

**ANALISIS PERBANDINGAN TOPIC MODELLING *LATENT*  
*DIRICHLET ALLOCATION* DAN *LATENT SEMANTIC*  
*ANALYSIS* PADA BERITA DETIKFINANCE**

Skripsi

untuk memenuhi sebagian persyaratan mencapai derajat Sarjana S-1

Program Studi Informatika



Disusun Oleh:

**Desy Portuna**  
20106050035

**PROGRAM STUDI INFORMATIKA  
FAKULTAS SAINS DAN TEKNOLOGI  
UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA  
YOGYAKARTA**

**2024**



## PENGESAHAN TUGAS AKHIR

Nomor : B-2366/Un.02/DST/PP.00.9/12/2024

Tugas Akhir dengan judul : Analisis Perbandingan Metode Topic Modelling Latent Dirichlet Allocation Dan Latent Semantic Analysis Pada Berita Detikfinance

yang dipersiapkan dan disusun oleh:

Nama : DESY PORTUNA  
Nomor Induk Mahasiswa : 20106050035  
Telah diujikan pada : Kamis, 12 Desember 2024  
Nilai ujian Tugas Akhir : A

dinyatakan telah diterima oleh Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta

### TIM UJIAN TUGAS AKHIR



Ketua Sidang

Ir. Muhammad Didik Rohmad Wahyudi, S.T., MT.

SIGNED

Valid ID: 676e0cc4b938f



Penguji I

Dr. Ir. Sumarsono, S.T., M.Kom.

SIGNED

Valid ID: 676d410a6e347



Penguji II

Nurochman, S.Kom., M.Kom

SIGNED

Valid ID: 6769251e92c5c



Yogyakarta, 12 Desember 2024

UIN Sunan Kalijaga

Dekan Fakultas Sains dan Teknologi

Prof. Dr. Dra. Hj. Khurul Wardati, M.Si.

SIGNED

Valid ID: 676e14cac79ca

## SURAT PERSETUJUAN SKRIPSI



Universitas Islam Negeri Sunan Kalijaga



FM-UINSK-BM-05-03/R0

### SURAT PERSETUJUAN SKRIPSI/TUGAS AKHIR

Hal : Persetujuan Skripsi / Tugas Akhir

Lamp : -

Kepada

Yth. Dekan Fakultas Sains dan Teknologi

UIN Sunan Kalijaga Yogyakarta

Di Yogyakarta

*Assalammu'alaikum wr. wb.*

Setelah membaca, meneliti, memberikan petunjuk dan mengoreksi serta mengadakan perbaikan seperlunya, maka saya selaku pembimbing berpendapat bahwa skripsi Saudari:

|               |   |                                                                                                                                |
|---------------|---|--------------------------------------------------------------------------------------------------------------------------------|
| Nama          | : | Desy Portuna                                                                                                                   |
| NIM           | : | 20106050035                                                                                                                    |
| Judul Skripsi | : | Analisis Perbandingan Metode Topic Modelling Latent Dirichlet Allocation Dan Latent Semantic Analysis Pada Berita Detikfinance |

sudah dapat diajukan kembali kepada Program Studi Informatika Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta sebagai salah satu syarat untuk memperoleh gelar Sarjana Strata Satu dalam Program Studi Informatika.

Dengan ini kami mengharap agar skripsi/tugas akhir Saudari dapat segera dimunaqasyahkan. Atas perhatiannya saya ucapkan terima kasih.

*Wassalammu'alaikum wr. wb.*

Yogyakarta, 29 November 2024

Pembimbing

Muhammad Didik Rohmad Wahyudi, S.T., MT

NIP. 19841115 201903 1 003

## PERNYATAAN KEASLIAN SKRIPSI

### PERNYATAAN KEASLIAN SKRIPSI

Yang bertanda tangan di bawah ini:

Nama : Desy Portuna  
NIM : 20106050035  
Program Studi : Informatika  
Fakultas : Sains dan Teknologi

Dengan ini menyatakan bahwa skripsi yang berjudul “Analisis Perbandingan Metode Topic Modelling Latent Dirichlet Allocation Dan Latent Semantic Analysis Pada Berita Detikfinance” merupakan hasil penelitian saya sendiri, tidak terdapat pada karya yang pernah diajukan untuk memperoleh sarjana di suatu perguruan tinggi, dan bukan plagiasi karya orang lain kecuali yang secara tertulis yang menjadi acuan dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 29 November 2024

