



Pemanfaatan Teknik Topic Modeling dalam Analisis Bio Informatika untuk Digital dengan LDA

M. Fajar Hidayah Kembaren¹, M. Rafif Syauqi Al-Hadi², Putri Kurni Wati³, Widiya⁴,
Suci Wulandari⁵

¹⁻⁵Sains dan Teknologi, Ilmu Komputer, Universitas Islam Negeri Sumatera Utara
E-mail: fajarkembaren01@gmail.com¹, rafifsyauqi2016@gmail.com², kurniawatiputri61@gmail.com³,
widiawidia856@gmail.com⁴, suci46931@gmail.com⁵

Abstrak

Kemajuan yang cepat dalam bioinformatika, terutama karena perkembangan teknologi sekuensing generasi berikutnya, telah menciptakan volume data biologis digital yang sangat besar dan rumit. Data ini mencakup informasi tentang genom, transkriptom, proteom, dan publikasi ilmiah yang bertambah setiap tahun. Situasi ini memunculkan tantangan besar dalam mengolah, mengelompokkan, dan menarik informasi berarti dari data biologis digital yang ada. Salah satu metode yang dapat digunakan untuk mengatasi tantangan tersebut adalah teknik pemodelan topik. Pemodelan topik adalah metode pembelajaran mesin tanpa pengawasan yang berfokus pada menemukan pola atau tema tersembunyi dalam kumpulan data. Dalam penelitian ini, kami menggunakan algoritma Latent Dirichlet Allocation (LDA) untuk menganalisis data bioinformatika digital yang berupa abstrak jurnal ilmiah. LDA mendeskripsikan setiap dokumen sebagai kombinasi dari beberapa topik dan tiap topik sebagai distribusi probabilistik dari kata-kata tertentu, sehingga dapat mengungkap struktur tersembunyi dalam data teks ilmiah. Hasil dari penelitian ini menunjukkan bahwa penggunaan LDA pada data bioinformatika digital dapat mengidentifikasi tujuh topik utama yang mencerminkan tren penelitian terbaru, termasuk penerapan machine learning dalam analisis kanker, metabolomik, pemodelan protein, dan juga penemuan obat berbasis data. Dengan begitu, teknik pemodelan topik yang menggunakan LDA dapat menjadi alat yang bermanfaat dalam mendukung analisis dan pemetaan pengetahuan bioinformatika di zaman digital ini.

Kata kunci: Bioinformatika, Topic Modeling, Latent Dirichlet Allocation, Big Data Biologis, Analisis Teks.

PENDAHULUAN

Kemajuan di bidang biologi molekuler, seperti teknologi sekuensing generasi berikutnya, telah menciptakan jumlah data biologis yang sangat besar, termasuk DNA, RNA, protein, dan data multi-omics. Peningkatan data ini menghadirkan tantangan besar dalam analisis dan pengolahan karena sifat data yang kompleks, beragam, dan besar. Oleh karena itu, dibutuhkan metode komputasi yang dapat mengelola serta mengekstrak informasi penting dari data biologis yang ada.

Bioinformatika merupakan bidang interdisipliner yang mengalami perkembangan pesat seiring dengan meningkatnya ketersediaan data biologis dalam bentuk digital. Salah satu metode yang bisa digunakan dalam hal ini adalah pemodelan topik, yang merupakan pendekatan pembelajaran tanpa pengawasan untuk menemukan pola atau topik tersembunyi dalam data. Metode ini awalnya banyak digunakan dalam analisis teks, tetapi selama perkembangannya, metode ini juga diterapkan dalam analisis data bioinformatika digital.

Latent Dirichlet Allocation adalah salah satu algoritma pemodelan topik yang paling sering digunakan. Algoritma LDA menganggap setiap dokumen sebagai kombinasi dari beberapa topik, di mana setiap topik diwakili oleh distribusi kata tertentu. Penggunaan LDA dalam bioinformatika telah berhasil dalam mengelompokkan pola ekspresi gen, menganalisis publikasi ilmiah, serta mendukung pencarian biomarker. Karena itu, penelitian tentang penggunaan teknik pemodelan topik dengan LDA dalam analisis bioinformatika digital menjadi sangat penting.

