INOTEK, Vol. 9
ISSN: 2580-3336 (Print) / 2549-7952 (Online)
Url: https://proceeding.unpkediri.ac.id/index.php/inotek/

Optimalisasi Latent Dirichlet Allocation untuk Ekstraksi Topik Utama dalam Teks Dongeng

[1*]Yosan Okta Odhianto, [2]Daniel Swanjaya, [3]Julian Sahertian
Teknik Informatika, Universitas Nusantara PGRI Kediri
E-mail: [1]*yosanokta12@gmail.com,* [2]*daniel@unpkediri.ac.id,* [3]*Juliansahertian@unpkediri.ac.id*
Penulis Korespondens : Yosan Okta Odhianto

Abstrak— *Latent Dirichlet Allocation* (LDA) adalah algoritma topic modeling yang bekerja tanpa label data dan sangat dipengaruhi oleh pra-pemrosesan dan pengaturan parameter. Penelitian ini bertujuan mengoptimalkan LDA untuk mengekstraksi topik utama dari 100 teks dongeng berbahasa Indonesia. Teks diproses menggunakan berbagai kombinasi teknik pra-pemrosesan seperti tokenisasi, *stopword removal*, *stemming*, dan normalisasi. Eksperimen dilakukan dengan memvariasikan jumlah topik (K) serta parameter *alpha* dan *eta*. Evaluasi menggunakan *coherence score* untuk menilai konsistensi semantik topik. Hasil terbaik diperoleh pada kombinasi pra-pemrosesan kedua dengan 15 topik, menghasilkan *coherence score* tertinggi sebesar 0,4885. Temuan ini menunjukkan bahwa pemilihan pra-pemrosesan dan parameter yang tepat dapat meningkatkan kualitas topik secara signifikan. Penelitian ini diharapkan mendukung pengembangan analisis topik pada teks naratif Indonesia.
Kata Kunci— *Coherence Score*, Dongeng, *Latent Dirichlet Allocation*, Pemodelan Topik, Pra-pemrosesan Teks Bahasa Indonesia

Abstract— *Latent Dirichlet Allocation (LDA) is a topic modeling algorithm that works without labeled data and is greatly influenced by preprocessing and parameter settings. This study aims to optimize LDA to extract the main topics from 100 Indonesian fairy tale texts. The texts were processed using various combinations of preprocessing techniques such as tokenization, stopword removal, stemming, and normalization. Experiments were conducted by varying the number of topics (K) as well as the alpha and eta parameters. Evaluation was performed using the coherence score to assess the semantic consistency of the topics. The best results were obtained with the second preprocessing combination and 15 topics, producing the highest coherence score of 0.4885. These findings indicate that proper selection of preprocessing and parameters can significantly improve topic quality. This study is expected to support the development of topic analysis for Indonesian narrative texts.*
Keywords— *Coherence Score, Fairy Tale Latent Dirichlet Allocation,Topic Modeling, Preprocessing, Indonesian Text*

I. PENDAHULUAN

Teks digital berbahasa Indonesia, termasuk cerita rakyat dan dongeng, semakin mudah diakses. Namun, pengorganisasian dan pemahaman topik utama teks ini menjadi tantangan. Topic modeling menjadi solusi populer dalam NLP, dengan *Latent Dirichlet Allocation* (LDA) sebagai metode paling umum. LDA adalah model probabilistik generatif untuk menemukan topik laten dalam teks besar melalui pembelajaran tanpa pengawasan. Meskipun banyak digunakan, aplikasi LDA pada teks dongeng Indonesia masih terbatas. Performa LDA sangat

bergantung pada pra-pemrosesan teks, pemilihan jumlah topik (K), dan parameter model. Eksplorasi yang kurang terhadap aspek ini dapat menghasilkan model yang tidak representatif.

Penelitian ini bertujuan mengoptimalisasi model LDA untuk ekstraksi topik utama dari teks dongeng Indonesia. Optimalisasi dilakukan dengan mengevaluasi berbagai kombinasi pra-pemrosesan dan jumlah topik, serta menilai hasilnya menggunakan *coherence score*. *Topic coherence* adalah metrik kuantitatif untuk mengevaluasi kualitas topik, dengan visualisasi menggunakan *pyLDAvis* untuk eksplorasi interaktif. Penelitian ini diharapkan berkontribusi pada pengembangan sistem analisis topik berbasis topic modeling dan memperluas aplikasi LDA pada teks berbahasa Indonesia.

