BAB II
TINJAUAN PUSTAKA

Dengan meningkatnya kecanggihan kecerdasan buatan, semakin banyak penelitian yang membahas tentang bidang yang difokuskan pada perkembangan pemodelan pembelajaran yang umumnya dikenal dengan sebutan *deep learning* dan *machine learning*. AI *(artificial intelligence)* atau kecerdasan buatan merupakan bidang dari ilmu komputer diedikasikan guna pengembangan sistem yang dapat menyerupai cara berpikir manusia seperti logika, pengembangan pengetahuan, dan penyesuaian. Memudahkan peneliti untuk melakukan analisis data tingkat lanjut, termasuk tugas seperti kategorisasi, klasterisasi, regresi, perkiraan, dan visualisasi [6]. Sehingga menarik minat para peneliti dari berbagai belahan dunia untuk mengembangkan hal tersebut. Maka dari itu, penting menganalisis tren topik pada penelitian yang telah dilakukan supaya bisa mengetahui perkembangan penelitian di bidang ini.

2.1 Studi Literature

Menganalisis tren topik penelitian dengan menggunakan metode LDA dan BERTopic sudah diterapkan oleh beberapa penelitian. Sebelumnya, ada penelitian yang dilakukan oleh **Liu dan Wan (2024)** meneliti tren topik penelitian dari publikasi jurnal yang relevan dengan topik *precision agriculture* menggunakan BERTopic [7]. Hasil penelitian ini menunjukan bahwa nilai koherensi BERTopic 0.7. **Akan tetapi,** pada penelitian ini tidak mencakup semua penelitian dari daerah yang teknologinya masih terbatas, akibatnya hasil dari penelitian ini dapat terjadi bias. Lalu, ada penelitian yang dilakukan oleh **Atzeni et al (2022)** membahas *trend topic* penelitian tentang interaksi antara *WiFi* dan *Machine Learning* dari publikasi seperti scopus, *web of science,* dan IEEE Xplore menggunakan BERTopic menghasilkan 8 topik utama [8]. Selanjutnya, terdapat penelitian oleh **Erna dan Jailani (2024)** membahas tentang identifikasi topik penelitian judul skripsi teknik informatika di Universitas

Dipa Makassar dari tahun 2022 sampai bulan maret pada tahun 2024 menggunakan LDA [9]. Hasil penelitian ini menjelaskan 13 topik utama.

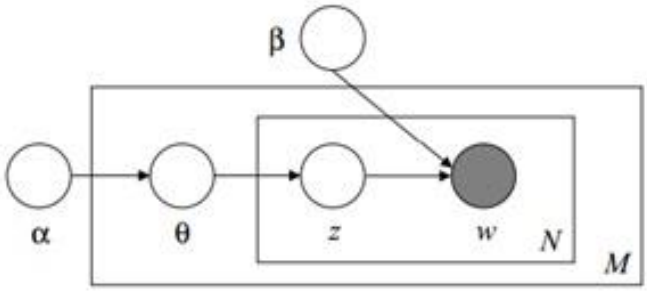
Selain itu, penelitian yang dilakukan oleh **Adewunmi et al (2022)** membahas identifikasi *trend topic* kesenjangan dalam penanganan penyakit kanker dari jurnal PubMed menggunakan BERTopic, dan hasil penelitian tersebut mendapatkan 32 topik [10]. Disisi lain, penelitian oleh **Sahria dan Fudholi (2020)** menganalisis tren peneltian pada bidang kesehatan dari publikasi jurnal Sinta menggunakan LDA, hasil analisis ini menunjukan bahwa pada topik ke 8 mendapat *coherence* paling tinggi dan penelitian ini juga dievaluasi dengan survei kepada narasumber seperti ilmuwan, petugas medis, ahli Pendidikan bahwa kebanyakan narasumber menilai pemodelan topik yang sudah dilakukan 94.1% baik sekali, dan 5.9% baik [11]. Tidak hanya itu, penelitian oleh **Gupta et al (2022)** mengkaji identifikasi tren penelitian dari *Journal of Applied Intelligence* menggunakan LDA dengan menerapkan BOW dan TF-IDF. Dari hasil penelitian ini, penggunaan teknik TF-IDF dan turning parameter bisa meningkatkan evaluasi coherence pada model LDA [12]. **Namun,** pada penelitian tersebut masih membutuhkan campur tangan manusia pada saat proses identifikasi topik, sehingga membuat proses lebih lama dan kemungkinan meningkatnya bias.