Berdasarkan konteks tersebut, studi ini memiliki tujuan untuk mengeksplorasi penggunaan metode pemodelan topik dengan algoritma Latent Dirichlet Allocation (LDA)

dalam menganalisis data bioinformatika dalam format abstrak jurnal riset. Penelitian ini terfokus pada penemuan pola dan tema penelitian yang tidak tampak dalam kumpulan teks ilmiah yang berkaitan dengan bioinformatika. Dengan pendekatan ini, diharapkan dapat diraih wawasan yang terstruktur tentang perkembangan penelitian dalam bioinformatika digital dan mendukung inisiatif digitalisasi pengetahuan ilmiah di sektor bioinformatika.

TINJAUAN TEORITIS

Bioinformatika adalah disiplin yang mengintegrasikan biologi molekuler dengan komputer, matematika, dan statistik untuk mengatur data biologis yang berukuran besar. Berdasarkan penelitian oleh Yunita et al. (2016), bioinformatika berkembang karena adanya kebutuhan yang semakin tinggi untuk sistem komputasi yang dapat mengelola data dari eksperimen biologi modern. Bioinformatika memainkan peran yang sangat penting dalam kesehatan, bioteknologi, dan lingkungan karena kemampuannya untuk mendukung analisis data genetik dan molekuler dengan cara yang menyeluruh.

Big data biologis memiliki ciri-ciri seperti volume besar, variasi, kecepatan, dan akurasi tinggi. Valle et al. (2025) mengemukakan bahwa dibutuhkan pendekatan komputasi yang lebih maju untuk mendapatkan informasi yang tersembunyi dalam data transkriptomik yang rumit. Tantangan ini telah mendorong penggunaan teknik analisis yang berbasis pembelajaran mesin untuk mendeteksi pola yang tidak dapat ditemukan secara manual.

Pemodelan topik adalah metode pembelajaran tanpa pengawasan yang bertujuan untuk menemukan tema tersembunyi dalam data. Subedi et al. (2020) mengungkapkan bahwa terdapat kesamaan dalam struktur statistik antara data teks dan data biologis, yang memungkinkan penerapan pemodelan topik dalam bioinformatika. LDA, yang diperkenalkan oleh Blei et al. (2021), menjadi metode yang banyak digunakan karena dapat mengungkap pola tersembunyi dengan pendekatan probabilistik. Penelitian oleh Pancheva et al. (2022) dan Sahria et al. (2020) menunjukkan bahwa LDA efektif untuk menganalisis data biologis dan publikasi ilmiah, sehingga memperkuat dasar teoritis dari penelitian ini.

Selanjutnya, penggunaan LDA dalam bioinformatika juga didukung oleh kemampuannya untuk menangani data besar dan tidak terstruktur. Valle et al. (2025) mengindikasikan bahwa pemodelan topik dapat digunakan agar dapat menjelajahi ruang latent dari data transkriptomik, sehingga dapat membedakan kondisi biologis yang rumit, seperti jaringan normal dan kanker. Hal ini menunjukkan bahwa LDA sangat relevan tidak hanya untuk analisis teks, tetapi juga dapat mengungkap struktur biologis yang tersembunyi.

Penelitian lain oleh Pancheva et al. (2022) menekankan penerapan LDA dalam analisis single-cell RNA sequencing (scRNA-seq) untuk menemukan interaksi antar sel. Metode ini memungkinkan untuk mengidentifikasi pola ekspresi gen yang sulit ditemukan dengan cara analisis biasa. Hasil tersebut menunjukkan bahwa LDA menawarkan perspektif baru dalam memahami dinamika seluler di tingkat molekuler.

