E-ISSN 2808-5841
P-ISSN 2808-7283

Perspektif Wisatawan Mancanegara (Wisman) Terhadap Pariwisata Indonesia menggunakan *Latent Dirichlet Allocation* (LDA)

Rafik Septiana[1], Muhammad Hanif[2], Christeigen Theodore Suhalim[3], Faqih Syamil[4], Arya Duta Gordra Sumitro Putra[5], Maryamah Maryamah[6]

[1, 2, 3, 4, 5, 6] *Teknologi Sains Data, Universitas Airlangga*
[1] rafik.septiana-2021@ftmm.unair.ac.id
[2] muhammad.hanif-2021@ftmm.unair.ac.id
[3] christeigen.theodore.suhalim-2021@ftmm.unair.ac.id
[4] faqih.syamil-2021@ftmm.unair.ac.id
[5] arya.duta.gor-2021@ftmm.unair.ac.id
[6] maryamah@ftmm.unair.ac.id
Corresponding author email: maryamah@ftmm.unair.ac.id

A Abstract: *The 2022 G20 Bali summit in Indonesia provides an opportunity to revitalize the tourism industry after the Covid-19 pandemic. The perspectives of foreign tourists visiting Indonesia, including the advantages and disadvantages of their tourism experience are an important aspect to improve tourism in Indonesia. In this paper, we propose the perspective of foreign tourists on Indonesian tourism using Latent Dirichlet Allocation (LDA). The first step was scrapping data comments by foreign tourists regarding tourism in Indonesia from Twitter. The next step of this paper is data preprocessing consisting of cleaning data, tokenization, stopwords, and stemming. The LDA method is used to analyze and identify the main topics from the comments until obtaining relevant insights. The results of the research reveal three main topics: tourist satisfaction with Indonesian tourism, foreign tourist activities during their tour, and plans for foreign tourists to visit Indonesia. These findings provide valuable insights for the government in optimizing the tourism sector in the future and supporting post-pandemic economic recovery. In the context of recovery, understanding the strengths and weaknesses identified by foreign tourists can help the government take appropriate steps to increase Indonesia's attractiveness as a tourist destination.*

Keywords: *Perspective of International Tourists, Indonesia Tourism, Topic Modelling, Latent Dirichlet Allocation (LDA)*

Abstrak: Konferensi Tingkat Tinggi (KTT) G20 2022 di Bali-Indonesia memberikan peluang untuk merevitalisasi industri pariwisata pasca pandemi Covid-19. Penelitian ini menganalisis perspektif wisatawan mancanegara (wisman) yang telah mengunjungi Indonesia, termasuk kelebihan dan kekurangan pengalaman pariwisata mereka. Pada paper ini, kami mengusulkan perspektif wisatawan mancanegara (wisman) terhadap pariwisata indonesia menggunakan Latent Dirichlet Allocation (LDA). Langkah-langkah yang dilakukan meliputi *scrapping* data dari *platform* Twitter yang berisi komentar wisman terkait pariwisata di Indonesia. Selanjutnya menggunakan Metode LDA untuk menganalisis mengidentifikasi topik utama yang muncul dari komentar. Langkah-langkah pengolahan data dan analisis dilakukan untuk memperoleh wawasan yang relevan. Hasil penelitian mengungkap tiga topik utama: kepuasan wisman terhadap pariwisata Indonesia, aktivitas wisman selama berwisata, dan rencana kunjungan wisman ke Indonesia. Temuan ini memberikan wawasan berharga bagi pemerintah dalam mengoptimalkan sektor pariwisata di masa depan dan mendukung pemulihan ekonomi pasca pandemi. Dalam konteks pemulihan, pemahaman kelebihan dan kekurangan yang diidentifikasi oleh wisman dapat membantu pemerintah mengambil langkah-langkah yang tepat untuk meningkatkan daya tarik Indonesia sebagai tujuan wisata yang diinginkan oleh wisman.

Kata kunci: Perspektif Wisatawan Mancanegara, Pariwisata Indonesia, Pemodelan Topik, *Latent Dirichlet Allocation (LDA)*

E-ISSN 2808-5841
P-ISSN 2808-7283

I. PENDAHULUAN

Pariwisata merupakan sektor yang sangat potensial untuk memberikan kontribusi bagi kemakmuran ekonomi suatu negara [1]. Pariwisata merupakan salah satu bidang kritis di Indonesia yang mendapat perhatian besar dari pemerintah, karena berpotensi memberikan kontribusi yang signifikan terhadap perekonomian nasional. Pemerintah Indonesia telah melakukan beberapa upaya untuk meningkatkan sektor pariwisata, antara lain memasarkan tempat wisata, menyediakan infrastruktur yang memadai, dan meningkatkan pelayanan pariwisata. Namun, memahami perspektif dan persepsi wisatawan mancanegara (wisman) terhadap lokasi pariwisata Indonesia sangat penting dalam mendongkrak pariwisata. Pengunjung asing memainkan peran penting dalam perluasan industri pariwisata karena mereka sangat berkontribusi terhadap pengeluaran pariwisata, penciptaan lapangan kerja, dan peningkatan pendapatan masyarakat lokal [2]. Oleh karena itu, pemahaman menyeluruh tentang persepsi wisatawan internasional terhadap pariwisata Indonesia dapat menjadi landasan penting untuk mengembangkan tujuan wisata. Untuk mengetahui persepsi pengunjung asing terhadap pariwisata Indonesia, penelitian ini memanfaatkan metode *Latent Dirichlet Allocation* (LDA). LDA adalah metode untuk mengidentifikasi topik dan tema dalam teks yang telah banyak digunakan dalam analisis dan penelitian teks. LDA akan digunakan dalam penelitian ini untuk mengevaluasi evaluasi pengunjung internasional terhadap pariwisata Indonesia untuk menentukan topik utama yang muncul dalam ulasan mereka.

