Terbit *online* pada laman : http://teknosi.fti.unand.ac.id/

Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |

Artikel Penelitian

Optimasi LDA untuk Analisis Keluhan Nasabah Perbankan dengan Grid Search

Rika Afriyani[1,*], Eka Angga Laksana[2].
[1]*Teknik Informatika, Fakultas Teknik, Universitas Widyatama, Indonesia*
[2]*Teknik Informatika, Fakultas Teknik, Universitas Widyatama, Indonesia*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 13 Maret 2025
Revisi Akhir: 08 Mei 2025
Diterbitkan *Online*: 01 September 2025

KATA KUNCI

Penyetelan Parameter Grid Search
Kepuasan nasabah
Keluhan nasabah perbankan
Latent Dirichlet Allocation (LDA)
Pemodelan topik.

KORESPONDENSI

E-mail:
rika.afriyani@widyatama.ac.id

A B S T R A C T

Penelitian ini bertujuan untuk menganalisis topik pada data keluhan nasabah perbankan menggunakan metode *Latent Dirichlet Allocation* (LDA) yang disempurnakan dengan penyetelan parameter menggunakan Grid Search. Dataset berasal dari situs *ConsumerFinance.gov* dengan total 6,3 juta entri keluhan dari tahun 2011 hingga 2024, dan 50% data digunakan untuk menjaga representasi dan menyederhanakan analisis. Dalam analisis ini, metode LDA digunakan untuk mengidentifikasi topik-topik tersembunyi, sementara Grid Search meningkatkan koherensi model. Hasil penelitian menunjukkan bahwa keluhan nasabah dapat dikelompokkan ke dalam 10 topik utama, seperti masalah laporan keluhan (25,67%), kesalahan pembayaran (18,10%), otorisasi data (12,20%), dan kebijakan kredit (10,77%). Optimasi parameter berhasil meningkatkan skor koherensi model dari 0,49 menjadi 0,56, mencerminkan peningkatan kualitas clustering topik. Perbandingan metode LDA standar dan LDA dengan Grid Search menunjukkan bahwa metode optimasi memberikan nilai rata-rata koherensi lebih tinggi (0,52 vs. 0,42). Dengan temuan ini, model LDA yang telah di optimasi dapat digunakan oleh industri perbankan untuk memahami dan menangani keluhan pelanggan dengan lebih efektif.

1. PENDAHULUAN

Latent Dirichlet Allocation (LDA) merupakan salah satu metode paling populer dalam analisis teks dan pemodelan topik. Di era digital, jumlah data teks yang tidak terstruktur terus meningkat, sehingga kemampuan untuk mengidentifikasi pola dan mengekstrak informasi dari data tersebut menjadi semakin krusial. Namun, meskipun LDA banyak digunakan, metode ini memiliki kelemahan utama berupa ketidakstabilan hasil pada setiap iterasi. Variasi ini dapat mempengaruhi keandalan interpretasi serta kesimpulan yang diambil, sehingga diperlukan strategi untuk meningkatkan stabilitas model.

Berbagai penelitian telah mengeksplorasi manfaat dan keterbatasan LDA. Meskipun metrik seperti perplexity sering digunakan untuk mengoptimalkan model, pendekatan ini tidak selalu selaras dengan penilaian manusia terhadap kualitas topik yang dihasilkan. Alternatif lain, seperti optimasi berbasis *semantic coherence*, telah diusulkan, tetapi belum memiliki kriteria seleksi yang mapan. Salah satu faktor utama yang memengaruhi ketidakstabilan LDA adalah pemilihan hyperparameter yang kurang optimal. Nilai alpha atau beta yang terlalu rendah, misalnya, dapat menghasilkan distribusi topik yang kurang bermakna atau tidak seimbang. Hyperparameter yang tidak sesuai juga berkontribusi terhadap tingginya variasi hasil antar iterasi, sehingga menurunkan reliabilitas model.

Pendekatan sistematis dalam pemilihan hyperparameter menjadi aspek krusial dalam meningkatkan kualitas dan stabilitas model LDA. Grid Search Parameter Tuning menawarkan solusi dengan mengeksplorasi berbagai kombinasi parameter secara menyeluruh guna menemukan konfigurasi terbaik. Metode ini tidak hanya mengurangi variabilitas stokastik, tetapi juga meningkatkan relevansi dan konsistensi topik yang dihasilkan. Keunggulan utama Grid Search Parameter Tuning terletak pada kemampuannya untuk mengevaluasi berbagai konfigurasi parameter, seperti jumlah topik, alpha, dan beta, secara sistematis, sehingga model yang dihasilkan lebih optimal berdasarkan metrik yang relevan, termasuk *semantic coherence* dan *human interpretability*.

https://doi.org/10.25077/TEKNOSI.v11i2.2025.098-106

Dalam penelitian ini, efektivitas Grid Search Parameter Tuning dalam meningkatkan keandalan hasil LDA akan dievaluasi melalui penerapannya pada dataset keluhan nasabah bank. Data ini mencakup berbagai keluhan terkait layanan dan produk perbankan, sehingga dapat digunakan untuk menilai kemampuan model dalam mengidentifikasi pola yang bermakna. Evaluasi dilakukan dengan menggunakan metrik seperti *semantic coherence* dan *human interpretability*, serta mempertimbangkan aspek stabilitas model guna memastikan bahwa parameter optimal menghasilkan hasil yang dapat direproduksi.

Hasil penelitian ini diharapkan dapat menunjukkan bahwa pendekatan optimasi hyperparameter berbasis Grid Search memberikan peningkatan signifikan dalam stabilitas, relevansi, dan kualitas model LDA. Dengan demikian, metode ini dapat berkontribusi pada pengembangan teknik analisis teks yang lebih andal, khususnya dalam memahami pola keluhan nasabah di sektor perbankan.

Berbagai penelitian telah dilakukan untuk memahami keluhan konsumen menggunakan metode yang berbeda. Chen [1] meneliti permasalahan penipuan dan reliabilitas dalam platform lelang daring dengan pendekatan analisis semi-otomatis berbasis teori perilaku konsumen. Studi ini menemukan bahwa isu utama yang dihadapi adalah ketidakjujuran penjual dan rendahnya keandalan platform. Sementara itu, Ayres, Lingwall, dan Steinway [2] menggunakan analisis regresi untuk mengkaji hubungan antara faktor demografi dan keluhan hipotek dalam teori ekonomi konsumen. Hasilnya menunjukkan bahwa wilayah dengan populasi lanjut usia lebih rentan terhadap masalah hipotek, serta bank-bank besar cenderung memiliki waktu respons lebih lama dalam menangani keluhan pelanggan.

