Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 924−933
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10242

Pengembangan Model Deteksi Isu Publik Berbasis Latent Dirichlet Allocation Dengan Pendekatan Tren Waktu dan Analisis Sentimen pada Berita Online Nasional

Dhimas Bagus Prasetyo, Indah Susilawati

Faskultas Teknologi Informasi, Prodi Informatika, Universitas Mercu Buana Yogyakarta, Yogyakarta, Indonesia
Jl.Ring Road Utara, Ngropoh, Condongcatur, Kecamatan Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta, Indonesia
Email: [1,*]221110014@student.mercubuana-yogya.ac.id, [2]indah@mercubuana-yogya.ac.id
Email Penulis Korespondensi: 221110014@student.mercubuana-yogya.ac.id
Submitted: **10/06/2026**; Accepted: **04/07/2026**; Published: **05/07/2026**

Abstrak−Perkembangan media digital dan berita online di Indonesia menghasilkan volume informasi yang sangat besar dan terus bertambah setiap hari. Kondisi ini menimbulkan kesulitan dalam mengidentifikasi isu publik secara cepat dan akurat karena proses pemantauan berita secara manual membutuhkan waktu, tenaga, dan sumber daya yang besar. Selain itu, banyaknya sumber berita dengan fokus pemberitaan yang berbeda menyebabkan informasi tersebar dan sulit dianalisis secara menyeluruh. Oleh karena itu, diperlukan pendekatan otomatis untuk mendeteksi dan memantau isu publik dari data berita online dalam jumlah besar. Penelitian ini bertujuan membangun model deteksi isu publik pada berita online nasional menggunakan metode Latent Dirichlet Allocation (LDA). Data penelitian diperoleh melalui web scraping dari CNBC Indonesia, Detik.com, Kompas.com, dan Liputan6.com selama Januari–Desember 2025 sebanyak 149.335 judul berita, yang setelah preprocessing menghasilkan 146.557 data bersih. Pemodelan topik dilakukan menggunakan LDA, kemudian dianalisis melalui tren waktu, spike detection, perbandingan media, dan analisis sentimen berbasis kamus InSet. Hasil penelitian menunjukkan bahwa model LDA berhasil mengidentifikasi 16 topik utama yang merepresentasikan berbagai isu publik. Analisis menunjukkan adanya perbedaan fokus pemberitaan antar media, beberapa lonjakan isu pada periode tertentu, serta dominasi sentimen negatif pada sebagian besar topik. Temuan ini menunjukkan bahwa pendekatan yang diusulkan mampu mendukung pemantauan isu publik secara otomatis dan terstruktur.

Kata Kunci: Latent Dirichlet Allocation; Deteksi Isu Publik; Analisis Sentimen; Berita Online; Topic Modeling.

Abstract−The growth of digital media and online news in Indonesia has generated a massive volume of information that continues to expand daily. This situation makes it difficult to identify public issues quickly and accurately, as manual news monitoring requires significant time, effort, and resources. Furthermore, the multitude of news sources with varying editorial focuses results in fragmented information that is challenging to analyze comprehensively. Consequently, an automated approach is needed to detect and monitor public issues within large datasets of online news. This study aims to develop a public issue detection model for national online news using the Latent Dirichlet Allocation (LDA) method. Research data was obtained via web scraping from CNBC Indonesia, Detik.com, Kompas.com, and Liputan6.com between January and December 2025, yielding 149,335 news headlines; after preprocessing, 146,557 clean data points remained. Topic modeling was performed using LDA, followed by analysis involving temporal trends, spike detection, media comparisons, and sentiment analysis based on the InSet dictionary. The results demonstrate that the LDA model successfully identified 16 key topics representing various public issues. The analysis revealed differences in reporting focus across media outlets, spikes in specific issues during certain periods, and a predominance of negative sentiment across most topics. These findings indicate that the proposed approach is capable of supporting the automated and structured monitoring of public issues.

Keywords: Latent Dirichlet Allocation; Public Issue Detection; Sentiment Analysis; Online News; Topic Modeling.

1. PENDAHULUAN

Perkembangan teknologi informasi telah meningkatkan produksi dan distribusi berita melalui berbagai portal berita daring. Setiap hari, ribuan berita dipublikasikan dengan beragam isu yang mencerminkan kondisi sosial, ekonomi, politik, teknologi, kesehatan, maupun keamanan masyarakat. Besarnya volume informasi tersebut menyulitkan proses identifikasi isu secara manual sehingga diperlukan pendekatan otomatis yang mampu mengelompokkan dan menganalisis informasi secara efisien[1][2]. Selain jumlahnya yang terus meningkat, kecepatan penyebaran informasi juga menyebabkan perubahan isu berlangsung sangat dinamis sehingga diperlukan metode yang mampu memantau perkembangan isu secara berkelanjutan dan menghasilkan informasi yang mudah dipahami oleh pengambil keputusan maupun masyarakat.

Salah satu metode yang banyak digunakan untuk mengidentifikasi topik pada data teks adalah Latent Dirichlet Allocation (LDA)[3][4]. Metode ini mampu menemukan pola topik tersembunyi berdasarkan distribusi probabilitas kata dalam dokumen[5]. Berbagai penelitian menunjukkan bahwa LDA efektif diterapkan pada dokumen akademik[6]. media sosial[7], [8]. ulasan pengguna[9][10], dan berita daring[1], [11][1] Selain digunakan secara mandiri, LDA juga dikombinasikan dengan metode lain seperti BiLSTM[12], LSTM[13], TF-IDF, dan K-Means[14] untuk meningkatkan kualitas analisis topik maupun sentimen. Kemampuan LDA dalam memodelkan kumpulan dokumen berukuran besar menjadikannya salah satu metode yang banyak dipilih untuk mengeksplorasi informasi tersembunyi dari data teks yang terus bertambah setiap waktu.

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 924−933
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10242

Selain LDA, pemodelan topik juga dapat dilakukan menggunakan metode seperti Non-negative Matrix Factorization (NMF), TF-IDF, maupun BERTopic. TF-IDF lebih berfokus pada pembobotan kata sehingga tidak secara langsung membentuk topik, sedangkan NMF menghasilkan topik melalui proses dekomposisi matriks dan BERTopic memanfaatkan representasi embedding untuk menangkap hubungan semantik yang lebih kompleks. Namun demikian, penelitian ini memilih LDA karena mampu menghasilkan topik yang mudah diinterpretasikan, stabil dalam menganalisis kumpulan dokumen berukuran besar, serta telah banyak diterapkan pada penelitian pemodelan topik berita sehingga sesuai dengan karakteristik data yang digunakan dalam penelitian ini [4][5][9]

Meskipun demikian, sebagian besar penelitian terdahulu masih berfokus pada pemodelan topik atau analisis sentimen secara terpisah [15][13]. Penelitian berbasis media sosial umumnya menggunakan data yang memiliki tingkat kebisingan tinggi [7][8] sedangkan penelitian pada berita daring umumnya hanya mengidentifikasi topik tanpa menganalisis perkembangan isu dari waktu ke waktu maupun kecenderungan sentimen yang menyertainya [1][2] Selain itu, sebagian besar penelitian masih menggunakan sumber data yang terbatas pada domain atau platform tertentu [9][11][10].

