Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 959−971
DOI 10.47065/bits.v8i2.10951
https://ejurnal.seminar-id.com/index.php/bits

Implementasi Algoritma BERTopic Berbasis IndoBERT untuk Pemodelan dan Analisis Trending Topik Berita Online pada Dashboard Website

Implementation of IndoBERT-Based BERTopic Algorithm for Modeling and Analysis of Trending Topics in Online News on a Website Dashboard

M. Wirayuda*, Ari Muzakir, Wydyanto, Andri

Fakultas Sains Teknologi, Program Studi Teknik Informatika, Universitas Bina Darma, Palembang, Indonesia
Email: 1,*mwirayudayuda@gmail.com, 2arimuzakir@binadarma.ac.id, 3wydyanto@binadarma.ac.id, 4andri@binadarma.ac.id
Email Penulis Korespondensi: mwirayudayuda@gmail.com
Received: **09/08/2026**, Revised: **17/09/2026**, Accepted: **21/09/2026**, Published**: 22/09/2026**

Abstrak−Volume berita online yang terus meningkat menyulitkan tim redaksi Sumatera Ekspres memantau topik populer secara manual dan tepat waktu. Penelitian ini mengimplementasikan algoritma BERTopic berbasis IndoBERT untuk memodelkan trending topik berita online secara otomatis, dengan tahapan CRISP-DM meliputi business understanding, data understanding, data preparation, modeling, evaluation, dan deployment. Data dikumpulkan melalui web scraping dari delapan portal berita nasional dan YouTube, kemudian diproses melalui case folding, cleaning, tokenization, dan penghapusan stopword sebelum dimodelkan menggunakan IndoBERT sebagai embedding, UMAP untuk reduksi dimensi, HDBSCAN untuk clustering, dan c-TF-IDF untuk ekstraksi kata kunci, dilengkapi metrik trend_score untuk mengidentifikasi topik yang benar-benar sedang meningkat intensitas pemberitaannya. Kontribusi utama penelitian ini adalah integrasi data multi-sumber yang lebih heterogen dibandingkan penelitian sebelumnya serta metrik trend_score sebagai penanda tren yang lebih akurat, yang diwujudkan dalam sistem dashboard siap pakai bagi redaksi. Pengujian pada empat periode data (1, 7, 14, dan 30 hari) menunjukkan bahwa model secara konsisten menghasilkan topik dengan kualitas koherensi dan diversitas yang baik. Hasil penelitian diimplementasikan ke dalam dashboard berbasis website menggunakan FastAPI dan SQLite yang menyajikan visualisasi trending topik, word cloud, dan rekomendasi judul berita, serta telah diuji dan dinyatakan berfungsi sesuai kebutuhan, sehingga layak digunakan sebagai alat bantu redaksi dalam memantau isu populer secara cepat, efisien, dan berbasis data.

Kata Kunci: BERTopic; IndoBERT; HDBSCAN; Topic Modeling; Trending Topik Berita

Abstract−The growing volume of online news makes it difficult for the Sumatera Ekspres editorial team to manually track popular topics in a timely manner. This study implements the BERTopic algorithm based on IndoBERT to automatically model trending topics of online news, following the CRISP-DM stages of business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Data were collected through web scraping from eight national news portals and YouTube, then processed through case folding, cleaning, tokenization, and stopword removal before being modeled using IndoBERT as the embedding, UMAP for dimensionality reduction, HDBSCAN for clustering, and c-TF-IDF for keyword extraction, complemented by a trend_score metric to identify topics that are genuinely gaining momentum. The main contribution of this study lies in integrating more heterogeneous multi-source data than prior work and introducing the trend_score metric as a more accurate indicator of momentum, realized as a ready-to-use dashboard for the editorial team. Testing across four data periods (1, 7, 14, and 30 days) shows that the model consistently produces topics of good coherence and diversity quality. The results were implemented into a website-based dashboard built with FastAPI and SQLite that presents trending topic visualizations, word cloud, and news headline recommendations, which has been tested and confirmed to function as required, making it suitable as a decision-support tool for the editorial team to monitor popular issues quickly, efficiently, and in a data-driven manner.