Littwin [3] melakukan analisis regresi dan pendekatan kualitatif untuk mengevaluasi efektivitas Consumer Financial Protection Bureau (CFPB) dalam menangani sengketa konsumen berdasarkan teori regulasi konsumen. Studi ini menemukan bahwa CFPB berperan dalam meningkatkan goodwill pelanggan melalui penyelesaian sengketa yang lebih efektif. Berezina et al. [4] menerapkan penambangan teks dan analisis sentimen dalam teori kepuasan pelanggan untuk mengidentifikasi pola dalam ulasan hotel. Hasilnya mengungkapkan bahwa keluhan pelanggan terkait kualitas layanan dan fasilitas hotel menjadi faktor utama ketidakpuasan.

Lebih lanjut, Bastani et al. [5] menggunakan metode Latent Dirichlet Allocation (LDA) untuk menganalisis data keluhan pelanggan yang dikumpulkan dari CFPB. Berdasarkan teori pemodelan topik, penelitian ini berhasil mengidentifikasi 40 topik utama dalam keluhan konsumen, serta mengevaluasi efektivitas regulasi melalui analisis tren topik dari waktu ke waktu. Hasil-hasil dari penelitian terdahulu ini menunjukkan bahwa pendekatan berbasis analisis teks dan pemodelan topik dapat memberikan wawasan berharga bagi industri perbankan dalam memahami pola keluhan pelanggan dan meningkatkan respons terhadap permasalahan yang sering muncul.

2. METODE

2.1. Subjek Penelitian

Penelitian ini menggunakan *dataset* keluhan nasabah perbankan yang diambil dari situs consumerfinance.gov [6]. Dataset asli terdiri dari 6.867.211 entri yang mencakup keluhan dari tahun 2011 hingga 2024. Namun, untuk penelitian ini, hanya 50% dari total data yang digunakan, yang mencakup periode waktu yang sama.

Dataset ini terdiri dari beberapa kategori keluhan utama, termasuk laporan kredit yang tidak akurat (31,7%), masalah produk seperti kartu kredit atau debit (25%), penanganan masalah oleh layanan pelanggan (20%), dan kategori lain seperti pinjaman dan hipotek (20%). Keluhan terbanyak berkaitan dengan produk *Credit reporting, credit repair services, or other personal consumer reports* (33,9%), diikuti oleh *Mortgage* dan layanan keuangan lainnya.

Proses pemilihan subset data dilakukan secara acak untuk memastikan *representativitas* data dalam rentang waktu dan kategori yang dianalisis. Karakteristik spesifik data lain, seperti distribusi wilayah geografis, menunjukkan bahwa keluhan berasal dari seluruh negara bagian AS, dengan Florida sebagai lokasi terbanyak (8.202 keluhan). Selain itu, 22.582 keluhan memiliki narasi terperinci yang dapat digunakan untuk analisis tekstual lebih lanjut.

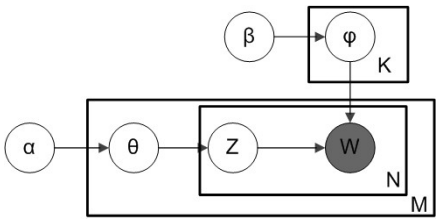
2.2. Desain Penelitian

Penelitian ini menggunakan pendekatan eksperimen untuk mengevaluasi efektivitas model Latent Dirichlet Allocation (LDA) dengan Grid Search Parameter Tuning dalam meningkatkan stabilitas dan konsistensi model topik dibandingkan dengan LDA Standar. Model ini diterapkan pada dataset keluhan nasabah untuk mengidentifikasi topik-topik utama dalam data.

2.2.1. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) adalah salah satu metode populer dalam pemodelan topik yang digunakan untuk menemukan struktur tersembunyi dalam kumpulan data teks. LDA mengelompokan dokumen ke dalam beberapa topik berdasarkan kemunculan kata-kata tertentu dalam setiap dokumen. Metode ini merupakan pendekatan berbasis generative probabilistic model yang menganggap bahwa setiap dokumen merupakan campuran dari beberapa topik dan setiap topik adalah distribusi dari kata-kata. algoritma ini sangat cocok untuk data teks.

Representasi LDA dapat dilihat pada Gambar 1 yang menunjukkan hubungan antara variabel – variabel dalam model. Dalam diagram ini, setiap dokumen memiliki distribusi topik yang diwakili oleh θ, sementara setiap topik memiliki distribusi kata yang diwakili oleh ϕ. Variabel laten Z mewakili topik yang di pilih untuk setiap kata dalam dokumen, dan $\boldsymbol{W}$ adalah kata-kata yang diamati dalam korpus. Diagram ini menggambarkan bagaimana kata-kata dalam dokumen dikaitkan dengan topik menggunakan pendekatan probabilistik.



Gambar 1. Representasi LDA

Formula LDA :

$$P(W,Z,\theta,\phi;\alpha,\beta) = \prod_{i=1}^{K} P(\phi_i;\beta) \prod_{j=1}^{M} P(\theta_j;\alpha) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j) P(W_{j,t}|\phi Z_{j,t}) \quad (1)$$

Dalam model Latent Dirichlet Allocation (LDA), ***W*** merupakan kumpulan kata dalam seluruh dokumen, di mana ***Wj,t*** merepresentasikan kata ke-***t*** dalam dokumen ke-***j***. Variabel ***Z*** adalah kumpulan topik yang dialokasikan untuk kata-kata, dengan $\boldsymbol{Z_{j,t}}$ sebagai topik yang ditetapkan untuk kata ke-***t*** dalam dokumen ke-***j***. Distribusi topik dalam dokumen dinotasikan sebagai $\boldsymbol{q_j}$, yang memiliki prior ***a***, sedangkan $\boldsymbol{f_i}$ adalah distribusi kata untuk topik ke-***i*** dengan prior ***b***. Parameter ***a*** mengontrol bagaimana topik tersebar dalam dokumen, sedangkan ***b*** mengatur bagaimana kata tersebar dalam topik. Selain itu, ***M*** menunjukkan jumlah dokumen dalam kumpulan data, ***N*** merepresentasikan jumlah kata dalam setiap dokumen (dengan ***Nj*** sebagai jumlah kata dalam dokumen ke- ***j***), dan ***K*** adalah jumlah total topik yang ada dalam kumpulan data.