Isu-isu yang dibahas dalam penelitian ini meliputi bagaimana pengunjung internasional memandang pariwisata Indonesia, elemen apa yang membuat mereka khawatir, dan apakah ada perbedaan sikap di antara kelompok turis asing yang berbeda dari negara yang berbeda. Studi ini akan menganalisis penelitian sebelumnya tentang masalah yang sama untuk menjawab pertanyaan-pertanyaan ini. Studi sebelumnya telah dilakukan dalam upaya untuk memahami perspektif wisatawan mancanegara terhadap pariwisata Indonesia, namun penelitian yang menggunakan pendekatan LDA dalam menilai evaluasi wisatawan masih terbatas. Beberapa penelitian sebelumnya telah menemukan bahwa unsur-unsur seperti kualitas layanan, keindahan alam, keragaman budaya, dan aksesibilitas berdampak pada pendapat wisatawan internasional tentang pariwisata. Namun, belum ada penelitian yang mengidentifikasi subjek utama yang muncul dalam ulasan pengunjung internasional tentang pariwisata Indonesia dengan menggunakan pendekatan LDA.

Beberapa studi tambahan menemukan fitur budaya dan keragaman menjadi faktor penarik wisatawan internasional memilih Indonesia sebagai tujuan wisata. Daya tarik terbesar wisatawan internasional adalah kekayaan budaya Indonesia, seperti kehidupan adat, kesenian tradisional, dan upacara keagamaan [3]. Hal ini menunjukkan pentingnya faktor budaya dalam menarik perhatian dan menghasilkan pengalaman unik bagi wisatawan internasional. Meskipun studi ini memberikan wawasan awal yang bermanfaat, mengidentifikasi masalah utama yang muncul dalam penilaian pengunjung internasional terhadap pariwisata Indonesia membutuhkan pendekatan yang lebih sistematis dan menyeluruh. Maka, metode *Latent Dirichlet Allocation* (LDA) digunakan dalam pekerjaan ini sebagai alat analisis yang dapat memberikan wawasan yang lebih dalam[4]. Pendekatan LDA akan membantu akademisi untuk secara otomatis menemukan tren topik dalam evaluasi wisatawan mancanegara. Penelitian ini akan menggunakan pendekatan ini untuk menentukan isu-isu utama yang sering muncul dalam review pengunjung asing, seperti "keindahan alam", "budaya dan tradisi", "kuliner", "aktivitas rekreasi", dan sebagainya. Teknik ini dimaksudkan untuk memberikan pengetahuan yang lebih menyeluruh tentang apa yang dianggap vital dan menarik oleh pengunjung internasional dalam pariwisata Indonesia.

Penelitian ini memiliki implikasi penting bagi pengembangan destinasi pariwisata di Indonesia karena memberikan pengetahuan yang lebih baik tentang perspektif pengunjung internasional terhadap

pariwisata Indonesia. Temuan penelitian ini dapat membantu pemerintah, organisasi pariwisata, dan pelaku sektor pariwisata meningkatkan pengalaman pengunjung internasional selama kunjungan mereka ke Indonesia. Mengetahui hal apa yang paling diapresiasi dan disukai wisatawan mancanegara memungkinkan upaya pengembangan dan promosi pariwisata lebih tepat sasaran dan efektif. Temuan penelitian ini diproyeksikan dapat memberikan kontribusi yang signifikan terhadap pertumbuhan pariwisata Indonesia. Dengan pemahaman yang lebih baik tentang selera dan pendapat wisatawan internasional, pemerintah dan sektor pariwisata dapat mengidentifikasi area yang memungkinkan untuk perbaikan, seperti pembangunan infrastruktur, layanan pariwisata, promosi destinasi, dan pelestarian budaya lokal. Selain itu, penelitian ini dapat membantu mengembangkan rencana pemasaran yang lebih efisien untuk menarik wisatawan mancanegara dan meningkatkan daya saing Indonesia sebagai tujuan wisata internasional. Penelitian ini juga dapat berfungsi sebagai sumber bagi akademisi lain yang tertarik dengan pariwisata dan analisis evaluasi pengunjung. Temuan penelitian ini dapat digunakan untuk memandu penelitian di masa depan tentang subjek yang muncul dalam evaluasi wisatawan mancanegara dari berbagai tempat wisata di seluruh dunia. Hal ini dapat membantu memperluas pemahaman kami tentang selera dan perspektif turis internasional, serta memberikan wawasan yang berguna tentang inisiatif pengembangan pariwisata lain.

Untuk mencapai tujuan penelitian, peneliti mengusulkan perspektif wisatawan mancanegara (wisman) terhadap pariwisata indonesia menggunakan LDA. Hal pertama yang dilakukan adalah mengumpulkan data ulasan pengunjung internasional dari media sosial Twitter yang mengandung kata kunci. Penelitian ini berfokus pada beberapa pusat lokasi pariwisata Indonesia, yaitu Bali, Pulau Komodo, Labuan Bajo, dan Lombok. Rentang waktu unggahan data tweet yaitu pada bulan Mei tahun 2022. Data akan dianalisis menggunakan pendekatan LDA untuk menentukan subjek utama yang disebutkan dalam evaluasi wisatawan internasional. Selanjutnya, temuan tersebut akan diperiksa secara kualitatif dan statistik untuk mengkarakterisasi dan memahami sudut pandang pengunjung internasional terhadap pariwisata Indonesia. Dengan adanya penelitian ini, para pemangku kepentingan pariwisata, masyarakat lokal, dan wisatawan mancanegara dapat menumbuhkan sinergi dalam pembangunan pariwisata Indonesia yang berkelanjutan, budaya, dan menguntungkan bagi semua pihak yang berkepentingan.

