Analisis Topik Ulasan Wisatawan Digital: Pendekatan Latent Dirichlet Allocation (LDA) pada Destinasi Wisata Mendukung Wisata Cerdas

[12]Hartatik, [1]R Rizal Isnanto , [1]Budi Warsito
[1] School of Postgraduate Studies, Universitas Diponegoro, Semarang, Indonesia
[2]Department of Informatics Engineering, Vocational School, Universitas Sebelas Maret, Surakarta, Indonesiaemail: hartatik@students,undip.ac.id

Abstract

The exponential growth of user-generated tourist reviews on digital platforms holds significant potential for understanding traveler experiences and perceptions of destinations. However, the vast and unstructured nature of this textual data requires reliable methods for systematic theme extraction. This study aims to map the dominant topics embedded in digital tourist reviews using the Latent Dirichlet Allocation (LDA) approach. Employing an exploratory descriptive research design, a total of 2,000 reviews from Indonesian tourism destinations were collected and processed through tokenization, normalization, and stemming, before being modeled using Gensim's LDA. The model was evaluated using the coherence score (c_v), resulting in an optimal model with ten topics (K=10). The identified topics reflect various tourism aspects, including natural beauty, historical heritage, family activities, culinary experiences, and popular landmarks. Topic distribution analysis revealed the spread and proportion of themes across the entire corpus, offering rich contextual insights into tourist narratives. These findings contribute to data-driven destination management and form a foundation for the development of smart tourism applications, while also providing new insights into tourist preferences in Indonesia. The LDA-based thematic mapping demonstrates strong potential in supporting intelligent destination management and enhancing personalized tourism recommendation systems.

Keywords: Tourist Reviews; Topic Modeling; LDA; Topic Coherence; Destination Management; Indonesia.

Abstrak

Pertumbuhan eksponensial ulasan wisata berbasis pengguna pada platform digital menyimpan potensi informasi yang kaya untuk memahami pengalaman wisatawan dan persepsi terhadap destinasi. Namun, banyaknya data teks yang tidak terstruktur membutuhkan metode yang andal untuk mengekstrak tema secara sistematis. Penelitian ini bertujuan untuk memetakan topik-topik dominan yang terkandung dalam ulasan wisatawan digital dengan pendekatan Latent Dirichlet Allocation (LDA). Menggunakan desain penelitian eksploratif deskriptif, sebanyak 2.000 ulasan dari destinasi wisata Indonesia (alam dan budaya) dikumpulkan dan diproses melalui tahap tokenisasi, normalisasi, dan stemming sebelum dimodelkan dengan LDA Gensim. Evaluasi model dilakukan menggunakan skor koherensi (c_v), dan diperoleh model optimal pada sepuluh topik (K=10). Hasil penelitian menunjukkan topik-topik yang dihasilkan mencerminkan berbagai aspek wisata, seperti keindahan alam, warisan sejarah, aktivitas keluarga, kuliner, serta landmark populer. Analisis distribusi topik menunjukkan sebaran dan proporsi tema dalam korpus secara keseluruhan, memberikan pemahaman kontekstual yang kaya terhadap narasi wisatawan. Temuan ini berkontribusi pada pengelolaan destinasi berbasis data dan menjadi landasan bagi pengembangan aplikasi pariwisata cerdas dan memberikan wawasan baru terkait preferensi wisatawan di Indonesia. Pemetaan tematik berbasis LDA ini menunjukkan potensinya dalam mendukung manajemen destinasi berbasis data serta pengembangan sistem rekomendasi wisata yang personal.

Keywords : Ulasan Wisatawan; Pemodelan Topik; LDA; Koherensi Topik; Manajemen Destinasi; Indonesia

1. PENDAHULUAN

Transformasi digital dalam sektor pariwisata telah mendorong terjadinya pergeseran signifikan dalam cara wisatawan merespons dan mengevaluasi pengalaman kunjungan mereka. Platform digital seperti TripAdvisor, Google Review, dan Booking.com memungkinkan wisatawan untuk berbagi ulasan secara terbuka dan real-time, yang kemudian menjadi sumber informasi penting bagi calon wisatawan maupun pengelola destinasi (Xiang et al., 2022). Ulasan digital tersebut mencerminkan persepsi pengguna secara langsung terhadap kualitas layanan, fasilitas, hingga daya tarik suatu destinasi. Sayangnya, ulasan tersebut umumnya bersifat tidak terstruktur dan jumlahnya sangat besar, sehingga menyulitkan proses analisis manual.