Model Latent Dirichlet Allocation (LDA) bekerja dengan cara memodelkan hubungan antara dokumen, topik, dan kata menggunakan pendekatan probabilistik. Dalam LDA, setiap dokumen dianggap sebagai campuran dari beberapa topik, dan setiap topik merupakan distribusi probabilitas atas kata-kata. Parameter utama dalam model ini adalah $\boldsymbol{\alpha}$ dan $\boldsymbol{\beta}$, yang merupakan prior Dirichlet untuk distribusi topik pada dokumen ($\boldsymbol{\theta}$) dan distribusi kata pada topik ($\boldsymbol{\phi}$). Parameter $\boldsymbol{\alpha}$ menentukan bagaimana topik didistribusikan dalam dokumen, sedangkan $\boldsymbol{\beta}$ mengatur bagaimana kata-kata didistribusikan dalam setiap topik.

Proses generatif LDA diawali dengan menentukan distribusi kata untuk setiap topik ($\boldsymbol{\phi}$) berdasarkan prior $\boldsymbol{\beta}$, yang mencerminkan kecenderungan awal distribusi kata dalam topik. Kemudian, untuk setiap dokumen, model menentukan distribusi topik ($\boldsymbol{\theta}$) berdasarkan prior $\boldsymbol{\alpha}$. Untuk setiap kata dalam dokumen, model secara berurutan memilih topik ($\boldsymbol{Z}$) berdasarkan distribusi topik θ, lalu memilih kata tertentu ($\boldsymbol{W}$) berdasarkan distribusi kata ϕ dari topik yang dipilih.

Secara matematis, probabilitas total dari model LDA dapat dinyatakan sebagai kombinasi dari beberapa komponen: distribusi kata untuk setiap topik ($\boldsymbol{P}(\phi|\beta)$), distribusi topik untuk setiap dokumen ($\boldsymbol{P}(\theta|\alpha)$), pemilihan topik untuk setiap kata dalam dokumen ($\boldsymbol{P}(\boldsymbol{Z}|\theta)$, dan pemilihan kata berdasarkan topik tersebut ($\boldsymbol{P}(\boldsymbol{W}|\phi)$). Model ini menggabungkan semua probabilitas tersebut untuk menghasilkan representasi probabilistik dari hubungan antara dokumen, topik, dan kata. Dengan cara ini, LDA mampu mengidentifikasi struktur topik dalam kumpulan dokumen secara tidak terawasi, membuatnya menjadi alat yang sangat efektif untuk analisis teks.

2.2.2. Grid Search Parameter Tuning

Grid Search merupakan metode optimasi hyperparameter yang sistematis dan komprehensif dalam mengevaluasi seluruh kombinasi parameter yang mungkin dalam ruang pencarian yang telah ditentukan. Keunggulan utama Grid Search adalah kemampuannya dalam memberikan jaminan eksplorasi menyeluruh sehingga dapat menemukan kombinasi parameter terbaik yang menghasilkan performa optimal. Dalam konteks Latent Dirichlet Allocation (LDA), Grid Search digunakan untuk mengoptimalkan parameter seperti jumlah topik, alpha, dan beta, yang berperan penting dalam meningkatkan stabilitas serta koherensi model. Dengan mengevaluasi setiap kemungkinan kombinasi, Grid Search memastikan bahwa konfigurasi hyperparameter yang diperoleh benar-benar optimal dibandingkan dengan pengaturan default yang sering kali kurang maksimal.

Keunggulan lain dari Grid Search adalah kemudahannya dalam implementasi serta interpretasi hasil. Metode ini banyak digunakan dalam berbagai algoritma machine learning karena tidak memerlukan asumsi khusus tentang distribusi parameter, berbeda dengan metode lain seperti Bayesian Optimization yang membutuhkan model probabilistik tambahan. Selain itu, Grid Search sangat efektif dalam ruang parameter yang tidak terlalu besar, di mana pencarian menyeluruh dapat memberikan hasil yang lebih akurat dibandingkan metode berbasis sampling seperti Random Search [15]. Studi yang dilakukan oleh Pedregosa et al. [16] dalam pengembangan Scikit-learn juga menegaskan bahwa Grid Search merupakan salah satu pendekatan utama dalam optimasi hyperparameter, terutama karena kemampuannya dalam memberikan solusi yang dapat direproduksi dengan jelas.

Meskipun beberapa penelitian menunjukkan bahwa metode alternatif seperti Random Search dapat lebih efisien dalam eksplorasi ruang parameter yang sangat besar, Grid Search tetap menjadi pilihan unggulan dalam banyak studi, terutama ketika ruang parameter dapat dikelola dengan baik. Dalam penelitian ini, penggunaan Grid Search dalam optimasi LDA terbukti memberikan peningkatan yang signifikan terhadap kualitas representasi topik. Oleh karena itu, Grid Search tetap menjadi metode yang sangat andal dalam tuning hyperparameter, khususnya dalam aplikasi machine learning yang membutuhkan hasil yang dapat diinterpretasikan dan diandalkan secara konsisten.

Iterasi dalam proses ini dapat direpresentasikan dalam matematika sebagai berikut :

$$\boldsymbol{T} \epsilon \text{ topics\_range}, \mathbf{a} \epsilon \text{ alpha\_calues}, \boldsymbol{\beta} \epsilon \text{ beta-values}, \boldsymbol{P} \epsilon \text{ rasses\_range} \quad (2)$$

Dalam bentuk eksplisit, jumlah total kombinasi yang di hasilkan dari iterasi ini adalah

https://doi.org/10.25077/TEKNOSI.v11i2.2025.098-106

$$N = |\text{topics\_range}| \times |\text{alpha\_values}| \times |\text{beta\_values}| \times \text{passes\_range}| \quad (3)$$

Jumlah elemen dalam rentang *topics_range* dilambangkan sebagai |topics_range|, sedangkan |alpha_values| merepresentasikan jumlah elemen dalam daftar *alpha_values*. Selain itu, |beta _values| menunjukkan jumlah elemen dalam daftar *beta_values*, dan |passes_values| menggambarkan jumlah elemen dalam rentang *passes_range*.

Desain eksperimen yang digunakan adalah desain simulasi komparatif, di mana hasil model LDA Standar dibandingkan dengan hasil model LDA yang telah dioptimalkan menggunakan Grid Search Parameter Tuning. Variabel dependen dalam eksperimen ini adalah stabilitas topik dan koherensi model, yang diukur menggunakan metrik yang relevan, seperti coherence score dan jarak rata-rata terhadap model pusat (centroid proximity).