II. METODE

A. Subjek Penelitian

Penelitian ini menggunakan 100 dokumen teks dongeng berbahasa Indonesia dari situs web dan repositori cerita rakyat daring. Dataset terdiri dari 100 dokumen teks naratif yang masing-masing mengandung satu cerita utuh. Seluruh teks berbentuk digital (format .pdf) dan tidak memiliki label topik sebelumnnya, sehingga sesuai untuk pendekatan *unsupervised learning*.

B. Alat dan Perangkat Lunak

Penelitian ini menggunakan *Python* (v3.10) dengan pustaka NLTK, Sastrawi (pra-pemrosesan), *gensim* (model LDA), *Matplotlib*, dan *pyLDAvis* (visualisasi).

C. Desain Eksperimental

Digunakan desain eksperimental eksploratif untuk menguji pengaruh pra-pemrosesan dan jumlah topik terhadap kualitas model LDA. Variasi meliputi:

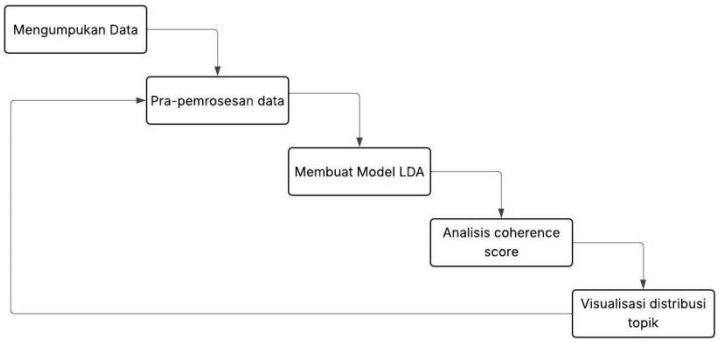
1. Jumlah topik (K) : [5,10,15,20]
2. Parameter LDA : *Alpha* [*asymmetric, symetric*, 0.1, 0.5] , *Eta* : [*auto* , 0.1, 0.01]
3. Pra-pemrosesan teks : dilakukan dengan tiga kombinasi berbeda, yaitu :
   a. (1)Tokenisasi + *case folding* + *stopword removal*
   b. (2)kombinasi (1) + *stemming*
   c. (3)kombinasi (2) + menghapus kata kosong (“”) dan *non-alfanumerik*
   d. (4) kombinasi (3) + normalisasi teks

D. Variabel Penelitian dan Evaluasi

1. Variabel bebas : Kombinasi pra-pemrosesan, jumlah topik (K), dan parameter LDA.
2. Variabel terikat : Kualitas topik yang diukur dengan *coherence score* (c_v) dari pustaka *gensim*.

E. Prosedur Penelitian

Prosedur mencangkup : koleksi data, pra-pemrosesan, pemodelan LDA, evaluasi *coherence score*, dan visualisasi topik.



Gambar 1 Prosedur Penelitian

Diagram yang disajikan merepresentasikan alur kerja sekuensial yang menyerupai model waterfall, dimulai dari Mengumpulkan Data dan dilanjutkan ke Pra-pemrosesan Data untuk persiapan analisis. Setelah itu, Model LDA dibuat atau dilatih, diikuti oleh Analisis *Coherence Score* untuk evaluasi kualitas topik yang dihasilkan. Meskipun alur utama bersifat linear, panah kembali dari tahapan analisis dan pra-pemrosesan mengindikasikan adanya iterasi untuk mengoptimalkan model, sebelum berakhir pada Visualisasi Distribusi Topik

III. HASIL DAN PEMBAHASAN

A. *Latent Dirichlet Allocation*

Pemodelan topik telah muncul sebagai solusi NLP yang populer, dengan *Latent Dirichlet Allocation*(LDA) menjadi metode yang paling banyak digunakan[1]. LDA adalah teknik unsupervised learning yang digunakan untuk mengidentifikasi topik tersembunyi dalam kumpulan dokumen dengan memodelkan distribusi probabilitas kata-kata dan topik[2]. Sebagai metode pembelajaran tanpa pengawasan (*unsupervised learning*), LDA bekerja pada data tanpa label, mengidentifikasi struktur topik tersembunyi dari suatu korpus dokumen[3]. Algoritma ini mengasumsikan bahwa setiap dokumen adalah campuran dari beberapa topik, dan setiap topik adalah distribusi probabilitas atas kata-kata tertentu[4]. Proses utama LDA melibatkan inferensi untuk mengestimasi distribusi topik per dokumen dan distribusi kata per topik. Performa model LDA sangat bergantung pada beberapa aspek penting seperti teknik pra-pemrosesan teks dan pemilihan parameter yang digunakan. Implementasi LDA yang umum sering menggunakan pustaka seperti gensim.