Salah satu pendekatan yang relevan untuk menangani tantangan tersebut adalah pemodelan topik, khususnya melalui metode Latent Dirichlet Allocation (LDA). LDA merupakan metode unsupervised learning yang digunakan untuk mengungkap struktur tematik laten dalam kumpulan dokumen teks dengan memodelkan distribusi kata dalam dokumen sebagai campuran dari beberapa topik (Blei, Ng, & Jordan, 2003). Dalam konteks pariwisata, pendekatan ini telah digunakan untuk mengidentifikasi tema-tema umum yang muncul dalam ulasan wisatawan seperti

kebersihan, aksesibilitas, harga, keramahan, dan fasilitas (García-Pablos, Cuadros, & Rigau, 2017). Pendekatan ini memberikan peluang untuk menangkap dimensi semantik dari pengalaman wisatawan secara lebih mendalam, yang dapat digunakan untuk mendukung pengambilan keputusan berbasis data.

Namun demikian, penelitian yang menerapkan LDA untuk menganalisis ulasan wisatawan digital di Indonesia masih relatif terbatas. Studi yang tersedia umumnya masih bersifat deskriptif dan belum secara sistematis mengkaji kualitas model topik yang dihasilkan, misalnya melalui evaluasi koherensi topik. Selain itu, konteks Indonesia sebagai negara dengan keragaman budaya, bahasa, dan jenis destinasi yang luas, belum banyak dieksplorasi secara mendalam dalam studi pemodelan topik. Hal ini menunjukkan adanya celah penelitian (research gap) terkait minimnya eksplorasi tematik terhadap ulasan wisatawan Indonesia menggunakan pendekatan LDA yang tervalidasi, serta belum optimalnya pemanfaatan pemetaan topik untuk mendukung pengelolaan destinasi berbasis data (Mukminin, Santoso, & Nuraini, 2021).

Berdasarkan latar belakang dan celah penelitian tersebut, studi ini bertujuan untuk melakukan analisis tematik terhadap ulasan wisatawan digital pada berbagai destinasi wisata di Indonesia menggunakan pendekatan LDA. Studi ini tidak hanya membangun model LDA untuk mengidentifikasi topik-topik dominan, tetapi juga mengevaluasi kualitas model dengan menggunakan metrik koherensi (*coherence score*) sebagai bentuk validasi intrinsik. Temuan dari penelitian ini diharapkan dapat memberikan kontribusi pada pengembangan sistem informasi pariwisata yang lebih adaptif, serta memberikan masukan praktis bagi pengelola destinasi dalam menyusun strategi peningkatan kualitas layanan dan promosi dalam mendukung wisata cerdas.

2. KERANGKA TEORI

2.1 Ulasan Digital dan Perilaku Wisatawan

Perkembangan teknologi digital telah mendorong perubahan signifikan dalam perilaku wisatawan, khususnya dalam cara mereka mencari, menilai, dan membagikan informasi tentang destinasi wisata. Platform seperti TripAdvisor, Google Review, dan Booking.com menjadi sumber utama bagi wisatawan untuk memperoleh informasi langsung dari pengguna lain, yang dianggap lebih autentik dan relevan dibandingkan informasi promosi formal (Xiang, Du, Ma, & Fan, 2022).

Meningkatnya penggunaan perangkat digital dan internet di kalangan wisatawan juga berdampak pada meningkatnya literasi digital wisatawan, yaitu kemampuan individu untuk mengakses, memahami, dan memanfaatkan informasi digital secara kritis. Literasi ini mendorong partisipasi wisatawan sebagai produsen konten (content creator), bukan sekadar konsumen informasi. Studi Lee dan Jan (2015) menunjukkan bahwa wisatawan dengan literasi digital tinggi cenderung lebih aktif dalam membaca dan menulis ulasan daring.

