

ABSTRAK

Abstrak skripsi Teknik Informatika memiliki karakteristik teks pendek dengan kosakata yang beragam, kombinasi istilah teknis bahasa Indonesia dan Inggris, serta informasi multidimensi yang mencakup metode, domain, dan objek penelitian dalam satu paragraf. Karakteristik tersebut menyebabkan proses klasterisasi dokumen secara otomatis menjadi menantang karena pendekatan berbasis kemiripan kata tradisional belum mampu menangkap hubungan semantik secara kontekstual. Penelitian ini bertujuan untuk mengidentifikasi dan menganalisis topik-topik penelitian pada abstrak skripsi Teknik Informatika menggunakan BERTopic, serta mengevaluasi beberapa konfigurasi parameter untuk memperoleh model dengan kualitas topik terbaik berdasarkan *Topic Coherence* (Cv) dan *Topic Diversity* (TD). Data yang digunakan terdiri atas 14.629 abstrak skripsi dari 10 perguruan tinggi di Indonesia periode 2013–2025. Penelitian mengikuti kerangka CRISP-DM yang meliputi preprocessing teks dua lapis, pembentukan *sentence embedding* menggunakan model multilingual *paraphrase-MiniLM*, reduksi dimensi dengan UMAP, klasterisasi menggunakan HDBSCAN, serta ekstraksi kata kunci menggunakan c-TF-IDF. Eksplorasi parameter dilakukan secara sistematis melalui dua tahap, yaitu *coarse search* dan *fine-tuning OFAT*. Hasil penelitian menunjukkan bahwa konfigurasi parameter terbaik diperoleh pada UMAP ($n_neighbors=30$, $min_dist=0.0$) dan HDBSCAN ($min_cluster_size=25$, $min_samples=5$), yang menghasilkan 106 topik optimal. Model ini berhasil mengklasterkan 71,75% dokumen (10.496 abstrak) dengan nilai *Topic Coherence* (Cv) sebesar 0,7003 dan *Topic Diversity* (TD) sebesar 0,9179. Berdasarkan analisis temporal, teridentifikasi 11 topik yang menunjukkan tren meningkat (*emerging*) dan 8 topik yang menunjukkan tren menurun (*declining*). Penelitian ini juga menghasilkan visualisasi interaktif untuk mendukung eksplorasi hasil klasterisasi dan tren penelitian..

Kata Kunci: abstrak skripsi, BERTopic, klasterisasi semantik, topic modeling, tren penelitian

ABSTRACT

Undergraduate thesis abstracts in Informatics Engineering are characterized by short texts with diverse vocabularies, a combination of Indonesian and English technical terms, and multidimensional information covering research methods, domains, and research objects within a single paragraph. These characteristics make automatic document clustering challenging because traditional word similarity-based approaches are unable to capture contextual semantic relationships effectively. This study aims to identify and analyze research topics in Informatics Engineering undergraduate thesis abstracts using BERTopic and to evaluate several parameter configurations to obtain the model with the best topic quality based on Topic Coherence (Cv) and Topic Diversity (TD). The dataset consists of 14,629 undergraduate thesis abstracts collected from 10 universities in Indonesia during the 2013–2025 period. The study follows the CRISP-DM framework, including two-layer text preprocessing, sentence embedding generation using the multilingual paraphrase-MiniLM model, dimensionality reduction with UMAP, clustering using HDBSCAN, and keyword extraction using c-TF-IDF. Parameter exploration was conducted systematically through two stages: coarse search and OFAT fine-tuning. The results show that the optimal parameter configuration was achieved with UMAP (n_neighbors=30, min_dist=0.0) and HDBSCAN (min_cluster_size=25, min_samples=5), producing 106 optimal topics. This model successfully clustered 71.75% of the documents (10,496 abstracts) with a Topic Coherence (Cv) score of 0.7003 and a Topic Diversity (TD) score of 0.9179. Based on temporal analysis, 11 topics were identified as emerging trends and 8 topics as declining trends. This study also produces interactive visualizations to support the exploration of clustering results and research trends.

Keywords: undergraduate thesis abstracts, BERTopic, semantic clustering, topic modeling, research trends