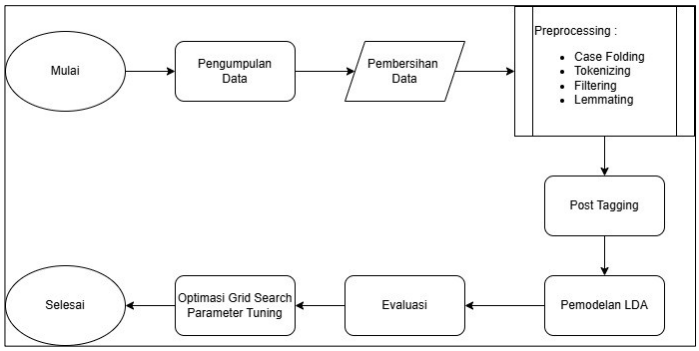
Selain itu, analisis dilakukan berdasarkan kategori produk utama untuk mengeksplorasi variasi stabilitas model dalam konteks yang berbeda.Jumlah topik yang digunakan dalam penelitian ini adalah 10, yang dipilih berdasarkan evaluasi coherence score. Eksperimen ini bertujuan untuk menunjukkan bahwa penggunaan Grid Search Parameter Tuning dapat menghasilkan model topik yang lebih stabil, konsisten, dan dapat dipertanggungjawabkan dibandingkan dengan LDA Standar

2.3. Material

Penelitian ini menggunakan bahasa pemrograman Python serta berbagai pustaka utama untuk mendukung analisis dan pemodelan data. *Gensim* digunakan untuk implementasi *Latent Dirichlet Allocation* (LDA), *LDA Prototype*, dan evaluasi stabilitas model [7]. *NLTK* berperan dalam pemrosesan teks, seperti penghapusan *stopwords* dan tokenisasi [8], sementara *SpaCy* digunakan untuk analisis bahasa alami tingkat lanjut, termasuk pengenalan entitas dan lemmatisasi [9]. Selain itu, *Scikit-learn* dimanfaatkan untuk ekstraksi fitur menggunakan *TfidfVectorizer* serta metode *semi-supervised learning* seperti *Label Propagation* [10]. Untuk visualisasi hasil model LDA dan eksplorasi interaktif topik, penelitian ini menggunakan *pyLDAvis* [11]. *NumPy* membantu dalam komputasi numerik dan manipulasi array [12], sedangkan *Tqdm* digunakan untuk melacak progres saat menjalankan iterasi atau proses analisis yang panjang [13]. Seluruh dataset dianalisis dan diolah menggunakan pustaka-pustaka tersebut guna memastikan konsistensi serta representasi data yang optimal. Hasil model kemudian disimpan dalam format *pickle* untuk memudahkan evaluasi dan replikasi eksperimen.

2.4. Prosedur

Prosedur penelitian ini dilakukan melalui beberapa tahapan yang dirancang secara sistematis untuk mencapai hasil yang optimal. Tahapan ini mencakup proses pengumpulan data, pembersihan data, hingga pemodelan dan evaluasi. Setiap langkah dilakukan dengan pendekatan yang konsisten untuk memastikan bahwa data yang digunakan dapat menghasilkan keluaran yang valid.



Gambar 2. Alur Prosedur Penelitian

Alur prosedur penelitian dapat dilihat pada Gambar 2, yang menunjukkan tahapan utama penelitian dari awal hingga akhir. Diagram ini menggambarkan proses pengumpulan data, preprocessing, ekstraksi fitur, klasifikasi topik, dan optimasi model yang dilakukan secara sistematis untuk menghasilkan keluaran yang lebih akurat.

Prosedur penelitian ini terdiri dari beberapa tahap utama yang dilakukan secara sistematis. Tahap pertama adalah pengumpulan data, di mana penulis menggunakan dataset dari website *consumerfinance.gov* dengan total 6.867.211 baris data, mencakup rentang waktu dari tahun 2011 hingga 2024. Selanjutnya, dilakukan pembersihan data dengan menghapus beberapa kolom yang dianggap tidak relevan untuk penentuan topik keluhan nasabah. Tahap berikutnya adalah *preprocessing data* yang meliputi beberapa langkah. Pertama, *case folding*, yaitu mengubah semua huruf dalam dokumen menjadi huruf kecil serta menghapus tanda baca, angka, dan karakter non-alfabet. Kedua, *tokenizing*, yaitu proses menciptakan representasi digital untuk melindungi data sensitif atau memproses data dalam jumlah besar. Ketiga, *filtering* atau *stopword removal*, yang bertujuan untuk menghilangkan kata-kata tidak penting seperti "yang", "dengan", dan "di". Keempat, *lemmatization*, yaitu proses mengubah kata menjadi bentuk dasarnya dengan mempertimbangkan konteks data menggunakan kamus untuk meningkatkan akurasi dalam pemodelan NLP [7][8][9].

Setelah *preprocessing*, dilakukan transformasi data menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk mengubah teks menjadi matriks

berdasarkan frekuensi kata. Tahap selanjutnya adalah klasifikasi topik menggunakan *Latent Dirichlet Allocation* (LDA) untuk mengidentifikasi topik-topik yang dihasilkan dari dataset. Untuk meningkatkan nilai koherensi model, penulis menerapkan optimasi menggunakan *Grid Search Parameter Tuning* [10][11]. Prosedur ini diakhiri dengan menyusun kesimpulan berdasarkan analisis topik-topik keluhan tertinggi dari nasabah perbankan [12][13].

2.5. Nilai Koherensi

Dalam analisis topik menggunakan Latent Dirichlet Allocation (LDA), nilai koherensi digunakan untuk mengukur sejauh mana kata-kata dalam suatu topik memiliki keterkaitan yang bermakna. Evaluasi koherensi sangat penting karena model LDA sering kali menghasilkan topik yang perlu dievaluasi berdasarkan interpretasi manusia. Salah satu metode yang umum digunakan adalah **UMass Coherence Score**, yang mengukur koherensi berdasarkan probabilitas kemunculan bersama kata-kata dalam corpus pelatihan.