Ulasan digital terbukti memiliki pengaruh kuat dalam pengambilan keputusan perjalanan. Calon wisatawan sering kali mengandalkan pengalaman wisatawan lain sebelum memilih destinasi, hotel, atau atraksi, karena ulasan tersebut dianggap sebagai bentuk electronic word-of-mouth (eWOM) yang kredibel dan kontekstual (Filieri & McLeay, 2014). Oleh karena itu, data ulasan menjadi sumber penting untuk mengkaji persepsi, preferensi, dan kepuasan pengunjung secara real-time. Hal ini di era transformasi digital sangat penting mengetahui perilaku wisatawan.

2.2 Pemodelan Topik dan Latent Dirichlet Allocation (LDA)

Pemodelan topik merupakan metode dalam pemrosesan bahasa alami (Natural Language Processing/NLP) yang digunakan untuk mengidentifikasi struktur laten dalam dokumen teks yang tidak berlabel. Salah satu pendekatan yang paling banyak digunakan adalah Latent Dirichlet Allocation (LDA). LDA memodelkan setiap dokumen sebagai kombinasi dari beberapa topik, dan setiap topik sebagai distribusi probabilistik atas kata-kata (Blei, Ng, & Jordan, 2003).

Model ini bekerja berdasarkan asumsi generatif, di mana dokumen dianggap dihasilkan melalui proses acak berdasarkan dua parameter utama: α (alpha) dan η (eta). Parameter α mengontrol distribusi topik dalam dokumen; nilai α kecil akan menghasilkan dokumen yang cenderung fokus pada sedikit topik. Sementara itu, η mengatur distribusi kata dalam topik; nilai η kecil menghasilkan topik yang lebih spesifik. Nilai-nilai ini dapat diatur secara manual atau ditentukan secara otomatis selama proses pelatihan model.

Dalam konteks pariwisata, LDA telah digunakan untuk menganalisis ulasan wisatawan dan mengungkap tema-tema seperti pelayanan, fasilitas, harga, dan pengalaman lokal (He, Zha, & Li, 2013; García-Pablos, Cuadros, & Rigau, 2017). Meskipun demikian, penerapan LDA dalam studi pariwisata di Indonesia masih terbatas, terutama pada destinasi non-urban yang tersebar di berbagai wilayah Nusantara. Hal ini menjadi peluang untuk mengeksplorasi penggunaan LDA pada konteks data lokal yang kaya dan beragam.

2.3 Evaluasi Topik: Koherensi dan Stabilitas

Evaluasi model topik sangat penting untuk memastikan bahwa topik yang dihasilkan tidak hanya bermakna secara statistik, tetapi juga dapat diinterpretasikan secara semantik. Salah satu metrik utama yang digunakan adalah koherensi topik (topic coherence). Koherensi mengukur tingkat keterkaitan semantik antar kata dalam satu topik, dan salah satu varian yang paling populer digunakan adalah metrik c_v, yang menggabungkan statistik co-occurrence dengan model semantik berbasis sliding window dan confirmation measure (Röder, Both, & Hinneburg, 2015).

Koherensi yang tinggi menunjukkan bahwa kata-kata dalam suatu topik sering muncul bersama dalam dokumen yang sama, dan secara semantik membentuk satu kesatuan makna yang dapat dipahami oleh manusia. Oleh karena itu, pemilihan jumlah topik (K) dalam LDA sering kali didasarkan pada nilai koherensi terbaik dari beberapa model uji coba.

Selain koherensi, aspek stabilitas topik juga perlu diperhatikan. Stabilitas merujuk pada konsistensi topik yang dihasilkan ketika model dilatih ulang dengan data yang sedikit bervariasi, misalnya dengan teknik bootstrapping atau data shuffling. Salah satu cara umum untuk mengukur stabilitas adalah dengan menggunakan Jaccard similarity, yaitu rasio antara irisan dan gabungan dari daftar kata penting yang muncul dalam topik yang sama dari dua model berbeda (Syed & Spruit, 2017). Nilai Jaccard yang tinggi menunjukkan bahwa topik bersifat robust dan tidak terlalu dipengaruhi oleh noise dalam data.

