KLUSTERISASI TOPIK SKRIPSI MAHASISWA PROGRAM STUDI INFORMATIKA MENGGUNAKAN METODE LDA (LATENT DIRICHLET ALLOCATION) DAN BERT (BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS)

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi S1 Informatika

disusun oleh

YAHYA HANDARESTANTO

21.11.4365

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

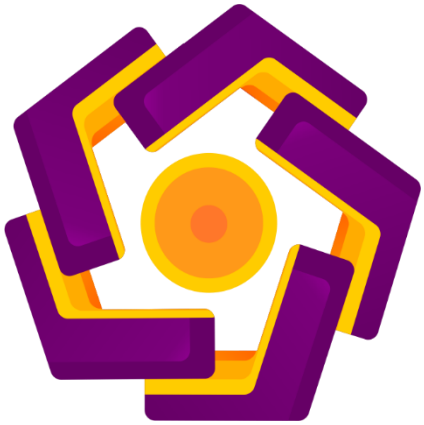
2025

KLUSTERISASI TOPIK SKRIPSI MAHASISWA PROGRAM STUDI INFORMATIKA MENGGUNAKAN METODE LDA (LATENT DIRICHLET ALLOCATION) DAN BERT (BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS)

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi *S1* Informatika



disusun oleh

YAHYA HANDARESTANTO

21.11.4365

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2025

HALAMAN PERSETUJUAN

SKRIPSI

KLUSTERISASI TOPIK SKRIPSI MAHASISWA PROGRAM STUDI INFORMATIKA MENGGUNAKAN METODE LDA (LATENT DIRICHLET ALLOCATION) DAN BERT (BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS)

yang disusun dan diajukan oleh

Yahya Handarestanto

21.11.4365

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 20 Januari 2025

Dosen Pembimbing,

Hanafi, S.Kom., M.Eng., Ph.D

NIK. 190302024

HALAMAN PENGESAHAN

SKRIPSI

KLUSTERISASI TOPIK SKRIPSI MAHASISWA PROGRAM STUDI INFORMATIKA MENGGUNAKAN METODE LDA (LATENT DIRICHLET ALLOCATION) DAN BERT (BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS)

yang disusun dan diajukan oleh

Yahya Handarestanto

21.11.4365

Telah dipertahankan di depan Dewan Penguji
pada tanggal 20 Januari 2025

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Dr. Ferry Wahyu Wibowo, S.Si., M.Cs. **NIK. 190302235**	
Dina Maulina, S.Kom., M.Kom. **NIK. 190302250**	
Hanafi, S.Kom., M.Eng., Ph.D **NIK. 190302024**	

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 20 Januari 2025

DEKAN FAKULTAS ILMU KOMPUTER

Hanif Al Fatta,S.Kom., M.Kom., Ph.D.
NIK. 190302096

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Yahya Handarestanto
NIM : 21.11.4365

Menyatakan bahwa Skripsi dengan judul berikut:

Klusterisasi topik skripsi mahasiswa program studi Informatika menggunakan metode LDA (Latent Dirichlet Allocation) Dan BERT (Bidirectional Encoder Representations from Transformers)

Dosen Pembimbing: Hanafi, S. Kom., M.Eng., Ph.D.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 20-01-2025

Yang Menyatakan.