Berdasarkan kondisi tersebut, masih terdapat peluang untuk mengembangkan pendekatan yang mengintegrasikan pemodelan topik, analisis tren waktu, dan analisis sentimen dalam satu kerangka analisis. Integrasi tersebut diharapkan mampu memberikan gambaran yang lebih komprehensif mengenai isu publik, perkembangan isu, serta kecenderungan opini yang muncul di masyarakat. Penggunaan data dari beberapa portal berita nasional juga memungkinkan representasi isu publik yang lebih luas dibandingkan penelitian yang hanya memanfaatkan satu sumber data [1][2].

Dengan demikian, penelitian ini tidak hanya berfokus pada identifikasi topik sebagaimana penelitian terdahulu, tetapi juga mengintegrasikan analisis tren waktu, spike detection, analisis sentimen berbasis InSet Lexicon, dan perbandingan antar media. Pendekatan tersebut diharapkan mampu memberikan gambaran yang lebih komprehensif mengenai perkembangan isu publik berdasarkan data berita daring nasional [5][15][16].

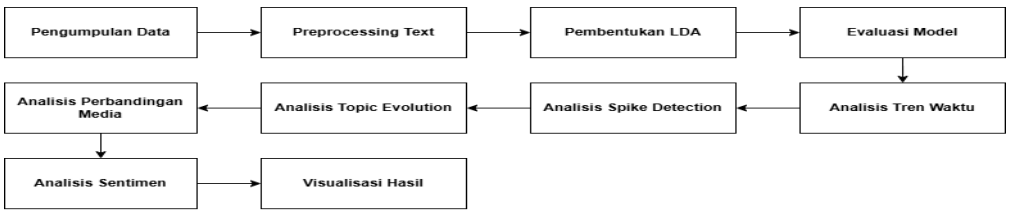
Penelitian ini menerapkan metode Latent Dirichlet Allocation (LDA) untuk mengidentifikasi isu publik pada berita nasional yang diperoleh dari beberapa portal berita daring di Indonesia. Data diproses melalui tahapan preprocessing yang meliputi case folding, cleaning, tokenizing, stopword removal, normalisasi, dan stemming [4], [17]. Kualitas model dievaluasi menggunakan coherence score, kemudian hasil pemodelan dianalisis menggunakan tren waktu dan analisis sentimen [5][15]. Penelitian ini bertujuan mengidentifikasi isu publik dominan, menganalisis perkembangan isu [18][19], serta mengetahui kecenderungan sentimen pada setiap topik sehingga dapat memberikan gambaran yang lebih komprehensif mengenai dinamika isu publik di Indonesia [16] [20]. Pendekatan serupa juga telah diterapkan untuk menilai kesiapan pembelajaran tatap muka terbatas melaluianalisis sentimenTwitter menggunakan metode Inset Lexicon dan Levenshtein Distance [21][22].

Kontribusi utama penelitian ini adalah pengembangan model deteksi isu publik yang mengintegrasikan pemodelan topik menggunakan LDA dengan analisis tren waktu, spike detection, analisis sentimen, dan perbandingan antar media pada data berita daring nasional. Integrasi tersebut diharapkan mampu menghasilkan informasi yang lebih komprehensif mengenai dinamika isu publik serta memberikan dukungan bagi proses pemantauan isu dan pengambilan keputusan berbasis data [4][5][15].

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini diawali dengan proses pengumpulan data menggunakan teknik web scraping untuk memperoleh data berita dari beberapa portal berita daring [1]. Selanjutnya dilakukan preprocessing teks yang meliputi case folding, cleaning, tokenizing, stopword removal, normalization, dan stemming untuk menghasilkan data yang lebih bersih dan seragam sebelum dilakukan pemodelan topik [4][17], Data hasil preprocessing kemudian dianalisis menggunakan metode Latent Dirichlet Allocation (LDA) guna mengidentifikasi topik-topik utama dalam kumpulan dokumen [4][7] . Kualitas model dievaluasi menggunakan coherence score untuk memperoleh jumlah topik yang optimal [5][15]. Hasil pemodelan selanjutnya dianalisis melalui analisis tren waktu, spike detection, topic evolution, analisis sentimen berbasis InSet Lexicon, serta perbandingan antar media sehingga mampu memberikan gambaran yang lebih komprehensif mengenai dinamika isu publik [18][23]. Tahap akhir dilakukan visualisasi hasil untuk memudahkan interpretasi dan penyajian temuan penelitian. Alur penelitian secara lengkap ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 924−933
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10242

2.2 Pengumpulan Data

Data penelitian diperoleh melalui teknik web scraping menggunakan Python pada beberapa portal berita daring. Metode ini memanfaatkan struktur HTML untuk mengekstrak informasi secara otomatis. Informasi yang dikumpulkan meliputi judul berita, tanggal publikasi, dan sumber berita, yang selanjutnya disimpan dalam format CSV sebagai dataset penelitian [1].

2.3 Preprocessing Text

Preprocessing teks bertujuan membersihkan dan menyeragamkan data sebelum analisis. Tahap ini meliputi case folding, cleaning, tokenizing, stopword removal, normalization, dan stemming untuk menghilangkan elemen yang tidak relevan serta mengubah kata ke bentuk yang lebih konsisten [4][17]. Keluaran dari tahap ini berupa teks bersih yang selanjutnya menjadi masukan bagi pemodelan topik menggunakan Latent Dirichlet Allocation (LDA).

2.4 Pemodelan Topik Menggunakan LDA

Pemodelan topik pada penelitian ini menggunakan metode Latent Dirichlet Allocation (LDA), yaitu model probabilistik yang merepresentasikan setiap dokumen sebagai kombinasi beberapa topik dan setiap topik sebagai distribusi kata [4][7]. Implementasi dilakukan menggunakan library Gensim pada Python.

