BAB II
TINJAUAN PUSTAKA

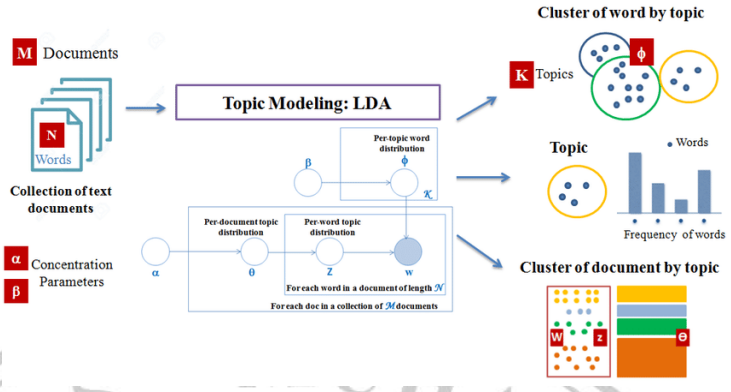
Penelitian ini bertujuan untuk mengidentifikasi tren-topik penelitian di bidang Islam dengan menggunakan teknik pemodelan topik berbasis NLP. Seiring dengan pesatnya perkembangan teknologi pemrosesan bahasa alami (NLP) dan algoritma pemodelan topik, pemahaman terhadap teknik-teknik ini sangat penting untuk memahami cara mengolah dan menginterpretasikan data teks dalam skala besar. Bab ini mengulas beberapa teori dasar yang relevan, mencakup pemodelan topik, serta aplikasi metode LDA dan BERTopic yang menjadi fokus utama dalam penelitian ini.

2.1. Landasan Teori

2.1.1. Topik Modeling

Topic modeling adalah teknik komputasional dalam machine learning dan pemrosesan bahasa alami yang bertujuan mengidentifikasi, mengekstraksi, dan mengklasifikasikan struktur tematik tersembunyi dalam kumpulan dokumen [5]. Menurut Blei (2012), teknik ini berasumsi bahwa setiap dokumen terdiri dari kombinasi beberapa topik, dengan setiap topik diwakili oleh distribusi probabilitas kata-kata tertentu [6]. Metode ini membantu mengidentifikasi pola makna yang tersembunyi dalam data tekstual yang besar, memudahkan pengorganisasian dan interpretasi informasi. Boyd-Graber et al. (2017) menyoroti keunggulan utama pendekatan ini, yaitu sifatnya yang unsupervised dan tidak memerlukan data berlabel, sehingga berguna dalam eksplorasi data teks di berbagai bidang [7]. Dua metode utama yang digunakan dalam penelitian ini adalah Latent Dirichlet Allocation (LDA) dan BERTopic.

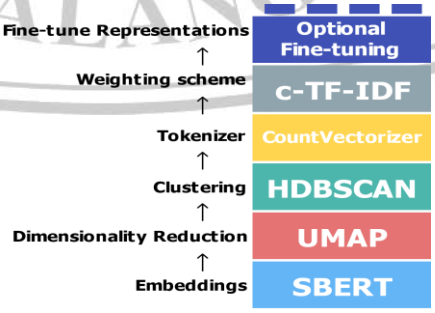
2.1.2. *Latent Dirichlet Allocation* (LDA)



Gambar 2.1 Model Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA), yang diperkenalkan oleh Blei, Ng, dan Jordan pada tahun 2003, merupakan metode unsupervised learning yang dilakukan dengan pendekatan pemodelan topik berbasis probabilistik yang berfungsi untuk mengungkap struktur tematik yang tidak terlihat dalam korpus dokumen. Menurut Gambar 2.1, model ini bekerja dengan asumsi dasar bahwa setiap dokumen merupakan kombinasi dari beberapa topik dengan proporsi yang bervariasi, dimana setiap topik direpresentasikan oleh distribusi probabilitas terhadap kosakata tertentu [8]. Dengan menggunakan LDA, peneliti dapat mengidentifikasi struktur topik tersembunyi dalam kumpulan dokumen secara otomatis, sehingga sangat berguna untuk analisis teks skala besar.