Penulis



Desy Portuna

NIM. 20106050035

STATE ISLAMIC UNIVERSITY  
**SUNAN KALIJAGA**  
YOGYAKARTA

## KATA PENGANTAR

Puji Syukur kehadiran Allah SWT dengan anugerah dan petunjuk-Nya, sehingga penulis dapat menyelesaikan tugas akhir yang berjudul “Analisis Perbandingan Metode Topic Modelling Latent Dirichlet Allocation Dan Latent Semantic Analysis Pada Berita Detikfinance”. Sholawat dan Salam penulis persembahkan ke hadapan ikutan kita Muhammad Rasulullah SAW yang telah mengeluarkan umat-Nya dari kegelapan kepada yang terang benderang dan semoga kita semua mendapatkan syafaat di akhirat kelak.

Skripsi ini disusun dan diajukan untuk memenuhi persyaratan mendapatkan gelar sarjana pada Program Studi Informatika UIN Sunan Kalijaga Yogyakarta. Selama proses penyusunan tugas akhir ini penulis mendapat banyak bantuan, saran, dan dukungan yang diberikan dari berbagai pihak secara langsung maupun tidak langsung. Oleh karena itu pada kesempatan ini, izinkan penulis mengucapkan terima kasih kepada:

1. Prof. Noorhaidi Hasan, S.Ag., M.A., M.Phil., Ph.D., selaku Rektor UIN Sunan Kalijaga Yogyakarta.
2. Prof. Dr. Dra. Hj. Khurul Wardati, M.Si., selaku Dekan Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta.
3. Dr. Muhammad Mustakim, S.T. M.T., selaku Ketua Program Studi Informatika Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta.
4. Mandahadi Kusuma, M.Eng., selaku Dosen Penasihat Akademik.
5. Muhammad Didik Rohmad Wahudi, S.T., MT., selaku Dosen Pembimbing Skripsi yang telah membimbing serta memberikan masukan perbaikan kepada penulis selama pengerjaan skripsi ini.
6. Dosen peminatan sistem cerdas dan seluruh dosen serta staf Program Studi Informatika Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta.

7. Kedua orang tua, saudara-saudari dan keponakan penulis atas dukungan spiritual, emosional dan materiilnya serta menjadi penyemangat selama penelitian ini.
8. Saudara sekandung dalam agama atas dukungannya.
9. Bapak MH yang berkenan mendengarkan, mengingatkan untuk selalu bersyukur dan mengajarkan afirmasi positif.
10. Teman-teman peminatan sistem cerdas yang saling mendukung.
11. Teman-teman Program Studi Informatika UIN Sunan Kalijaga 2020.
12. Teman badminton sekaligus senior di Program Studi Informatika yang berkenan sharing mengenai tugas akhir dan hal lainnya.
13. Semua pihak yang terlibat dalam penyusunan tugas akhir yang tidak dapat disebutkan satu persatu.

Semoga Allah SWT membalas kebaikan yang lebih baik. Penulis menyadari dalam penulisan skripsi ini masih banyak kekurangan dan jauh dari sempurna. Oleh karena itu penulis memohon maaf dan selalu terbuka untuk menerima kritik dan saran demi proses pembelajaran penulis menjadi lebih baik. Semoga penelitian ini dapat bermanfaat bagi pembaca. Akhir kata penulis ucapkan terimakasih.

STATE ISLAMIC UNIVERSITY  
SUNAN KALIJAGA  
YOGYAKARTA

Yogyakarta, 29 November 2024

Penulis

Desy Portuna

NIM. 20106050035

## HALAMAN PERSEMBAHAN

Tugas akhir ini saya persembahkan kepada diri sendiri sebagai *part of Habit Development and Human Learning Process*



## HALAMAN MOTTO

*Pikir Itu Pelita Hati*

*Tidak Dipikir Merusak Diri*

*Terlalu Dipikir Binasakan Diri*



## **LEMBAR PEDOMAN PENGGUNAAN SKRIPSI**

Tugas Akhir ini tidak dipublikasikan, tetapi tersedia di perpustakaan dalam lingkungan UIN Sunan Kalijaga Yogyakarta, diperkenankan dipakau sebagai referensi kepustakaan, tetapi pengutipan harus seizin penyusun, dan harus menyebutkan sumbernya sesuai dengan kebiasaan ilmiah. Dokumen Tugas Akhir ini merupakan hak milik UIN Sunan Kalijaga Yogyakarta.



