

**ANALISIS PEMODELAN TOPIK TERJEMAH INDONESIA
HADIST BUKHORI DENGAN BERTOPIC DAN *LATENT*
DIRECHLET ALLOCATION (LDA)**

Proposal Tesis

**Diajukan untuk memenuhi syarat memperoleh gelar Magister Komputer pada
Program Studi Teknik Informatika S-2**



OLEH:

**ASEP RIDWAN HIDAYAT
231012050036**

**PROGRAM STUDI TEKNIK INFORMATIKA S-2
PROGRAM PASCASARJANA
UNIVERSITAS PAMULANG
TANGERANG SELATAN
2025**

ABSTRAK

Penelitian ini bertujuan untuk menganalisis performa dua metode pemodelan topik, yaitu *BerTopic* dan *Latent Dirichlet Allocation* (LDA), dalam mengelompokkan dan mengidentifikasi topik-topik dari terjemahan Indonesia Hadist Bukhari. Pemodelan topik merupakan salah satu pendekatan penting dalam pemrosesan bahasa alami yang memungkinkan pemahaman lebih mendalam terhadap konten teks dengan mengekstraksi tema-tema utama. *BerTopic* adalah metode yang memanfaatkan representasi teks berbasis transformer dan *clustering* hierarkis untuk menghasilkan topik yang lebih akurat dan interpretatif. LDA merupakan model probabilistik yang memodelkan dokumen sebagai campuran dari beberapa topik dan mengasumsikan distribusi kata pada topik tertentu. Proses pengumpulan data dilakukan melalui *Scraping* dari sumber penerjemah yang relevan, kemudian diproses menggunakan teknik preprocessing seperti *Lowercasing*, *Cleansing*, *Stemming Sastrawi*, dan Normalisasi. Kedua algoritma, LDA dan *BERTopic*, akan diterapkan untuk mengidentifikasi dan mengklasifikasikan topik-topik utama. Dalam penelitian ini, kedua metode tersebut diterapkan pada dataset terjemahan Hadist Bukhari, dan dibandingkan berdasarkan beberapa metrik seperti koherensi topik, akurasi klasifikasi, dan kemudahan interpretasi. Harapan dari penelitian ini menunjukkan bahwa setiap metode memiliki kelebihan dan kekurangan yang berbeda. juga bisa memberikan wawasan tentang metode terbaik untuk pemodelan topik teks religius dengan bahasa yang kompleks.

Kata Kunci: Topik Model, *BerTopic*, *Latent Dirichlet Allocation*, Hadist Bukhari

ABSTRACT

This study aims to analyze and compare topic modeling methods, namely BerTopic and Latent Dirichlet Allocation (LDA), in clustering and identifying topics from the Indonesian translation of Hadith Bukhari. Topic modeling is one of the important approaches in natural language processing that allows a deeper understanding of text content by extracting key themes. BerTopic is a method that utilizes transformer-based text representation and hierarchical clustering to generate more accurate and interpretative topics, LDA is a probabilistic model that models documents as a mixture of several topics and assumes the distribution of words on a particular topic. The data collection process is carried out through Scraping from relevant translator sources, then processed using preprocessing techniques such as Lowercasing, Cleansing, Stemming Sastrawi, and Normalization. Both algorithms, LDA and BERTopic, will be applied to identify and classify key topics. In this study, both methods are applied to the Hadith Bukhari translation dataset, and compared based on several metrics such as topic coherence, classification accuracy, and ease of interpretation. The hope of this study is that each method has different advantages and disadvantages. It can also provide insight into the best method for topic modeling of religious texts with complex language.

Keywords: Topic Model, BerTopic, Latent Dirichlet Allocation, Hadith Bukhari