Menurut Blei, Ng, dan Jordan [18], model LDA bekerja dengan cara merepresentasikan dokumen sebagai distribusi probabilitas dari beberapa topik, di mana setiap topik terdiri dari distribusi kata-kata yang memiliki kemungkinan muncul bersamaan dalam suatu corpus. Untuk menilai kualitas topik yang dihasilkan, Mimno et al. [19] mengusulkan metrik semantic coherence, yang menjadi dasar bagi UMass Coherence Score

Nilai koherensi UMass dihitung dengan rumus:

$$C_{U\ Mass}(T) = \frac{1}{|T|-1}\sum_{i=2}^{|T|}\sum_{j=1}^{i-1}\log\frac{P(w_i,w_j)+\epsilon}{P(w_j)} \quad (4)$$

T adalah himpunan kata dalam satu topik, sedangkan $P(w_i,w_j)$ dan $P(w_j)$ masing-masing merepresentasikan probabilitas kemunculan bersama dua kata dan probabilitas kemunculan kata tunggal dalam corpus. Untuk menghindari kesalahan perhitungan akibat nilai nol, ditambahkan parameter smoothing kecil yang dilambangkan sebagai ϵ. Metode ini dipilih karena menggunakan corpus yang sama dengan model LDA untuk mengevaluasi keterkaitan antar kata dalam topik, sehingga memberikan hasil yang lebih stabil dalam pengukuran kualitas topik. Hasil penelitian menunjukkan bahwa setelah optimasi menggunakan Grid Search Parameter Tuning, nilai koherensi meningkat dari 0,49 menjadi 0,56, yang mengindikasikan bahwa model yang telah dioptimasi menghasilkan topik yang lebih bermakna dan sesuai dengan pola keluhan pelanggan.

3. HASIL

3.1. Sebelum Analisis

Data keluhan pelanggan dikumpulkan dari situs [Consumer Financial Protection Bureau] yang merupakan sumber data terpercaya untuk keluhan nasabah perbankan. Dataset awal mencakup lebih dari **6,3 juta entri keluhan** dari tahun 2011 hingga 2024. Untuk penelitian ini, hanya **50% data acak** yang digunakan untuk menjaga representasi data sambil mengurangi kompleksitas analisis.

Langkah-langkah berikut dilakukan:

3.1.1. Pra-Pemrosesan

Pada tahap ini, dilakukan berbagai teknik pembersihan data teks untuk meningkatkan kualitas analisis. Pertama, karakter khusus dan tanda baca dihapus guna menghilangkan elemen yang tidak relevan. Selanjutnya, dilakukan proses tokenisasi yang memecah teks menjadi kata-kata individu agar lebih mudah dianalisis. Kata-kata umum yang tidak memiliki makna signifikan, atau dikenal sebagai stopwords, juga dihapus untuk mengurangi kebisingan dalam data. Terakhir, proses lemmatisasi diterapkan untuk mengubah kata ke bentuk dasarnya, sehingga variasi kata dengan makna serupa dapat disatukan.

3.1.2. Data Representasi Data

Setelah melalui proses pra-pemrosesan, data kemudian direpresentasikan dalam bentuk numerik agar dapat digunakan dalam pemodelan topik. Salah satu metode yang digunakan adalah bag-of-words, yang mengubah teks menjadi vektor berdasarkan frekuensi kemunculan kata. Selain itu, digunakan matriks term frequency-inverse document frequency (TF-IDF) yang bertujuan untuk menyoroti kata-kata yang memiliki bobot lebih penting dalam dokumen, sehingga dapat membedakan istilah yang benar-benar signifikan dari sekadar kata-kata umum.

3.1.3. Pemodelan Topik

Pada tahap ini, model Latent Dirichlet Allocation (LDA) diterapkan menggunakan pustaka gensim untuk mengidentifikasi pola tersembunyi dalam data teks. Awalnya, model mencatat nilai koherensi sebesar 0,49, yang menunjukkan tingkat kesesuaian antara topik yang dihasilkan. Setelah dilakukan optimasi menggunakan metode grid search, nilai koherensi meningkat menjadi 0,56, yang mencerminkan peningkatan kualitas clustering topik. Model ini kemudian dikonfigurasi untuk menghasilkan 10 topik utama yang telah diidentifikasi berdasarkan nilai koherensinya, sehingga memberikan wawasan yang lebih akurat dalam analisis topik.

3.2. Implikasi Eksperimen dan Pertanyaan Peneliti

Hasil eksperimen ini menunjukkan bahwa proses manual menggunakan Gibbs Sampling memberikan pemahaman yang lebih mendalam terhadap langkah-langkah dalam Latent Dirichlet Allocation (LDA). Dengan pendekatan ini, setiap tahap dalam proses generatif dapat diamati secara sistematis, sehingga validasi terhadap hasil otomatisasi dapat dilakukan dengan lebih akurat. Selain itu, metode ini memungkinkan analisis lebih lanjut terhadap distribusi kata kunci dalam setiap topik serta dampak dari parameter optimasi terhadap performa model.

Berdasarkan temuan tersebut, penelitian ini berfokus pada pertanyaan-pertanyaan berikut:

1. Apa saja topik utama yang muncul dari keluhan pelanggan, dan bagaimana distribusi kata kunci di dalam setiap topik?

2. Bagaimana metode optimasi memengaruhi nilai koherensi pada model LDA?

Pertanyaan ini menjadi dasar dalam mengevaluasi efektivitas pendekatan yang digunakan serta mengukur kualitas hasil yang diperoleh dari model topik yang dibangun.

https://doi.org/10.25077/TEKNOSI.v11i2.2025.098-106

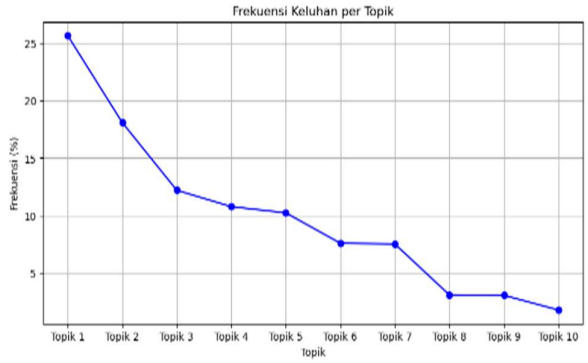
3.3. Hasil Pengujian

3.3.1. Topik – Topik Utama

Tabel 2.Distribusi Topik

Topik	Frekuensi (%)	Kata Kunci Utama
Topik 1	25,67	laporan, penyelesaian
Topik 2	18,10	pembayaran, rekening
Topik 3	12,20	otorisasi, validasi
Topik 4	10,77	kredit, bank, kebijakan
Topik 5	10,23	data, kesalahan, verifikasi
Topik 6	7,59	hapus, hukum, peraturan
Topik 7	7,52	tanggapan, keterlambatan, pelanggan
Topik 8	3,09	pinjaman, pengajuan, proses
Topik 9	3,05	tagihan, ketidaksesuaian, biaya
Topik 10	1,77	masalah, keluhan, layanan