## DAFTAR ISI

|                                                 |      |
|-------------------------------------------------|------|
| HALAMAN PENGESAHAN .....                        | ii   |
| SURAT PERSETUJUAN SKRIPSI.....                  | iii  |
| PERNYATAAN KEASLIAN SKRIPSI .....               | iv   |
| KATA PENGANTAR.....                             | v    |
| HALAMAN PERSEMBAHAN .....                       | vii  |
| HALAMAN MOTTO .....                             | viii |
| LEMBAR PEDOMAN PENGGUNAAN SKRIPSI .....         | ix   |
| DAFTAR ISI .....                                | x    |
| DAFTAR TABEL .....                              | xii  |
| DAFTAR GAMBAR.....                              | xiii |
| DAFTAR KODE .....                               | xiv  |
| INTISARI .....                                  | xv   |
| ABSTRACT .....                                  | xvi  |
| BAB I PENDAHULUAN .....                         | 17   |
| 1.1 Latar Belakang .....                        | 17   |
| 1.2 Rumusan Masalah.....                        | 18   |
| 1.3 Batasan Masalah .....                       | 18   |
| 1.4 Tujuan Penelitian.....                      | 19   |
| 1.5 Manfaat Penelitian .....                    | 19   |
| BAB II TINJAUAN PUSTAKA DAN LANDASAN TEORI..... | 20   |
| 2.1 Tinjauan Pustaka.....                       | 20   |
| 2.2 Landasan Teori .....                        | 25   |
| 2.2.1 Text Mining.....                          | 25   |
| 2.2.2 Topic Modeling.....                       | 26   |
| 2.2.3 Latent Dirichlet Allocation.....          | 26   |
| 2.2.4 Latent Semantic Analysis.....             | 31   |
| 2.2.5 Prapemrosesan.....                        | 34   |
| 2.2.6 Evaluasi Pemodelan Topik.....             | 35   |
| 2.2.6.1 Koherensi Topik .....                   | 36   |
| 2.2.6.2 Keberagaman Topik .....                 | 36   |
| 2.2.6.3 Waktu Pemrosesan .....                  | 37   |
| 2.2.7 Gensim .....                              | 37   |
| BAB III METODE PENELITIAN .....                 | 38   |

|                                    |    |
|------------------------------------|----|
| 3.1 Alat dan Bahan Penelitian..... | 38 |
| 3.2 Metode Penelitian .....        | 38 |
| BAB IV HASIL DAN PEMBAHASAN.....   | 41 |
| 4.1 Hasil .....                    | 41 |
| 4.1.1 Pengumpulan Data .....       | 41 |
| 4.1.2 Data Preprocessing.....      | 48 |
| 4.1.3 Pemodelan Topik.....         | 61 |
| 4.1.4 Evaluasi .....               | 62 |
| 4.2 Pembahasan .....               | 64 |
| BAB V PENUTUP .....                | 67 |
| 5.1 Kesimpulan .....               | 67 |
| 5.2 Saran .....                    | 67 |
| DAFTAR PUSTAKA.....                | 69 |
| LAMPIRAN .....                     | 73 |



## DAFTAR TABEL

|                                                    |    |
|----------------------------------------------------|----|
| Tabel 2.1 Ringkasan Tinjauan Pustaka .....         | 22 |
| Tabel 2.2 Perbedaan LDA dan LSA .....              | 34 |
| Tabel 3.1 Alat-Alat Penelitian .....               | 38 |
| Tabel 4.1 Sampel Hasil Scraping .....              | 44 |
| Tabel 4.2 sample setelah preprocessing .....       | 49 |
| Tabel 4.3 Sample Hasil Seleksi Pos Tagging .....   | 57 |
| Tabel 4.4 Nilai koherensi LDA dan LSA .....        | 64 |
| Tabel 4.5 nilai keberagaman LDA dan LSA .....      | 65 |
| Tabel 4.6 waktu pemrosesan model LDA dan LSA ..... | 66 |



## DAFTAR GAMBAR

|                                                             |    |
|-------------------------------------------------------------|----|
| Gambar 2.1 Intuisi dibalik LDA.....                         | 27 |
| Gambar 2.2 Representasi model grafik dari LDA.....          | 30 |
| Gambar 3.1 Tahapan Penelitian.....                          | 39 |
| Gambar 3. 2 Rangkaian Proses Pelatihan dan Evaluasi .....   | 40 |
| Gambar 4.1 Perbandingan nilai koherensi LDA dan LSA .....   | 62 |
| Gambar 4.2 Perbandingan nilai keberagaman LDA dan LSA ..... | 63 |
| Gambar 4.3 Perbandingan waktu pemrosesan LDA dan LSA .....  | 64 |