2.4.1 Pembentukan Dictionary dan Corpus

Data hasil preprocessing dikonversi ke bentuk numerik melalui pembentukan dictionary dan corpus. Dictionary merupakan kumpulan kata unik yang terdapat pada seluruh dokumen, sedangkan corpus dibentuk menggunakan representasi Bag-of-Words (BoW) melalui fungsi doc2bow() pada Gensim [4].

$$V = \{w_1, w_2, w_3, \dots, w_n\} \quad (1)$$

Pada persamaan tersebut, (V) merupakan vocabulary yang berisi seluruh kata unik dalam corpus, (w_i) menyatakan kata unik ke-(i), dan (n) menunjukkan jumlah total kata unik yang terdapat pada corpus.

$$d = \{(id_1, f_1), (id_2, f_2), \dots, (id_n, f_n)\} \quad (2)$$

Pada persamaan tersebut, (d) merupakan representasi dokumen dalam bentuk Bag-of-Words, (id_i) menunjukkan ID atau indeks kata pada dictionary, sedangkan (f_i) menyatakan frekuensi kemunculan kata tersebut dalam dokumen.

Untuk mempertahankan makna frasa yang sering muncul, penelitian ini juga menerapkan pembentukan N-gram, bigram dan trigram sehingga kata-kata yang memiliki keterkaitan kuat dapat diperlakukan sebagai satu kesatuan [7].

2.4.2 Pembentukan LDA

Model LDA dibangun menggunakan fungsi LdaModel() pada Gensim dengan parameter utama num_topics, passes, alpha, dan eta. Nilai num_topics ditentukan berdasarkan hasil evaluasi model, sedangkan parameter lainnya digunakan untuk mengoptimalkan proses pembelajaran [5].

$$LDA = (D, T, W) \quad (3)$$

Pada persamaan tersebut, (D) merupakan kumpulan dokumen, (T) merupakan kumpulan topik, dan (W) merupakan kumpulan kata dalam corpus. Metode LDA memodelkan distribusi probabilitas topik pada dokumen ((P(T|D))) dan distribusi probabilitas kata pada topik ((P(W|T))). Hasil proses pemodelan berupa distribusi topik pada setiap dokumen dan kumpulan kata yang merepresentasikan masing-masing topik.

2.4.3 Evaluasi Model

Kualitas model dievaluasi menggunakan coherence score dan perplexity. Coherence score mengukur tingkat keterkaitan antar kata dalam suatu topik, sedangkan perplexity digunakan untuk menilai kemampuan model dalam merepresentasikan data [5][15] Model yang baik memiliki nilai coherence yang tinggi dan perplexity yang rendah.

$$\text{Coherence} = \sum\{i < j\}\text{Score}(w_i, w_j) \quad (4)$$

Pada persamaan tersebut, (w_i) dan (w_j) merupakan pasangan kata dalam suatu topik, sedangkan (Score(w_i,w_j)) menyatakan tingkat keterkaitan antara kedua kata tersebut. Nilai skor yang lebih tinggi menunjukkan hubungan yang lebih kuat antar kata dalam merepresentasikan topik yang terbentuk.

$$\text{Perplexity}(D) = \backslash\exp(-\backslash\text{frac}\{\sum\backslash\log P(w_d)\}\{N\}) \quad (5)$$

Pada persamaan tersebut, (P(w_d)) merepresentasikan probabilitas suatu kata dalam dokumen, sedangkan (N) menyatakan jumlah total kata yang terdapat pada corpus. Probabilitas tersebut digunakan untuk menunjukkan peluang kemunculan suatu kata berdasarkan distribusi kata dalam keseluruhan data teks yang dianalisis.

2.5 Analisis Tren Waktu dan Identifikasi Lonjakan

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 924−933
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10242

Analisis tren waktu dilakukan untuk mengamati perubahan kemunculan topik pada berbagai periode publikasi berita. Data dikelompokkan ke dalam interval bulanan, mingguan, dan harian, kemudian setiap dokumen diberi label topik dominan berdasarkan probabilitas tertinggi yang dihasilkan oleh model LDA [18][19]. Selanjutnya, frekuensi kemunculan setiap topik dihitung dan dikonversi menjadi proporsi untuk menggambarkan dinamika isu dari waktu ke waktu. Hasil analisis divisualisasikan menggunakan line chart guna memudahkan identifikasi pola dan perubahan tren topik. Proporsi kemunculan topik pada setiap periode dihitung menggunakan persamaan berikut:

$$P_{\{t,k\}} = \backslash frac\{F_{\{t,k\}}\}\{\sum\{k = 1\}^{\{K\}F}_{\{t,k\}}\} \times 100\backslash\% \quad (6)$$

Pada persamaan tersebut, (P_{t,k}) menyatakan proporsi topik ke-(k) pada periode (t), (F_{t,k}) adalah jumlah dokumen pada topik ke-(k), dan (K) merupakan jumlah total topik. Hasil analisis divisualisasikan menggunakan line chart untuk mengidentifikasi pola dan perubahan tren topik.

2.6 Deteksi Lonjakan Isu (Spike Detection)nna

Deteksi lonjakan isu dilakukan untuk mengidentifikasi topik yang mengalami peningkatan perhatian secara signifikan dalam periode tertentu. Pada penelitian ini, lonjakan isu diukur menggunakan metode Z-Score terhadap proporsi kemunculan topik pada setiap periode waktu [16][20]. Nilai Z-Score yang tinggi menunjukkan bahwa suatu topik muncul jauh di atas kondisi normal sehingga dapat dianggap sebagai indikasi terjadinya lonjakan isu.

$$Z = \backslash frac\{x - \mu\}\{\sigma\} \quad (7)$$

Pada persamaan tersebut, (Z) menyatakan nilai z-score, (x) adalah proporsi topik pada suatu periode, (\mu) merupakan rata-rata proporsi topik, dan (\sigma) adalah standar deviasi. Persamaan ini digunakan untuk mengidentifikasi lonjakan atau penurunan proporsi topik yang signifikan dibandingkan pola normalnya.

2.7 Analisis Topic Evolution

Analisis topic evolution dilakukan untuk mengidentifikasi perubahan fokus isu publik dari waktu ke waktu berdasarkan topik dominan pada setiap periode [16][20] Topik dominan ditentukan dari frekuensi kemunculan tertinggi pada masing-masing periode, kemudian perubahan topik diamati untuk menggambarkan dinamika isu selama periode penelitian. Hasil analisis divisualisasikan dalam bentuk diagram transisi antar topik.

$$T_t = \backslash arg\backslash max_{\{k\}F_{\{t,k\}}} \quad (8)$$

Pada persamaan tersebut, (T_t) menyatakan topik dominan pada periode (t), sedangkan (F_{t,k}) adalah frekuensi kemunculan topik ke-(k). Operator (\arg\max) digunakan untuk memilih topik dengan frekuensi tertinggi pada setiap periode sehingga perubahan fokus isu dari waktu ke waktu dapat dianalisis.