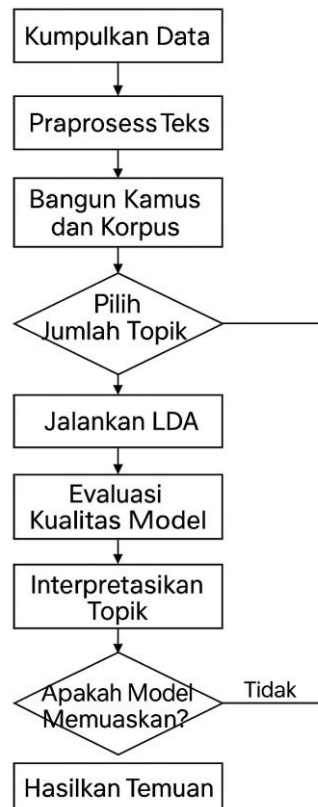
Di tingkat nasional, Sahria et al. (2020) menerapkan LDA untuk memetakan tren penelitian keperawatan di Indonesia melalui repositori jurnal SINTA. Keberhasilan dari studi ini membuktikan bahwa LDA efektif dalam mengelompokkan literatur ilmiah ke dalam tema penelitian yang sistematis. Metode yang serupa bisa juga diterapkan dalam bioinformatika untuk mengelompokkan publikasi serta data biologis digital ke dalam topik-topik yang sesuai.

Dari berbagai studi tersebut, dapat disimpulkan bahwa pemodelan topik dengan LDA memiliki dasar teoritis dan empiris yang kuat untuk digunakan dalam analisis bioinformatika digital. Kemampuannya dalam menangani big data biologis, menemukan pola tersembunyi, serta mendukung digitalisasi ilmu pengetahuan mengindikasikan bahwa LDA adalah metode yang ideal untuk penelitian ini.

METODE PENELITIAN

Flowchart Penelitian

Untuk menjelaskan langkah-langkah dari penelitian yang dilaksanakan, proses penelitian dipresentasikan dalam format diagram alir. Diagram alir ini menunjukkan tahapan penelitian dengan cara yang sistematis, yang mencakup mulai dari pengumpulan data abstrak jurnal, proses awal pengolahan teks, pemodelan topik melalui algoritma LDA, evaluasi model, hingga interpretasi hasil yang didapat.



Gambar 1. Flowchart penelitian

Berdasarkan diagram alir yang terdapat pada Gambar X, tahap awal penelitian melibatkan pengumpulan data ringkasan dari jurnal bioinformatika melalui basis data PubMed dengan memanfaatkan API resmi NCBI. Setelah mendapatkan data, langkah berikutnya adalah melakukan pra-pemrosesan teks untuk menghilangkan gangguan dan menyiapkan data agar siap untuk pemodelan. Setelah itu, dibuatlah kamus dan korpus yang berfungsi sebagai representasi data dalam model LDA. Proses pemodelan dilakukan dengan variasi jumlah topik untuk menemukan model yang paling baik berdasarkan skor koherensi. Model terbaik kemudian akan dianalisis dan ditafsirkan untuk mengidentifikasi tema-tema penting dalam studi bioinformatika digital.

Algoritma Latent Dirichlet Allocation (LDA) dipilih sebagai pendekatan utama dalam studi ini karena kemampuannya dalam mengidentifikasi struktur tersembunyi pada data yang tidak terstruktur tanpa memerlukan pelabelan manual. LDA adalah model probabilistik generatif yang memungkinkan bahwa satu dokumen dapat mencakup lebih dari satu topik, sehingga lebih mencerminkan sifat kompleks dan multidimensional dari data bioinformatika. Jika dibandingkan dengan metode klastering tradisional, LDA dianggap lebih adaptif dan lebih

mudah dipahami dalam mengelompokkan teks ilmiah, terutama pada volume data besar dan beragam seperti publikasi bioinformatika digital.

