202210370311299
Muhammad Eka Nur Arief
Prodi Informatika

BAB I
PENDAHULUAN

1.1. Latar Belakang

Penelitian ilmiah merupakan instrumen fundamental dalam kemajuan peradaban manusia, yang memungkinkan terjadinya penemuan-penemuan inovatif di berbagai disiplin ilmu guna mendorong pembangunan global. Namun, pada era informasi saat ini, dunia akademik menghadapi fenomena ledakan data di mana jumlah publikasi ilmiah tumbuh secara eksponensial setiap tahunnya. Pertumbuhan volume literatur yang masif ini menimbulkan hambatan serius bagi para peneliti untuk mengidentifikasi tema-tema baru, melacak perkembangan tren riset yang sedang berkembang, serta menentukan prioritas kajian secara akurat dan efisien [1], [2], [3]. Proses tinjauan literatur yang dilakukan secara manual, yang sebelumnya menjadi standar riset, kini dianggap tidak lagi efisien dan rentan terhadap bias subjektivitas manusia ketika harus berhadapan dengan jutaan artikel yang terbit setiap tahunnya.

Kondisi tersebut menciptakan urgensi bagi pengembangan metode analisis otomatis berbasis Natural Language Processing (NLP) dan penambangan teks untuk melakukan penemuan pengetahuan (knowledge discovery) secara kuantitatif dan prediktif [4]. Tinjauan literatur otomatis diperlukan untuk memetakan arah penelitian masa depan secara komprehensif [5], [6]. Untuk mencapai hasil analisis yang akurat, tahapan awal pra-pemrosesan data menjadi sangat krusial. Tahapan seperti tokenisasi—yang memecah kalimat menjadi unit kata yang granular—menjadi langkah awal yang menentukan keberhasilan model dalam memahami pola bahasa yang kompleks [7], [8]. Selain itu, penggunaan teknik lematisasi juga berperan penting untuk mereduksi variasi kata menjadi bentuk dasarnya, sehingga model dapat menangkap hubungan semantik antar kata dengan lebih konsisten [9].

Dalam sejarahnya, ekstraksi informasi dari teks telah mengalami evolusi dari pendekatan statistik tradisional menuju paradigma baru yang berbasis deep learning [10], [11]. Metode klasik seperti Latent Dirichlet Allocation (LDA) cenderung memiliki keterbatasan karena mengandalkan representasi frekuensi kata (bag-of-words) yang mengabaikan konteks dan urutan kata dalam kalimat [12], [13]. Sebaliknya, munculnya arsitektur transformer telah membawa perubahan signifikan melalui penggunaan contextual embeddings yang mampu memahami nuansa makna sebuah kata berdasarkan konteks sekitarnya secara lebih mendalam [14], [15]. Salah satu rujukan utama dalam domain ini adalah penelitian oleh Wijanto et al. yang memberikan tolok ukur (benchmark) berharga mengenai penggunaan model berbasis transformer yang dioptimalkan untuk artikel ilmiah [16].

Saat ini, terdapat berbagai strategi dalam menerapkan teknologi transformer untuk ekstraksi topik. Strategi pertama adalah membangun pipeline modular kustom yang memberikan kontrol penuh pada setiap tahapannya, seperti penggunaan SBERT untuk representasi vektor dan UMAP untuk reduksi dimensi [17], [18]. Strategi kedua adalah menggunakan kerangka kerja terintegrasi seperti BERTopic yang mengautomasi interaksi antara komponen-komponen tersebut dan memperkenalkan mekanisme class-based TF-IDF (c-TF-IDF) untuk mengekstraksi kata kunci yang paling deskriptif [19]. Pemilihan komponen modular dibandingkan pendekatan terintegrasi diyakini memberikan dampak signifikan pada koherensi topik yang dihasilkan [18], [19], [20].

Meskipun model terintegrasi seperti BERTopic menawarkan kemudahan, perbandingan kinerjanya dengan pipeline modular kustom pada korpus literatur ilmiah masih memerlukan kajian mendalam. Penelitian ini hadir untuk melakukan analisis komparatif yang komprehensif antara dua pendekatan berbeda: pipeline kustom SBERT-UMAP-HDBSCAN dan model terintegrasi BERTopic. Analisis dilakukan menggunakan metrik koherensi topik C_v untuk menentukan pendekatan yang paling optimal bagi komunitas akademik.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka permasalahan utama dalam penelitian ini dapat dirumuskan dalam bentuk pertanyaan-pertanyaan berikut:

a. Bagaimana perbandingan kinerja koherensi topik, yang diukur dengan metrik C_v, antara pipeline kustom SBERT-UMAP-HDBSCAN dan model terintegrasi BERTopic pada korpus abstrak literatur ilmiah?
b. Di antara pipeline yang diuji, manakah yang mampu menghasilkan nilai koherensi topik (C_v) tertinggi dan dapat dianggap sebagai pendekatan paling optimal untuk analisis tematik pada dataset yang digunakan?
c. Bagaimana pengaruh kombinasi spesifik dari model embedding (SBERT), teknik reduksi dimensi (UMAP), dan algoritma clustering (HDBSCAN) terhadap kualitas dan jumlah topik yang dihasilkan?

1.3. Tujuan Penelitian

Sejalan dengan rumusan masalah di atas, tujuan dari penelitian ini adalah sebagai berikut:

a. Menganalisis dan membandingkan secara kuantitatif skor koherensi topik (C_v) yang dicapai oleh dua pipeline topic modeling berbasis transformer yang berbeda.
b. Mengidentifikasi dan merekomendasikan pipeline pemodelan topik yang paling efektif dan unggul untuk menghasilkan topik yang sangat koheren dari kumpulan abstrak penelitian ilmiah.
c. Mengevaluasi dan memberikan pemahaman praktis mengenai kelebihan serta keterbatasan dari setiap kombinasi komponen (embedding, reduksi dimensi, dan clustering) dalam tugas pemodelan topik pada teks ilmiah.

1.4. Batasan Masalah

Adapun batasan-batasan dalam penelitian ini adalah sebagai berikut:

a. Penelitian ini berfokus pada analisis komparatif dua pipeline topic modeling berbasis transformer untuk mengevaluasi koherensi topik yang dihasilkan.
b. Dataset yang digunakan adalah korpus publik berbahasa Inggris dari Kaggle, berisi 20.972 judul dan abstrak artikel ilmiah dari berbagai bidang keilmuan.
c. Kinerja dan efektivitas pipeline diukur secara kuantitatif hanya berdasarkan skor koherensi C_v.
d. Implementasi dan eksperimen dilakukan menggunakan bahasa pemrograman Python pada google colab dengan memanfaatkan pustaka utama seperti bertopic, sentence-transformers, scikit-learn, dan nltk.

1.5. Jadwal Pengerjaan

Proses penelitian dan penyusunan Laporan Tugas Akhir ini dilaksanakan sesuai dengan jadwal pengerjaan yang dirangkum dalam tabel di bawah ini.

Tabel 1.1 Jadwal Pengerjaan Tugas Akhir

No.	Kegiatan	Bulan 1	Bulan 2	Bulan 3	Bulan 4	Bulan 5
1	Studi Literatur dan Perumusan Masalah					
2	Pengumpulan dan Pra-pemrosesan Dataset					
3	Implementasi dan Eksperimen Pipeline					
4	Analisis Hasil dan Pembahasan					
5	Penulisan Laporan Tugas Akhir					

6	Bimbingan dan Revisi					