BAB I

PENDAHULUAN

1.1 Latar Belakang

Dalam beberapa tahun terakhir, dunia akademik mengalami lonjakan besar dalam jumlah publikasi ilmiah. Setiap harinya, ribuan artikel dari berbagai disiplin ilmu diterbitkan, menciptakan tantangan tersendiri dalam proses pencarian dan pengelolaan informasi yang relevan [1]. Di tengah derasnya arus informasi ini, kata kunci (keywords) memainkan peran penting sebagai pintu masuk untuk memahami isi artikel secara cepat [2]. Kata kunci yang tepat dan representatif tidak hanya membantu pembaca menemukan literatur yang sesuai, tetapi juga sangat menentukan apakah sebuah artikel dapat terindeks dengan baik di repositori digital atau mesin pencari ilmiah [3].

Sayangnya, hingga saat ini proses pembangkitan kata kunci masih banyak dilakukan secara manual. Penulis diminta menentukan sendiri kata kunci artikel mereka, yang tentu saja sangat bergantung pada persepsi pribadi [4]. Akibatnya, sering terjadi inkonsistensi antara kata kunci yang dipilih dengan isi sebenarnya dari artikel tersebut. Praktik ini cukup umum ditemukan, terutama pada jurnal nasional atau sistem repository institusi pendidikan tinggi yang belum memiliki dukungan teknologi otomatisasi yang memadai [5].

Sementara itu, pendekatan otomatis seperti TF-IDF memang sudah lama digunakan, namun masih memiliki keterbatasan. Metode ini cenderung mengangkat kata-kata yang paling sering muncul, tanpa memahami konteks atau makna yang lebih dalam dari teks [6]. Padahal dalam banyak kasus, kata kunci yang paling penting justru tidak selalu muncul secara eksplisit dalam abstrak artikel [7].

Untuk mengatasi masalah ini, berbagai metode evaluasi juga dikembangkan. Salah satu pendekatan sederhana yang umum digunakan adalah Jaccard Similarity, yang menghitung kesamaan berbasis tumpang tindih kata (lexical overlap) [8]. Namun, Jaccard hanya mengukur kesamaan bentuk kata, sehingga kurang mampu menangkap makna semantik yang lebih dalam [9]. Sebaliknya, pendekatan berbasis model seperti BERT dapat mengukur kesamaan makna antar frasa, bahkan ketika bentuk katanya berbeda [10]. Oleh karena itu, membandingkan peran Jaccard dan BERT sebagai filter hasil pembangkitan kata kunci menjadi relevan, terutama untuk memahami sejauh mana evaluasi berbasis leksikal dan semantik memberikan hasil yang berbeda.

Kemajuan teknologi pemrosesan bahasa alami (Natural Language Processing/NLP), khususnya dengan hadirnya model-model berbasis transformer, membuka peluang baru untuk mengatasi masalah ini [11]. Model seperti T5 [12], BART [13], dan BERT [14] kini mampu menangkap makna dan struktur kalimat dengan lebih baik. Tidak hanya sekadar menghitung frekuensi kata, model-model ini belajar dari konteks dan hubungan antar kata dalam kalimat, sehingga lebih cermat dalam menyarikan informasi penting, termasuk untuk tugas pembangkitan kata kunci.

Dalam penelitian ini, penulis mencoba mengeksplorasi bagaimana model T5 dan gabungan BART dengan BERT dapat dimanfaatkan untuk menghasilkan kata kunci secara otomatis dari abstrak artikel ilmiah. T5 dipilih karena kemampuannya dalam mengubah berbagai tugas NLP menjadi format text-to-text yang fleksibel, sementara BART dikenal efektif dalam menghasilkan teks baru berdasarkan pemahaman konteks, dan BERT digunakan untuk menyaring serta mengevaluasi relevansi hasil tersebut. Kombinasi model generatif dan evaluatif ini diharapkan mampu menjawab tantangan dalam pembangkitan kata kunci yang tidak hanya akurat, tetapi juga efisien.

Penelitian ini juga berupaya menempatkan pendekatan tersebut dalam konteks yang lebih nyata. Sistem yang dikembangkan nantinya diarahkan untuk dapat diterapkan pada jurnal-jurnal nasional, repository kampus, atau manajemen metadata publikasi ilmiah yang masih mengandalkan proses manual. Dengan hadirnya sistem otomatis yang cerdas, proses indexing dapat menjadi lebih cepat, artikel lebih mudah ditemukan, dan pengelolaan repositori menjadi lebih efisien.