Kedua metrik ini —koherensi dan stabilitas— dianggap sebagai kombinasi yang memadai untuk mengevaluasi performa intrinsik model LDA, khususnya dalam konteks eksplorasi topik berbasis ulasan wisatawan yang cenderung memiliki gaya bahasa informal dan keragaman konteks yang tinggi.

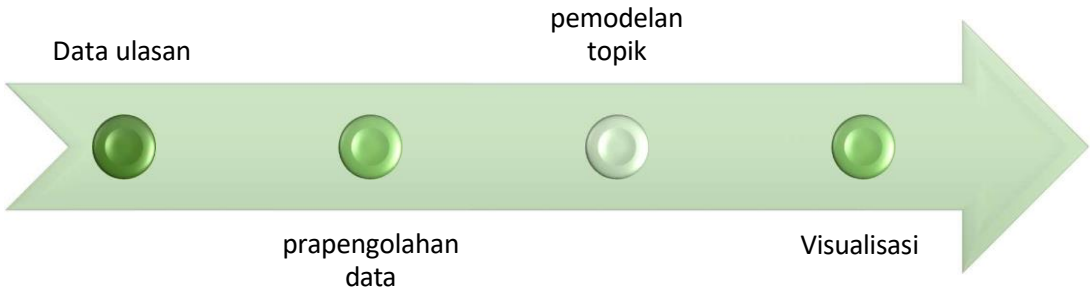
3. METODOLOGI

Penelitian ini menggunakan pendekatan eksploratif deskriptif, dengan tujuan untuk menggali pola tematik dalam data teks ulasan wisatawan secara tidak berlabel. Metode yang digunakan adalah pemodelan topik berbasis Latent Dirichlet Allocation (LDA) untuk mengidentifikasi tema-tema tersembunyi yang terkandung dalam kumpulan ulasan digital.

Secara umum, tahapan penelitian ini meliputi:

1. Akuisisi data ulasan wisatawan digital;
2. Pra-pemrosesan teks (tokenisasi, normalisasi, dll.);
3. Pembangunan model LDA dengan berbagai jumlah topik (K);
4. Evaluasi kualitas model berdasarkan *coherence score* dan *stability*;
5. Visualisasi hasil model dalam bentuk wordcloud, grafik distribusi topik, dan pyLDAvis.

Diagram alur proses penelitian disajikan sebagai berikut:



Gambar 1. Alur penelitian

Pra-pemrosesan dilakukan untuk mengubah teks mentah menjadi bentuk yang siap dianalisis oleh model LDA. Langkah-langkah yang dilakukan meliputi:

1. Tokenisasi: memecah kalimat menjadi satuan kata.

2. Normalisasi teks: termasuk mengubah semua huruf menjadi huruf kecil, menghapus tanda baca, angka, dan karakter non-alfabet.
3. Stopword Removal: menghapus kata-kata umum yang tidak memiliki nilai semantik tinggi (misalnya: "yang", "dan", "juga"). Digunakan daftar stopword Bahasa Indonesia yang disesuaikan dengan domain pariwisata.
4. Pembuatan Bigram *(optional)*: mendeteksi frasa dua kata yang sering muncul, seperti “tiket masuk”, “spot foto”.
5. Stemming: menggunakan algoritma Sastrawi untuk mengembalikan kata ke bentuk dasarnya, misalnya “berkunjung” menjadi “kunjung”.