Yahya Handarestanto

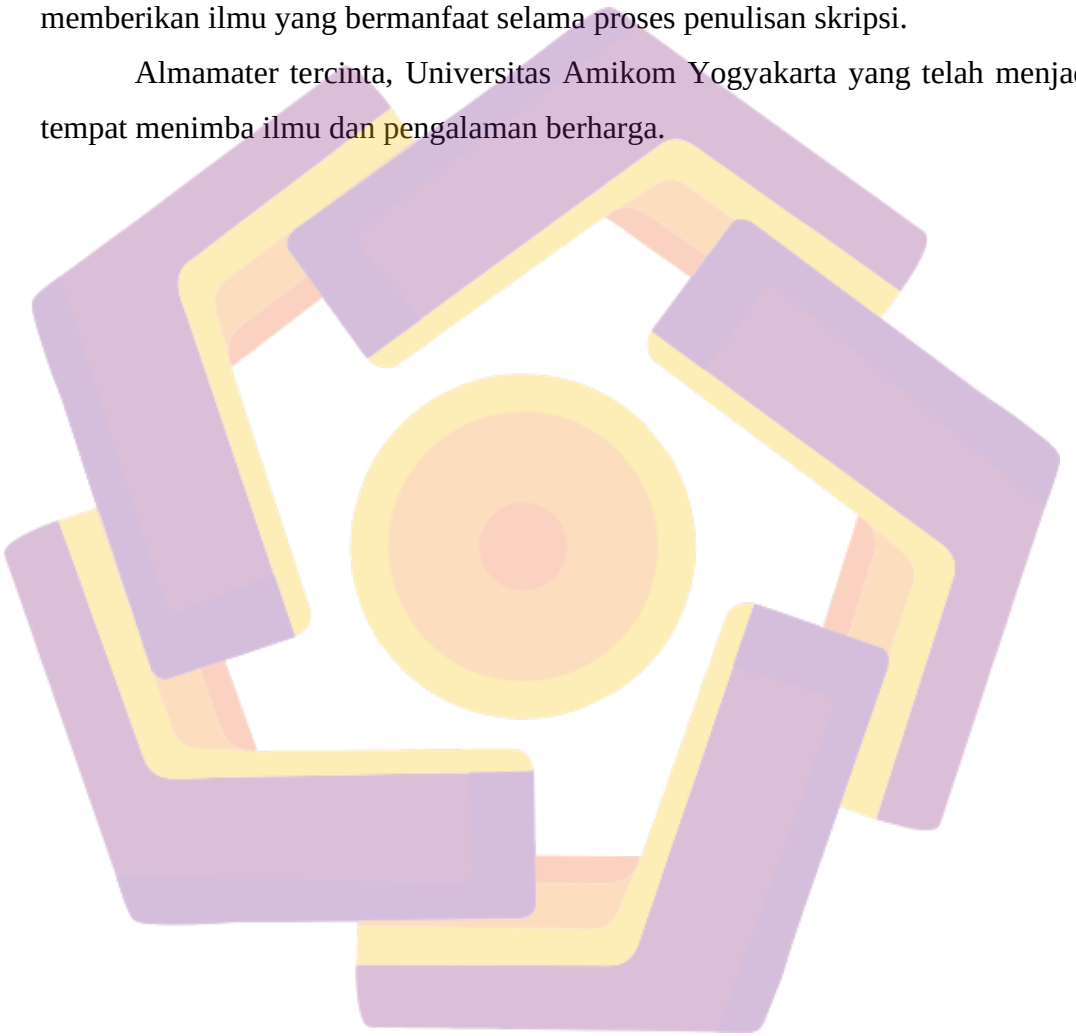
HALAMAN PERSEMBAHAN

Dengan penuh rasa syukur kepada Allah SWT, karya ilmiah ini saya persembahkan kepada:

Kedua orang tua saya tercinta, Bapak Radiyo dan Ibu Kustrini yang telah memberikan dukungan, doa, dan pengorbanan yang tak terhingga.

Para dosen pembimbing yang telah dengan sabar membimbing dan memberikan ilmu yang bermanfaat selama proses penulisan skripsi.

Almamater tercinta, Universitas Amikom Yogyakarta yang telah menjadi tempat menimba ilmu dan pengalaman berharga.



