UNIVERSITAS
GADJAH MADA
Analisis Topik Pemberitaan Teknologi Bersih Menggunakan Latent Dirichlet Allocation dengan Pendekatan Statistik dan Semantik
Priskilla Novianna Puteri Br Silalahi, Prof. Dr. Azhari, MT.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

Analisis Topik Pemberitaan Teknologi Bersih Menggunakan Latent Dirichlet Allocation dengan Pendekatan Statistik dan Semantik

Oleh
Priskilla Novianna Puteri Br Silalahi
22/493324/PA/21176

Penelitian ini bertujuan untuk menganalisis topik utama dalam pemberitaan media teknologi bersih (*cleantech*) menggunakan *Latent Dirichlet Allocation* (LDA). Data yang digunakan adalah *Cleantech Media Dataset* dari Kaggle yang berisi lebih dari 20.000 artikel media bertema *cleantech* pada periode 2022–2024. Penelitian ini membandingkan dua strategi representasi teks sebelum LDA, yaitu TF-IDF sebagai representasi statistik dan KeyBERT sebagai representasi berbasis *semantic embedding*. Evaluasi dilakukan menggunakan *coherence C_v* dan *topic diversity*. Hasil penelitian menunjukkan bahwa metode representasi berbasis KeyBERT menghasilkan kualitas topik terbaik pada K = 6 dengan *coherence C_v* sebesar 0,6558 dan *topic diversity* sebesar 0,9333. Sementara itu, metode TF-IDF terbaik diperoleh pada K = 40 dengan *coherence C_v* sebesar 0,6043 dan *topic diversity* sebesar 0,7050. Dengan demikian, representasi berbasis KeyBERT mampu menghasilkan topik yang lebih ringkas, koheren, beragam, dan lebih mudah diinterpretasikan dibandingkan representasi TF-IDF.

Kata Kunci: *Cleantech, Pemodelan Topik, Latent Dirichlet Allocation, TF-IDF, KeyBERT*

UNIVERSITAS
GADJAH MADA
Analisis Topik Pemberitaan Teknologi Bersih Menggunakan Latent Dirichlet Allocation dengan Pendekatan Statistik dan Semantik
Priskilla Novianna Puteri Br Silalahi, Prof. Dr. Azhari, MT.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

Topic Analysis of Clean Technology Media Using Latent Dirichlet Allocation with Statistical and Semantic Approachesk

Oleh
Priskilla Novianna Puteri Br Silalahi
22/493324/PA/21176

This study aims to analyze the main topics in clean technology (cleantech) media coverage using Latent Dirichlet Allocation (LDA). The dataset used in this study is the Cleantech Media Dataset from Kaggle, which consists of more than 20,000 media articles related to cleantech published between 2022 and 2024. This study compares two text representation strategies before LDA, namely TF-IDF as a statistical representation and KeyBERT as a semantic embedding-based representation. Model evaluation is conducted using coherence C_v and topic diversity. The results show that the KeyBERT-based representation produces the best topic quality at K=6, with a coherence C_v score of 0.6558 and a topic diversity score of 0.9333. Meanwhile, the best TF-IDF-based model is obtained at K=40, with a coherence C_v score of 0.6043 and a topic diversity score of 0.7050. Therefore, the KeyBERT-based representation is able to produce topics that are more concise, coherent, diverse, and easier to interpret compared to the TF-IDF representation.

Keywords: Cleantech, Topic Modeling, Latent Dirichlet Allocation, TF-IDF, KeyBERT