Penelitian ini menerapkan metode kuantitatif deskriptif dan mengangkat pendekatan text mining serta topic modeling. Fokus utama dari penelitian ini adalah untuk menemukan pola dan tema dalam penelitian di bidang bioinformatika digital dengan menggunakan algoritma LDA. Desain penelitian terdiri dari beberapa langkah, termasuk pengumpulan data, pra-pengolahan teks, pemodelan LDA, evaluasi model, dan interpretasi topik.

Data yang digunakan dalam penelitian diambil dari 100 abstrak jurnal bioinformatika yang diambil dari database PubMed melalui API resmi NCBI. Pengumpulan data dilakukan dengan mencari kata kunci yang berhubungan dengan bioinformatika dan pembelajaran mesin, kemudian data disimpan dalam format CSV untuk mempermudah analisis. Langkah-langkah pra-pengolahan mencakup lowercasing, tokenisasi, penghapusan stopword, dan pembersihan teks agar data siap untuk dimodelkan.

Topic modeling dilakukan dengan bantuan library Gensim, dengan variasi jumlah topik dari satu hingga delapan. Untuk menilai kualitas topik yang dihasilkan, evaluasi model menggunakan coherence score. Model terbaik didapatkan pada jumlah tujuh topik dengan nilai coherence mencapai 0,323, yang kemudian digunakan untuk analisis lebih mendalam.

Teknik Pengumpulan Data

Untuk mendapatkan informasi penelitian, proses ekstraksi abstrak dari jurnal bioinformatika dilakukan secara otomatis dengan menggunakan skrip pemrograman.

```
from Bio import Entrez
import pandas as pd
import time

Entrez.email = "rafifsyauqi2016@gmail.com"
query = ("bioinformatics AND (topic modeling OR LDA OR text mining OR data analysis OR machine learning)")
max_results = 100
handle = Entrez.esearch(db="pubmed", term=query, retmax=max_results, sort="relevance")
record = Entrez.read(handle)
handle.close()
id_list = record["idlist"]

print(f"✓ Ditemukan {len(id_list)} artikel yang cocok dengan query '{query}'")

articles = []
for i, pmid in enumerate(id_list, 1):
    try:
        handle = Entrez.efetch(db="pubmed", id=pmid, rettype="abstract", retmode="xml")
        records = Entrez.read(handle)
        handle.close()

        for article in records["PubMedArticle"]:
            title = article["MedlineCitation"]["Article"]["ArticleTitle"]
            try:
                abstract = article["MedlineCitation"]["Article"]["Abstract"]["AbstractText"][0]
            except KeyError:
                abstract = "(tidak ada abstrak)"

            articles.append({
                "PMID": pmid,
                "Title": title,
                "Abstract": abstract
            })

    print(f"✓ [{i}/{len(id_list)}] (title:{60})..."
    time.sleep(0.5)
except Exception as e:
    print(f"⚠️ Gagal mengambil data untuk PMID {pmid}: {e}")
    continue

df = pd.DataFrame(articles)
df.to_csv("pubmed_bioinformatics_general_abstracts.csv", index=False, encoding="utf-8")
print("\n📄 Abstrak berhasil disimpan ke pubmed_bioinformatics_general_abstracts.csv")
print("\n👁️ Contoh hasil:")
print(df.head())
```

Gambar 2. Source code scraping data abstrak jurnal

Gambar 2 menunjukkan segmen kode yang digunakan dalam pengambilan data ringkasan jurnal bioinformatika dari database PubMed. Aktivitas pengambilan data dijalankan memakai API resmi NCBI (Entrez E-utilities) dengan menggunakan bahasa pemrograman Python. Kode ini bertujuan untuk mengajukan permintaan berdasarkan kata kunci tertentu, mengambil daftar PMID, dan menyaring informasi yang mencakup judul serta ringkasan artikel ilmiah. Data yang berhasil diambil selanjutnya disimpan dalam format yang terstruktur agar analisis berikutnya menjadi lebih mudah.

