202210370311299
Muhammad Eka Nur Arief
Prodi Informatika

BAB II
TINJAUAN PUSTAKA

2.1. Topic Modelling

Pemodelan topik adalah sebuah teknik statistik dan unsupervised machine learning yang dirancang untuk menganalisis sekumpulan dokumen teks dan menemukan tema-tema abstrak, yang dikenal sebagai "topik," yang terkandung di dalamnya [4], [5], [6]. Tujuan utamanya adalah untuk secara otomatis mengorganisasi, memahami, dan merangkum korpus teks yang besar tanpa memerlukan intervensi manual atau data berlabel sebelumnya. Dalam praktiknya, setiap topik direpresentasikan sebagai sebuah distribusi probabilitas atas sejumlah kata, di mana kata-kata dengan probabilitas tertinggi dalam sebuah topik adalah kata-kata yang paling representatif untuk tema tersebut [5]. Pemodelan topik telah menjadi alat analisis data tekstual yang krusial dalam berbagai domain, mulai dari analisis media sosial hingga penambangan literatur ilmiah [6].

2.2. Evolusi Topic Modelling

Metodologi pemodelan topik telah mengalami evolusi yang signifikan. Pendekatan klasik generasi awal seperti Latent Semantic Analysis (LSA) dan Latent Dirichlet Allocation (LDA) beroperasi di bawah asumsi representasi bag-of-words (BoW), di mana dokumen dianggap sebagai kumpulan kata tanpa memperhatikan urutan atau struktur gramatikalnya [10], [11]. Meskipun efektif pada masanya, pendekatan berbasis BoW memiliki keterbatasan fundamental dalam menangkap makna kontekstual dan gagal menangkap hubungan semantik yang halus, seperti pada kasus kata yang memiliki banyak makna [12]. Hal ini dipertegas oleh Kalepalli dkk. yang menunjukkan bahwa meskipun LDA memiliki keunggulan dalam kecepatan, ia seringkali memiliki keterbatasan dalam menjaga konsistensi topik jika dibandingkan dengan metode berbasis semantik [21].

Sebuah perubahan paradigma ditandai oleh kemunculan deep learning dan arsitektur transformer [13]. Model-model berbasis transformer

memperkenalkan contextual embeddings, di mana representasi sebuah kata dibentuk berdasarkan konteks kata-kata di sekitarnya. Hal ini memungkinkan mesin untuk menangkap nuansa makna yang jauh lebih dalam, sehingga menghasilkan representasi semantik yang lebih kaya dibandingkan pendekatan statistik tradisional [13], [14]. Pemanfaatan embedding berbasis transformer kini menjadi fondasi utama bagi pengembangan pipeline pemodelan topik modern yang lebih koheren dan akurat [15]. Seiring perkembangan tersebut, evolusi pemodelan topik terus berlanjut ke arah model hibrida yang mencoba mengkombinasikan keunggulan struktur statistik dengan kekuatan representasi neural guna menutupi kelemahan yang ada pada model generasi sebelumnya [22].

2.3. Komponen Inti dalam Pipeline Topic Modeling Modern

Dalam pemodelan topik tingkat lanjut, proses ekstraksi informasi biasanya melibatkan tiga komponen utama yang saling berinteraksi:

1. Document Embedding: Tahap ini mengubah dokumen teks menjadi vektor padat (dense vectors) yang merepresentasikan makna semantik dokumen tersebut. Model seperti SBERT (Sentence-BERT) menggunakan struktur siamese network untuk menghasilkan vektor tingkat kalimat yang efisien [14]. Selain itu, model RoBERTa juga sering digunakan karena prosedur pra-pelatihannya yang lebih optimal dalam menangkap representasi teks akademik [16].
2. Reduksi Dimensi: Vektor yang dihasilkan oleh model transformer umumnya memiliki dimensi yang sangat tinggi (misalnya 768 dimensi), yang dapat menghambat efisiensi algoritma klasterisasi. Teknik linear seperti Principal Component Analysis (PCA) bekerja dengan memaksimalkan varians data [16], namun teknik non-linear seperti Uniform Manifold Approximation and Projection (UMAP) sering kali dianggap lebih unggul karena kemampuannya dalam mempertahankan struktur lokal dan global data dalam ruang dimensi rendah [17].