II. METODE PENELITIAN

Tahapan metodologi yang dilakukan pada penelitian ini terdiri dari beberapa proses yaitu pengumpulan data, preprocessing data, pemodelan topik, data prosessing dan analisis dari topik yang dihasilkan. Metodologi yang digunakan pada penelitian ini mengikuti proses *text mining* atau proses menambang data berupa teks yang diperoleh melalui dokumen untuk menemukan kata-kata yang berhubungan dengan isi dokumen sehingga dilakukan analisis terhadap hubungan dari sumber teks dapat dilakukan [4]. Penambangan teks ini menganalisis sejumlah data teks tidak terstruktur sehingga dapat mengidentifikasi pola-pola tertentu dan kata kunci dalam data [5].

Penelitian ini menggunakan data yang diperoleh melalui pengambilan data melalui situs media sosial Twitter. Pemrosesan data teks yang diambil dari data Twitter digunakan untuk mengambil kata-kata yang terkait dengan topik pariwisata Indonesia, yang kemudian dilakukan analisis pola-pola tertentu untuk memperoleh informasi yang bermakna dari data tersebut. Data yang diambil berupa *tweet* dari para pengguna aplikasi Twitter. Bahasa yang diambil untuk penelitian ini adalah bahasa inggris berdasarkan topik-topik terkait yang ditentukan oleh keyword yang telah ditetapkan. keyword yang digunakan adalah beberapa tempat pariwisata yaitu Bali, Pulau Komodo, Labuan Bajo, dan Lombok. Data diambil berdasarkan dalam rentang bulan Mei 2022 dikarenakan pada bulan tersebut adalah angka kunjungan tertinggi pada tahun 2022 kuartal pertama yaitu sebesar 212332 kunjungan.

Data yang digunakan pada penelitian ini diambil menggunakan teknik *web scraping* menggunakan bahasa pemrograman Python dan *library* Selenium. *Library* Selenium akan digunakan untuk

melakukan otomatisasi pada halaman web, seperti melakukan *scrolling* dan melakukan pencarian topik. Setelah itu, halaman web akan diekstrak ke dalam file berekstensi CSV dengan memanfaatkan struktur halaman web, seperti tag HTML dan elemen CSS di dalam elemen unggahan. *Web scraping* atau ekstraksi web merupakan suatu teknik untuk mengekstrak data dari internet dan menyimpan hasilnya ke dalam *file* seperti database, CSV, atau *file* berekstensi lainnya [6]. *Web scraping* adalah proses mendapatkan dokumen tidak terstruktur dari website dalam bentuk *markup language* (HTML). Hasil ekstraksi data akan dianalisis untuk mengambil data tertentu dari halaman tersebut [7]. *Web scraping* adalah teknik mendapatkan informasi secara otomatis dari sebuah website tanpa menyalinnya secara manual [8]. Program akan menganalisis dokumen HTML dari internet dan mendapatkan data berdasarkan tag HTML untuk mengapit informasi yang ingin diambil (membuat *scraping template*). Setelah itu, informasi tersebut akan disimpan ke dalam tabel database, CSV, atau format *file* lainnya. Pengambilan data ini akan menggunakan bahasa pemrograman Python, yakni dengan mengambil data melalui tag HTML dan menyimpannya ke dalam *file* CSV untuk diproses ke tahap selanjutnya.

Setelah mendapatkan data dengan menggunakan *web scraping*, proses selanjutnya adala *data cleaning,* yaitu proses mengidentifikasi dan menghilangkan kesalahan dalam data[12]. Hal ini untuk memastikan bahwa data yang diambil memiliki kualitas yang baik. Pembersihan data dapat digunakan untuk membersihkan data *string* untuk menghilangkan karakter yang tidak relevan, salah eja, dan kesalahan serupa lainnya dalam teks. *Data cleaning* dalam pengertian umum adalah proses menyelidiki data dari ketidakakuratan dan membuat data yang dapat dikelola untuk analisis. Dalam penelitian ini, peneliti akan melakukan pembersihan data pada tipe data *string*, yang mana akan dilakukan pembersihan terhadap kata- kata yang tidak relevan dengan topik, menghapus kesalahan ejaan, menormalkan data, atau teknik pembersihan lainnya. Hal ini penting untuk menjamin data yang telah diperoleh dari Twitter dapat dianalisis dan diekstraksi dengan lebih mudah dan terstruktur. Beberapa teknik yang dilakukan dalam *data cleaning* adalah *case folding* (mengubah huruf kapital menjadi kecil), *stemming* (mengubah suatu kata menjadi kata dasar), dan *tokenization* (memecah kalimat menjadi kumpulan kata).

Data yang telah dikumpulkan akan dilakukan proses pembersihan (*cleaning*) melalui *library* Pandas dan Pyspark. Pyspark merupakan *library* Python yang menggunakan bantuan *tools* berupa Apache Spark. *Library* tersebut digunakan untuk meningkatkan performa dalam melakukan pembersihan data. Hal ini disebabkan oleh besarnya data yang akan digunakan dan proses eksekusi akan memakan waktu yang sangat lama apabila hanya menggunakan bahasa Python saja. Proses pembersihan data ini meliputi penghilangan data yang kosong (*missing data*), penghapusan tanda baca di kalimat, *case folding* (pengubahan huruf kapital menjadi huruf kecil), penghapusan *stopwords* pada kalimat (misalnya of, or, dll), *stemming* (pengubahan suatu kata menjadi kata dasar), dan proses *tokenization* (pemecahan kalimat menjadi kata per kata).