Proses pengumpulan data ringkasan jurnal dari database PubMed dilakukan melalui sejumlah langkah yang sistematis.

```
PS C:\Users\user\Documents\S7\bioinform> py scrape_abstrak_bioinformatika.py
✓ Ditemukan 100 artikel yang cocok dengan query 'bioinformatics AND (topic modeling OR LDA OR text mining OR data analy
sis OR machine learning)'
✓ [1/100] Bioinformatics Methods for Mass Spectrometry-Based Proteomic...
✓ [2/100] Deep Learning for healthcare: review, opportunities and chal...
✓ [3/100] Machine learning and complex biological data....
✓ [4/100] A deep look into radiomics....
✓ [5/100] Past and future uses of text mining in ecology and evolution...
✓ [6/100] Machine learning and graph analytics in computational biomed...
✓ [7/100] Radiogenomics and IR...
✓ [8/100] Data-Driven Modeling of Pregnancy-Related Complications....
✓ [9/100] Anti-tumor mechanism of Artemisia annua based on bioinforma...
✓ [10/100] A Survey of Data Mining and Deep Learning in Bioinformatics....
✓ [11/100] Unsupervised and self-supervised deep learning approaches fo...
✓ [12/100] Statistical and Machine-Learning Analyses in Nutritional Gen...
✓ [13/100] Revisit of Machine Learning Supported Biological and Biomed...
✓ [14/100] Applications of deep learning for the analysis of medical da...
✓ [15/100] Text Mining for Translational Bioinformatics....
```

Gambar 3. Proses pengambilan abstrak

Gambar 3 memperlihatkan langkah-langkah otomatis dalam mendapatkan abstrak jurnal dari PubMed. Prosedur ini dimulai dengan pengiriman kata kunci pencarian, kemudian dilanjutkan dengan pengambilan metadata artikel melalui endpoint efetch. Data yang berhasil diperoleh mencakup judul dan abstrak dari jurnal ilmiah yang berkaitan dengan bidang bioinformatika. Tujuan dari tahapan ini adalah untuk mengumpulkan dataset teks ilmiah yang mewakili sebagai sumber utama untuk pemodelan topik dengan menggunakan algoritma LDA.

Tahapan Pra-Pengolahan Data (Praprocessing)

Sebelum memulai pemodelan topik, data teks yang telah diperoleh harus menjalani proses pra-pemrosesan untuk memperbaiki mutu analisis.

```
import pandas as pd
import re
from nltk.corpus import stopwords
import nltk

nltk.download('stopwords')
df = pd.read_csv("pubmed_bioinformatics_general_abstracts.csv")
df["Text"] = df["Title"].fillna('') + " " + df["Abstract"].fillna('')

stop_words = set(stopwords.words('english'))
def clean_text(text):
    text = text.lower()
    text = re.sub(r'[^a-z\s]', '', text)
    tokens = [w for w in text.split() if w not in stop_words and len(w) > 3]
    return " ".join(tokens)

df["Clean_Text"] = df["Text"].apply(clean_text)
print(df.head())

df.to_csv("cleaned_bioinformatics_abstracts.csv", index=False)
print("✓ Data teks sudah dibersihkan dan disimpan.")
```

	PMID	Title \
0	32326049	Bioinformatics Methods for Mass Spectrometry-B...
1	28935226	Machine learning and graph analytics in comput...
2	30992073	Machine learning and complex biological data.
3	34653224	Text mining-based word representations for bio...
4	25703199	Text mining of biomedical literature: doing we...

Gambar 4. Pre-processing data yang dijalankan pada Google Colab

Gambar 4 menunjukkan langkah-langkah awal dalam pengolahan data teks yang dilakukan di platform Google Colab. Proses awal ini bertujuan untuk memperbaiki mutu data sebelum diolah menggunakan metode LDA. Langkah-langkah yang dilakukan mencakup pengubahan huruf menjadi huruf kecil, pemisahan kata (tokenisasi), penghapusan simbol dan angka, penghilangan kata-kata umum, serta pembersihan dari karakter khusus. Hasil dari langkah-langkah ini adalah teks yang bersih dan siap dipakai untuk membuat kamus dan kumpulan data.

