PEMODELAN TOPIK MELALUI MEDIA X MENGGUNAKAN APACHE SPARK (STUDI KASUS COVID-19 DI INDONESIA TAHUN 2020-2022)

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi Sistem Informasi

disusun oleh

HANIF TOFA DARUSSALAM

18.12.0751

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

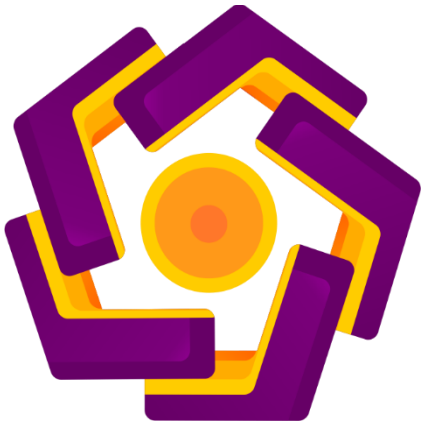
2025

PEMODELAN TOPIK MELALUI MEDIA X MENGGUNAKAN APACHE SPARK (STUDI KASUS COVID-19 DI INDONESIA TAHUN 2020-2022)

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi Sistem Informasi



disusun oleh

HANIF TOFA DARUSSALAM

18.12.0751

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2025

HALAMAN PERSETUJUAN

HALAMAN PERSETUJUAN

SKRIPSI

PEMODELAN TOPIK MELALUI MEDIA X MENGGUNAKAN APACHE SPARK (STUDI KASUS COVID-19 DI INDONESIA TAHUN 2020-2022)

yang disusun dan diajukan oleh

Hanif Tofa Darussalam

18.12.0751

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 28 Februari 2025

Dosen Pembimbing,

Kusnawi, S.Kom., M.Eng.
NIK. 190302112

ii

HALAMAN PENGESAHAN

HALAMAN PENGESAHAN

SKRIPSI

PEMODELAN TOPIK MELALUI MEDIA X MENGGUNAKAN APACHE SPARK (STUDI KASUS COVID-19 DI INDONESIA TAHUN 2020-2022)

yang disusun dan diajukan oleh

Hanif Tofa Darussalam

18.12.0751

Telah dipertahankan di depan Dewan Penguji
pada tanggal 28 Februari 2025

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Eli Pujastuti, S.Kom., M.Kom. **NIK. 190302227**	
Ike Verawati, S.Kom., M.Kom. **NIK. 190302237**	
Kusnawi, S.Kom., M.Eng. **NIK. 190302112**	

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 28 Februari 2025

DEKAN FAKULTAS ILMU KOMPUTER

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

iii

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Hanif Tofa Darussalam
NIM : 18.12.0751

Menyatakan bahwa Skripsi dengan judul berikut:

PEMODELAN TOPIK MELALUI MEDIA X MENGGUNAKAN APACHE SPARK (STUDI KASUS COVID-19 DI INDONESIA TAHUN 2020-2022)

Dosen Pembimbing : Kusnawi, S.Kom., M.Eng.

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, 28 Februari 2025

Yang Menyatakan,

Hanif Tofa Darussalam

HALAMAN PERSEMBAHAN

Skripsi ini saya persembahkan untuk:

Bapak Muhammad Saleh dan Ibu Sundari,

Terima kasih atas dukungan, kasih sayang, materi, dan doa yang telah diberikan sepanjang perjalanan menyelesaikan masa studi dan skripsi ini.

Kakak Noorlia Dharmawati, Nurissa Fatmawati, Fauzan Thoriq Perdana Kusuma

Terima kasih selalu memberi semangat dan doa yang diberikan sepanjang perjalanan menyelesaikan masa studi dan skripsi ini.

Kakak Ipar Anjar, Candra

Terima kasih selalu memberi semangat, doa, dan dukungan pada saat mengerjakan skripsi ini.

Teman

Terima kasih telah memberi semangat yang tidak pernah putus dan meluangkan waktunya untuk membantu penulis dalam menyelesaikan skripsi ini.

Bapak Kusnawi, S.Kom., M.Eng.

Terima kasih selaku dosen pembimbing yang telah memberikan bimbingan serta arahan dalam proses menyelesaikan skripsi ini dengan baik.

KATA PENGANTAR

Assalaamu'alaikum wa rahmatullahi wa barakaatuh

Puji syukur kehadirat Allah SWT atas rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi dengan judul "PEMODELAN TOPIK MELALUI MEDIA X MENGGUNAKAN APACHE SPARK (STUDI KASUS COVID-19 DI INDONESIA TAHUN 2020-2022)". Skripsi ini diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer di Universitas Amikom Yogyakarta.