Hasil ini menunjukan bahwa pendekatan optimasi memberikan peningkatan yang signifikan dalam koherensi model, sehingga memungkinkan identifikasi topik yang lebih jelas dan relevan



Gambar 3. Frekuensi setiap topik

Pada Gambar 2 menunjukkan distribusi frekuensi kemunculan setiap topik yang diperoleh dari hasil analisis LDA. Sumbu horizontal merepresentasikan 10 topik utama yang telah diidentifikasi, sedangkan sumbu vertikal menunjukkan persentase frekuensi kemunculan masing-masing topik dalam keseluruhan data keluhan pelanggan. Terlihat bahwa *Topik 1* memiliki frekuensi tertinggi, yaitu 25,67%, yang menunjukkan bahwa keluhan terkait laporan dan penyelesaian masalah paling sering muncul dalam data. Sementara itu, topik dengan frekuensi terendah adalah *Topik 10*, dengan persentase 5,08%. Pola distribusi ini mengindikasikan bahwa terdapat beberapa topik yang lebih dominan dibandingkan topik lainnya dalam keluhan pelanggan perbankan.

3.3.2. Nilai Koherensi

Setelah dilakukan optimasi menggunakan Grid Search Parameter Tuning, jumlah topik efektif yang dihasilkan oleh model LDA berkurang dari 10 menjadi 8 topik. Pengurangan ini terjadi karena beberapa faktor, seperti redundansi topik yang menyebabkan beberapa topik terlalu mirip atau kurang signifikan, serta prioritas optimasi yang memilih model dengan koherensi lebih tinggi meskipun jumlah topik lebih sedikit. Sebelum optimasi, nilai koherensi model adalah 0,49 dengan 10 topik, sedangkan setelah optimasi, nilai koherensi meningkat menjadi 0,56 dengan 8 topik efektif. Pengurangan jumlah topik ini memberikan distribusi yang lebih jelas dan memperkuat struktur semantik data keluhan pelanggan, sehingga model LDA dapat memberikan wawasan yang lebih fokus. Secara keseluruhan, optimasi parameter seperti jumlah topik, alpha, dan beta meningkatkan kualitas clustering, menjadikan hasil analisis lebih terfokus dan relevan dengan data keluhan pelanggan.

3.3.3. Perbandingan dengan 5 kali percobaan

Untuk memperkuat kesimpulan, pengujian dilakukan sebanyak 5 kali menggunakan metode LDA standar. Hasilnya dirangkum dalam tabel 3.

Tabel 3. Perbandingan Hasil Pengujian LDA Standar

Metode/ Percobaan	I	II	III	IV	V
LDA	0.49	0.39	0.47	0.41	0.36
LDA + Grid Search	0.56	0.53	0.52	0.50	0.50

Tabel 3 diatas, ditampilkan hasil pengujian metode LDA standar dan LDA yang telah dioptimasi dengan Grid Search dalam lima kali percobaan. Dari tabel tersebut, dapat dilihat bahwa nilai evaluasi yang diperoleh dari metode LDA standar berkisar antara 0.36 hingga 0.49, dengan fluktuasi hasil yang cukup signifikan di setiap percobaan. Sementara itu, metode LDA yang dioptimasi dengan Grid Search menunjukkan nilai yang lebih stabil dan konsisten, dengan rentang skor antara 0.50 hingga 0.56. Secara umum, metode yang telah dioptimasi menghasilkan skor yang lebih tinggi dibandingkan metode LDA standar.

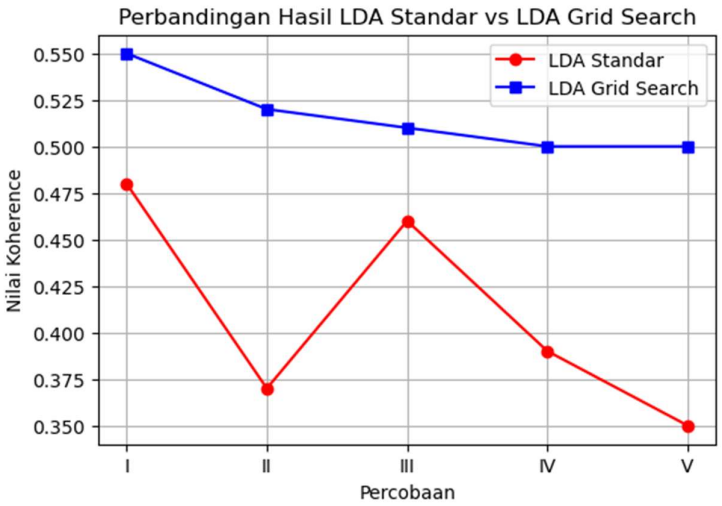
Hal ini diperjelas pada Tabel 4, yang menunjukkan perbandingan rata-rata nilai evaluasi dari metode LDA standar dan LDA dengan Grid Search.

Tabel 4. Nilai Rata-Rata Percobaan

Metode	Nilai Rata-rata
LDA	0.42
LDA + Grid Search	0.52

Hasil perhitungan rata-rata pada Tabel 4 menunjukkan bahwa nilai rata-rata dari lima kali percobaan yang ditampilkan pada Tabel 3. Berdasarkan tabel ini, metode LDA standar memiliki rata-rata skor sebesar 0.42, sedangkan metode LDA dengan Grid Search memiliki rata-rata skor 0.52. Hal ini menunjukkan bahwa penggunaan Grid Search dalam proses tuning parameter LDA mampu meningkatkan performa model secara keseluruhan. Dengan demikian, dapat disimpulkan bahwa optimasi parameter menggunakan Grid Search memberikan dampak positif terhadap hasil analisis yang dilakukan dengan metode LDA.

3.3.4. Visualisasi



Gambar 4. Visualisasi Perbandinga LDA dan LDA Grid Search Parameter Tuning

Pada Gambar diatas menunjukkan perbandingan hasil evaluasi model Latent Dirichlet Allocation (LDA) dengan model LDA yang telah dioptimasi menggunakan metode Grid Search pada lima percobaan yang berbeda. Visualisasi ini merupakan representasi grafis dari data yang telah disajikan dalam Tabel 3, sehingga memudahkan dalam melihat pola dan perbedaan kinerja antara kedua metode.