Pemodelan Topik Modeling dengan LDA

Penilaian mutu model LDA dilakukan dengan menilai nilai coherence score pada berbagai variasi jumlah tema.

```
from gensim.models import CoherenceModel

coherence_model_lda = CoherenceModel(model=lda_model, texts=texts, dictionary=dictionary, coherence='c_v')
coherence_score = coherence_model_lda.get_coherence()
print(f"Coherence Score: {coherence_score:.3f}")

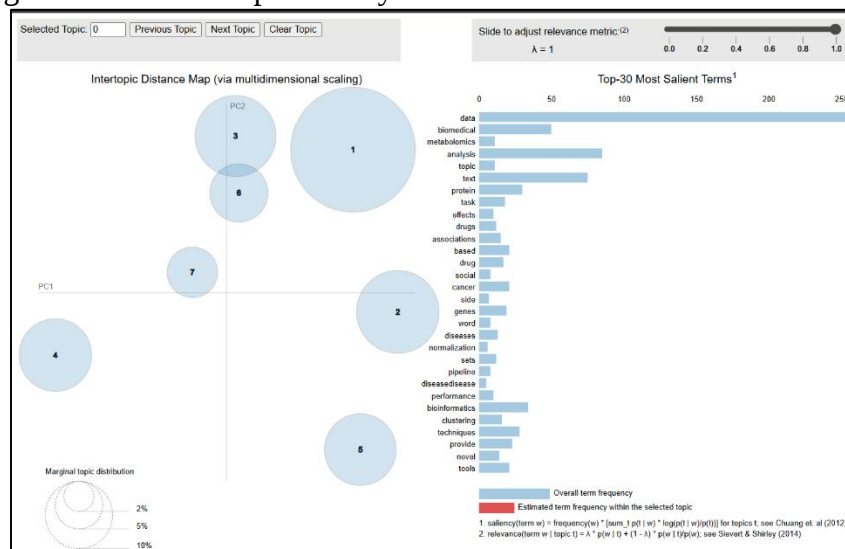
Coherence Score: 0.323
```

Gambar 5. Hasil dari cohenrence score

Gambar 5 memperlihatkan hasil penilaian pada model LDA dengan menggunakan metrik skor koherensi untuk variasi jumlah topik. Penilaian ini bertujuan untuk menemukan jumlah topik yang paling ideal, yang dapat memberikan interpretasi topik yang paling jelas. Melihat grafik yang ada, model yang memiliki tujuh topik memperoleh skor koherensi tertinggi, yaitu 0,323. Dengan demikian, jumlah topik tersebut dipilih sebagai model terunggul untuk analisis selanjutnya.

Evaluasi Model

Untuk mendukung pemahaman hasil dari pemodelan topik, digunakan visualisasi yang interaktif dengan memanfaatkan pustaka PyLDAvis.



Gambar 6. Visualisasi model dengan PyLDAvis

Gambar 6 menunjukkan tampilan hasil dari pemodelan topik yang dilakukan dengan pustaka PyLDAvis. Setiap lingkaran melambangkan satu topik, di mana besar lingkaran mencerminkan persentase topik dalam dataset keseluruhan, sementara jarak antara lingkaran menunjukkan sejauh mana topik-topik tersebut saling berhubungan. Visualisasi ini berfungsi untuk memahami bagaimana topik-distribusi tertentu serta memastikan bahwa topik yang muncul dari model LDA memiliki pemisahan yang jelas dan tidak terlalu tumpang tindih satu sama lain.