## DAFTAR KODE

|                                        |    |
|----------------------------------------|----|
| Kode 4.1 membuat daftar tanggal.....   | 41 |
| Kode 4.2 Web scraping .....            | 42 |
| Kode 4.3 Preprocessing .....           | 48 |
| Kode 4.4 Inisialisasi POS Tagging..... | 56 |
| Kode 4.5 Ekstraksi kata .....          | 57 |
| Kode 4.6 Membuat Corpus.....           | 61 |
| Kode 4.7 Pemodelan LDA.....            | 61 |
| Kode 4.8 Pemodelan LSA .....           | 62 |
| Kode 4.9 Evaluasi model.....           | 62 |



# **ANALISIS PERBANDINGAN TOPIC MODELLING LATENT DIRICHLET ALLOCATION DAN LATENT SEMANTIC ANALYSIS PADA BERITA DETIKFINACE**

**Desy Portuna**  
**20106050035**

## **INTISARI**

Pemodelan topik menjadi teknik penting dalam analisis teks seiring dengan meningkatnya volume data teks tidak terstruktur. Teknik ini dapat mengidentifikasi tema tanpa perlu label kategori yang ditentukan sebelumnya. Metode yang digunakan dalam penelitian ini adalah *Latent Dirichlet Allocation* (LDA) dan *Latent Semantic Analysis* (LSA). Sebelumnya, penelitian komparatif pada kedua metode ini hanya menggunakan metrik nilai koherensi untuk evaluasi, namun penelitian ini menambahkan dua metrik tambahan, yaitu nilai keberagaman dan waktu pemrosesan, guna memberikan analisis performa yang lebih komprehensif.

Penelitian ini bertujuan untuk membandingkan performa LDA dan LSA dalam pemodelan topik pada korpus bahasa Indonesia. Data yang digunakan adalah 1.837 artikel DetikFinance yang diambil melalui *web scraping* dalam rentang waktu Mei hingga Juli 2024. Tahapan penelitian meliputi pengumpulan data, prapemrosesan lalu pemodelan topik menggunakan LDA dan LSA. Selanjutnya, evaluasi dilakukan berdasarkan tiga metrik utama yaitu koherensi topik, keberagaman, dan waktu pemrosesan.

Hasil penelitian menunjukkan bahwa LDA menghasilkan nilai koherensi dan keberagaman lebih tinggi dan stabil dibandingkan LSA. Hasil tersebut menunjukkan bahwa topik yang dihasilkan LDA relevan secara semantik serta bervariasi. Namun, dalam hal waktu pemrosesan LSA menunjukkan hasil yang lebih cepat. Dengan demikian, pemilihan metode pemodelan topik perlu disesuaikan dengan kebutuhan analisis, apakah lebih mementingkan struktur topik atau efisiensi waktu pemrosesan.

Kata kunci: pemodelan topik, LDA, LSA, artikel berita, nilai koherensi, nilai keberagaman

# COMPARATIVE ANALYSIS OF LATENT DIRICHLET ALLOCATION AND LATENT SEMANTIC ANALYSIS TOPIC MODELING ON DETIKFINANCE NEWS

**Desy Portuna**  
**20106050035**

## ABSTRACT

Topic modeling is becoming an important technique in text analysis as the volume of unstructured text data increases. This technique can identify themes without the need for predefined category labels. The methods used in this study are Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA). Previously, the comparative study on these two methods only used the coherence value metric for evaluation, but this study added two additional metrics, namely diversity value and processing time, to provide a more comprehensive performance analysis.

This study aims to compare the performance of LDA and LSA in topic modeling in the Indonesian corpus. The data used is 1,837 DetikFinance articles taken through *web scraping* in the period from May to July 2024. The research stages include data collection, preprocessing, and topic modeling using LDA and LSA. Furthermore, the evaluation was carried out based on three main metrics, namely topic coherence, diversity, and processing time.

The results showed that LDA produced higher and more stable coherence and diversity values than LSA. The results show that the topics generated by LDA are semantically relevant and varied. However, in terms of processing time, LSA shows faster results. Thus, the selection of topic modeling methods needs to be adjusted to the needs of analysis, whether it is more concerned with topic structure or processing time efficiency.