2.8 Analisis Perbandingan Antar Media (Multi Source Analysis)

Analisis perbandingan antar media dilakukan untuk mengidentifikasi perbedaan distribusi topik pada masing-masing portal berita. Setiap dokumen dikelompokkan berdasarkan sumber berita, kemudian proporsi kemunculan topik dihitung secara terpisah untuk setiap media. Hasil perbandingan divisualisasikan menggunakan grouped bar chart sehingga perbedaan kecenderungan pemberitaan antar media dapat diamati dengan lebih jelas [23][22].

$$P_{\{s,k\}} = \backslash frac\{F_{\{s,k\}}\}\{\sum\{k = 1\}^{\{K\}F}_{\{s,k\}}\} \times 100\backslash\% \quad (9)$$

Pada persamaan tersebut, (P_{s,k}) menyatakan proporsi topik ke-(k) pada sumber berita (s), (F_{s,k}) adalah jumlah dokumen pada topik ke-(k), dan (K) merupakan jumlah total topik. Persamaan ini digunakan untuk membandingkan distribusi topik antar sumber berita.

2.9 Analisis Sentimen

Analisis sentimen dilakukan untuk mengidentifikasi kecenderungan pemberitaan terhadap isu yang telah diperoleh dari pemodelan topik. Penelitian ini menggunakan pendekatan berbasis kamus (lexicon-based approach) dengan memanfaatkan kamus InSet yang dirancang untuk Bahasa Indonesia [21]. Setiap kata pada dokumen dicocokkan dengan kamus InSet untuk memperoleh nilai polaritas, kemudian seluruh nilai polaritas dijumlahkan untuk menghasilkan skor sentimen dokumen. Skor sentimen dihitung menggunakan persamaan berikut:

$$S_d = \sum\{i = 1\}^{\{n\}w}_i \quad (10)$$

Pada persamaan tersebut, (S_d) menyatakan skor sentimen dokumen, (w_i) adalah nilai polaritas kata ke-(i), dan (n) merupakan jumlah kata dalam dokumen. Persamaan ini digunakan untuk menentukan kecenderungan sentimen dokumen berdasarkan akumulasi polaritas setiap kata.

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 924−933
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10242

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengumpulan Data

Sebanyak 149.335 judul berita berhasil diperoleh melalui teknik web scraping dari empat media daring nasional selama Januari–Desember 2025. Distribusi data dari masing-masing portal ditunjukkan pada Tabel 1.

Tabel 1. Jumlah Data Hasil Web Scraping per Portal

No	Sumber Portal	Jumlah Data
1	CNBC	3.642
2	Detik	36.935
3	Kompas	77.264
4	Liputan6	31.494
Total		149.335

Berdasarkan Tabel 1. kompas.com menjadi kontributor terbesar dalam korpus penelitian, diikuti oleh Detik.com dan Liputan6.com, sedangkan CNBC Indonesia memberikan kontribusi paling sedikit. Variasi sumber tersebut memungkinkan cakupan isu yang lebih beragam sehingga dapat memberikan gambaran yang lebih representatif terhadap dinamika pemberitaan nasional. Selanjutnya, korpus dibagi ke dalam tiga rentang waktu, yaitu Januari–April, Mei–Agustus, dan September–Desember 2025 untuk mendukung analisis perkembangan topik sepanjang periode penelitian.

3.2 Hasil Preprocessing Text

Tahap preprocessing dilakukan untuk membersihkan dan menyeragamkan data teks hasil web scraping sebelum digunakan pada proses pemodelan topik. Tahapan yang diterapkan meliputi case folding, cleaning, tokenizing, stopword removal, normalisasi, dan stemming untuk mengurangi noise serta meningkatkan kualitas data. Contoh hasil preprocessing ditunjukkan pada Tabel 2

Tabel 2. Hasil Preprocessing Text

Sebelum Preprocessing	Sesudah Preprocessing
Dunia Terancam Polusi Plastik Tahun 2040	dunia ancam polusi plastik tingkat

Berdasarkan Tabel 2. proses preprocessing, jumlah data berkurang dari 149.335 menjadi 146.557 judul berita akibat penghapusan data duplikat, data kosong, dan data yang tidak valid. Seperti ditunjukkan pada Tabel 5, teks berhasil dibersihkan dan diubah ke bentuk kata dasar sehingga menghasilkan data yang lebih seragam dan siap digunakan pada proses pemodelan topik menggunakan metode LDA.

3.3 Hasil Pembentukan Topik Menggunakan LDA

Setelah tahap preprocessing selesai dilakukan, data berita diproses menggunakan metode Latent Dirichlet Allocation (LDA) untuk mengidentifikasi topik-topik utama pada berita online nasional tahun 2025. Dataset hasil preprocessing dikonversi ke dalam bentuk representasi dokumen melalui pembentukan dictionary dan corpus sehingga dapat digunakan pada proses pelatihan model LDA. Proses tersebut menghasilkan 149.277 dokumen dengan 19.568 kata unik (vocabulary) yang digunakan dalam pemodelan topik. Hasil pembentukan topik ditunjukkan pada Tabel 3.

Tabel 3. Hasil Pembentukan Topik LDA

Topik	Interpretasi Bidang	Topik	Interpretasi Bidang
Topik 1	Politik	Topik 9	Bencana
Topik 2	Energi	Topik 10	Olahraga
Topik 3	Pariwisata	Topik 11	Hiburan
Topik 4	Hukum	Topik 12	Ekonomi
Topik 5	Pendidikan	Topik 13	Internasional
Topik 6	Kesehatan	Topik 14	Sosial
Topik 7	Teknologi	Topik 15	Transportasi
Topik 8	Lingkungan	Topik 16	Pertahanan

Berdasarkan Tabel 3, Model LDA menghasilkan 16 topik utama yang merepresentasikan berbagai isu publik nasional. Topik yang terbentuk mencakup bidang politik, energi, pariwisata, hukum, pendidikan, kesehatan, teknologi, lingkungan, bencana, olahraga, hiburan, ekonomi, internasional, sosial, transportasi, dan pertahanan. Hasil ini menunjukkan bahwa metode LDA mampu mengelompokkan berita ke dalam tema-tema yang relevan sehingga dapat digunakan untuk analisis distribusi topik, tren waktu, dan sentimen pada tahap selanjutnya.