2.1.3. BERTopic



Gambar 2.2 Model BERTopic

BERTopic adalah teknik pemodelan topik kontemporer yang diperkenalkan oleh Grootendorst (2022) yang mengintegrasikan kekuatan BERT dengan algoritma *cluster*ing dan TF-IDF untuk mengekstrak topik dari dokumen [9] yang digambarkan pada Gambar 2.2. Berbeda dengan metode pemodelan topik tradisional, BERTopic memanfaatkan embedding kontekstual dari BERT untuk menghasilkan representasi semantik dokumen yang lebih mendalam, sehingga mampu menangkap nuansa makna dan hubungan tematik yang lebih kompleks, terutama untuk dokumen-dokumen yang memiliki struktur linguistik yang beragam dan konteks yang bervariasi.

2.1.4. Coherence Score

Coherence score merupakan metrik evaluasi kuantitatif yang mengukur tingkat kesamaan semantik antar kata dalam suatu topik hasil pemodelan. Nilai *coherence* dihitung berdasarkan *pointwise mutual information* (PMI), dimana:

$$pmi(x,y) = log_2 \frac{p(x,y)}{p(x)p(y)} = log_2 \frac{p(x|y)}{p(y)} \dots (1)$$

Persamaan (1) menunjukkan bahwa P(x,y) adalah probabilitas kemunculan bersama kata x dan y, sedangkan P(x) dan P(y) adalah probabilitas kemunculan individual masing-masing kata. Semakin tinggi nilai *coherence*, semakin kecil kesamaan antar topik, menandakan bahwa model berhasil mengidentifikasi topik-topik yang distinktif dan bermakna [10]. Dalam perbandingan algoritma topic modeling seperti LDA dan BERTopic, *coherence* score menjadi parameter kunci untuk mengevaluasi kualitas dan interpretabilitas topik yang dihasilkan.

2.2. Kajian Penelitian Terdahulu

Dalam konteks kajian teks berbahasa Arab, penerapan Latent Dirichlet Allocation (LDA) telah menunjukkan potensinya untuk

menggali tema-tema utama dari sumber literatur keislaman. Alhawarat (2015) membuktikan bahwa penggunaan LDA pada Surah Yusuf menghasilkan topik yang lebih koheren daripada LSA serta lebih efektif untuk pencarian semantik berbasis makna [11]. Di sisi lain, penerapan LDA pada abstrak ilmiah telah menjadi fokus berbagai penelitian. Zhao et al. (2023) mengidentifikasi lima topik utama dari 4.311 artikel terkait keamanan data dan privasi [12]. Sugiantoro et al. (2023) juga menggunakan LDA untuk menganalisis 666 abstrak tesis sarjana dan menemukan enam topik, termasuk penambangan data dan keamanan komputer [13]. Masinde (2023) menerapkan LDA pada 11.731 publikasi 4IR, menemukan tren teknologi seperti IoT dan AI [14].

Sementara itu, penggunaan BERTopic dalam konteks Islam masih terbatas. Aouichaty et al. (2024) memanfaatkan BERTopic untuk dokumen hukum Arab yang berkaitan dengan budaya Islam [15]. Alhaj et al. (2022), meskipun tidak berfokus pada konteks Islam, menggunakan BERTopic untuk pemodelan teks Arab secara umum [16]. Dengan demikian, penelitian terkait penggunaan BERTopic untuk kajian Islam di Indonesia masih minim dan potensial untuk dieksplorasi lebih lanjut.

Dalam ranah analisis abstrak ilmiah, BERTopic telah diterapkan secara luas. Meitei et al. (2024) menemukan sepuluh topik utama dari 50.261 abstrak terkait kompresi gambar [17]. Madrid-García et al. (2024) mengidentifikasi tren utama dari 96.004 abstrak rheumatologi, termasuk JAK inhibitors dan COVID-19 [18]. Kumar et al. (2024) menganalisis 3.482 artikel PubMed tentang neuropsikiatri terkait mikrobiota usus [19]. Hal ini menunjukkan bahwa BERTopic efektif dalam menggali tren penelitian ilmiah, namun aplikasinya pada abstrak ilmiah terkait Islam di Indonesia masih belum banyak dikaji.