KATA PENGANTAR

Puji syukur penulis panjatkan kepada Allah Yang Maha Esa atas limpahan rahmat, hidayah, dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "Klusterisasi topik skripsi mahasiswa program studi Informatika menggunkaan metode LDA dan BERT" dengan lancar dan tepat waktu. Skripsi ini disusun guna memenuhi salah satu syarat untuk memperoleh gelar S.Kom pada Program Studi S1 Informatika di Universitas Amikom Yogyakarta.

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Bapak Hanafi, S.Kom., M.Eng,. Ph.D., selaku dosen pembimbing yang telah memberikan bimbingan, arahan, dan motivasi selama proses penyusunan skripsi ini.
2. Tim Dosen Penguji yang telah memberikan masukan dan saran yang sangat berharga untuk perbaikan skripsi ini.
3. Kedua orang tua yang senantiasa memberikan dukungan moril dan materiil serta doa yang tiada henti.
4. Universitas Amikom Yogyakarta yang telah memberikan kesempatan dan dukungan dalam penelitian ini.

Penulis menyadari bahwa skripsi ini masih jauh dari sempurna. Oleh karena itu, penulis mengharapkan kritik dan saran yang membangun dari berbagai pihak untuk perbaikan di masa mendatang.

Akhir kata, semoga skripsi ini dapat memberikan manfaat bagi pembaca dan pihak-pihak yang berkepentingan.

Yogyakarta, 20 Januari 2025

Penulis

Yahya Handarestanto

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

DAFTAR LAMBANG DAN SINGKATAN

AI	Artificial Intelligence
BoW	Bag-Of-Words
CPU	Central Processing Unit
GPU	Graphics Processing Unit
IDF	Inverse Document Frequency
LDA	Latent Dirichlet Allocation
BERT	Bidirectional Encoder Representations from Transformers
NLP	Natural Languange Procesing
NLTK	Natural Language Toolkit
OOP	Objek-oriented programing
PCA	Principal Component Analysis
TF	Term Frekuensi
TF-IDF	Term Frequency-Inverse Document Frequency
TPU	Tensor Processing Unit
t-SNE	t-distributed Stochastic Neighbor Embedding
URL	Uniform Resource Locator
VAE	Variational Autoencoders
N	Jumlah Kemunculan Kata
θ	Distribusi topik
ω	Probabilitas Kata pada Topik

DAFTAR ISTILAH

Alpha	distribusi topik per dokumen
Beta	distribusi kata per topik
Case folding	Mengubah semua huruf dalam suatu teks menjadi huruf kecil semua
Corpus	kumpulan teks yang digunakan untuk analisis atau pelatihan model.
Gensim	Pustaka python unutk pemodelan topik
IDF	Mengukur seberapa jarang kata tersebut muncul di seluruh dokumen dalam korpus.
Lemmatization	Mengubah kata ke bentuk dasarnya
Machine learning	Cabang dari kecerdasan buatan (Artificial Intelligence/AI) yang memungkinkan komputer untuk belajar dari data dan membuat prediksi atau keputusan tanpa diprogram secara eksplisit.
Matplotlib	Modul untuk visualiasi data secara 2d dan 3d
Meaningless_words	Daftar kata tidak relevan
Missing value	Nilai yang hilang atau kosong
Num Topics	Jumlah topik yang ingin dihasilkan
Overfitting	Model terlalu hafal data pelatihan sehingga sulit menangkap pola umum untuk data lain
Packages	Sekumpulan modul Python
Pandas	Modul Python untuk analisis data.
Preprocessing	Srangkaian langkah untuk mempersiapkan data mentah agar siap digunakan dalam proses pelatihan model.

