ABSTRAK

Text mining, sebagai bagian dari Knowledge Discovery in Databases, berperan penting dalam ekstraksi informasi dari data tidak terstruktur, khususnya teks. Salah satu teknik utamanya, topic modeling, digunakan untuk mengidentifikasi struktur topik dalam koleksi teks berjumlah besar. Teknik ini sangat relevan untuk menganalisis tren topik penelitian, termasuk di bidang Ilmu Komputer yang berkembang pesat dengan jumlah publikasi meningkat signifikan dalam beberapa tahun terakhir. Penelitian ini bertujuan untuk menganalisis topik dan tren penelitian di bidang Ilmu Komputer dengan menggunakan dua metode topic modeling, yaitu Latent Dirichlet Allocation (LDA) dan BERTopic. Data yang digunakan terdiri dari 4.892 metadata artikel penelitian dari situs Emerald Insight untuk periode 2019-2023. Metode BERTopic berbasis embedding menghasilkan coherence score tertinggi sebesar 0,49 dengan kombinasi TruncatedSVD-KMeans yang mengidentifikasi 13 topik, sementara LDA dengan teknik Bag-of-Words menghasilkan coherence score tertinggi sebesar 0,42 dengan 11 topik. BERTopic unggul dalam menghasilkan topik yang lebih koheren dan relevan, berkat kemampuannya mempertahankan konteks semantik antar kata. Namun, LDA menunjukkan keunggulan dalam akurasi prediksi pada perangkat dengan sumber daya terbatas, dengan akurasi 100% dibandingkan dengan 82,17% pada BERTopic. Temuan ini menyoroti trade-off antara kualitas representasi topik dan akurasi prediksi. Analisis tren penelitian Ilmu Komputer periode 2019-2023 menunjukkan pergeseran dari topik konvensional seperti manajemen proyek ke teknologi mutakhir seperti blockchain, IoT, dan kecerdasan buatan, dipengaruhi oleh kemajuan teknologi, kebutuhan industri, dan peristiwa global seperti pandemi COVID-19. Beberapa topik menunjukkan pertumbuhan yang konsisten, sementara yang lain mengalami fluktuasi minat dari tahun ke tahun. Penelitian ini memberikan kontribusi penting dalam memahami perkembangan tren penelitian di bidang Ilmu Komputer dan dapat menjadi acuan dalam perencanaan penelitian di masa depan, serta menyoroti kekuatan dan keterbatasan dari dua metode topic modeling yang berbeda dalam konteks analisis tren penelitian.

Kata Kunci: Text Mining, LDA, BERTopic, Tren Penelitian, Ilmu Komputer

ABSTRACT

Text mining, as part of Knowledge Discovery in Databases, plays a crucial role in extracting information from unstructured data, particularly text. One of its main techniques, topic modeling, is used to identify topic structures in large text collections. This technique is highly relevant for analyzing research trends, especially in the rapidly growing field of Computer Science, which has seen a significant increase in publications in recent years. This study aims to analyze research topics and trends in Computer Science using two topic modeling methods: Latent Dirichlet Allocation (LDA) and BERTopic. The data consists of 4,892 research article metadata from the Emerald Insight website, covering the period 2019-2023. The embedding-based BERTopic method achieved the highest coherence score of 0.49, with the TruncatedSVD-KMeans combination identifying 13 topics, while LDA with the Bag-of-Words technique produced a highest coherence score of 0.42 with 11 topics. BERTopic outperformed LDA in generating more coherent and relevant topics due to its ability to preserve the semantic context between words. However, LDA showed superiority in prediction accuracy on resource-limited devices, achieving 100% accuracy compared to BERTopic's 82.17%. These findings highlight a trade-off between topic representation quality and prediction accuracy. The analysis of Computer Science research trends from 2019 to 2023 reveals a shift from conventional topics like project management to emerging technologies such as blockchain, IoT, and artificial intelligence, driven by technological advances, industry needs, and global events like the COVID-19 pandemic. Some topics have shown consistent growth, while others have fluctuated in interest over the years. This study makes a valuable contribution to understanding the evolution of research trends in Computer Science and serves as a reference for future research planning, while also highlighting the strengths and limitations of the two topic modeling methods used in trend analysis.

Keywords: Text Mining, LDA, BERTopic, Research Trends, Computer Science