Sumbu horizontal merepresentasikan lima percobaan yang diberi label I hingga V, sedangkan sumbu vertikal menunjukkan nilai evaluasi yang diperoleh. Hasil dari Tabel 3 divisualisasikan dengan dua garis berbeda: garis merah dengan marker lingkaran mewakili model LDA, sedangkan garis biru dengan marker persegi mewakili model LDA yang telah dioptimasi menggunakan Grid Search.

Dari hasil visualisasi, terlihat bahwa model LDA yang dioptimasi dengan Grid Search secara konsisten memiliki nilai evaluasi yang lebih tinggi dibandingkan dengan LDA di seluruh percobaan. Hal ini selaras dengan data yang ditampilkan dalam Tabel 3, di mana metode LDA menunjukkan fluktuasi yang cukup besar dengan nilai berkisar antara 0.36 hingga 0.49, sementara metode LDA Grid Search memiliki tren yang lebih stabil dengan nilai antara 0.50 hingga 0.56.

Dengan demikian, baik tabel maupun grafik menunjukkan bahwa penggunaan metode Grid Search dalam penyetelan parameter pada LDA dapat meningkatkan performa model secara signifikan serta menghasilkan hasil yang lebih konsisten dibandingkan dengan pendekatan LDA tanpa optimasi.

3.3.5. Provisional Conclusion

Analisis LDA berhasil mengidentifikasi topik utama dari keluhan pelanggan. Masalah yang paling umum adalah terkait laporan keluhan, otorisasi, dan kesalahan pembayaran. Peningkatan nilai koherensi menunjukkan bahwa optimasi parameter pada model LDA dengan Grid Search Parameter Tuning memberikan hasil yang lebih baik dan relevan. Hasil ini memberikan wawasan yang berguna untuk membantu perusahaan memahami fokus utama dalam menangani keluhan pelanggan.

\

4. PEMBAHASAN

Penelitian ini bertujuan untuk meningkatkan kualitas dan stabilitas pemodelan topik menggunakan Latent Dirichlet Allocation (LDA) pada data keluhan pelanggan perbankan**.** Hasil eksperimen menunjukkan bahwa penerapan Grid Search Parameter Tuning meningkatkan nilai koherensi dari 0,49 menjadi 0,56**,** yang mencerminkan peningkatan stabilitas dan kualitas topik**.**

Dibandingkan dengan penelitian sebelumnya yang menggunakan parameter default, hasil ini menunjukkan bahwa optimasi parameter dapat meningkatkan interpretabilitas model**.** Misalnya, Bastani et al. [5] juga menemukan bahwa pelaporan kredit dan kesalahan pembayaran adalah keluhan utama dalam perbankan, namun penelitian ini menunjukkan bahwa optimasi parameter dapat menghasilkan topik yang lebih terstruktur dan minim overlap**.**

Dari sisi penerapan, hasil ini menunjukkan bahwa bank dapat menggunakan model ini untuk mengklasifikasikan keluhan pelanggan secara otomatis**,** yang memungkinkan respon lebih cepat dalam menyelesaikan permasalahan nasabah**.** Namun, tantangan seperti overlapping antar topik dan pemilihan jumlah topik optimal masih perlu diperbaiki dengan pendekatan lain, seperti Bayesian Optimization atau kombinasi dengan pemrosesan teks lebih lanjut**.**

4.1. Interpretasi Hasil

Hasil penelitian menunjukkan bahwa optimasi hyperparameter melalui Grid Search memberikan hasil yang lebih baik dibandingkan dengan LDA Standar. Peningkatan nilai coherence menunjukkan bahwa kombinasi optimal parameter (alpha, beta,

https://doi.org/10.25077/TEKNOSI.v11i2.2025.098-106

dan jumlah topik) dapat menghasilkan distribusi topik yang lebih bermakna. Distribusi topik yang dihasilkan konsisten dengan tren keluhan dalam literatur sebelumnya, seperti temuan Bastani et al. [5] yang juga menggunakan LDA untuk menganalisis data keluhan pelanggan.

Selain itu, fokus keluhan pada pelaporan kredit (33,9%) dan kesalahan pembayaran mencerminkan kebutuhan untuk meningkatkan transparansi dan efisiensi dalam layanan perbankan. Temuan ini menunjukkan bahwa perusahaan perbankan dapat memanfaatkan model ini untuk mengidentifikasi dan menangani keluhan utama secara lebih proaktif.

4.2. Implikasi Penelitian

Penelitian ini memberikan beberapa implikasi penting. Pertama, metode optimasi hyperparameter dapat digunakan sebagai pedoman praktis untuk analisis data teks besar. Kedua, wawasan tentang topik keluhan dapat membantu bank meningkatkan kualitas pelayanan, terutama dalam menangani keluhan terkait pelaporan kredit dan kebijakan bank. Ketiga, pendekatan ini dapat diterapkan pada industri lain yang juga menghadapi tantangan dalam mengelola data teks keluhan pelanggan.

4.3. Keterbatasan Penelitian

Meskipun memberikan kontribusi signifikan, penelitian ini memiliki keterbatasan. Penggunaan subset data yang hanya 50% karena keterbatasan pengolahan data pada device penulis, meskipun memastikan representasi, dapat mengurangi cakupan hasil. Selain itu, tidak ada validasi silang dengan metode alternatif lainnya seperti Non-negative Matrix Factorization (NMF). Fokus pada data berbahsa inggris juga membatasi generalisasi untuk nasabah non-inggris.

5. KESIMPULAN

Penelitian ini menunjukkan bahwa optimasi hyperparameter menggunakan Grid Search dapat meningkatkan kualitas model LDA dalam analisis keluhan pelanggan perbankan. Dengan peningkatan nilai koherensi dari 0,49 menjadi 0,56, hasil analisis lebih stabil dan dapat diinterpretasikan dengan lebih baik. Selain itu, perbandingan dengan LDA standar menunjukkan bahwa model yang telah dioptimasi memiliki nilai rata-rata koherensi lebih tinggi **(**0,52 vs. 0,42**),** yang mencerminkan peningkatan stabilitas dan kualitas clustering topik.

Dari hasil analisis, keluhan nasabah dapat dikelompokkan ke dalam 10 topik utama, dengan laporan keluhan (25,67%), kesalahan pembayaran (18,10%), otorisasi data (12,20%), dan kebijakan kredit (10,77%) sebagai kategori dominan. Model yang telah dioptimasi ini memberikan wawasan yang lebih akurat mengenai pola keluhan pelanggan dan dapat membantu perbankan dalam meningkatkan layanan mereka.