Setelah dilakukan pembersihan data, tahap selanjutnya adalah pencarian topik-topik tersembunyi pada kumpulan tweet dengan menggunakan Latent Dirichlet Allocation (LDA). Penelitian ini menggunakan metode Latent Dirichlet Allocation (LDA) yang merupakan model dalam *machine learning* yang menggunakan probabilistik korpus yang diwakili oleh distribusi data untuk setiap kata dalam dokumen [9]. Latent Dirichlet Allocation (LDA) adalah suatu metode untuk mengolah data dalam jumlah besar dengan asumsi bahwa satu dokumen terdiri dari berbagai topik yang merupakan distribusi kosakata [10]. Latent Dirichlet Allocation (LDA) adalah suatu metode dalam *unsupervised learning* yang digunakan untuk mengelompokkan data menjadi beberapa topik, merangkum, dan mengolah data yang besar [11]. Pada penelitian ini, akan dilakukan pengelompokan data Twitter yang telah diperoleh menjadi beberapa topik berdasarkan sebaran kata-kata tertentu dalam *tweet* pengguna media sosial Twitter sehingga ditemukan informasi terkait perspektif wisatawan mancanegara (wisatawan mancanegara) dari topik tersebut. Latent Dirichlet Allocation dapat dikalkulasi menggunakan persamaan (1):

$$P(z = j|z_t, w_t, d_t, \cdot) \infty = \frac{(C_{W_{tj}}^{WT} + \beta)}{\sum_{w=1}^{w}(C_{W_{tj}}^{WT} + W\beta)} x \frac{(C_{d_{tj}}^{DT} + \alpha)}{\sum_{w=1}^{w}(C_{dj}^{DT} + T\alpha)} \quad (1)$$

$C_{W_{tj}}^{WT}$= Frekuensi sebuah kata muncul dalam topik 1 dan topik 2

β = distribusi kata berdasarkan topik (parameter konsentrasi)

W = Panjang kosakata (jumlah token/kata unik dalam dokumen lengkap)

C_{dj}^{DT} = Frekuensi dokumen muncul sebagai topik 1 dan topik 2, ketika terasi dimulai

α = Distribusi topik setiap dokumen

T = Jumlah topik

Regex (*Regular Expression*) merupakan gabungan dari dua jenis karakter, yaitu karakter literal dan meta. Literal adalah karakter yang mewakili dirinya sendiri seperti keseluruhan alfabet, huruf kapital, huruf kecil, atau karakter lainnya [13]. Regex juga merupakan salah satu implementasi operasi pencocokan pola untuk tipe data teks atau *string* [14]. Regex adalah kunci untuk melakukan pemrosesan teks secara kuat, fleksibel, dan efisien dengan menggunakan pola notasi umum seperti bahasa pemrograman [15]. Regex memudahkan tahap pra pemrosesan data, khususnya dalam pencarian teks. Hal ini menjadi dasar penggunaan regex untuk tokenisasi sebuah teks karena informasi terkait letak angka, tanda baca, dan karakter di luar *string* dapat diketahui. Setelah letak karakter diketahui, karakter tersebut akan dihapus. Regex juga dapat memisahkan kalimat menjadi kelompok kata dengan memisahkannya dengan spasi. Hal tersebut dapat memudahkan tokenisasi dokumen saat tahap pra pemrosesan data.

Pada analisis ini, dibutuhkan dua *hyperparameter*, yaitu α dan β. Untuk mengendalikan distribusi topik pada data, dibutuhkan hyperparameter α, yang mana semakin kecil nilai α maka data *tweet* akan cenderung memiliki satu topik, dan semakin besar nilai α maka semakin seragam topik-topik pada data tersebut. Di sisi lain, nilai β berperan dalam mengendalikan distribusi kata-kata pada suatu topik. Nilai β yang semakin kecil menunjukkan bahwa topik cenderung memiliki sedikit kata, sedangkan nilai yang semakin besar menunjukkan bahwa topik-topik memiliki variasi kata yang banyak. Luaran dari LDA adalah probabilitas kata-kata pada tiap topik. Suatu topik pada LDA (k) memiliki semua kosakata dari kumpulan dokumen yang diolah. Masing-masing kata pada topik k tersebut akan memiliki nilai dengan rentang 0 hingga 1. Nilai ini menunjukkan tingkat probabilitas suatu kata untuk muncul pada topik tertentu. Selain probabilitas dari masing-masing kata, LDA juga menampilkan probabilitas dari masing-masing topik pada suatu dokumen. Dengan nilai probabilitas ini, hasil dari penelitian dapat direpresentasikan dalam suatu dokumen atau topik berdasarkan kata-kata yang memiliki probabilitas paling tinggi. Dengan demikian, seluruh informasi dari data unggahan para pengguna Twitter dapat diekstrak untuk mendapatkan topik-topik tersembunyi, yang mana hasil akan direpresentasikan berdasarkan kata-kata yang memiliki probabilitas paling tinggi. Pada penelitian ini, algoritma LDA juga akan dibandingkan dengan algoritma PLSA yang memiliki kemampuan yang sama dalam mengekstrak topik sehingga kita dapat membandingkan keduanya untuk mendapatkan hasil yang terbaik.

III. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan *Latent Dirichlet Allocation* (LDA) untuk menemukan topik-topik tersembunyi (laten) beserta distribusi kata di dalam topik tersebut. Sebagai pembanding, algoritma *Latent Semantic Analysis* (LSA) dan *Probabilistic Latent Semantic Analysis* (pLSA) juga digunakan dalam analisis ini. LDA, LSA, dan pLSA merupakan beberapa model *machine learning* yang umum digunakan pada kasus analisis topik [16,17]. Pada algoritma LSA, matriks dokumen yang berisi istilah-istilah dipecah ke dalam tiga matriks menggunakan proses *singular value decomposition* (SVD) [17]. Algoritma LSA membangun ruang semantik dari korpus (kumpulan dokumen) yang berupa ruang vektor, yang mana ruang semantik ini kemudian dapat digunakan untuk menghitung kesamaan antar kata, kalimat, paragraf, ataupun seluruh dokumen [17,18]. Akan tetapi, algoritma LSA tidak memiliki

model probabilistik untuk kemunculan kata atau istilah sehingga pengelompokan topik lebih sulit untuk diinterpretasikan [16]. Selain itu, algoritma LSA memiliki kekurangan dari sisi interpretabilitas *embeddings* yang dapat menimbulkan subjektivitas pada hasil [19]. Untuk mengatasi limitasi pada algoritma LSA, digunakan algoritma pemodelan topik yang memanfaatkan model probabilistik, yaitu pLSA dan LDA. Kedua model tersebut memodelkan topik laten dan memanfaatkan konsep probabilitas untuk memberikan bobot atau tingkat kepentingan dari setiap kata yang terkelompok pada suatu topik.

Analisis pada penelitian ini difokuskan pada model LDA. Sebagai pembanding, model LSA dan pLSA juga digunakan. Hal ini disebabkan oleh keunggulan model LDA yang dapat mengeneralisasikan dokumen (kumpulan kata) baru dengan lebih baik [16]. Model pLSA merupakan pengembangan dari model LSA yang menyelesaikan masalah interpretabilitas dan subjektivitas [17]. Akan tetapi, model pLSA termasuk *incomplete* karena tidak menyediakan model probabilistik pada tingkat dokumen. Setiap dokumen direpresentasikan sebagai *list* angka (proporsi campuran topik), dan tidak ada model probabilistik generatif untuk angka-angka tersebut. Hal ini menimbulkan jumlah parameter yang dibutuhkan dalam model semakin besar seiring dengan semakin besarnya korpus, yang mana dapat menyebabkan *overfitting* pada model. Selain itu, proses menetapkan probabilitas ke dokumen di luar *training set* juga belum jelas [16]. Singkatnya, model PLSA bukanlah model generatif yang tepat apabila terdapat dokumen baru. Oleh karenanya, model LDA digunakan sebagai solusi dari kelemahan model pLSA. Model LDA memiliki kemampuan pertukaran (*exchangeability*) dari kata dan dokumen menggunakan distribusi Dirichlet, yang memungkinkan proses generatif yang koheren untuk data uji [16].

Tabel 1 merujuk kepada kata-kata penting pada setiap topik yang diperoleh menggunakan metode LSA, sedangkan Tabel 2 merujuk kepada kata-kata penting pada setiap topik yang diurutkan berdasarkan bobot probabilitas atau tingkat kepentingannya, yang diperoleh menggunakan metode pLSA dan LDA.

Tabel 1. Hasil Keseluruhan Skenario Topik (Model LSA)

topic_id	word
0	bali; photo; posted; indonesia; com; go; love; speedway; irene; island
1	posted; photo; speedway; com; indonesia; gods; courtesy; peace; island; cd
2	go; want; lets; wanna; let; posted; bad; photo; lombok; need

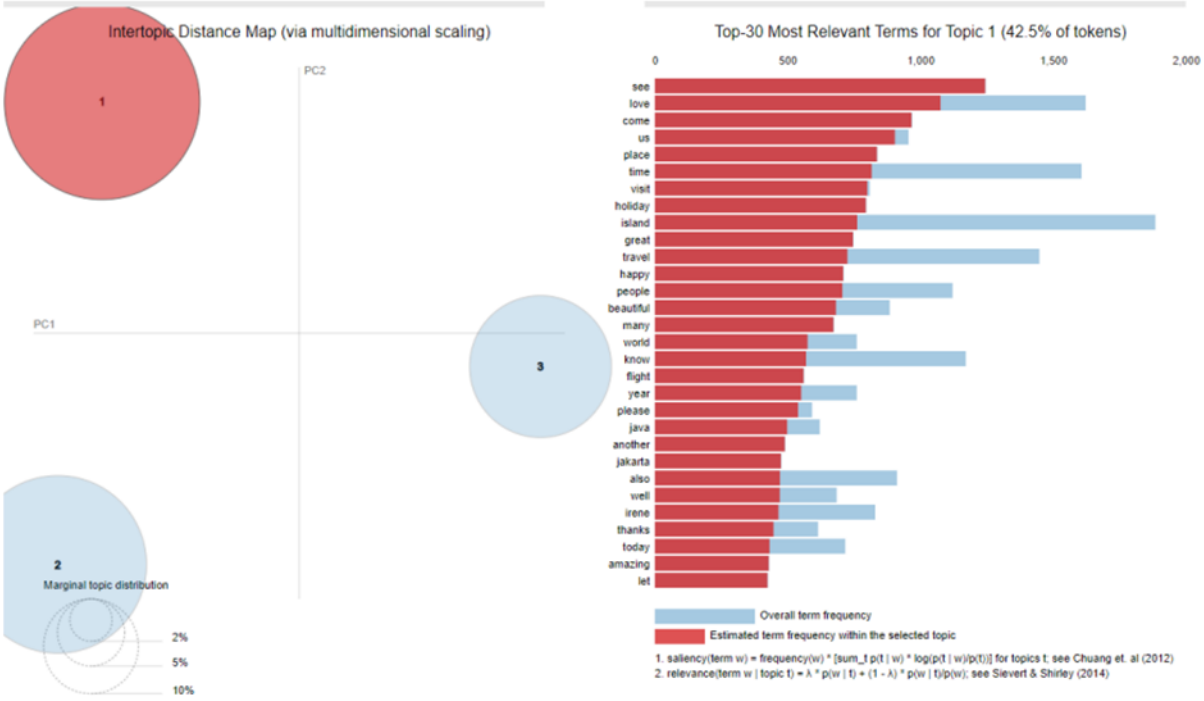
Tabel 2. Perbandingan Hasil Keseluruhan Skenario Topik (Model pLSA dan LDA)

pLSA				LDA		
no	**word**	**importance**	**topic_id**	**word**	**importance**	**word_count**
1	flight	0.013002	0	lombok	0.015760	809
2	bali	0.011489	0	like	0.010584	887
3	man	0.009647	0	love	0.007022	668
4	photo	0.009001	0	island	0.006647	553
5	com	0.008833	0	day	0.006442	529
6	cliff	0.008426	0	back	0.006229	530

