# BAB I
# PENDAHULUAN

## 1.1 Latar Belakang Masalah

Perkembangan jurnal ilmiah terus mengalami peningkatan secara signifikan, baik di tingkat nasional maupun internasional. Berbagai bidang keilmuan menghasilkan ribuan publikasi setiap tahunnya, mencerminkan pesatnya perkembangan penelitian dan inovasi. Berdasarkan data dari Scimago Journal & Country Rank, Indonesia menempati peringkat ke-38 dunia dengan jumlah publikasi sebanyak 376.908 artikel ilmiah dalam rentang tahun 1996–2023 (SCImago, n.d.). Tingginya angka publikasi ini menciptakan tantangan bagi akademisi, peneliti, dan mahasiswa dalam memahami literatur ilmiah yang terus bertambah.

Salah satu tantangan utama dalam memahami jurnal ilmiah adalah kompleksitas struktur teks dan penggunaan bahasa akademik yang padat serta penuh dengan istilah teknis. Jurnal ilmiah umumnya memiliki format standar yang terdiri dari abstrak, pendahuluan, metode penelitian, hasil, pembahasan, serta kesimpulan. Untuk memperoleh informasi yang relevan, seorang peneliti harus membaca dan memahami keseluruhan isi jurnal, yang sering kali memerlukan waktu yang lama dan usaha yang besar.

Secara umum, akademisi sering kali mengatasi tantangan ini dengan melakukan peringkasan secara manual, yakni dengan membaca keseluruhan teks dan menuliskan ulang poin-poin penting dalam bentuk ringkasan yang lebih singkat. Namun, metode ini memiliki keterbatasan yang signifikan. Pertama, proses membaca dan memahami jurnal ilmiah membutuhkan waktu yang cukup lama, terutama ketika harus membandingkan banyak jurnal sekaligus untuk mendapatkan informasi yang sesuai dengan kebutuhan penelitian. Kedua, tingkat kesulitan dalam memahami istilah teknis dan gaya bahasa akademik yang berbeda di setiap jurnal dapat memperlambat proses peringkasan. Dengan semakin banyaknya publikasi yang harus dikaji dalam berbagai disiplin ilmu, metode manual menjadi tidak

efisien dalam praktiknya. Oleh karena itu, dibutuhkan solusi otomatis yang dapat membantu dalam menyederhanakan proses pemahaman jurnal tanpa mengurangi esensi informasinya. Dalam beberapa tahun terakhir, teknologi kecerdasan buatan (Artificial Intelligence - AI) telah digunakan dalam berbagai bidang, termasuk dalam pemrosesan bahasa alami (Natural Language Processing - NLP). Salah satu aplikasi NLP yang potensial untuk mengatasi masalah ini adalah text summarization, yaitu teknik yang memungkinkan sistem untuk meringkas teks secara otomatis dengan tetap mempertahankan informasi inti dari dokumen.

Berbagai penelitian telah dilakukan untuk mengembangkan sistem text summarization. Namun, mayoritas penelitian sebelumnya lebih berfokus pada data berita atau artikel populer, bukan jurnal ilmiah yang memiliki struktur lebih kompleks dan bahasa yang lebih formal. Selain itu, banyak sistem yang ada masih menggunakan pendekatan ekstraktif, yaitu metode yang hanya mengambil kalimat-kalimat penting dari teks asli tanpa melakukan parafrase atau analisis lebih lanjut. Meskipun pendekatan ini dapat menghasilkan ringkasan dengan cepat, sering kali hasilnya kurang koheren dan terasa seperti kumpulan kutipan yang tidak menyatu dengan baik.

Sejumlah penelitian sebelumnya telah mengembangkan sistem peringkasan otomatis berbasis pendekatan ekstraktif. Sebagai contoh, Halimah et al., mengembangkan sistem peringkas otomatis menggunakan metode LexRank pada data artikel dengan berbagai topik, mulai dari agama hingga teknologi. Sementara itu, Saputra dan Al Maki menggunakan pendekatan abstraktif berbasis Long Short-Term Memory (LSTM) pada data artikel berita daring berbahasa Indonesia yang mencakup kategori seperti hiburan, olahraga, dan teknologi.

Namun, terdapat beberapa keterbatasan utama dalam penelitian sebelumnya. Pertama, sebagian besar penelitian masih berfokus pada artikel berita atau teks umum, bukan pada jurnal ilmiah yang memiliki struktur lebih kompleks dan membutuhkan pemahaman konteks yang lebih dalam. Kedua, banyak penelitian masih mengandalkan metode ekstraktif, yang sering kali menghasilkan ringkasan kurang koheren dan terasa seperti kumpulan kalimat yang dipilih dari teks asli. Ketiga, sebagian besar model dievaluasi hanya menggunakan metrik

ROUGE, yang mengukur kesamaan antara ringkasan model dengan referensi berdasarkan kemiripan kata atau frasa, tanpa mempertimbangkan aspek keterbacaan dan koherensi dari ringkasan yang dihasilkan.

Penelitian ini bertujuan untuk mengatasi keterbatasan tersebut dengan mengembangkan model text summarization berbasis pendekatan abstraktif menggunakan arsitektur Text-to-Text Transfer Transformer (T5). Model ini akan di *fine-tuning* secara khusus pada dataset jurnal ilmiah agar dapat menghasilkan ringkasan yang lebih koheren, informatif, dan sesuai dengan gaya bahasa akademik. Dengan memanfaatkan model berbasis transformer, penelitian ini diharapkan mampu menghasilkan ringkasan yang lebih alami dan mudah dipahami dibandingkan pendekatan sebelumnya.

