INTISARI

ANALISIS KOMPARATIF PEMODELAN TOPIK BERBASIS PROBABILISTIK, NON-PROBABILISTIK, DAN *NEURAL EMBEDDING* SERTA INTEGRASI *GENERATIVE AI* UNTUK PELABELAN TOPIK

Oleh

Rahma Nur Annisa

22/497022/PA/21386

Penelitian ini bertujuan untuk melakukan analisis komparatif terhadap metode pemodelan topik dari berbagai paradigma serta mengevaluasi integrasi *Generative AI* pada tahap pelabelan topik. Meningkatnya volume data teks tidak terstruktur menuntut metode analisis yang mampu mengekstraksi struktur tematik laten secara sistematis. Perbedaan paradigma pemodelan topik menghasilkan karakteristik topik yang berbeda dari sisi koherensi, keberagaman, dan interpretabilitas. Data yang digunakan berupa 10.000 ulasan pengguna aplikasi Shopee berbahasa Indonesia yang diperoleh dari Google Play Store hingga 20 Oktober 2025 dan telah melalui prapemrosesan teks. Metode yang diimplementasikan meliputi *Latent Dirichlet Allocation* (LDA), Non-*Negative Matrix Factorization* (NMF), *Contextualized Topic Model* (CTM), dan *Bidirectional Encoder Representations from Transformers Topic Modeling* (BERTopic). Evaluasi performa dilakukan menggunakan metrik *topic coherence* dan *topic diversity*. Selanjutnya, pelabelan topik dilakukan menggunakan model *Generative AI* Gemini Flash untuk menghasilkan label topik yang konsisten secara semantik dan kontekstual. Hasil penelitian menunjukkan bahwa model berbasis *neural embedding*, khususnya BERTopic, menghasilkan keseimbangan performa terbaik antara *topic coherence* dan *topic diversity*, sementara LDA dan NMF memiliki keterbatasan dalam menangkap relasi semantik kompleks meskipun tetap unggul dari sisi interpretabilitas struktural. Integrasi *Generative AI* terbukti meningkatkan keterbacaan interpretasi topik tanpa memodifikasi struktur topik yang dihasilkan oleh model dasar.

ABSTRACT

COMPARATIVE ANALYSIS OF PROBABILISTIC, NON-PROBABILISTIC, AND NEURAL-EMBEDDING-BASED TOPIC MODELING AND THE INTEGRATION OF GENERATIVE AI FOR TOPIC LABELING

By

Rahma Nur Annisa

22/497022/PA/21386

This study aims to conduct a comparative analysis of topic modeling methods across these paradigms and to evaluate the integration of Generative AI at the topic labeling stage. The increasing volume of unstructured textual data necessitates analytical methods capable of systematically extracting latent thematic structures. Differences in topic modeling paradigms lead to variations in topic characteristics in terms of coherence, diversity, and interpretability. The dataset consists of 10,000 Indonesian-language user reviews of the Shopee application collected from Google Play Store up to 20 October 2025 and preprocessed through standard text cleaning procedures. The implemented methods include Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), Contextualized Topic Model (CTM), and Bidirectional Encoder Representations from Transformers Topic Modeling (BERTopic). Model performance is evaluated using topic coherence and topic diversity metrics. Subsequently, topic labeling is performed using the Gemini Flash Generative AI model to generate semantically consistent and context-aware topic labels. The results indicate that neural embedding–based models, particularly BERTopic, achieve the best balance between topic coherence and topic diversity, while LDA and NMF exhibit limitations in capturing complex semantic relationships despite their strong structural interpretability. The integration of Generative AI is shown to improve the readability of topic interpretation without modifying the underlying topic structures produced by the base models.