3.4 Hasil Evaluasi Model

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 924−933
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10242

Evaluasi model dilakukan untuk mengukur kualitas topik yang dihasilkan oleh metode Latent Dirichlet Allocation (LDA). Pada penelitian ini, evaluasi menggunakan dua metrik utama, yaitu Perplexity Score dan Coherence Score. Perplexity digunakan untuk mengukur kemampuan model dalam merepresentasikan dokumen, sedangkan coherence digunakan untuk menilai keterkaitan semantik antar kata dalam suatu topik. Hasil evaluasi model ditunjukkan pada Tabel 4.

Tabel 4. Hasil Evaluasi Model

Parameter	Hasil
Jumlah Dokumen	149.277
Jumlah Vocabulary	19.568
Jumlah Topik	16
Perplexity Score	-15.5903
Coherence Score	0.3071

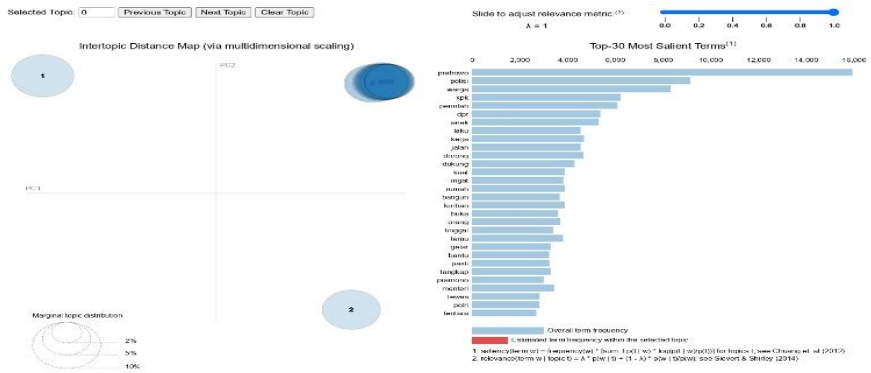
Berdasarkan Tabel 4, model LDA dibangun menggunakan 149.277 dokumen dengan 19.568 kata unik dan menghasilkan 16 topik utama. Hasil evaluasi menunjukkan nilai Perplexity Score sebesar -15,5903 dan Coherence Score sebesar 0,3071. Nilai tersebut menunjukkan bahwa model mampu menghasilkan topik yang cukup representatif dan memiliki keterkaitan kata yang memadai untuk mendukung analisis isu publik. Oleh karena itu, model dengan 16 topik digunakan pada tahap analisis topik, tren waktu, dan sentimen pada penelitian ini.

3.5 Visualisasi Topik LDA

Untuk memberikan gambaran mengenai distribusi kata-kata yang paling dominan pada keseluruhan dataset berita, dilakukan visualisasi menggunakan WordCloud sebagaimana ditunjukkan pada Gambar 2.

Gambar 2. Wordcloud Global Topik LDA

WordCloud menampilkan kata-kata yang paling sering muncul pada keseluruhan dataset berita. Kata seperti polisi, warga, prabowo, anak, korban, dan kerja memiliki ukuran lebih besar karena frekuensi kemunculannya lebih tinggi dibandingkan kata lainnya. Kemunculan kata-kata tersebut menunjukkan bahwa isu pemerintahan, keamanan, sosial, dan masyarakat menjadi tema yang dominan dalam pemberitaan selama periode penelitian. Untuk mengetahui hubungan dan tingkat kedekatan antar topik yang dihasilkan oleh model LDA, dilakukan visualisasi menggunakan PyLDAVis sebagaimana ditunjukkan pada Gambar 3.



Gambar 3. Visualisasi Intertopic Distance Map Hasil Pemodelan LDA Menggunakan PyLDAvis

Visualisasi PyLDAVis menunjukkan hubungan antar topik yang dihasilkan oleh model LDA. Sebagian besar topik membentuk kelompok yang terpisah sehingga memiliki karakteristik yang berbeda satu sama lain. Kondisi ini mengindikasikan bahwa model mampu mengidentifikasi tema berita secara cukup jelas dengan tingkat tumpang tindih yang relatif rendah antar topik.

3.6 Hasil Analisis Tren Waktu

Analisis tren waktu dilakukan untuk mengamati perubahan distribusi topik pemberitaan sepanjang tahun 2025 berdasarkan hasil pemodelan LDA. Dari total 149.277 dokumen hasil pemodelan, sebanyak 72.027 dokumen

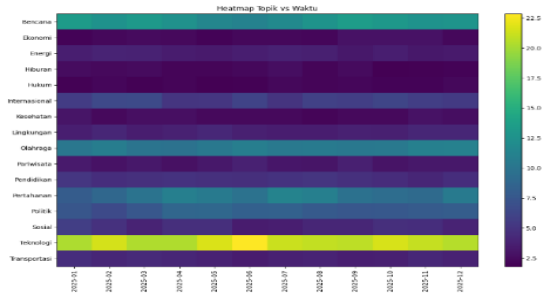
Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 924−933
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10242

memiliki informasi tanggal yang valid dan digunakan dalam analisis bulanan. Distribusi valid per bulan ditunjukkan pada Tabel 5.

Tabel 5. Dokumen Valid Per Bulan

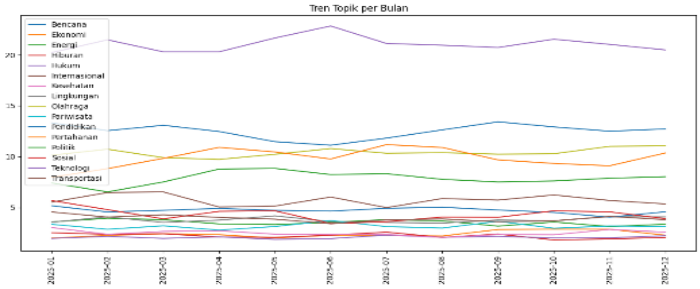
Bulan	Total Dokumen	Bulan	Total Dokumen
2025-01	6.035	2025-07	6.117
2025-02	5.986	2025-08	6.277
2025-03	6.033	2025-09	5.938
2025-04	5.580	2025-10	6.022
2025-05	6.460	2025-11	5.558
2025-06	6.017	2025-12	6.004
Total			**72.027**

Berdasarkan Tabel 5, sebanyak 72.027 dokumen valid digunakan dalam analisis tren waktu. Jumlah dokumen yang tersedia pada setiap bulan relatif merata, dengan rentang 5.558–6.460 dokumen. Kondisi ini menunjukkan bahwa data yang digunakan memiliki cakupan yang konsisten sepanjang periode pengamatan, sehingga mampu merepresentasikan pola pemberitaan secara memadai. Sebaran data yang relatif seimbang juga mendukung analisis perubahan proporsi topik dari waktu ke waktu tanpa dipengaruhi oleh perbedaan jumlah dokumen yang terlalu besar antarbulan. Untuk mengetahui distribusi intensitas proporsi setiap topik pada masing-masing bulan, dilakukan visualisasi menggunakan heatmap sebagaimana ditunjukkan pada Gambar 4.