3. Klasterisasi: Setelah dimensi data dikurangi, algoritma klasterisasi diterapkan untuk mengelompokkan dokumen dengan kemiripan semantik. Algoritma berbasis pusat seperti K-Means memerlukan penentuan jumlah klaster (K) di awal [16], sementara algoritma berbasis kepadatan seperti HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) mampu menemukan klaster dengan berbagai bentuk tanpa perlu menentukan jumlahnya terlebih dahulu, serta mampu mengidentifikasi data pencilan sebagai noise [18], [19].

Penggunaan kombinasi komponen-komponen di atas menjadi sangat krusial terutama saat berhadapan dengan korpus abstrak ilmiah. Karakteristik abstrak yang memiliki jumlah kata terbatas sering kali menimbulkan masalah kelangkaan fitur (sparsity), sehingga diperlukan teknik pemodelan topik khusus yang mampu menangkap hubungan semantik secara mendalam meskipun dalam teks yang pendek dan padat [23].

2.4. BERTopic: Integrated Pipeline

BERTopic merupakan sebuah kerangka kerja pemodelan topik modern yang secara sistematis mengintegrasikan komponen transformer embeddings, reduksi dimensi UMAP, dan klasterisasi HDBSCAN ke dalam satu pipeline yang kohesif [19]. Keunggulan utama BERTopic terletak pada mekanisme uniknya yang disebut class-based TF-IDF (c-TF-IDF). Mekanisme ini memperlakukan seluruh dokumen dalam satu klaster sebagai satu entitas besar untuk mengekstraksi kata-kata kunci yang paling representatif bagi topik tersebut. Kombinasi ini membuat BERTopic sangat fleksibel dan mampu menghasilkan struktur tematik yang sangat koheren dan mudah diinterpretasikan, bahkan pada korpus teks yang kompleks seperti literatur ilmiah [19], [20].

2.5. Penelitian Terkait

Penelitian rujukan utama dalam studi ini adalah karya Wijanto et al. [16]. Dalam penelitian tersebut, dilakukan eksplorasi mendalam untuk menemukan penyetelan hyperparameter yang optimal dalam pemodelan topik pada artikel ilmiah menggunakan model berbasis transformer. Wijanto et al. membangun sebuah pipeline standar menggunakan kombinasi RoBERTa untuk embedding, PCA untuk reduksi dimensi, dan K-Means untuk klasterisasi. Hasil penelitian tersebut menjadi acuan penting bagi studi ini, khususnya dalam mereplikasi arsitektur tersebut sebagai pembanding (benchmark) untuk mengevaluasi efektivitas pipeline modular lainnya.

Penelitian pendukung lainnya adalah studi oleh Shafiyah [24] dan Shin & Yang [25], yang secara konsisten menunjukkan bahwa model berbasis transformer seperti BERTopic memiliki kemampuan yang lebih dinamis dalam mendeteksi tren tematik dibandingkan metode tradisional seperti LDA. Keandalan BERTopic juga telah divalidasi pada domain spesifik, seperti pada identifikasi urgensi postingan forum pendidikan oleh Khodeir dan Elghannam [26] serta pemetaan tren riset dalam jurnal kesehatan oleh Madrid-García et al. [27].

Lebih lanjut, analisis komparatif oleh Zengul dkk. [28] terhadap berbagai metode seperti LDA, Top2Vec, dan LSA memberikan wawasan bahwa model berbasis embedding cenderung menghasilkan topik yang lebih padat. Fleksibilitas arsitektur berbasis BERT juga dibuktikan oleh Atagun dkk. [29] dalam menangani berbagai variasi bahasa. Selain itu, Kun dkk. [30] membuktikan bahwa penggunaan BERTopic untuk analisis bibliometrik mampu memberikan visualisasi yang lebih baik dalam memetakan potensi riset masa depan. Tantangan teknis dalam pemodelan topik neural juga banyak dibahas oleh Vayansky dan Kumar [5] serta Hankar dkk. [6], yang menegaskan bahwa efektivitas model sangat bergantung pada sinergi antar komponen pipeline. Selain itu, kualitas pra-pemrosesan data tetap menjadi faktor krusial dalam keberhasilan model deep learning sebagaimana ditekankan dalam studi Aditya dkk. [7], [8] dan Rodiyah [9]. Berdasarkan

rujukan-rujukan tersebut, penelitian ini memposisikan diri untuk mengevaluasi secara komprehensif berbagai kombinasi komponen pipeline guna menemukan solusi optimal bagi analisis literatur ilmiah berskala besar menggunakan dataset publik dari Kaggle [31].