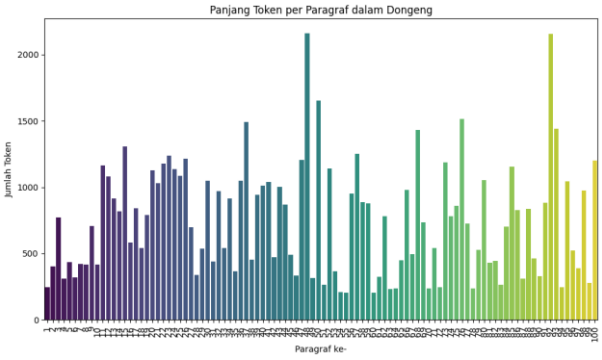
B. *Coherence Score*

Coherence score merupakan metrik penting yang digunakan untuk mengevaluasi kualitas topik yang dihasilkan oleh algoritma topic modeling seperti LDA[5]. Metrik ini mengukur konsistensi semantik antar kata-kata yang membentuk suatu topik, dengan tujuan untuk menilai seberapa mudah topik tersebut dapat diinterpretasikan dan dipahami oleh manusia[6]. Sebuah *coherence score* yang lebih tinggi mengindikasikan topik yang lebih koheren dan bermakna[7]. Dalam penelitian ini, evaluasi model LDA dilakukan menggunakan *coherence score* (c_v), yang dinilai efektif dalam mengukur interpretasi manusia terhadap topik[8]. Visualisasi seperti *pyLDAvis* juga dapat digunakan untuk

mendukung evaluasi koherensi topik secara interaktif dan meningkatkan interpretabilitas hasil LDA.

C. Deskripsi Dataset

Digunakan 100 teks dongeng berbahasa Indonesia dengan panjang bervariasi. Visualisasi data yang ditampilkan pada Gambar 1 diperoleh setelah setiap teks dongeng melalui proses tokenisasi.



Gambar 2 Distribusi Kata Sebelum Pra-pemrosesan

D. Hasil Pra-pemrosesan teks

Analisis dilakukan terhadap hasil pra-pemrosesan data guna meningkatkan kualitas data dan mengevaluasi performa model secara lebih akurat.

1. *Stopword removal*

Stopword removal adalah langkah praproses penting yang menyaring kata-kata yang mengandung sedikit informasi (misalnya, preposisi, konjungsi) untuk meningkatkan kinerja model topik[9].Menghapus *stopwords* secara drastis mengurangi jumlah token. Hasil pra-pemrosesan dapat dilihat pada *Tabel 1*.

Tabel 1 *Stopword Removal*

Sebelum Proses	Sesudah Proses
['aku', 'sendirian', ',', 'aku', 'tersesat', 'karena', 'aku', 'tidak', 'mendengar', 'perkataan', 'ibuku.', '"', '"', 'wahai', 'anak', 'domba', 'yang', 'manis', ',', 'ketahuilah', 'apa', 'yang', 'dilarang', 'ibumu', 'itu', 'adalah', 'bentuk', 'dari', 'rasa', 'sayangnya', 'kepadamu', '"', 'lalu', ',', 'sang', 'monyet', 'mengantar', 'anak', 'domba', 'ke', 'tepi', 'sungai', 'agar', 'anak', 'domba', 'bisa', 'kembali', 'bersama', 'ibunya', '.']	['tersesat', 'mendengar', 'perkataan', 'ibuku.', '"', '"', 'anak', 'domba', 'manis', ',', 'ketahuilah', 'dilarang', 'ibumu', 'bentuk', 'sayangnya', 'kepadamu', '"', ',', 'sang', 'monyet', 'mengantar', 'anak', 'domba', 'tepi', 'sungai', 'anak', 'domba', 'ibunya', '.']

2. Stemming

Stemming adalah proses dalam *Natural Language Processing* (NLP) untuk mengubah kata berimbuhan atau turunan menjadi bentuk dasarnya[10].

Tabel 2 *Stemming Token*

Sebelum Proses	Sesudah Proses
['anak', 'domba', 'manis', ',', 'ketahuilah', 'dilarang', 'ibumu', 'bentuk', 'sayangnya', 'kepadamu', '"', ',', 'sang', 'monyet', 'mengantar', 'anak', 'domba', 'tepi', 'sungai', 'anak', 'domba', 'ibunya', '.']	['anak', 'domba', 'manis', '', 'tahu', 'larang', 'ibu', 'bentuk', 'sayang', 'kepada', '', '', 'sang', 'monyet', 'antar', 'anak', 'domba', 'tepi', 'sungai', 'anak', 'domba', 'ibu', '']

Dari *Tabel 2* bisa dilihat jumlah token dari proses sebelumnnya (*stopword removal*) dibanding setelah proses stemming tidak berkurang signifikan, tetapi ada beberapa kata yang berubah, misalnya dari "'sayangnya"' menjadi "sayang".