Gambar 4. Heatmap vs Topik Waktu

Heatmap digunakan untuk menunjukkan intensitas proporsi setiap topik pada setiap bulan. Warna yang lebih terang menunjukkan proporsi yang lebih tinggi. Visualisasi ini memperlihatkan dominasi topik Teknologi sepanjang periode pengamatan, diikuti oleh topik Bencana dan Olahraga. Sebagian besar topik menunjukkan pola distribusi yang relatif konsisten tanpa perubahan yang signifikan. Untuk mengamati perubahan proporsi setiap topik dari waktu ke waktu, dilakukan visualisasi tren bulanan sebagaimana ditunjukkan pada Gambar 5.



Gambar 5. Tren Topik Per Bulan

Grafik tren menunjukkan bahwa distribusi proporsi topik pemberitaan selama tahun 2025 relatif stabil. Topik Teknologi tetap mendominasi sepanjang tahun, sedangkan Topik Pertahanan menunjukkan tren meningkat (slope 0,0765) dan Topik Sosial mengalami tren menurun (slope -0,0720). Sementara itu, sebagian besar topik memiliki nilai slope yang mendekati nol, mengindikasikan tidak adanya perubahan distribusi topik yang signifikan selama periode pengamatan.

3.7 Hasil Analisis Spike Detection

Hasil spike detection mengidentifikasi tujuh topik yang mengalami lonjakan signifikan sepanjang tahun 2025 (Z-Score ≥ 1,8). Lonjakan tertinggi terjadi pada topik Hiburan (Maret 2025; 2,58), diikuti Pariwisata (Juni 2025; 2,31) dan Ekonomi (November 2025; 2,17), serta lonjakan pada topik Sosial, Teknologi, Internasional, dan Kesehatan. Temuan ini menunjukkan bahwa meskipun sebagian besar topik relatif stabil, beberapa isu mengalami peningkatan perhatian media pada periode tertentu, sehingga pendekatan LDA yang dikombinasikan dengan analisis temporal mampu mengidentifikasi kemunculan isu yang memperoleh perhatian publik secara signifikan.

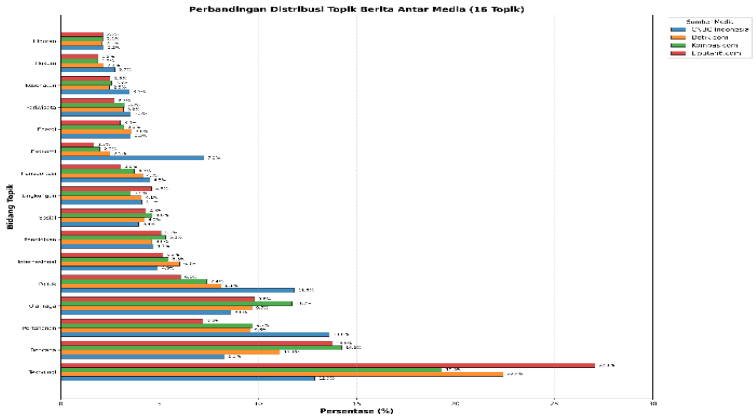
Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 924−933
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10242

3.8 Hasil Analisis Topic Evolution

Analisis *topic evolution* menunjukkan bahwa tidak terjadi pergeseran topik dominan yang signifikan sepanjang tahun 2025. Topik Teknologi, Bencana, dan Olahraga tetap menjadi fokus utama pemberitaan, sementara perubahan hanya terlihat pada fluktuasi intensitas beberapa topik, seperti meningkatnya Topik Pertahanan dan menurunnya Topik Sosial. Secara keseluruhan, evolusi topik berlangsung secara gradual dan relatif stabil.

3.9 Hasil Analisis Perbandingan Media

Analisis perbandingan media dilakukan untuk mengidentifikasi perbedaan fokus pemberitaan pada CNBC Indonesia, Detik.com, Kompas.com, dan Liputan6.com. Distribusi proporsi topik pada masing-masing portal berita ditampilkan pada Gambar 6.

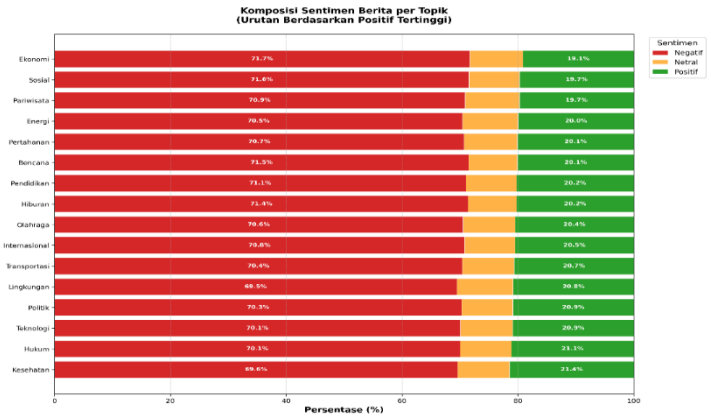


Gambar 6. Perbandinga Distribusi Topik Berita Antar Media

Terlihat bahwa Liputan6.com, Detik.com, dan Kompas.com didominasi oleh topik Teknologi, sedangkan CNBC Indonesia memiliki kecenderungan lebih besar pada topik Pertahanan. Perbedaan tersebut mencerminkan karakteristik dan segmentasi konten yang berbeda pada setiap portal berita.

3.10 Hasil Analisis Sentimen

Distribusi kategori positif, netral, dan negatif pada setiap bidang hasil pemodelan topik disajikan pada Gambar 7.



Gambar 7. Sentimen Berita Per Topik

Gambar 7 memperlihatkan dominasi kategori negatif pada seluruh bidang dengan proporsi sekitar 70%, sedangkan kategori positif berada pada kisaran 19–21%. Proporsi positif tertinggi ditemukan pada bidang Kesehatan, sementara Ekonomi, Sosial, dan Bencana memiliki kecenderungan negatif yang lebih besar dibandingkan bidang lainnya. Kondisi ini mengindikasikan bahwa pemberitaan media online nasional selama tahun 2025 lebih banyak berfokus pada isu yang berkaitan dengan permasalahan, risiko, dan tantangan publik. Di sisi lain, Kesehatan dan Teknologi memperlihatkan kecenderungan yang lebih positif, yang menunjukkan adanya perhatian media terhadap perkembangan dan capaian pada kedua bidang tersebut.