Keywords: BERTopic; IndoBERT; HDBSCAN; Topic Modeling; Trending News Topics

1. PENDAHULUAN

Perkembangan teknologi informasi dan internet telah mendorong pertumbuhan media berita *online* secara pesat. Portal berita digital memungkinkan masyarakat memperoleh informasi dengan cepat dan mudah melalui berbagai perangkat, seperti komputer maupun *smartphone* [1]. Banyaknya portal berita yang tersedia, seperti CNN Indonesia, Detik, Tribun, Liputan6, dan media lainnya, menyebabkan arus informasi yang sangat besar dan terus diperbarui setiap waktu. Kondisi ini membuat persaingan antarmedia berita menjadi semakin ketat dalam menyajikan informasi yang aktual, relevan, dan menarik bagi pembaca.

Dalam dunia jurnalistik digital, kemampuan mengidentifikasi *trending* topik yang sedang populer menjadi faktor penting bagi redaksi dalam menentukan agenda pemberitaan. Redaksi media perlu mengetahui isu apa yang sedang banyak dibahas oleh media lain maupun masyarakat agar dapat menghasilkan berita yang relevan dan berpotensi menarik perhatian pembaca. Namun, proses pemantauan *trending* berita secara manual dari berbagai portal berita membutuhkan waktu yang cukup lama dan kurang efisien, terutama ketika jumlah berita yang dipublikasikan setiap hari sangat banyak [2].

Salah satu pendekatan untuk mengatasi permasalahan tersebut adalah *topic modeling* dalam bidang *text mining*, yaitu teknik yang mengelompokkan dokumen berdasarkan kemiripan makna sehingga mampu mengidentifikasi topik

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 959−971
DOI 10.47065/bits.v8i2.10951
https://ejurnal.seminar-id.com/index.php/bits

yang sedang berkembang secara otomatis dari kumpulan berita [3]. Salah satu metode *topic modeling* yang banyak digunakan saat ini adalah BERTopic, karena memanfaatkan representasi semantik berbasis *transformer* sehingga mampu menghasilkan topik yang lebih kontekstual dibandingkan metode tradisional seperti *Latent Dirichlet Allocation* (LDA) [4]. Dalam penelitian ini, IndoBERT digunakan untuk menghasilkan representasi semantik teks berita berbahasa Indonesia, dilanjutkan dengan *clustering* menggunakan HDBSCAN berdasarkan kesamaan *embedding*, serta c-TF-IDF untuk mengekstraksi kata kunci dan memberikan label pada setiap topik yang terbentuk.

Beberapa penelitian terdahulu telah membuktikan efektivitas BERTopic pada berbagai domain data. Wahyuni dkk. menerapkan BERTopic pada ribuan komentar YouTube terkait program makan bergizi gratis dengan memanfaatkan fungsi *reduce_topics()* untuk menyederhanakan jumlah topik, dan membuktikan bahwa BERTopic mampu menangani data teks informal dengan baik [5]. Simanjuntak dkk. menerapkan fitur *Dynamic Topic Modeling* (DTM) pada BERTopic untuk memetakan evolusi opini publik terhadap kendaraan listrik secara kronologis [6]. Ramadhan & Fernando membandingkan BERTopic dan LDA pada ulasan pelanggan restoran, dan menemukan bahwa BERTopic unggul secara kuantitatif dengan nilai koherensi 0,5541 dibandingkan LDA sebesar 0,5226, meskipun LDA dinilai lebih interpretatif secara kualitatif untuk kebutuhan bisnis [7]. Aryani dkk. mengombinasikan BERTopic dengan K-Means untuk mengelompokkan 15.019 artikel Pemilu 2024 pada portal Detik.com menjadi lima klaster tema kampanye [8]. Muhammad Rayhan Nur dkk. menggabungkan *fine-tuning* IndoBERT untuk analisis sentimen dengan BERTopic untuk pemodelan topik pada 10.130 *post* di platform X, menghasilkan 16 klaster topik dengan rata-rata *coherence score* 0,5437 [9].