3. Hapus kata kosong dan *non-alfanumerik*

Karene dalam *library* ada beberapa kata yang tidak terfilter, maka perlu difilter ulang agar string kosong (‘’) dan beberapa tanda baca hilang.

Tabel 3 Penghapusan Kata Kosong

Sebelum proses	Sesudah proses
['anak', 'domba', 'malang', 'desa', 'pencil', 'hidup', 'domba', 'anak', '', 'senang', 'jalan', '-', 'jalan', 'pinggir', 'sungai', 'cari', 'makan', '', '', 'domba', 'larang', 'anakanaknya', 'menyebrangi', 'sungai']	['anak', 'domba', 'malang', 'desa', 'pencil', 'hidup', 'domba', 'anak', 'senang', 'jalan', 'jalan', 'pinggir', 'sungai', 'cari', 'makan', 'domba', 'larang', 'anakanaknya', 'menyebrangi', 'sungai',]

4. Normalisasi

Proses normalisasi disini dimaksuk untuk mengubah sebuah kata menjadi satu makna yang konsisten. Misalnya mengubah berbagai jenis hewan menjadi satu kata (“hewan”).

Tabel 4 Normalisasi token

Sebelum proses	Sesudah proses
['anak', 'domba', 'malang', 'desa', 'pencil', 'hidup', 'domba', 'anak', 'senang', 'jalan', 'jalan', 'pinggir', 'sungai', 'cari', 'makan', 'domba', 'larang']	['anak', 'hewan', 'malang', 'desa', 'pencil', 'hidup', 'hewan', 'anak', 'senang', 'jalan', 'jalan', 'pinggir', 'sungai', 'cari', 'makan', 'hewan', 'larang',]

Dari *Tabel 4* bisa dilihat bahwa beberapa kategori hewan berubah menjadi satu kata “hewan”.

E. Eksperimen model LDA

Sub-bab ini menjelaskan eksperimen penentuan jumlah topik (K) optimal dalam pemodelan topik LDA. Eksperimen melibatkan variasi K dan kombinasi pra-pemrosesan

untuk evaluasi kinerja model berdasarkan *coherence score*. *Coherence score* yang lebih tinggi menunjukkan topik yang lebih koheren.

1. Tabel Hasil *Coherence Score* dari Tiap Percobaan Pra-pemrosesan

Tabel 5 Hasil *Coherence Score*

Percobaan Pra-pemrosesan	Jumlah Topik (K)	Passes	Parameter Alpha Optimal	Parameter Eta Optimal	Coherence Score Tertinggi
Pertama	5	150	symmetric	auto	0.4040
	10	150	0.5	auto	0.3965
	15	150	asymmetric	auto	0.3903
	20	150	0.5	0.1	0.3644
Kedua	5	150	asymmetric	auto	0.4166
	10	150	0.5	auto	0.4016
	15	150	0.5	auto	0.4885
	20	150	0.5	0.1	0.4544
Ketiga	5	150	asymmetric	0.01	0.4426
	10	150	0.5	auto	0.4117
	15	150	asymmetric	0.01	0.3887
	20	150	0.1	auto	0.3609
Keempat	5	150	0.5	auto	0.4050
	10	150	0.5	auto	0.3965
	15	150	asymmetric	auto	0.3903
	20	150	0.5	0.1	0.3644

2. Analisis Hasil

Berdasarkan serangkaian eksperimen yang telah dilakukan, model *Latent Dirichlet Allocation* (LDA) menunjukkan kinerja yang bervariasi tergantung pada kombinasi jumlah topik (K) dan teknik pra-pemrosesan yang diterapkan. Analisis mendalam terhadap *coherence score* dari setiap konfigurasi memberikan wawasan penting mengenai parameter optimal untuk ekstraksi topik yang koheren dari teks dongeng berbahasa Indonesia.