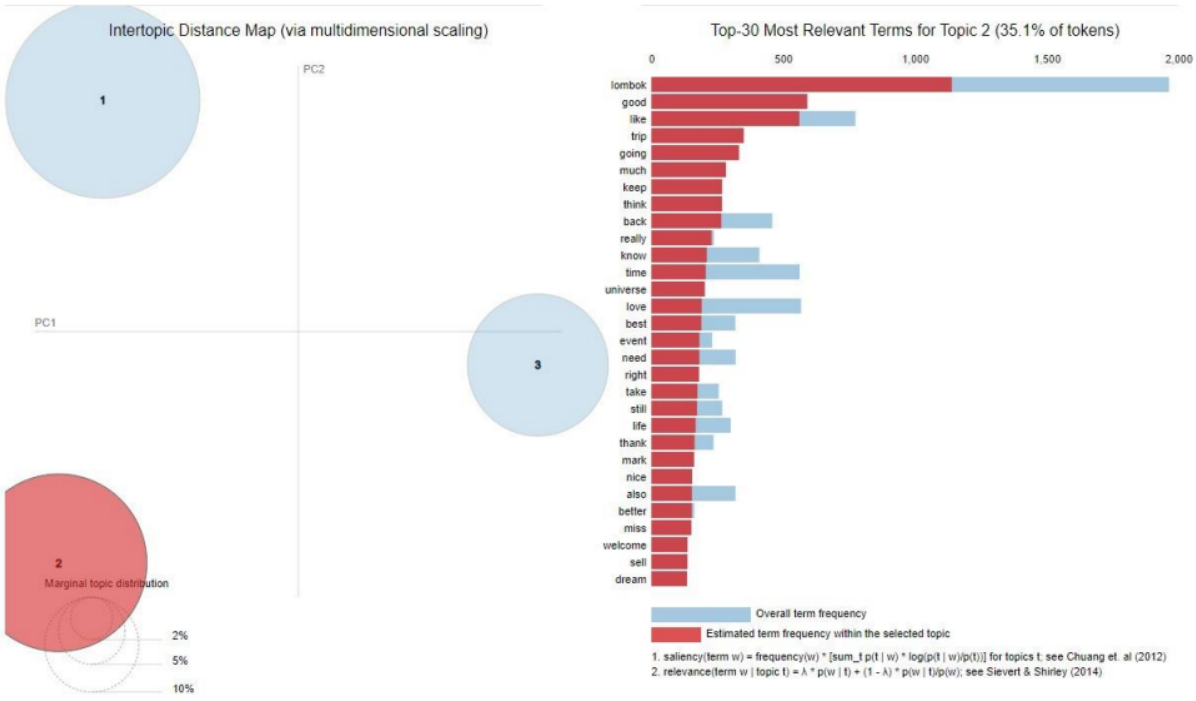
7	island	00.007442	0	time	0.006160	652
8	internet	0.007117	0	see	0.005527	469
9	plane	0.006988	0	know	0.005186	435
10	hotel	0.006988	0	trip	0.004864	404
11	bali	0.014843	1	lombok	0.021229	809
12	year	0.009379	1	photo	0.015976	533
13	com	0.008966	1	posted	0.011335	367
14	indonesia	0.008278	1	check	0.008028	313
15	trip	0.007553	1	good	0.007831	695
16	dom	0.007312	1	keep	0.007531	261
17	day	0.007296	1	universe	0.006517	123
18	friend	0.006948	1	event	0.005966	221
19	place	0.006929	1	listing	0.005816	222
20	week	0.006368	1	morning	0.005391	195
21	home	0.010192	2	lombok	0.004582	809
22	bali	0.009529	2	news	0.003762	225
23	position	0.007653	2	tourism	0.003532	157
24	family	0.7601	2	resort	0.003532	138
25	taste	0.006288	2	project	0.003227	132
26	visit	0.006254	2	flights	0.003020	147
27	kid	0.006074	2	lets	0.003009	149
28	life	0.005999	2	world	0.002936	298
29	beach	0.005961	2	video	0.002892	138

Tahap berikutnya adalah interpretasi topik yang sudah diekstraksi oleh LDA. Interpretabilitas merupakan kunci berharga dalam penentuan kualitas LDA yang diaplikasikan untuk masalah prediksi. Hal ini dikarenakan topik diskusi akan dikaitkan dengan risiko hasil. Matriks β dapat digunakan untuk mendapatkan pemahaman awal terkait jenis topik yang muncul dengan melihat kata-kata yang paling mewakili topik tersebut. Untuk menginterpretasikan topik tersebut dapat digunakan visualisasi berupa bar chart. Berikut merupakan visualisasi untuk penemuan kata dengan metode *Latent Dirichlet Allocation* LDA per topik.