Keywords: topic modeling, LDA, LSA, news articles, coherence value, diversity value

STATE ISLAMIC UNIVERSITY  
SUNAN KALIJAGA  
YOGYAKARTA

# BAB I

## PENDAHULUAN

### 1.1 Latar Belakang

Dengan meningkatnya *volume* data teks yang tidak terstruktur, pemodelan topik telah menjadi metode yang relevan dalam pengolahan data teks [1]. Berbeda dengan teknik seperti *clustering*, yang mengelompokkan dokumen berdasarkan kesamaan tanpa kategori yang ditentukan [2], atau klasifikasi yang memerlukan label kategori [3], *summarization* yang merangkum informasi [4], pemodelan topik menawarkan pendekatan yang dapat mengidentifikasi tema tanpa label kategori yang ditentukan sebelumnya.

Pemodelan topik merupakan metode analisis teks yang bertujuan untuk mengidentifikasi dan menemukan topik atau tema yang terkandung dalam kumpulan data teks yang berukuran besar [5]. Topik dapat diartikan sebagai sekumpulan kata yang sering muncul dan merepresentasikan suatu ide atau konsep tertentu [6]. Pemodelan topik didasarkan pada prinsip bahwa setiap dokumen dapat dipandang sebagai gabungan dari berbagai topik. [7]. Setiap topik terdiri dari kumpulan kata yang saling berhubungan dan setiap kata dalam sebuah dokumen dapat dikaitkan dengan salah satu topik tersebut[8].

Dalam upaya memahami tema-tema yang muncul dalam kumpulan teks besar, beragam teknik pemodelan topik telah dikembangkan. Salah satu metode yang paling sering digunakan adalah *Latent Dirichlet Allocation* (LDA), yang bekerja dengan anggapan bahwa setiap dokumen merupakan kombinasi dari beberapa topik, sementara setiap topik terdiri dari kumpulan kata berdasarkan frekuensi kemunculannya dalam dokumen [9]. Di samping itu, *Latent Semantic Analysis* (LSA) menggunakan dekomposisi nilai tunggal (SVD) dan dekomposisi matriks untuk memodelkan struktur semantik, yang secara efektif mengurangi korelasi kata dan menyederhanakan vektor teks [10]. Metode lain, seperti *Non-Negative Matrix Factorization* (NMF) merupakan teknik yang digunakan untuk mengekstrak topik

dari kumpulan data yang besar dengan cara menguraikannya menjadi dua matriks non-negatif yang merepresentasikan hubungan antara kata dan topik [11]. Adapun metode lainnya seperti Dynamic Topic Models (DTM) dapat memahami perubahan topik dari waktu ke waktu [12], sedangkan Correlated Topic Model (CTM) dapat mengeksplorasi hubungan antar topik yang ada [13].

Pada penelitian sebelumnya yang menganalisis perbandingan antara metode pemodelan topik Latent Dirichlet Allocation dan Latent Semantic Analysis hanya menggunakan metrik coherence score untuk mengevaluasi performa kedua metode ini [14], [15]. Meskipun coherence score memberikan gambaran tentang seberapa baik topik yang dihasilkan, metrik ini belum cukup untuk mengevaluasi efektivitas metode pemodelan topik secara menyeluruh [16]. Oleh karena itu, penelitian ini menambahkan dua metrik evaluasi tambahan, yaitu *diversity score* dan waktu pemrosesan, untuk memberikan analisis yang lebih komprehensif mengenai metode pemodelan topik.

## **1.2 Rumusan Masalah**

Berdasarkan penjelasan latar belakang tersebut, permasalahan yang muncul adalah bagaimana membandingkan hasil evaluasi dari penerapan algoritma Latent Dirichlet Allocation dan Latent Semantic Analysis dalam pemodelan topik, dengan memperhatikan nilai koherensi topik, nilai keberagaman topik, dan durasi waktu pemrosesan.

## **1.3 Batasan Masalah**

Pada penelitian ini ada beberapa batasan masalah agar penyusunan dan pembahasan penelitian ini dapat dilakukan secara terarah, sebagai berikut:

1. Penerapan algoritma Latent Dirichlet Allocation dan Latent Semantic Analysis untuk melakukan pemodelan topik.

2. Data yang akan digunakan adalah isi artikel dari DetikFinance.
3. Bahasa pemrograman yang akan digunakan adalah Python, beserta berbagai library yang tersedia di dalamnya.