Oleh karena itu, dari penjelasan penelitian sebelumnya. Belum ada peneltian yang menganalisis tren topik penelitian dari Open Review, terutama ICLR (International Conference on Learning Representations) menggunakan topic modeliing LDA (Latent Dirichlet Allocation) dan BERTopic. Dengan demikian, penelitian ini diharapkan bisa membagikan perspektif baru tentang penelitian yang sering dibahas dari tahun 2019 sampai 2023.

2.2 Pembuatan Corpus & Dictionary

Pada proses pembuatan corpus dan pembuatan kamus (dictionary) untuk mempersingkat proses, hal ini berguna untuk mengumpulkan kata-kata unik lalu dihubungkan pada ID dan melihat frekuensi munculnya kata-kata pada dokumen yang biasanya dalam bentuk numerik [13].

2.3 Pemodelan LDA



Gambar 2. 1. Pemodelan LDA(sumber Blei et al, 2003)

LDA adalah model generatif yang menerapkan proses probabilistik untuk menghasilkan dokumen dari variabel laten atau tersembunyi. Metode ini juga digunakan untuk mengidentifikasi topik-topik pada teks, bisa dengan mudah dilatih secara daring atau *online* [15]. Gagasan inti dari metode tersebut bahwa setiap teks memiliki banyak topik, dan setiap kata pada teks yang muncul mempunyai probabilitas tertentu yang terkait dengan salah satu topik [16].

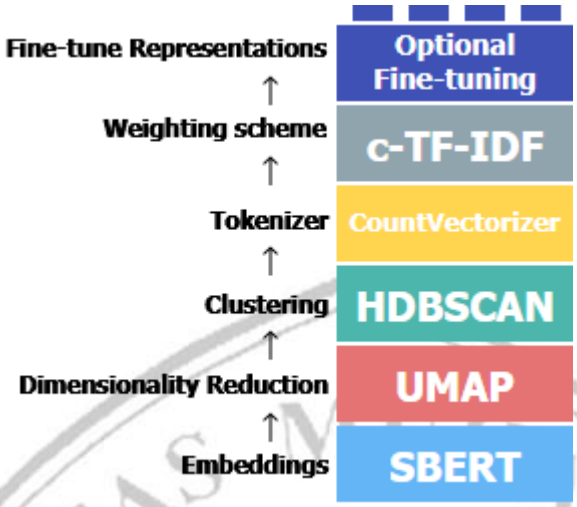
Proses LDA dijabarkan pada berikut ini :

- D merupakan banyaknya dokumen yang dimiliki
- N merupakan frekuensi setiap kata pada dokumen
- Masing-masing topik dilambangkan $\beta 1{:}K$, dan pada βk menjabarkan distribusi tiap kata dalam topik
- Dokumen ke -d, skala topik dilambangkan dengan θd
- $\theta d{,}k$ merupakan frekuensi topik K muncul pada dokumen d
- $wd{,}n$ merupakan kata atau frasa ke-n pada dokumen d
- Wd merupakan setiap kata pada dokumen ke - d, dimana kata atau frasa ke- n pada dokomen dilambangkan $wd{,}n$
- α dan η adalah parameter yang diterapkan pada model LDA

  Berikut persamaan yang digunakan [17], [18]:

  $$p(w, z, \theta, \beta | \alpha\, \eta) = p\,(\beta|\eta)p(\theta|\alpha)p(z|\theta)p(w|\beta_k) \quad (1)$$