4. KESIMPULAN

Penelitian ini membuktikan bahwa penerapan metode Latent Dirichlet Allocation (LDA) mampu mendukung proses identifikasi isu publik dari data berita online nasional secara sistematis dan efisien. Melalui tahapan pengumpulan data, pembersihan teks, pemodelan topik, analisis tren waktu, perbandingan media, serta analisis

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 924−933
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10242

sentimen, diperoleh gambaran yang lebih terstruktur mengenai dinamika isu yang berkembang sepanjang tahun 2025. Hasil yang diperoleh menunjukkan bahwa berita dapat dikelompokkan ke dalam sejumlah topik yang merepresentasikan berbagai bidang pembahasan sehingga memudahkan proses pemantauan informasi dalam skala besar. Analisis yang dilakukan juga memperlihatkan adanya variasi fokus pemberitaan pada setiap portal berita, yang mencerminkan karakteristik dan prioritas masing-masing media dalam menyampaikan informasi kepada publik. Selain itu, pengamatan terhadap perubahan intensitas topik dari waktu ke waktu mampu memberikan indikasi mengenai meningkatnya perhatian terhadap isu tertentu pada periode tertentu. Hasil analisis sentimen menunjukkan bahwa pemberitaan cenderung didominasi oleh sentimen negatif, yang mengindikasikan banyaknya isu yang berkaitan dengan tantangan, permasalahan, atau peristiwa yang mendapat perhatian luas dari masyarakat. Secara keseluruhan, pendekatan yang digunakan mampu menghasilkan informasi yang lebih mudah dipahami dan dapat dimanfaatkan sebagai pendukung pengambilan keputusan, pemantauan opini publik, serta evaluasi perkembangan isu berdasarkan data berita yang tersedia secara berkelanjutan. Temuan penelitian ini juga diharapkan dapat menjadi referensi bagi penelitian selanjutnya dalam mengembangkan metode deteksi isu publik yang lebih akurat, adaptif, dan mampu memanfaatkan sumber data yang lebih beragam sehingga menghasilkan informasi yang semakin komprehensif.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada dosen pembimbing yang telah memberikan arahan, masukan, dan dukungan selama proses penelitian ini. Terima kasih juga kepada Program Studi Teknik Informatika Universitas Mercu Buana Yogyakarta yang telah memberikan fasilitas dan kesempatan sehingga penelitian dapat terlaksana dengan baik. Selain itu, penulis menyampaikan apresiasi kepada keluarga, teman, dan semua pihak yang telah memberikan bantuan serta semangat selama proses penyusunan penelitian hingga artikel ini dapat diselesaikan.

REFERENCES

[1] Welianus Yohanes Yehuda Mramra and Yessica Nataliani, “Penentuan bidang unggulan akademik universitas melalui metode topik Linear Discriminant Analysis (LDA),” IT-Explore: Jurnal Penerapan Teknologi Informasi dan Komunikasi, vol. 4, no. 1, pp. 1–13, Feb. 2025, doi: https://doi.org/10.24246/itexplore.v4i1.2025.pp93-105.

[2] Ervan Yogi Arifianto, “Analisis Topik Data Tindak Kriminal pada Media Sosial Twitter Menggunakan Metode LDA (Latent Dirichlet Allocation),” pp. 1–204, 2020, Accessed: Apr. 02, 2026. [Online]. Available: https://dspace.uii.ac.id/bitstream/handle/123456789/28540/14917209%20Ervan%20Yogi%20Arifianto.pdf?sequence=1

[3] Siti Mutmainah, Dhomas Hatta Fudholi, and Syarif Hidayat, “Analisis Sentimen dan Pemodelan Topik Aplikasi Telemedicine Pada Google Play Menggunakan BiLSTM dan LDA,” JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 7, no. 1, pp. 1–12, Jan. 2023, doi: 10.30865/mib.v7i1.5486.

[4] Saikin, Dosiansyah Fadli, Jihadul Akbar, and Hairul Fahmi, “TOPIC MODELING JUDUL PENELITIAN MENGGUNAKAN METODE LATENT DIRICHLET ALLOCATION (LDA) DAN FAKTORISASI MATRIKS NON-NEGATIF (NMF),” vol. 9, no. Vol. 9 No. 2 (2025): JATI Vol. 9 No. 2, pp. 1–7, Apr. 2025, doi: https://doi.org/10.36040/jati.v9i2.12998.

[5] Chairullah Naury, Dhomas Hatta Fudhol, and Ahmad Fathan Hidayatullah, “Topic Modelling pada Sentimen Terhadap Headline Berita Online Berbahasa Indonesia Menggunakan LDA dan LSTM,” JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 5, no. 1, pp. 1–10, Jan. 2021, doi: 10.30865/mib.v5i1.2556.

[6] Al Farhan Irhanusa Ridat and Auliya Rahman Isnain, “PEMODELAN TOPIK DENGAN LDA UNTUK MEMAHAMI PERSEPSI PUBLIK TERHADAP APPLE VISION PRO MELALUI ANALISIS BERITA TWITTER,” JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), vol. 10, no. 4, pp. 1–14, Dec. 2025, doi: 10.29100/jipi.v10i4.6500.

[7] Evi Puspita, Diqy Fakhrun Shiddieq, and Fikri Fahru Roj, “Pemodelan Topik pada Media Berita Online Menggunakan Latent Dirichlet Allocation (Studi Kasus Merek Somethinc),” MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 4, no. 2, pp. 481–489, Apr. 2024, doi: https://doi.org/10.57152/malcom.v4i2.1204.

[8] Fajar Maula Hidayat, Cahyadi, Hafidz Sanjaya, Dwi Purnomo, and Heri Wiranto, “PEMODELAN TOPIK BERITA NASIONAL INDONESIA MENGGUNAKAN LATENT DIRICHLET ALLOCATION,” INFOTECH journal, vol. 12, no. 1, pp. 1–8, Jan. 2026, doi: https://doi.org/10.31949/infotech.v12i1.16926.

[9] Yayang Matiara, Junaidi, and Iman Setiawan, “Pemodelan Topik pada Judul Berita Online Detikcom Menggunakan Latent Dirichlet Allocation,” Estimasi: Journal of Statistics and Its Application, vol. 4, no. Vol. 4, No. 1, Januari, 2023 : Estimasi, pp. 53–63, Jan. 2023, doi: https://doi.org/10.20956/ejsa.vi.24843.