#### **1.4 Tujuan Penelitian**

Berdasarkan rumusan masalah yang telah disebutkan, tujuan dari penelitian ini adalah membandingkan pemodelan topik antara LDA dan LSA menggunakan korpus berbahasa Indonesia pada artikel berita DetikFinance. Selanjutnya akan dievaluasi kualitasnya berdasarkan nilai koherensi topik, nilai keberagaman topik, dan durasi waktu pemrosesan.

#### **1.5 Manfaat Penelitian**

Berdasarkan poin sebelumnya, adapun manfaat dari penelitian ini antara lain, sebagai berikut:

1. Menambah wawasan dan pemahaman dalam text mining.
2. Menyediakan landasan dalam pemilihan metode yang tepat untuk analisis berita.

## **BAB V**

### **PENUTUP**

#### **5.1 Kesimpulan**

Berdasarkan penelitian yang sudah dilakukan dapat disimpulkan analisis perbandingan pemodelan topik menggunakan metode Latent Dirichlet Allocation (LDA) dan Latent Semantic Analysis (LSA) pada berita ekonomi di DetikFinance berhasil dilakukan. Dari hasil analisis, dapat disimpulkan bahwa metode LDA memiliki keunggulan dalam mempertahankan koherensi dan keberagaman topik yang lebih baik dibandingkan dengan LSA pada jumlah topik yang lebih besar. Meskipun nilai koherensi LDA tidak setinggi LSA untuk jumlah topik yang sedikit, LDA menunjukkan stabilitas yang lebih baik seiring bertambahnya jumlah topik. Di sisi lain, LSA lebih efisien dalam waktu proses, namun mengalami penurunan koherensi dan keberagaman topik yang signifikan saat jumlah topik meningkat. Oleh karena itu, LDA lebih cocok digunakan untuk analisis yang memerlukan interpretabilitas dan diferensiasi topik yang baik. Sementara itu, LSA dapat menjadi alternatif yang efektif jika waktu pemrosesan menjadi prioritas utama.

#### **5.2 Saran**

Pada penelitian ini terdapat beberapa kekurangan. Oleh karena itu, penulis menyarankan beberapa hal untuk penelitian di masa mendatang, di antaranya:

1. Penelitian selanjutnya bisa menggunakan data yang lebih bervariasi, baik dari segi jenis berita maupun sumber berita. Penelitian hanya ini menggunakan berita dari situs DetikFinance dengan subtema ekonomi bisnis.
2. Penelitian selanjutnya dapat mengeksplorasi metode lainnya guna menemukan metode yang relevan untuk korpus *Long-text* berbahasa Indonesia.

3. Penelitian selanjutnya dapat membandingkan pengaruh tahapan *preprocessing* terhadap hasil pemodelan topik.



## DAFTAR PUSTAKA

- [1] M. M. Mohd Sharif, R. Maskat, Z. Baharum, and K. Maskat, "A scoping review of topic modelling on online data," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, no. 3, p. 1633, Sep. 2023, doi: 10.11591/ijeecs.v31.i3.pp1633-1641.
- [2] F. Andriyani and Y. Puspitarani, "Performance Comparison of K-Means and DBScan Algorithms for Text Clustering Product Reviews," *Sinkron*, vol. 7, no. 3, pp. 944–949, Jul. 2022, doi: 10.33395/sinkron.v7i3.11569.
- [3] M. Veziroğlu, E. Veziroğlu, and İ. Ö. Bucak, "Performance Comparison between Naive Bayes and Machine Learning Algorithms for News Classification," in *Bayesian Inference - Recent Trends*, IntechOpen, 2024. doi: 10.5772/intechopen.1002778.
- [4] Dr. Vidyagouri B H and BibiSadiqa M D, "Text Summarization Using Machine Learning Algorithm," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 167–173, Jul. 2022, doi: 10.32628/CSEIT228421.
- [5] R. Muthusami, N. Mani Kandan, K. Saritha, B. Narenthiran, N. Nagaprasad, and K. Ramaswamy, "Investigating topic modeling techniques through evaluation of topics discovered in short texts data across diverse domains," *Sci Rep*, vol. 14, no. 1, p. 12003, 2024, doi: 10.1038/s41598-024-61738-4.
- [6] J. Rashid, S. M. A. Shah, and A. Irtaza, "An Efficient Topic Modeling Approach for Text Mining and Information Retrieval through K-means Clustering," *Mehran University Research Journal of Engineering and Technology*, vol. 39, no. 1, pp. 213–222, Jan. 2020, doi: 10.22581/muet1982.2001.20.
- [7] D. C. Zhang and H. W. Lauw, "Topic Modeling on Document Networks with Dirichlet Optimal Transport Barycenter," in *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, 2024, pp. 5743–5744. doi: 10.1109/ICDE60146.2024.00503.
- [8] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012, doi: 10.1145/2133806.2133826.
- [9] , Philip Obiorah, F. Onuodu, and B. Eke, "Topic Modeling Using Latent Dirichlet Allocation & Multinomial Logistic Regression," *Advances in Multidisciplinary and scientific Research Journal Publication*, vol. 10, no. 4, pp. 99–112, Dec. 2022, doi: 10.22624/AIMS/DIGITAL/V10N4P11a.
- [10] S. Li, "A Cross Language Information Retrieval Model Based on Latent Semantic Analysis," 2024. doi: 10.3233/ATDE231290.
- [11] N. F. Casillano and K. Cantilang, "Exploring Information Technology Industry Programming Language Trends with Non-Negative Matrix Factorization Topic Modelling," *EDUCATUM Journal of Science, Mathematics and Technology*, vol. 10, no. 1, pp. 35–44, Jun. 2023, doi: 10.37134/ejsmt.vol10.1.5.2023.
- [12] C. James, M. Nagda, N. H. Ghassemi, M. Kloft, and S. Fellenz, "Evaluating Dynamic Topic Models," Sep. 2023.

