengan jumlah besar menjadi bentuk yang lebih ringkas tanpa kehilangan esensi atau konteks asli dari teks tersebut.

1.3 Pembatasan Masalah

Dalam Upaya mengatasi permasalahan yang telah diidentifikasi sebelumnya, peneliti perlu memastikan adanya batasan dalam cakupan penelitian agar lebih terfokus. Berikut merupakan batasan penelitian yang dapat dipertimbangkan.

a. Penelitian ini berfokus meneliti kinerja model *pre-trained* IndoBERT dan PEGASUS untuk menghasilkan ringkasan teks abstraktif menggunakan *dataset* berita artikel berbahasa Indonesia berdasarkan hasil metrik evaluasi ROUGE dan BERTScore.
b. Model yang digunakan dalam penelitian ini adalah varian dari Transformer yaitu IndoBERT+GPT-2 *decoder* (*checkpoint*: cahya/bert2gpt-indonesian summarization) dan PEGASUS-*Large* (google/pegasus-large). Tujuan menambahkan GPT-2 *decoder* pada model IndoBERT karena penelitian ini dilakukan dalam bentuk seq2seq model.
c. *Dataset* yang digunakan dalam penelitian ini adalah "Indonesian News Dataset" yang diambil dari *website* Kaggle oleh pengguna 'iqbalmaulana' (iqbalmaulana. (2023). Indonesian News Dataset [Dataset]. Kaggle. Retrieved Agustus 21, 2024, from https://www.kaggle.com/datasets/iqbalmaulana/indonesian-news dataset) yang kemudian nantinya akan dilakukan *pre-processing* dengan jumlah data asli sebanyak 32.148 x 11 rows yang di *scrapping* oleh pengguna dari 7 sumber berita online berbeda (Tempo, CNN Indonesia, CNBC Indonesia, Okezone, Suara, Kumparan, dan JawaPos). *Dataset* ini terdiri dari beberapa kategori yang berbeda beda seperti politik, pendidikan dan sosial kemasyarakatan. Namun, penelitian ini tidak mempertimbangkan kategori tersebut, artinya semua kategori dianggap sama.
d. Tahapan *pre-processing* yang dilakukan pada *dataset* "Indonesian News Dataset" yaitu meliputi *text cleaning*, *normalization*, *tokenization*, dan *filtering*. Jumlah *dataset* setelah dilakukan *pre-processing* adalah sebanyak 29.985 x 2

(‘*content*’ dan ‘*summary*’) kemudian dibagi menjadi 3 dengan proporsi 70% untuk *training*, 15% untuk *testing*, dan 15% untuk *validation*.

e. Evaluasi performa model dibatasi pada metrik ROUGE (ROUGE-1, ROUGE-2, ROUGE-L) dan BERTScore saja tanpa melibatkan proses *human-evaluator*. Proses evaluasi dilakukan dengan melakukan perhitungan kecocokan ringkasan teks yang dihasilkan oleh model IndoBERT dan PEGASUS dengan *ground truth* atau ringkasan referensi yang sudah disediakan oleh penyedia *dataset* (Iqbal Maulana) dan kemudian di evaluasi menggunakan metrik (ROUGE dan BERTScore).

1.4 Rumusan Masalah

Berdasarkan pada uraian tersebut, masalah yang akan dikaji pada penelitian ini adalah sebagai berikut.

a. Bagaimana cara implementasi model IndoBERT+GPT-2 *decoder* dan PEGASUS-*Large* dalam menghasilkan teks ringkasan abstraktif?
b. Bagaimana perbandingan performa dari model IndoBERT+GPT-2 *decoder* dan PEGASUS-*Large* dalam menghasilkan teks ringkasan abstraktif berdasarkan hasil skor metrik evaluasi ROUGE dan BERTScore?

1.5 Tujuan Penelitian

Adapun beberapa tujuan yang peneliti targetkan dalam penelitian ini adalah sebagai berikut.

a. Melakukan implementasi pada model IndoBERT+GPT-2 *decoder* dan PEGASUS-*Large* dalam menghasilkan teks ringkasan abstraktif.
b. Membandingkan performa dari model IndoBERT+GPT-2 *decoder* dan PEGASUS-*Large* dalam menghasilkan teks ringkasan abstraktif berdasarkan hasil skor metrik evaluasi ROUGE dan BERTScore.

1.6 Manfaat Hasil Penelitian

Adapun manfaat yang diharapkan dapat dicapai melalui penelitian ini adalah sebagai berikut.

a. Manfaat Teoritis

Harapan dari dilakukannya penelitian ini adalah untuk ikut berkolaborasi dan memperluas khazanah ilmu pengetahuan dalam bidang *Natural Language Processing* (NLP) khususnya dalam tugas peringkasan teks berbahasa Indonesia.

b. Manfaat Praktis

Memberikan kemudahan pekerjaan pengguna, dalam kasus ini adalah menghasilkan ringkasan teks abstraktif dari teks asli dengan bantuan kedua model *pre-trained* IndoBERT dan PEGASUS.