Hingga saat ini, belum banyak studi yang secara langsung membandingkan pendekatan T5 dengan gabungan BART dan BERT dalam konteks pembangkitan kata kunci ilmiah. Padahal, perbandingan sistematis antara kedua pendekatan ini sangat penting untuk menjawab kebutuhan nyata dalam dunia akademik. Dari sisi teoritis, penelitian semacam ini berkontribusi terhadap pengembangan ilmu di bidang Natural Language Processing (NLP), khususnya dalam eksplorasi model transformer untuk tugas ekstraksi informasi. Dari sisi praktis, hasil penelitian dapat mendukung pengelolaan repositori ilmiah dan sistem indeksasi jurnal agar lebih efisien, akurat, dan mudah diakses. Dengan demikian, penelitian ini diharapkan mampu menjembatani kesenjangan antara kebutuhan akademik dan penerapan teknologi cerdas dalam manajemen publikasi ilmiah.

1.2 Rumusan Masalah

Meskipun pembangkitan kata kunci otomatis sudah mulai banyak diterapkan dalam bidang NLP, masih terdapat kesenjangan dalam hal efektivitas dan efisiensi, terutama ketika dihadapkan pada kebutuhan sistem indeksasi jurnal berskala besar. Dua pendekatan populer, yaitu model T5 dan kombinasi BART dengan BERT, menawarkan solusi berbasis transformer yang menjanjikan, namun belum banyak penelitian yang secara langsung membandingkan keduanya dalam konteks ilmiah multidisipliner. Oleh karena itu, penelitian ini bertujuan untuk menjawab pertanyaan inti: bagaimana performa model T5 dibandingkan dengan kombinasi BART+BERT dalam membangkitkan kata kunci ilmiah secara otomatis, baik dari segi kualitas hasil (akurat, relevan, dan kontekstual), maupun dari segi efisiensi proses?

1.3 Tujuan

Penelitian ini bertujuan untuk membandingkan performa model T5 dan kombinasi BART+BERT dalam menghasilkan kata kunci otomatis dari abstrak artikel ilmiah. Perbandingan dilakukan berdasarkan metrik evaluasi seperti akurasi, precision, recall, dan F1-score, sekaligus menilai kualitas hasil pembangkitan kata kunci dalam menangkap istilah non-eksplisit serta bentuk bigram atau multigram yang kontekstual. Selain itu, penelitian ini juga menganalisis efisiensi waktu proses (runtime) dari kedua pendekatan, serta mengidentifikasi pengaruh karakteristik teks, seperti panjang abstrak dan distribusi kata kunci, terhadap efektivitas masing-masing model. Dengan demikian, penelitian ini diharapkan dapat memberikan dasar yang kuat bagi pengembangan sistem rekomendasi pembangkitan kata kunci otomatis yang lebih akurat dan efisien untuk mendukung pengelolaan repositori ilmiah.

1.4 Manfaat

Manfaat dari penelitian ini adalah sebagai berikut:

1. Memberikan wawasan akademik dan praktis mengenai keunggulan dan keterbatasan pendekatan T5 dan BART+BERT dalam tugas pembangkitan kata kunci dari abstrak ilmiah.
2. Mendukung pengembangan sistem indexing otomatis yang lebih akurat dan efisien, terutama untuk institusi yang memiliki keterbatasan sumber daya dalam proses tagging dan pengelolaan metadata artikel.

3. Menawarkan alternatif pendekatan pembangkitan kata kunci yang adaptif untuk berbagai jenis teks ilmiah, termasuk artikel dengan topik multidisipliner dan format abstrak yang bervariasi.
4. Menjadi referensi dalam pengambilan keputusan terkait pemilihan model NLP pada sistem manajemen publikasi ilmiah di tingkat institusi, nasional, maupun internasional.

1.5 Ruang Lingkup

Untuk menjaga fokus dan kedalaman analisis, penelitian ini dibatasi pada ruang lingkup berikut:

1. Cakupan Data: Penelitian hanya menggunakan bagian abstrak dan kata kunci dari artikel ilmiah yang dipublikasikan dalam jurnal internasional bereputasi. Total data sebanyak 100 artikel dari berbagai bidang ilmu akan digunakan.
2. Format Dataset: Dataset disusun dalam format Abstrak – Kata Kunci sebagai acuan pembelajaran dan evaluasi. Kata kunci asli dari artikel digunakan sebagai ground truth untuk evaluasi model.
3. Ruang Aplikasi: Fokus penelitian adalah pada potensi penerapan sistem di lingkungan repository institusi, jurnal nasional, dan sistem manajemen publikasi ilmiah internal kampus, bukan platform komersial besar seperti Scopus.
4. Metodologi: Penelitian ini menerapkan model T5 (generatif murni) dan kombinasi BART (sebagai generator) + BERT (sebagai ranker atau filter), yang masing-masing dievaluasi berdasarkan performa dan efisiensi waktu.
5. Batasan Analisis: Penelitian tidak mencakup fine-tuning lanjutan terhadap model di luar pretraining yang tersedia, serta tidak menganalisis aspek linguistik non-teknis seperti struktur wacana artikel.