2.4 Implementasi BERTopic



Gambar 2. 2. Pemodelan BERTopic

Topic modelling yang bisa digunakan untuk analisis adalah BERTopic, yaitu kerangka pemodelan topik yang menerapkan *Bi-directional Encoder Representation transformer* terhadap pembentukan topik. Meskipun demikian, pemodelan topik ini tidak semata-mata bergantung pada BERT [19]. BERTopic menyusun topik dari sekumpulan dokumen melalui tiga tahap yang beragam. Awalnya, merubah dokumen menjadi representasi embedding. Selanjutnya, mengimplementasikan penyederhanaan dimensi pada embedding dan menggolongkan kedalam cluster. Setelah itu, pengambilan topik menerapkan hasil modifikasi dari TF-IDF [5]. Model ini juga bisa digunakan lebih dari 50 bahasa [20]. Berikut adalah tahapan implementasi BERTopic yang digunakan pada penelitian ini :

2.3.1. Embedding & UMAP

Langkah pertama dalam implementasi BERTopic yaitu dilakukan embedding untuk merubah dokumen menjadi numerik [5]. Proses embedding dengan menggunakan Sentence Transformer yaitu all-MiniLM-L6-v2 sebab kemampuannya dapat memberikan keseimbangan yang bagus antara performa tinggi dan proses yang praktis [21]. Menerapkan UMAP *(Uniform Manifold Approximation*

and Projection) dalam analisis berfungsi untuk pengurangan dimensi data. UMAP mempunyai kelebihan dibandingan teknik pengurangan dimensi yang lain, termasuk *T-Distributed Stochastic Neighbor Embedding* (t-SNE) dan *Principal Component Analysis* (PCA) [22].

2.3.2. HDBSCAN

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) berfungsi untuk mengklasifikasikan hasil data embedding. HDBSCAN dan UMAP merupakan kolaborasi yang baik dalam mempertahankan struktur data dimensi tinggi ketika klasifikasi sedang berlangsung dan tidak mengharuskan data outlier bergabung dengan kelompok [23].

2.3.3. C-TFIDF

Metode *class based TF-IDF* pada BERTopic diterapkan untuk menilai pentingnya setiap kata dalam *cluster.* Maka dari itu, metode ini dapat digunakan pada pengembangan distribusi topik setiap kata untuk *cluster* dokumen. Sehingga, dengan penggabungan yang dilakukan secara berulang representasi c-TF-IDF dari topik yang paling sedikit dibahas dengan topik yang memiliki kesamaan paling tinggi [5]. Berikut perhitungannya :

$$W_{t,c} = tf_{t,c} \times \log(1 + \frac{A}{t_{ft}}) \quad (2)$$

$Tf_{t,c}$ = jumlah kejadian pada istilah t dalam cluster c

t_{ft} = jumlah kejadian pada istilah t dalam semua cluster

A = rata-rata jumlah untuk kelas A

2.5 Coherence

Coherence adalah teknik yang diterapkan untuk meninjau seberapa baik model topik yang digunakan dan jika setiap frasa yang muncul dalam suatu topik terdapat kaitan atau hubungan pada sebuah konteks, maka topik tersebut bisa memiliki nilai coherence yang besar [4]. Pada masing-masing topik, frekuensi kata atau frasa yang paling tinggi dipilih guna mengkalkulasi koherensi dari topik model. koherensi topik dihitung dengan melihat frekuensi pasangan kata yang sering muncul bersamaan pada dokumen, lalu dilakukan penggabungan menghasilkan satu nilai atau skor koherensi. Persamaan dijelaskan sebagai berikut :

$$Coherence = \sum_{i<j} score(w_i, w_j) \quad (3)$$

2.6 API Gemini

API AI Gemini adalah suatu API yang diciptakan oleh Google untuk memberikan kemudahan dalam menggunakan AI (*artificial intelligence*). Pada API ini, bisa digunakan untuk analisis data, dan NLP (*Natural Language Processing)* [24].