E-ISSN 2808-5841
P-ISSN 2808-7283

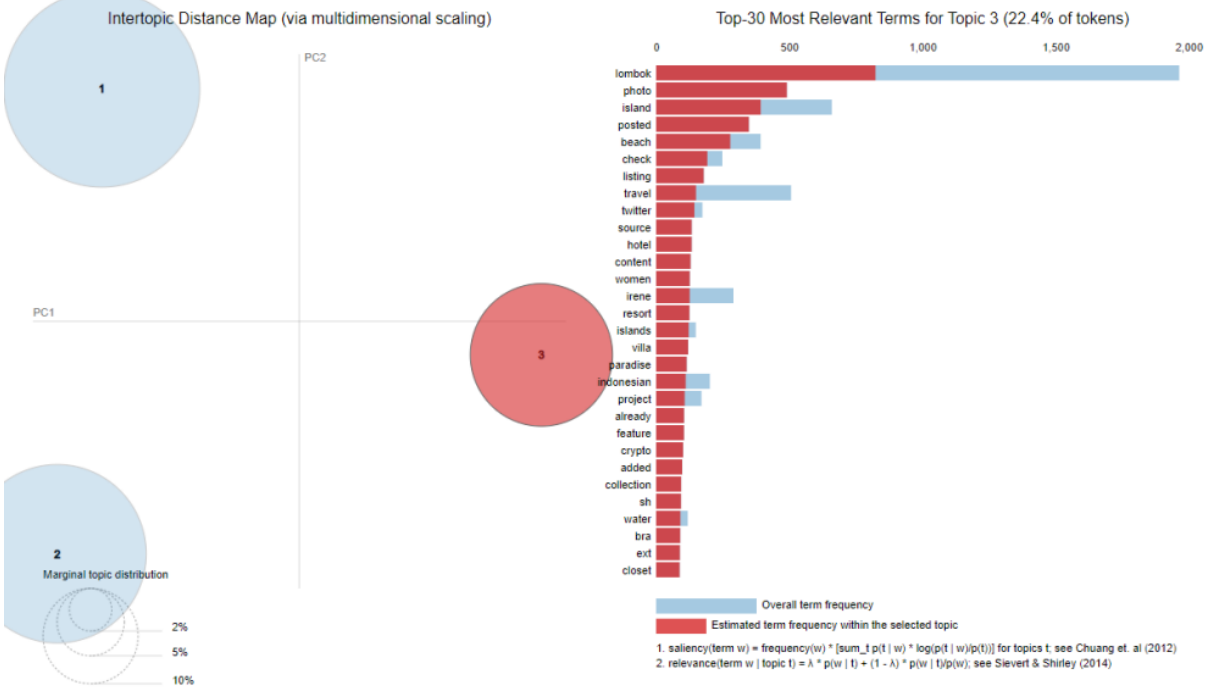


Gambar 1. *Topic 1 Intertopic Distance Location Map & its Top 30 Most Relevant Terms*



Gambar 2. *Topic 2 Intertopic Distance Location Map & its Top 30 Most Relevant Terms*

E-ISSN 2808-5841
P-ISSN 2808-7283



Gambar 3. *Topic 3 Intertopic Distance Location Map & its Top 30 Most Relevant Terms*

Topik dapat dilihat sebagai konstruksi dasar yang diukur dengan kombinasi istilah yang membentuk topik melalui logika probabilitas. Misalnya, topik 0 mengandung istilah seperti "like", "good", "love", "time", "back", "know", "see", dan "trip". Kata-kata tersebut mengacu pada kepuasan dan kecintaan para wisatawan mancanegara (wisman) saat berwisata di Indonesia. Demikian pula pada topik 2 yang berisi "photo", "posted", "check", "good", "keep", "event", dan "listing" yang merujuk kepada kegiatan-kegiatan yang dilakukan wisman di Indonesia. Terakhir, topik 3 mengandung istilah "news", "tourism", "resort", "project", "flights", "world", dan "video" yang merepresentasikan rencana wisman saat mendatangi Indonesia. Pola dari hasil tersebut dapat diidentifikasi dengan jelas melalui pemeriksaan setiap topik LDA.

Probabilitas dari topik yang dimodelkan memberikan informasi kuantitatif terkait tweet yang diunggah wisman mengenai pengalaman mereka ketika menikmati pariwisata Indonesia. Framework LDA memodelkan pariwisata Indonesia sebagai gabungan berkelanjutan dari beberapa topik. Dengan demikian, LDA menyediakan konteks tambahan dalam tingkat variabel terstruktur. Oleh sebab itu, topik LDA probabilitas meningkatkan prediksi serta pemahaman tentang apa pendapat wisman terhadap pariwisata di Indonesia.

IV. KESIMPULAN

Metode *Latent Dirichlet Allocation (LDA)* dapat membantu dalam memahami pandangan dan perspektif wisman. Metode ini memiliki keunggulan dapat menggeneralisasikan dokumen (kumpulan kata) baru dengan lebih baik daripada model pLSA yang termasuk *incomplete* karena tidak menyediakan model probabilistik pada tingkat dokumen. Dari hasil pengujian dengan metode LDA, didapati bahwa hasil dari ekstraksi topik LDA mengenai pendapat wisatawan mancanegara (wisman) terhadap pariwisata di Indonesia menunjukkan adanya pola berupa tiga topik. Tiga topik tersebut meliputi topik terkait kepuasaan dan kecintaan para wisman saat berwisata di Indonesia, kegiatan yang dilakukan saat berkunjung di Indonesia, dan rencana wisman saat mengunjungi Indonesia. Dari ketiga topik tersebut, didapati bahwa wisman yang berkunjung di Indonesia memiliki kesan yang positif

terhadap pariwisata Indonesia. Seluruh kata-kata di dalam topik menyatakan kepuasaan dan kecintaan wisman terhadap pariwisata di Indonesia. Wisman mengabadikan momen kunjungannya ke Indonesia melalui foto sehingga dapat diketahui kegiatan yang dilakukan selama di Indonesia. Kata-kata yang memiliki tingkat probabilitas tinggi menunjukkan tujuan wisman ke Indonesia, yaitu untuk berwisata ataupun melakukan projek pribadi. Dengan demikian, hasil dari pencarian topik-topik tersembunyi tersebut dapat digunakan untuk membantu pemerintah dan pihak terkait dalam potensi pariwisata di Indonesia terkhusus untuk membuat kebijakan atau membangun fasilitas yang memadai untuk meningkatkan daya tarik pariwisata Indonesia dalam rangka pemulihan ekonomi pasca pandemi Covid-19.