Selanjutnya Model LDA dibangun menggunakan library **Gensim** (Python) dan dibandingkan dengan versi berbasis **Mallet** (jika diperlukan untuk perbandingan kecepatan dan konsistensi hasil). Pemodelan dilakukan dengan langkah-langkah berikut:

- **Representasi vektor**: menggunakan metode *Bag-of-Words* (BoW), dengan memanfaatkan objek `Dictionary` dan `Corpus` dari Gensim.
- **Eksplorasi jumlah topik (K)**: menggunakan koherensi
- **Parameter**:
  - α = ‘asymmetric’
  - η = ‘auto’
  - Passes = 10
  - Iterations = 100
  - Random State = 42

4. HASIL DAN PEMBAHASAN

4.1 Model Optimal dan Evaluasi Kualitas Topik

Pemodelan topik menggunakan Latent Dirichlet Allocation (LDA) dilakukan terhadap korpus ulasan wisatawan digital yang telah melalui tahapan pra-pemrosesan. Eksperimen dilakukan dengan berbagai nilai topik (`K`), dan berdasarkan evaluasi metrik coherence score **,** diperoleh bahwa model terbaik dihasilkan pada **K = 10**, dengan nilai koherensi tertinggi.

Model ini dipilih sebagai model final karena menghasilkan topik-topik yang tidak hanya memiliki tingkat keterkaitan semantik yang baik (koheren), tetapi juga dapat diinterpretasikan secara tematis sesuai konteks pariwisata di Indonesia.

4.2 Deskripsi Topik yang Dihasilkan

Berikut adalah hasil identifikasi 10 topik utama yang diperoleh dari model LDA terbaik (`K = 10`). Masing-masing topik terdiri dari sepuluh kata kunci utama disertai bobot probabilitasnya, yang menunjukkan seberapa besar kontribusi kata tersebut dalam membentuk topik. Selanjutnya, berdasarkan interpretasi terhadap kata-kata tersebut dan kutipan dari ulasan yang termasuk dalam topik, dilakukan pemberian label manual untuk memudahkan pembacaan tematik. Sebagian besar topik mencerminkan jenis-jenis destinasi wisata populer di Indonesia**,** seperti wisata alam (pantai, gunung, air terjun), budaya (candi, museum), dan wisata keluarga (kereta, wahana anak). Topik 7 dan 9 menunjukkan kecenderungan ulasan yang berkaitan dengan pengalaman kuliner dan suasana kota pada malam hari.

Tabel 1. Hasil Identifikasi Topik Berdasarkan Pemodelan LDA

No	Hasil pemodelan topik
0	pantai (0.063), gua (0.010), orang, pasir, laut, besar, satu, indah, beberapa, jalan
1	foto (0.027), bagus, spot, jalan, tiket, waduk, datang, orang, hari, sekali
2	gunung (0.022), pemandangan, melihat, bukit, indah, merapi, menuju, jalan, perjalanan, buah
3	anak (0.076), kereta, api, rawa, ambarawa, barong, klenteng, stasiun, pening, danau
4	air (0.034), anak, wisata, terjun, keluarga, cocok, taman, sungai, sejuk, disini
5	kuda (0.023), cemara, outbond, mengadakan, bunker, menu, gedong, komunitas, dusun, diperbaiki
6	candi (0.093), prambanan, satu, kuil, jawa, terletak, salah, berada, kompleks, melihat
7	cafe (0.013), sultan, sidomukti, lezat, kraton, raja, wates, barang, profesional, tema
8	museum (0.042), batik, sejarah, rumah, koleksi, keraton, indonesia, baik, solo, menarik
9	kota (0.040), malam, jogja, hari, alun, wisata, semarang, taman, lampion, yogyakarta

Dari tabel tersebut, terlihat bahwa:

- Topik 0 dan 2 sangat dominan dalam ulasan wisata alam, menunjukkan kecenderungan wisatawan membicarakan pantai dan gunung dalam bentuk narasi visual dan deskriptif.
- Topik 6 dan 8 mendominasi ulasan wisata budaya, menandakan tingginya ketertarikan terhadap situs cagar budaya dan museum.
- Beberapa topik seperti Topik 1 (Spot Foto) dan Topik 9 (Aktivitas Malam) muncul dalam proporsi yang relatif merata, mengindikasikan bahwa kedua jenis destinasi sama-sama menyediakan pengalaman visual dan aktivitas malam yang menarik.