Penulisan skripsi ini tidak lepas dari bantuan dan dukungan berbagai pihak. Oleh karena itu, penulis ingin menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Bapak Kusnawi, S.Kom., M.Eng., selaku Dosen Pembimbing yang telah memberikan bimbingan, arahan, dan motivasi yang sangat berharga selama proses penyusunan skripsi ini.
2. Prof. Dr. M. Suyanto, M.M, selaku Rektor Universitas Amikom Yogyakarta
3. Kedua orang tua dan keluarga, yang sudah memberi semangat dan dukungan selama proses penyusunan skripsi ini.
4. Teman, yang telah memberikan dukungan, semangat, dan membantu dalam proses pengumpulan data dan penyelesaian skripsi ini.

Akhir kata, penulis berharap semoga skripsi ini dapat bermanfaat bagi pengembangan ilmu pengetahuan dan bagi para pembaca.

Yogyakarta, 28 Februari 2025
Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMBANG DAN SINGKATAN

LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing
RDD	Resilient Distributed Dataset
NLTK	Natural Language Toolkit
DM	Direct Message
TF-IDF	Term Frequency Inverse Document Frequency
AT	Author-Topic

INTISARI

Pandemi COVID-19 yang terjadi di Indonesia selama periode 2020–2022 telah mendorong berbagai diskusi serta pertukaran opini di media sosial, termasuk Media X. Platform ini menghasilkan data dalam jumlah besar dan bersifat tidak terstruktur, sehingga dibutuhkan metode yang efektif untuk menganalisis serta mengidentifikasi pola topik yang muncul. Penelitian ini berfokus pada pemodelan topik dari data Media X dengan memanfaatkan *Apache Spark* sebagai kerangka kerja pemrosesan big data.

Metode yang digunakan dalam penelitian ini adalah *Latent Dirichlet Allocation* (LDA), yang diterapkan dalam lingkungan *Apache Spark* untuk menangani data dalam jumlah besar secara paralel dan terdistribusi. Proses analisis meliputi pengumpulan data, penyimpanan data, pra-pemrosesan data, vektorisasi teks, pemodelan topik, evaluasi dan visualisasi.

Hasil penelitian menunjukkan bahwa metode ini mampu mengidentifikasi topik-topik utama yang berkembang selama pandemi COVID-19 di Indonesia, seperti kesehatan, vaksinasi, kebijakan pemerintah, serta persepsi masyarakat terhadap pandemi. Sebanyak 10 topik ditentukan untuk dianalisis dalam penelitian ini. Model dievaluasi menggunakan dua metrik utama, yaitu perplexity dan coherence score, dengan hasil masing-masing 5.5987700238050815 dan 0.6061784364959958. Hasil ini menunjukkan bahwa coherence score lebih relevan dalam menilai kualitas topik yang dihasilkan, karena memberikan gambaran yang lebih jelas tentang keterkaitan dan kebermaknaan kata-kata dalam setiap topik dibandingkan perplexity.

Kata kunci: Latent Dirichlet Allocation, Pemodelan Topik, Skor Koherensi, COVID-19, Perplexity.

ABSTRACT

The COVID-19 pandemic that occurred in Indonesia during the 2020–2022 period has driven various discussions and exchanges of opinions on social media, including Media X. This platform generates large amounts of unstructured data, requiring effective methods to analyze and identify emerging topic patterns. This study focuses on topic modeling from Media X data by utilizing Apache Spark as a big data processing framework.

The method used in this study is Latent Dirichlet Allocation (LDA), implemented in an Apache Spark environment to handle large-scale data in a parallel and distributed manner. The analysis process includes data collection, data storage, data preprocessing, text vectorization, topic modeling, evaluation, and visualization.

The study results indicate that this method successfully identifies key topics that emerged during the COVID-19 pandemic in Indonesia, such as health, vaccination, government policies, and public perception of the pandemic. A total of 10 topics were selected for analysis in this study. The model was evaluated using two primary metrics: perplexity and coherence score, yielding results of 5.5987700238050815 and 0.6061784364959958, respectively. These results suggest that the coherence score is more relevant in assessing the quality of the generated topics, as it provides a clearer representation of the relationships and meaningfulness of words within each topic compared to perplexity.

Keyword: *Latent Dirichlet Allocation, Topic Modeling, Coherence Score, COVID-19, Perplexity.*