HASIL DAN PEMBAHASAN

Hasil yang didapatkan menunjukkan bahwa metode pemodelan topik dengan LDA mampu mengelompokkan ringkasan jurnal bioinformatika ke dalam topik-topik penelitian yang relevan. Penemuan ini menandakan bahwa pendekatan yang berfokus pada data serta pembelajaran mesin semakin memengaruhi penelitian bioinformatika di dunia digital. Temuan penelitian ini konsisten dengan penelitian sebelumnya yang menyatakan bahwa metode pemodelan topik efektif dalam mengungkap pola penelitian dan struktur tematik dalam data



ilmiah yang besar. Oleh karena itu, LDA dapat dimanfaatkan sebagai alat bantu dalam memetakan kemajuan riset bioinformatika secara terstruktur.

Hasil dari analisis model topik menunjukkan bahwa algoritma LDA berhasil menemukan tujuh tema utama dalam data bioinformatika digital. Tema-tema ini mencakup bidang penelitian seperti pembelajaran mesin untuk analisis kanker, metabolomik, pemodelan protein, analisis jaringan biologis, dan penemuan obat berbasis teks. Temuan ini menunjukkan bahwa LDA bisa mengelompokkan dokumen secara tematik meskipun dataset yang digunakan cukup terbatas.

Penilaian menggunakan skor koherensi menunjukkan angka 0,323 yang termasuk dalam kategori cukup baik. Ini menunjukkan bahwa topik yang dihasilkan masih dapat dengan baik diinterpretasikan. Visualisasi dengan PyLDAvis memperlihatkan pemisahan antar topik yang cukup jelas, sehingga memudahkan memahami hubungan antar tema penelitian.

Diskusi mengenai hasil menunjukkan bahwa saat ini, tren penelitian dalam bioinformatika digital lebih didominasi oleh pendekatan berbasis data dan pembelajaran mesin. Penemuan ini konsisten dengan penelitian sebelumnya yang mengungkapkan bahwa digitalisasi dan big data mendorong penerapan metode komputasi yang lebih canggih dalam bioinformatika. Dengan demikian, LDA terbukti bermanfaat sebagai alat analisis untuk mengidentifikasi tren dan pola dalam penelitian ilmiah.

KESIMPULAN DAN SARAN

Penelitian ini menyumbangkan dalam wujud penerapan metode pemodelan topik yang berlandaskan LDA untuk mengevaluasi data bioinformatika digital, terutama dalam publikasi akademik. Harapan dari kontribusi ini adalah agar dapat dijadikan acuan metodologi untuk penelitian mendatang yang berkaitan dengan analisis tren penelitian bioinformatika, serta membantu dalam proses pengembangan sistem analisis data biologi yang berbentuk digital.

Berdasarkan hasil studi, dapat dikatakan bahwa metode topic modeling dengan algoritma LDA mampu mengenali pola dan tema utama dalam data bioinformatika digital. Tujuh tema utama berhasil diidentifikasi dengan nilai koherensi mencapai 0,323, yang menandakan bahwa LDA bisa diaplikasikan untuk analisis tren penelitian meskipun hanya dengan dataset yang terbatas.

Studi ini juga menunjukkan bahwa tema-tema yang teridentifikasi mencerminkan perkembangan terbaru dalam bioinformatika, seperti penggunaan machine learning, analisis metabolomik, dan pemodelan protein. Ini menegaskan bahwa LDA dapat membantu dalam proses digitalisasi pengetahuan ilmiah di bidang bioinformatika.

Sebagai rekomendasi, pada penelitian berikutnya sebaiknya menggunakan dataset yang lebih besar serta menerapkan teknik pra-pengolahan yang lebih canggih untuk meningkatkan kualitas topik. Selain itu, akan baik juga mempertimbangkan penggunaan metode topic modeling lain sebagai perbandingan agar hasil analisis lebih menyeluruh.