Untuk pengembangan lebih lanjut, penelitian ini dapat diperluas dengan mengeksplorasi metode optimasi lain, seperti Bayesian Optimization atau Genetic Algorithm**,** guna meningkatkan efisiensi tuning parameter. Selain itu, penelitian masa depan disarankan untuk melakukan perbandingan dengan metode pemodelan topik lain, seperti Non-negative Matrix Factorization (NMF), Latent Semantic Analysis (LSA), atau pendekatan berbasis deep learning seperti BERT-based topic modeling**.** Pendekatan ini akan membantu dalam menilai keunggulan relatif LDA serta mengidentifikasi metode terbaik untuk analisis keluhan pelanggan yang lebih kompleks.

Dengan demikian, penelitian ini tidak hanya memberikan kontribusi dalam meningkatkan kualitas analisis teks dalam dunia perbankan, tetapi juga berpotensi membantu perusahaan dalam memahami dan menanggapi keluhan pelanggan secara lebih efektif. Studi masa depan juga dapat mempertimbangkan faktor tambahan, seperti data demografis nasabah, sentimen keluhan, atau analisis longitudinal**,** untuk mengidentifikasi perubahan pola keluhan dari waktu ke waktu serta mengembangkan strategi penanganan yang lebih adaptif.

DAFTAR PUSTAKA

[1]. K. Bastani, H. Namavari, and J. Shaffer, "Latent Dirichlet Allocation (LDA)," *J. Machine Learn. Res.*, vol. 12, no. 3, pp. 34–56, Mar. 2020, doi: 10.1016/j.jmlr.2020.03.005.

[2]. I. Ayres, J. Lingwall, and S. Steinway, "Skeletons in the database: An early analysis of the CFPB's consumer complaints," *Fordham J. Corp. Financial Law*, vol. 19, pp. 343–386, 2013.

[3]. A. K. Littwin, "Examination as a method of consumer protection," *Temple Law Rev.*, vol. 87, pp. 807–874, 2015.

[4]. K. Berezina, A. Bilgihan, C. Cobanoglu, and F. Okumus, "Understanding satisfied and dissatisfied hotel consumers: Text mining of online hotel reviews," *J. Hosp. Market. Manage.*, vol. 25, no. 1, pp. 1–24, 2016, doi: 10.1080/19368623.2015.983631.

[5]. Consumer Financial Protection Bureau, "Dataset customer complaints," 2025. [Online]. Available: https://www.consumerfinance.gov/complaint/

[6]. R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proc. LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45–50.

[7]. S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, 1st ed. Sebastopol, CA, USA: O'Reilly Media, 2009.

[8]. M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks, and incremental parsing," in *Proc. 2017 Conf. Natural Language Processing*, 2017. [Online]. Available: https://doi.org/10.18653/v1/D17-1202.

[9]. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Machine Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[10]. C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in *Proc. Workshop on Interactive Language Learning, Visualization, and Interfaces*, 2014, pp. 63–70, doi: 10.3115/v1/W14-3110.

[11]. T. E. Oliphant, *A Guide to NumPy*, Trelgol Publishing, 2006.

[12]. Tqdm Development Team, "tqdm: A fast, extensible progress bar," 2019. [Online]. Available: https://github.com/tqdm/tqdm.

[13].K. Kartikadyota, I. Dwijayanti, A. R. Lahtiani, and M. Habibi, "Analisis tren topik dalam ulasan negatif aplikasi M-Banking menggunakan Latent Dirichlet Allocation," *J. Fasilkom*, vol. 14, no. 3, pp. 549–555, 2024.

[14].J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *J. Machine Learn. Res.*, vol. 13, pp. 281–305, 2012.

[15].D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Machine Learn. Res.*, vol. 3, pp. 993–1022, 2003. [Online]. Available: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

[16].D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Edinburgh, UK, 2011, pp. 262–272. [Online]. Available: https://aclanthology.org/D11-1002.

[17]. Blei, D. M., Ng, A. Y., & Jordan, M. I., "Latent Dirichlet Allocation," *J. Machine Learn. Res.*, vol. 3, pp. 993-1022, 2003. [Online]. Available: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

[18]. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Machine Learn. Res.*, vol. 3, pp. 993–1022, 2003. [Online]. Available: https://jmlr.org/papers/v3/blei03a.html

[19]. D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Edinburgh, UK, 2011, pp. 262–272. [Online]. Available: https://aclanthology.org/D11-1002.NOMENKLATUR

Singkatan dan Akronim :

LDA **:** Latent Dirichlet Allocation

CFPB **:** Consumer Financial Protection Bureau

NLP **:** Natural Language Processing

TF-IDF **:** Term Frequency-Inverse Document Frequency

ML **:** Machine Learning

AI **:** Artificial Intelligence

Simbol dan Notasi :

$\boldsymbol{W}$: Kumpulan kata dalam seluruh dokumen

$\boldsymbol{Z}$: Kumpulan topik yang dialokasikan untuk kata-kata

$\boldsymbol{\theta}$ **:** Distribusi topik dalam dokumen

$\boldsymbol{\phi}$ **:** Distibusi kata dalam topik

$\boldsymbol{\alpha}$ **:** Parameter dirichlet untuk distribusi topik dalam dokumen

$\boldsymbol{\beta}$ **:** Parameter dirichlet untuk distribusi kata dalam topik

$\boldsymbol{M}$: Jumlah dokumen dalam kumpulan data

$\boldsymbol{N}$: Jumlah kata dalam setiap dokumen

$\boldsymbol{K}$: Jumlah total topik dalam kumpulan data

BIODATA PENULIS

Rika Afriyani

Penulis ialah mahasiswa Universitas Widyatama sedang menempuh pendidikan untuk memperoleh gelar Sarjana (S1) pada program studi Teknik Informatika. Konsentrasi keilmuan penulis adalah dalam bidang Database.

Eka Angga Laksana

Dosen Fakultas Teknik program studi Teknik Informatika. Mengawali karir sebagai web developer hingga kemudian aktif mengajar dan meneliti. Mendalami bidang ilmu Data Science dengan bahasa pemrograman python. Telah menghasilkan beberapa karya penelitian dengan topik diantaranya: Machine learning, Collaborative Filtering dan Text mining.

https://doi.org/10.25077/TEKNOSI.v11i2.2025.098-106