- [13] M. Khine Oo and M. Aye Khine, "Topic Extraction of Crawled Documents Collection using Correlated Topic Model in Mapreduce Framework," *International Journal on Natural Language Computing*, vol. 8, no. 6, pp. 11–23, Dec. 2019, doi: 10.5121/ijnlc.2019.8602.
- [14] U. Mulya, "Analisis Komparasi Pemodelan topik Latent Dirichlet Allocation dan Latent Semantic Analysis pada Ulasan Restoran di Yogyakarta," 2020. [Online]. Available: <http://digilib.uin-suka.ac.id/id/eprint/50297>
- [15] U. Malihatin S, Y. Findawati, and U. Indahyanti, "Topic Modeling In Covid-19 Vaccination Refusal Cases Using Latent Dirichlet Allocation And Latent Semantic Analysis," *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 5, pp. 1063–1074, Oct. 2023, doi: 10.52436/1.jutif.2023.4.5.951.
- [16] D. Aryani, I. Lucia Kharisma, A. Sujjada, and K. Kamdan, "Topic Modeling of the 2024 Election Using the BERTopic Method on Detik.com News Articles," *Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 9, no. 2, pp. 171–180, Aug. 2024, doi: 10.25139/inform.v9i2.8429.
- [17] A. I. Humam, "Pemodelan Topik Dokumen Skripsi Mahasiswa S1 Teknik Informatika Uin Sunan Kalijaga Yogyakarta Menggunakan Analisis Latent Dirichlet Allocation [Thesis (Skripsi), Uin Sunan Kalijaga Yogyakarta].," 2022. [Online]. Available: <https://digilib.uin-suka.ac.id/id/eprint/50970/>
- [18] A. Rachmawati, "Analisis Topic Modeling Pada Jurnal Menggunakan Metode Latent Dirichlet Allocation (LDA)," 2022. [Online]. Available: <https://digilib.uin-suka.ac.id/id/eprint/54439/>
- [19] A. D. Y. Nugroho, "Analisis Komparasi Pemodelan Topik Metode Latent Dirichlet Allocation (LDA) Dan Bertopic Pada Berita Berbahasa Indonesia," 2023. [Online]. Available: <http://digilib.uin-suka.ac.id/id/eprint/63479>
- [20] Ajay Jadhav, Pranjal Jagtap, Suraj Gurav, Shivani Jadhav, Nikita Jadhav, and Afsha Akkalkot, "A Survey on Text Mining - Techniques, Application," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 338–343, Jun. 2023, doi: 10.32628/CSEIT2390391.
- [21] G. A. S. Ribeiro, R. M. Barbosa, M. da Cunha Reis, and N. L. Costa, "From bibliometrics to text mining: exploring feature selection methods in microarray research," *Commun Stat Simul Comput*, pp. 1–17, Mar. 2024, doi: 10.1080/03610918.2024.2331083.
- [22] S. Vijay Gaikwad, A. Chaugule, and P. Patil, "Text Mining Methods and Techniques," *Int J Comput Appl*, vol. 85, no. 17, pp. 42–45, Jan. 2014, doi: 10.5120/14937-3507.
- [23] D. M. Blei, A. Y. Ng, and J. B. Edu, "Latent Dirichlet Allocation Michael I. Jordan," 2003.
- [24] K. B. Putra and R. P. Kusumawardani, "Analisis topik informasi publik media sosial di surabaya menggunakan pemodelan latent dirichlet allocation (LDA)," *Jurnal Teknik ITS*, vol. 6, no. 2, pp. A446–A450, 2017.
- [25] H. Jelodar *et al.*, "Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey."