PyLDAvis	Modul Python untuk visualisasi topik LDA
Remove punctuation	Menghapus tanda baca Dan karakter khusus
Sciket-learn	Modul digunakan untuk preprocesing data.
Seaborn	Modul python untuk membuat grafik.
Stopwords	Kata-kata yang dianggap tidak penting dalam pemrosesan teks/tidak memiliki makna.
TF	Mengukur frekuensi kemunculan suatu kata dalam dokumen untuk menunjukkan seberapa penting kata tersebut.
TF-IDF	Metode untuk mengukur pentingnya suatu kata dalam sebuah dokumen relatif terhadap seluruh koleksi dokumen
Tokenizing	Memecah teks menjadi unit-unit kecil.
Topic modeling	Menemukan dan mengelompokkan topik-topik tersembunyi dalam sekumpulan besar teks.
Wordcloud	Visualisasi berbentuk awan kata

INTISARI

Penelitian ini bertujuan untuk melakukan pemodelan topik pada kumpulan skripsi mahasiswa Studi Informatika di Universitas Amikom Yogyakarta dengan menggunakan metode Latent Dirichlet Allocation (LDA) dan BERTopic. Analisis dilakukan terhadap abstrak skripsi dari tahun 2020 hingga 2024, yang diolah melalui beberapa tahap preprocessing, seperti penghapusan nilai null, normalisasi teks, tokenisasi, dan penghapusan stopwords. Sebagai langkah awal, analisis TF-IDF dilakukan untuk mengidentifikasi kata-kata kunci yang paling berpengaruh dalam dataset.

Metode LDA digunakan untuk mengidentifikasi topik utama dalam dokumen dengan memandang setiap dokumen sebagai campuran dari beberapa topik tersembunyi. Sementara itu, BERTopic diterapkan dengan pendekatan berbasis BERT untuk menangkap representasi semantik yang lebih dalam dari teks. Evaluasi hasil pemodelan pada penilitian ini dilakukan dengan mencari parameter tebaik yaitu dengan melakukan pencarian skor koherensi tertinggi dengan rentang jumlah topik dari 2 hingga 20. Untuk model LDA menghasilkan skor koherensi sebesar 0.5933 dengan 5 topik, nilai Beta 0.6 dan nilai Alpha nya auto. Dan untuk model BERT/BERTopic menggunakan paraphrase all-MiniLM-L6-v2 menghasilkan skor koherensi 0.7502 dengan ukuran topik 16. Temuan ini menunjukkan bahwa model BERT/BERTopic lebih efektif dalam mengidentifikasi dan mengelompokkan topik pada dataset skripsi mahasiswa Studi Informatika di Universitas Amikom Yogyakarta. Hasil penelitian ini diharapkan dapat memberikan wawasan tentang arah penelitian dan minat mahasiswa, serta menjadi referensi bagi topik penelitian selanjutnya.

Kata kunci: LDA, BERTopic, TF-IDF, Skripsi, Topic modeling

ABSTRACT

This research aims to conduct topic modeling on a collection of Informatics Study student theses at Amikom University Yogyakarta using the Latent Dirichlet Allocation (LDA) and BERTopic methods. Analysis was carried out on thesis abstracts from 2020 to 2024, which were processed through several preprocessing stages, such as removing null values, text normalization, tokenization, and removing stop words. As a first step, TF-IDF analysis was carried out to identify the most influential keywords in the dataset.

The LDA method is used to identify the main topics in documents by viewing each document as a mixture of several hidden topics. Meanwhile, BERTopic is implemented with a BERT-based approach to capture deeper semantic representation of text. Evaluation of the modeling results in this research was carried out by looking for the best parameters, namely by searching for the highest coherence score with a range of topics from 2 to 20. The LDA model produced a coherence score of 0.5933 with 5 topics, a Beta value of 0.6 and an Alpha value of auto. And for the BERT/BERTopic model, using the all-MiniLM-L6-v2 paraphrase produces a coherence score of 0.7502 with a topic size of 16. These findings indicate that the BERT/BERTopic model is more effective in identifying and grouping topics in the thesis dataset of Informatics Study students at Amikom University, Yogyakarta. It is hoped that the results of this research will provide insight into research directions and student interests, as well as become a reference for further research topics.

Keyword: LDA, BERTopic, TF-IDF, Thesis, Topic modeling