Selain itu, dalam tahap evaluasi, penelitian ini tidak hanya berfokus pada pengukuran performa model menggunakan metrik ROUGE, tetapi juga akan mempertimbangkan aspek keterbacaan dan kekoherenan ringkasan. Dengan pendekatan ini, penelitian ini diharapkan dapat memberikan kontribusi bagi akademisi, peneliti, dan mahasiswa dalam memahami jurnal ilmiah dengan lebih efisien.

Dengan adanya model *summarization* yang lebih canggih dan spesifik untuk jurnal ilmiah, akademisi dapat menghemat waktu dalam memahami literatur ilmiah, mempercepat proses penelitian, serta memperoleh wawasan yang lebih luas dengan cara yang lebih sistematis. Selain berkontribusi dalam dunia akademik, penelitian ini juga memberikan sumbangan bagi bidang NLP dalam pengembangan model peringkasan teks yang lebih efektif untuk dokumen akademik yang kompleks.

## 1.2 Identifikasi Masalah

Berdasarkan latar belakang yang telah dipaparkan sebelumnya, terdapat beberapa permasalahan utama yang dapat diidentifikasi pada penelitian ini:

1. Peneliti dan akademisi kesulitan dalam menyaring informasi penting dari banyaknya jurnal ilmiah yang tersedia, terutama karena penggunaan bahasa akademik yang kompleks dan istilah teknis yang membingungkan.

2. Peringkasan jurnal ilmiah secara manual memerlukan waktu yang lama dan tidak efisien, terutama ketika harus menganalisis banyak jurnal sekaligus, sementara pendekatan ekstraktif yang umum digunakan cenderung menghasilkan ringkasan yang tidak koheren dan tidak informatif.
3. Diperlukan model text summarization berbasis abstraktif yang mampu memahami konteks jurnal ilmiah dan menghasilkan ringkasan yang lebih koheren, informatif, dan sesuai dengan gaya bahasa akademik yang lebih mudah dipahami.

## 1.3 Batasan Masalah

1. Penelitian ini berfokus pada pembangunan dan evaluasi model text summarization menggunakan arsitektur T5 serta implementasi model kedalam bentuk website. Model yang dikembangkan akan di fine-tune menggunakan dataset jurnal ilmiah.
2. Dataset yang digunakan terdiri dari artikel jurnal ilmiah dari berbagai bidang, termasuk sains, teknologi, dan sosial. Artikel dapat memuat angka, data statistik, maupun notasi matematis. Namun, untuk menjaga konsistensi pemrosesan, ekspresi atau rumus matematis kompleks tidak diolah secara langsung oleh model melainkan ditangani dengan teknik *masking* atau diberi penanda khusus.
3. Evaluasi yang digunakan menggunakan metrik ROUGE, yang berfokus pada kesamaan antara ringkasan yang dihasilkan oleh model dan ringkasan referensi dalam hal kesamaan leksikal.

## 1.4 Rumusan Masalah

Berdasarkan latar belakang, identifikasi masalah, serta batasan masalah yang telah dijelaskan sebelumnya, penelitian ini berusaha menjawab pertanyaan-pertanyaan berikut:

1. Apa saja atribut dalam dataset jurnal ilmiah yang digunakan untuk melatih model text summarization berbasis arsitektur T5?

2. Bagaimana membangun dan menyesuaikan arsitektur model T5 agar mampu menghasilkan ringkasan jurnal ilmiah yang koheren dan mudah dipahami?
3. Bagaimana kinerja model T5 dalam menghasilkan ringkasan jurnal ilmiah berdasarkan evaluasi menggunakan metrik ROUGE?

## 1.5 Tujuan Penelitian

Adapun tujuan yang ingin dicapai pada penilitian ini adalah sebagai berikut:

1. Mengidentifikasi atribut dalam dataset jurnal ilmiah yang digunakan untuk melatih model text summarization berbasis arsitektur T5.
2. Membangun dan menyesuaikan arsitektur model T5 agar mampu menghasilkan ringkasan jurnal ilmiah yang koheren dan mudah dipahami.
3. Mengevaluasi kinerja model T5 dalam menghasilkan ringkasan jurnal ilmiah berdasarkan metrik ROUGE.

## 1.6 Manfaat Penelitian

Adapun manfaat yang diharapkan pada penelitian ini adalah sebagai berikut:

### 1.6.1 Manfaat Teoritis

Secara teori, penelitian ini diharapkan berkontribusi dan menjadi referensi dalam bidang *Natural Language Processing* terkhuss pada penerapan model T5 dalam text summarization untuk jurnal ilmiah

### 1.6.2 Manfaat Praktis

a. Bagi peneliti

Penelitian ini diharapkan dapat meningkatkan pengetahuan terkait *Natural Language Processing* serta lebih dalam mengenai *Transformer* khususnya model *Text-to-Text Transfer Transformer* (T5). Selain itu, penelitian ini diharapkan sebagai tugas akhir untuk mendapatkan gelar Sarjana Komputer (S.Kom).

b. Bagi Pembaca

Hasil penelitian ini diharapkan dapat membantu pembaca dalam memahami jurnal ilmiah dengan lebih cepat melalui ringkasan yang dihasilkan oleh model T5.