- [26] J. C. Campbell, A. Hindle, and E. Stroulia, "Chapter 6 - Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data," in *The Art and Science of Analyzing Software Data*, C. Bird, T. Menzies, and T. Zimmermann, Eds., Boston: Morgan Kaufmann, 2015, pp. 139–159. doi: <https://doi.org/10.1016/B978-0-12-411519-4.00006-9>.
- [27] I. Vayansky and S. A. P. Kumar, "A review of topic modeling methods," *Inf Syst*, vol. 94, p. 101582, Dec. 2020, doi: 10.1016/j.is.2020.101582.
- [28] U. F. Njoku, A. Abelló, B. Bilalli, and G. Bontempi, "Impact of filter feature selection on classification: an empirical study," 2022. [Online]. Available: <https://www.openml.org>
- [29] Zulhanif, "Pemodelan Topik Dengan Latent Dirichlet Allocation," *Seminar Nasional Pendidikan Matematika*, 1–8., 2016.
- [30] T. K. Landauer and S. T. Dumais, "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge," 1997.
- [31] B. J. Jansen, K. K. Aldous, J. Salminen, H. Almerexhi, and S. Jung, "Data Preprocessing," 2024, pp. 65–75. doi: 10.1007/978-3-031-41933-1\_6.
- [32] D. Varma, A. Nehansh, and P. Swathy, "Data Preprocessing Toolkit : An Approach to Automate Data Preprocessing," *Interantional Journal Of Scientific Research In Engineering And Management*, vol. 07, no. 03, Mar. 2023, doi: 10.55041/IJSREM18270.
- [33] G. C. Banks, H. M. Woznyj, R. S. Wesslen, and R. L. Ross, "A Review of Best Practice Recommendations for Text Analysis in R (and a User-Friendly App)," *J Bus Psychol*, vol. 33, no. 4, pp. 445–459, 2018, doi: 10.1007/s10869-017-9528-3.
- [34] V. B. Kobayashi, S. T. Mol, H. A. Berkers, G. Kismihók, and D. N. Den Hartog, "Text Mining in Organizational Research," *Organ Res Methods*, vol. 21, no. 3, pp. 733–765, Aug. 2017, doi: 10.1177/1094428117722619.
- [35] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations," *Organ Res Methods*, vol. 25, no. 1, pp. 114–146, Jan. 2022, doi: 10.1177/1094428120971683.
- [36] J. H. Lau, D. Newman, and T. Baldwin, "Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, S. Wintner, S. Goldwater, and S. Riezler, Eds., Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 530–539. doi: 10.3115/v1/E14-1056.
- [37] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic Modeling in Embedding Spaces," *Trans Assoc Comput Linguist*, vol. 8, pp. 439–453, Jul. 2020, doi: 10.1162/tacl\_a\_00325.
- [38] A. Septiawan, "Analisis Metode Latent Dirichlet Allocation (LDA) Untuk Clustering Terjemahan Al-Qur'an Bahasa Indonesia," UIN Sunan Kalijaga Yogyakarta, 2020. Accessed: Jul. 07, 2021. [Online]. Available: <https://digilib.uin-suka.ac.id/id/eprint/42685/>
- [39] E. Anggraini, "Latent Dirichlet Allocation untuk Pemodelan Abstrak Dokumen Skripsi (Studi Kasus: Abstrak Dokumen Skripsi Mahasiswa Statistika UII Tahun Angkatan 2011-2015)," Universitas Islam Indonesia, 2020. Accessed: Apr. 03, 2020. [Online]. Available: <http://hdl.handle.net/123456789/23556>

- [40] M. Röder, A. Both, and A. Hinneburg, “Exploring the Space of Topic Coherence Measures,” in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, in WSDM ’15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 399–408. doi: 10.1145/2684822.2685324.
- [41] L. Li, L. Fu, and W. Zhang, “Impact of Text Diversity on Review Helpfulness: A Topic Modeling Approach,” *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 17, pp. 087–100, 2022, doi: 10.28945/4922.