Hasil ini memperkuat argumen bahwa preferensi wisatawan dapat diidentifikasi melalui distribusi topik yang muncul dalam ulasan. Dengan demikian, pendekatan berbasis topik ini memungkinkan pengelola destinasi untuk memahami persepsi wisatawan secara lebih granular.

5. KESIMPULAN

Penelitian ini bertujuan untuk mengeksplorasi struktur tematik yang terkandung dalam ulasan wisatawan digital terhadap destinasi wisata di Indonesia, menggunakan pendekatan Latent Dirichlet Allocation (LDA). Melalui proses eksploratif-deskriptif, diperoleh pemetaan 10 topik dominan yang mencerminkan keberagaman tema ulasan, seperti keindahan alam (pantai, gunung, air terjun), aktivitas keluarga, pengalaman budaya (candi, museum), hingga kuliner dan suasana kota malam. Evaluasi model menunjukkan bahwa nilai coherence terbaik diperoleh pada jumlah topik (K) = 10, dan Hasil penelitian menunjukkan topik-topik yang dihasilkan mencerminkan berbagai aspek wisata, seperti keindahan alam, warisan sejarah, aktivitas keluarga, kuliner, serta landmark populer. Analisis distribusi topik menunjukkan sebaran dan proporsi tema dalam korpus secara keseluruhan, memberikan pemahaman kontekstual yang kaya terhadap narasi wisatawan. Hal ini membuktikan bahwa pemodelan topik dapat mengungkap persepsi dan preferensi wisatawan secara tematis mendukung pariwisata cerdas.

Dari sisi implikasi praktis, hasil studi ini membuka peluang untuk menyediakan dashboard wawasan tematik berbasis ulasan bagi pengelola destinasi dan Dinas Pariwisata, Secara umum, pendekatan berbasis LDA terbukti efektif untuk menganalisis data tidak terstruktur dalam sektor pariwisata, khususnya dalam konteks Indonesia yang memiliki keanekaragaman destinasi. Ke depan, model ini dapat dikembangkan lebih lanjut dengan menggabungkan analisis sentimen, temporal (musiman), atau spasial (berbasis lokasi) untuk membentuk kerangka kerja Smart Tourism Intelligence yang lebih holistik.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) Universitas Sebelas Maret (UNS) dan Riset grup Applied Data Science and AI atas dukungan pendanaan melalui skema Program Penelitian Dosen (PDD) UNS Tahun 2025 dengan nomor kontrak: 369/UN27.22/PT.01.03/2025. Dukungan ini sangat berperan dalam kelancaran pelaksanaan penelitian, penulisan dan publikasi artikel ini.

DAFTAR PUSTAKA

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. Advances in Neural Information Processing Systems (NeurIPS), 22, 288–296. https://proceedings.neurips.cc/paper/2009/hash/38563a4a046052e6b9d8ed9640cab723-Abstract.html

Chuang, J., Ramage, D., Manning, C. D., & Heer, J. (2012). Interpretation and trust: Designing model-driven visualizations for text analysis. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 443–452. https://doi.org/10.1145/2207676.2207738

Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the world are you? Geolocation and language identification in Twitter. The Professional Geographer, 66(4), 568–578. https://doi.org/10.1080/00330124.2014.907699

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management, 33(3), 464–472. https://doi.org/10.1016/j.ijinfomgt.2013.01.001

Jelodar, H., Wang, Y., Yuan, C., et al. (2019). Latent Dirichlet Allocation (LDA) and topic modeling: Models, applications, a survey. Multimedia Tools and Applications, 78, 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Moraes, R., Valiati, J. F., & Gavião Neto, W. P. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications, 40(2), 621–633. https://doi.org/10.1016/j.eswa.2012.07.059

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Rahimi, M., & Moghaddam, M. (2022). A hybrid topic modeling and sentiment analysis approach for tourism experience analysis from online reviews. Tourism Management Perspectives, 44, 101037. https://doi.org/10.1016/j.tmp.2022.101037

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011). Comparing Twitter and traditional media using topic models. Advances in Information Retrieval. ECIR 2011, 338–349. https://doi.org/10.1007/978-3-642-20161-5_34