[10] Rimbun Siringoringo, Jamaluddin, and Resianta Perangin-Angi, “PEMODELAN TOPIK BERITA MENGGUNAKAN LATENT DIRICHLET ALLOCATION DAN K-MEANS CLUSTERING Universitas Methodist Indonesia,” Jurnal Informatika Kaputama(JIK), vol. 4, no. 2, pp. 216–222, Jul. 2020, Accessed: Apr. 03, 2026. [Online]. Available: https://jurnal.kaputama.ac.id/index.php/JIK/article/view/334

[11] D. L. Crispina Pardede and Muhammad Andrias Indra Waskita, “ANALISIS PEMODELAN TOPIK UNTUK ULASAN TENTANG PEDULI LINDUNGI,” Jurnal Ilmiah Informatika Komputer, vol. 28, no. 1, pp. 17–26, Apr. 2023, doi: https://doi.org/10.35760/ik.2023.v28i1.7925.

[12] Ahmad Syaifuddin, Teknologi Informasi iSTTS, Reddy Alexandro Harianto, Teknologi Informatika iSTTS, and Joan Santoso, “Analisis Trending Topik untuk Percakapan Media Sosial dengan Menggunakan Topic Modelling Berbasis Algoritme LDA,” JOURNAL OF INTELLIGENT SYSTEMS AND COMPUTATION, pp. 12–19, 2020, Accessed: Apr. 03, 2026. [Online]. Available:

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 924−933
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10242

https://scholar.google.co.id/scholar?hl=id&as_sdt=0%2C5&q=Analisis+Trending+Topik+untuk+Percakapan+Media+Sosial+dengan+Menggunakan+Topic+Modelling+Berbasis+Algoritme+LDA&btnG=

[13] Angga Reni Dwi Astuti and Nuri Cahyono, "Analisis Topic Modelling Persepsi Pengguna Internet Menggunakan Metode Latent Dirichlet Allocation," Indonesian Journal of Computer Science Attribution, vol. 12, no. 1, pp. 326–334, Feb. 2023, Accessed: Apr. 03, 2026. [Online]. Available: https://www.ijcs.net/ijcs/index.php/ijcs/article/view/3155

[14] Haris Tri Saputra, Alfath Damanik, M. Mayo Shaquille, Marini Adibah Rusydi, Muhammad Habib Riziq, and Naya Septia Zulva, "Analisis Modelling Pada Reviewes Lazada Indonesia Menggunakan Latent Dirichlet Allocation (LDA) Untuk Optimalisasi Strategi Bisnis," JEKIN - Jurnal Teknik Informatika, vol. 5, no. 1, pp. 361–371, Feb. 2025, doi: 10.58794/jekin.v5i1.1325.

[15] Uray Nur Khadijah and Nuri Cahyono, "Analisis Topic Modelling Pariwisata Yogyakarta Menggunakan Latent Dirichlet Allocation (LDA)," The Indonesian Journal of Computer Science, vol. 13, no. 4, pp. 6075–6086, Aug. 2024, doi: 10.33022/ijcs.v13i4.3816.

[16] Ibnu Surya Wibowo, Arita Witanti, and Indah Susilawati, "Keyword Extraction Judul Berita Online Di Indonesia Menggunakan Metode TF-IDF," Jurnal Teknik Informatika dan Sistem Informasi, vol. II, no. 1, pp. 1–13, Mar. 2024, doi: https://doi.org/10.35957/jatisi.v11i1.6718.

[17] Kadek Arya Budi Artana, Gade Aditra Pradnyana, and Gade Mahendra Darmawiguna, "ANALISIS SENTIMEN TWITTER UNTUK MENILAI KESIAPAN PEMBELAJARAN TATAP MUKA TERBATAS DENGAN INSET LEXICON DAN LEVENSHTEIN DISTANCE," Jurnal Pendidikan Teknologi dan Kejuruan, vol. 20, no. 2, pp. 200–208, Jul. 2023.

[18] Ari Muzakir and Uci Suriani, "Model Deteksi Berita Palsu Menggunakan Pendekatan Bidirectional Long Short-Term Memory (BiLSTM)," Journal of Computer and Information Systems Ampera, vol. 4, no. 2, pp. 93–105, May 2023, doi: 10.51519/journalcisa.v4i2.397.

[19] Aulia Cisatra, Daffa Daris Mahendra Ansori, Dara Nasywa Fathya Afiqah Akmal, Slamet Ramadhan, and Nur Aini Rakhmawati, "View of Identifikasi Topik Hangat di Media Berita Menggunakan Latent Dirichlet Allocation," JIEET (Journal of Information Engineering and Educational Technology), vol. 8, no. 1, pp. 14–17, Jun. 2024, doi: https://doi.org/10.26740/jieet.v8n1.p14-17.

[20] Ari Harsono and Aqila Mazi, "Representasi etnis Tionghoa dalam media: Analisis perbandingan di media berita daring tirto.id, republika.co.id, dan tempo.co," JEK Journal Earth Kingdom, vol. 1, no. 2, pp. 57–68, Jan. 2024, doi: https://doi.org/10.61511/jek.v1i1.2024.552.

[21] Vito Ramadhan, Asriyanik Asriyanik, and Agung Pambudi, "View of IMPLEMENTASI ALGORITMA CONVOLUTIONAL NEURAL NETWORK UNTUK MENGIDENTIFIKASI BERITA HOAKS BERBAHASA INDONESIA," JATI (Journal Mahasiswa Teknik Informatika), vol. 8, no. https://www.ejournal.itn.ac.id/jati/issue/view/343, pp. 10945–10952, Oct. 2024, doi: https://doi.org/10.36040/jati.v8i5.10889.

[22] Siti Nursaudah and Rinda Cahyana, "Tampilan Pemodelan Analisis Tren Topik Penelitian Sistem Informasi Menggunakan Latent Dirichlet Allocation," Jurnal Algoritma, vol. 22, no. 2, p. 12, Nov. 2025, doi: https://doi.org/10.33364/algoritma/v.22-2.2596.

[23] Saskia Khoirunnisa, Aulia Anggana Izzahra Tauzaman, and Resta Aviani, "Menganalisis Dan Meramalkan Tren Di Media Sosial Dengan Menerapkan Model Matematika," 2024. Accessed: Jun. 09, 2026. [Online]. Available: https://proceeding.unindra.ac.id/index.php/DPNPMunindra/article/view/7417/2667