a. Model terbaik keseluruhan

Dari seluruh percobaan, model LDA dengan 15 topik terbukti menjadi konfigurasi paling efektif dalam menghasilkan topik yang koheren. Model ini mencapai *coherence score* tertinggi yaitu 0.4885, yang merupakan indikator kuat dari kualitas topik yang dapat diinterpretasikan. Konfigurasi optimal ini dicapai dengan menggunakan kombinasi pra-pemrosesan kedua, yang meliputi tokenisasi, *case folding*, *stopword removal*, dan *stemming*. Parameter LDA yang berperan

dalam hasil optimal ini adalah *alpha* sebesar 0.5 dan *eta* yang diatur ke *'auto'*. Tingginya *coherence score* pada konfigurasi ini menunjukkan bahwa topik-topik yang dihasilkan memiliki keterkaitan semantik yang kuat antar kata-kata pembentuknya, sehingga lebih mudah untuk dipahami dan dianalisis secara kualitatif

b. Pengaruh pra-pemrosesan

Tahapan pra-pemrosesan memainkan peran krusial dalam menentukan kualitas topik yang diekstrak oleh model LDA. Hasil eksperimen secara konsisten menunjukkan bahwa penambahan stemming (seperti yang dilakukan pada kombinasi pra-pemrosesan kedua) dan penghapusan kata kosong/non-alfanumerik (pada kombinasi ketiga) cenderung meningkatkan *coherence score* secara signifikan dibandingkan dengan hanya melakukan tokenisasi, *case folding*, dan *stopword removal*. Proses *stemming* membantu mengurangi variasi bentuk kata ke akar katanya, yang mengkonsentrasikan bobot topik pada konsep inti. Sementara itu, penghapusan string kosong dan karakter *non-alfanumerik* lebih lanjut membersihkan data dari *noise* yang tidak relevan, memungkinkan model untuk fokus pada kata-kata yang bermakna. Di sisi lain, normalisasi (kombinasi pra-pemrosesan keempat), meskipun bertujuan untuk menyatukan makna kata (misalnya, semua jenis hewan menjadi "hewan"), tidak selalu memberikan peningkatan *coherence score* yang signifikan pada semua jumlah topik yang diuji. Hal ini mengindikasikan bahwa pada beberapa kasus, reduksi makna yang terlalu agresif melalui normalisasi dapat mengurangi informasi kontekstual yang diperlukan untuk membentuk topik yang sangat koheren. Pemilihan teknik pra-pemrosesan yang sesuai harus mempertimbangkan karakteristik spesifik dari dataset dan tujuan analisis topik.

IV. KESIMPULAN

Optimalisasi model LDA untuk ekstraksi topik utama dari teks dongeng berbahasa Indonesia dapat dicapai melalui kombinasi pra-pemrosesan teks yang tepat dan pemilihan jumlah topik yang optimal. Peningkatan *coherence score* secara signifikan menunjukkan bahwa model yang dioptimalkan menghasilkan topik yang lebih akurat dan mudah diinterpretasikan. Penelitian ini diharapkan berkontribusi pada pengembangan sistem analisis topik berbasis topic modeling dan perluasan aplikasi LDA dalam konteks teks berbahasa Indonesia.

DAFTAR PUSTAKA

[1] T. K. Landauer, F. Peter W., and D. and Laham, “An introduction to latent semantic analysis,” *Discourse Process*, vol. 25, no. 2–3, pp. 259–284, Jan. 2018, doi: 10.1080/01638539809545028.

[2] H. Jelodar *et al.*, “Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey,” *Multimed Tools Appl*, vol. 78, no. 11, pp. 15169–15211, 2019, doi: 10.1007/s11042-018-6894-4.

[3] Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch, “Handbook of Latent Semantic Analysis,” *Psychology Press*, 2019.

[4] K. F. Nurdin, T. E. Sutanto, and A. Santoso, "Analisa Pemodelan Topik Berita Daring Menggunakan Semi-Supervised dan Fully Unsupervised Latent Dirichlet Allocation," *jptam*, vol. 7, 2023, doi: https://doi.org/10.31004/jptam.v7i2.7506.

[5] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2018. doi: DOI: 10.1017/CBO9780511809071.

[6] 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.

[7] J. Wintrode and S. Khudanpur, "Can You Repeat That? Using Word Repetition to Improve Spoken Term Detection," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, K. Toutanova and H. Wu, Eds., Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 1316–1325. doi: 10.3115/v1/P14-1124.

[8] F. Zhang, J. Yao, and R. Yan, "On the Abstractiveness of Neural Document Summarization," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds., Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 785–790. doi: 10.18653/v1/D18-1089.

[9] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2013. doi: DOI: 10.1017/CBO9780511809071.

[10] G. Imin, M. Ablimit, H. Yilahun, and A. Hamdulla, "A Character String-Based Stemming for Morphologically Derivative Languages," *Information (Switzerland)*, vol. 13, no. 4, Apr. 2022, doi: 10.3390/info13040170.