DAFTAR PUSTAKA

- Basuki, S., & Fadillah, N. M. (2025). Implementasi algoritma topic modeling pada abstrak paper ilmiah untuk deteksi tren topik penelitian. *Repositor*, 7(1), 7–22..
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2021). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Bustami, D. K., & Noviaristanti, S. (2022). Service Quality Analysis of Tokopedia Application Using Text Mining Method. *International Journal of Management, Finance and Accounting*, 3(1), 1–21. <https://doi.org/10.33093/ijomfa.2022.3.1.1>

- C. L. Prawirosastro and E. P. A. Akhmad, “Pemodelan Topik Menggunakan Latent Dirichlet Allocatin Dan Pachinko Allocation Model Untuk Ekstraksi Berita Saham Online”, Program Diploma Pelayaran, Universitas Hang Tuah, 2022.
- Carbonetto, P., Sarkar, A., Wang, Z., & Stephens, M. (2025). Non-negative matrix factorization algorithms generally improve topic model fits. arXiv preprint arXiv:2105.13440v4.
- Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology*, 7, 886498. <https://doi.org/10.3389/fsoc.2022.886498>
- Firdaus, M. R., Rizki, F. M., Gaus, F. M., & Susanto, I. K. (2020). Analisis sentiment dan topic modelling dalam aplikasi ruangguru. *J-SAKTI (Jurnal Sains Komputer Informatika)*, 4(1), 66-76. <https://doi.org/10.30645/j=sakti.v4i1.188>
- Musliadi, K.H., Zainuddin, H., & Wabula, Y. (2022). Twiter Social Media Conversion Topik Treeding Analysis Using Latent Dirichlet Allocation Algorithm. *Journal of Applied Engineering and Technological Science (JAETS)*, 4(1), 390-399. <https://doi.org/10.37385/jeats.v4i1.1143>
- Muthusami, R., Kandan, N. M., Saritha, K., Narenthiran, B., Nagaprasad, N., & Ramaswamy, K. (2024). Investigating topic modeling techniques through evaluation of topics discovered in short texts data across diverse domains. *Scientific Reports*, 14, 12003. <https://doi.org/10.1038/s41598-024-61738-4>.
- Pancheva, A., Wheadon, H., Rogers, S., & Otto, T. D. (2022). Using topic modeling to detect cellular crosstalk in scRNA-seq. *PLOS Computational Biology*, 18(4), e1009975. <https://doi.org/10.1371/journal.pcbi.1009975>
- Rahmawati, D., & Shabrina, H. (2024). Analisis molekuler dan bioinformatika. SEAMEO BIOTROP. ISBN 978-979-8275-99-9.
- Sahria, Y., Febriarini, N. I., & Oktavianti, P. D. (2020). Pemodelan topik penelitian bidang keperawatan Indonesia pada repository jurnal Sinta menggunakan metode topic modelling LDA (Latent Dirichlet Allocation). *Prosiding Seminar Nasional Teknologi Informasi & Ilmu Komputer (SEMASTER)*, 1(1), 90–102. ISSN 1978-1520.
- Subedi, S., Chen, L., & Liu, J. (2020). A comprehensive review of topic modeling in biomedical data analysis. *Briefings in Bioinformatics*, 21(6), 2022–2037. <https://doi.org/10.1093/bib/bbz107>
- Valle, F., Caselle, M., & Osella, M. (2025). Exploring the latent space of transcriptomic data with topic modeling. *NAR Genomics and Bioinformatics*, 7(2), lqaf049. <https://doi.org/10.1093/nargab/lqaf049>
- Wang, J., & Zhang, X.-L. (2023). Deep NMF topic modeling. *Neurocomputing*, 515, 157-173. <https://doi.org/10.1016/j.neucom.2022.10.002>
- Widodo, A. O., Septiadi, F., & Rakhmawati, N. A. (2023). Analisis Tren Konten Pada Vtuber Indonesia Menggunakan Latent Dirichlet Allocation. *Jurnal Informatika Dan Rekayasa Elektronik*, 6(1), 56-63. <https://doi.org/10.3390/su14158992>
- Yunita, I., Tjandradiredja, K., & Hansun, S. (2016). Perkembangan bioinformatics dalam ruang lingkup ilmu komputer. *Ultimatics Jurnal Teknik Informatika*, 8(1), 65–72. <https://doi.org/10.31937/ti.v8i1.505>