Penelitian ini memberikan wawasan yang bermanfaat tentang pandangan dan perspektif wisman terhadap pariwisata Indonesia dengan menggunakan metode Latent Dirichlet Allocation (LDA). Meskipun demikian, masih ada beberapa area yang dapat dieksplorasi lebih lanjut (misalnya berfokus pada suatu daerah di Indonesia). Pertama, penelitian selanjutnya dapat memperluas cakupan data yang digunakan, dengan melibatkan jumlah sampel yang lebih besar, dari wisman yang berbeda asal dan usia. Dengan mengumpulkan data dari kelompok wisman yang berbeda, hasil dari ekstraksi topik LDA dapat menjadi lebih representatif. Kedua, penelitian selanjutnya dapat memperdalam analisis topik LDA untuk mengidentifikasi lebih banyak topik tersembunyi terkait pariwisata Indonesia yang belum terungkap dalam studi ini. Terakhir, penelitian masa depan dapat mengeksplorasi hubungan antara pandangan dan perspektif wisman dengan faktor-faktor seperti preferensi wisatawan terhadap destinasi wisata, tingkat kepuasan dan pengalaman wisatawan, dan dampak pariwisata pada lingkungan dan masyarakat setempat. Diharapkan penelitian ini dapat membantu lebih memahami pandangan dan perspektif wisman terhadap pariwisata Indonesia dan memberikan masukan untuk pengembangan pariwisata Indonesia.

REFERENSI

1. A. R. Rahma, "Potensi Sumber Daya Alam dalam Mengembangkan Sektor Pariwisata di Indonesia," Jurnal Nasional Pariwisata, vol. 12, April 2020.
2. O. Kaikara, "Tourism Development Strategy," International Journal Papier, vol. 1, no. 1, pp. 20-25, 27 Agustus 2020.
3. L. Asy'ari, R. D. Dienaputra, A. Nugraha, R. Tahir, C. U. Rakhman and R. R. Putra, "Kajian Konsep Ekowisata Berbasis Masyarakat Dalam Menunjang Pengembangan Pariwisata: Sebuah Studi Literatur," Jurnal Ilmiah Pariwisata Agama dan Budaya, vol. 6, 30 Desember 2021.
4. I. Sutherland, Y. Sim, S. K. Lee, J. Byun and K. Kiatkawsin, "Topic Modeling of Online Accommodation Reviews via Latent Dirichlet Allocation," Sustainability, vol. 12, 27 Februari 2020.
5. A. Udgave dan P. Kulkarni, "Text Mining and Text Analytics of Research Articles," PalArch's Journal of Archeology of Egypt/Egyptology, no. 17, p. 6, 2020.
6. M. L. C. Chilmi, "Latent Dirichlet Allocation (LDA) Untuk Mengetahui Topik Pembicaraan Warganet Twitter Tentang Omnibus Law," Universitas Islam Negeri Syarif Hidayatullah, Jakarta, 2021.
7. H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li and L. Zhao, "Latent Dirichlet Allocation (LDA) and Topic Modeling: models, applications, a survey," Multimedia Tools and Applications, vol. 78(11), pp. 15169-15211, 2019.
8. B. Zhao, Encyclopedia of Big Data, Viginia: Springer, 2017.
9. E. K. Putri dan T. Setiadi, "Penerapan Text Mining Pada Sistem Klasifikasi Email Spam Menggunakan Naive Bayes," Jurnal Sarjana Teknik Informatika, vol. 2, no. 3, 2014.
10. F. Fathonah dan A. Herliana, "Penerapan Text Mining Analisis Sentimen Mengenai Vaksin

Covid-19 Menggunakan Metode Naïve Bayes," Jurnal Sains dan Informatika, vol. 7, no. 2, 2021.
11. D. M. Blei, A. Y. Ng dan M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research 3, 2003.
12. M. Turland, php| architect's Guide to Web, 1 ed., Toronto: Marco Tabini, 2010.
13. D. D. Ayani, H. S. Pratiwi and H. Muhardi, "Implementasi Web Scraping Untuk Pengambilan Data Pada Situs Marketplace," Jurnal Sistem dan Teknologi Informasi, vol. 7(4), pp. 2460-3562, 2019.
14. R. R.Deshmukh and V. Wangikar, "Data Cleaning: Current Approches and Issues," Aurangabad, 2011.
15. S. Madya, Metodologi Pengajaran Bahasa dari Era Prametode Sampai Era Pascametode, Yogyakarta: UNY Press, 2013.
16. L. Gerald Andrew, "pLSA and LDA.".
17. L. . H. Anaya, "Comparing Latent Dirichlet Allocation and Latent Semantic Analysis as Classifiers," UNIVERSITY OF NORTH TEXAS, 2011.
18. T. Cvitanic, B. Lee, H. Ik Song, Katherine Fu, and D. Rosen, "LDA v. LSA: A Comparison of Two Computational Text Analysis Tools for the Functional Categorization of Patents," Georgia Institute of Technology.
19. J. Xu, "Topic Modeling with LSA, PSLA, LDA & lda2Vec ," *NanoNets*, Jul. 01, 2021.