ABSTRAK

Perbandingan algoritma *Latent Dirichlet Allocation* (LDA) dan *BERTopic* dilakukan untuk mengevaluasi efektivitas kedua metode dalam pemodelan topik pada kumpulan postingan media sosial X terkait Piala Asia U-23 2024. Penelitian ini bertujuan untuk mengidentifikasi algoritma yang lebih unggul dalam menghasilkan topik yang koheren, relevan, dan beragam berdasarkan metrik *Topic Coherence* dan *Topic Diversity*. Proses penelitian menggunakan pendekatan *Knowledge Discovery in Database* (KDD), meliputi pengumpulan 1.570 data cuitan dengan kata kunci #TimnasDay, pra-pemrosesan data, dan penerapan kedua algoritma pada data yang telah disiapkan. Hasil menunjukkan bahwa BERTopic unggul dalam *Topic Coherence* dengan skor 0,515 dibandingkan LDA yang mencapai skor tertinggi 0,439 pada variasi parameter tertentu. Selain itu, BERTopic menghasilkan *Topic Diversity* lebih tinggi (0,98) dibandingkan LDA (0,88–0,89). Berdasarkan evaluasi manual, *BERTopic* juga lebih efektif dalam menangkap konteks semantik dan menghasilkan topik yang lebih bermakna. Temuan ini mengindikasikan bahwa *BERTopic* lebih cocok untuk pemodelan topik pada data teks pendek seperti cuitan media sosial.

Kata kunci: X, Data Teks Pendek, LDA, BERTopic, Pemodelan Topik, Perbandingan Algoritma, Koherensi Topik, Keberagaman Topik.

ABSTRACT

A comparison of Latent Dirichlet Allocation (LDA) and BERTopic algorithms was conducted to evaluate the effectiveness of both methods in topic modeling on a collection of X social media posts related to the 2024 U-23 Asian Cup. This research aims to identify algorithms that are superior in generating coherent, relevant, and diverse topics based on Topic Coherence and Topic Diversity metrics. The research process used a Knowledge Discovery in Database (KDD) approach, including the collection of 1,570 tweet data with the keyword #TimnasDay, data pre-processing, and application of both algorithms on the prepared data. The results show that BERTopic excels in Topic Coherence with a score of 0.515 compared to LDA which achieves the highest score of 0.439 at certain parameter variations. In addition, BERTopic produces higher Topic Diversity (0.98) than LDA (0.88-0.89). Based on manual evaluation, BERTopic is also more effective in capturing semantic context and generating more meaningful topics. These findings indicate that BERTopic is more suitable for topic modeling on short text data such as social media tweets.

Keywords: X, Short Text Data, LDA, BERTopic, Topic Modeling, Algorithm Comparison, Topic Coherence, Topic Diversity.