Berdasarkan uraian penelitian terdahulu tersebut, dapat diidentifikasi sejumlah kesenjangan (*research gap*) yang mendasari penelitian ini. Pertama, dari segi sumber data, kelima penelitian tersebut — Wahyuni dkk. [5], Simanjuntak dkk. [6], Ramadhan & Fernando [7], Aryani dkk. [8], dan Muhammad Rayhan Nur dkk. [9] seluruhnya hanya menggunakan satu jenis sumber data tunggal, baik berupa komentar YouTube [5], [6], ulasan pelanggan pada satu platform bisnis [7], artikel dari satu portal berita saja yaitu Detik.com [8], maupun unggahan dari satu platform media sosial X [9], sehingga belum menguji kinerja BERTopic pada kondisi data multi-sumber yang lebih heterogen dari segi gaya penulisan, istilah, dan penekanan isu. Kedua, dari segi dimensi waktu, keempat penelitian selain [6] berhenti pada pemodelan dan evaluasi topik secara statis pada satu titik waktu [5], [7]–[9]; Simanjuntak dkk. [6] memang menyentuh aspek temporal melalui *Dynamic Topic Modeling*, namun sebatas memetakan evolusi opini secara kronologis tanpa metrik kuantitatif untuk membedakan topik yang benar-benar sedang meningkat intensitasnya dari topik yang hanya bervolume besar. Ketiga, dari segi implementasi, tidak satu pun dari kelima penelitian tersebut [5]–[9] mengintegrasikan hasil pemodelan topik ke dalam sistem *dashboard* yang dapat langsung digunakan oleh pengguna akhir, sehingga luarannya masih berhenti pada tahap analisis akademis dan belum menjawab kebutuhan operasional redaksi media secara langsung.

Ketiga kesenjangan tersebut menjadi dasar arah penelitian ini. Berbeda dengan penelitian-penelitian sebelumnya, penelitian ini mengintegrasikan data dari delapan portal berita nasional serta media sosial YouTube melalui proses *web scraping* yang lebih luas dan heterogen, mengusulkan metrik *trend_score* berbasis *split-period comparison* untuk menangkap dimensi waktu yang belum tersedia pada penelitian [5], [7]–[9], serta mengembangkan *dashboard* berbasis *website* menggunakan FastAPI yang tidak hanya menyajikan visualisasi *trending topic*, tetapi juga rekomendasi judul dan topik berita populer secara otomatis untuk membantu tim redaksi bekerja lebih cepat dan efisien. Pemilihan Sumatera Ekspres sebagai studi kasus juga didasari kebutuhan nyata redaksi media regional yang belum memiliki sistem pemantauan tren berbasis data, sehingga penelitian ini sekaligus menguji apakah pendekatan BERTopic berbasis IndoBERT dapat diandalkan pada kondisi data multi-sumber yang lebih heterogen dibandingkan studi-studi sebelumnya.

Untuk teks berbahasa Indonesia, IndoBERT sebagai model *embedding* dinilai lebih tepat karena dilatih pada korpus bahasa Indonesia sehingga menghasilkan representasi semantik yang lebih akurat. Penerapan IndoBERT pada deteksi berita berbahasa Indonesia terbukti mencapai akurasi 94,52% dan *F1-score* 94,53%, melampaui performa model multibahasa mBERT yang hanya mencapai akurasi 91,58% [10]. Temuan ini menjadi dasar pemilihan kombinasi BERTopic berbasis IndoBERT sebagai pendekatan yang tepat untuk mendukung identifikasi *trending topic* berita secara otomatis.

Untuk memastikan penelitian berjalan secara sistematis, terstruktur, dan sesuai dengan standar proyek *data mining*, penelitian ini mengadopsi metodologi CRISP-DM (*Cross-Industry Standard Process for Data Mining*), yang dipilih karena bersifat iteratif dan komprehensif, mencakup enam tahapan utama, yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment* [11]. Pendekatan ini sangat sesuai untuk penelitian yang menggabungkan proses *web scraping*, pemodelan topik dengan BERTopic, analisis tren, hingga implementasi *dashboard* berbasis *website*.

Sumatera Ekspres sebagai salah satu media berita di Indonesia saat ini masih mengandalkan proses pemantauan berita secara manual, di mana tim redaksi membaca dan memilih topik dari berbagai portal berita *online* tanpa dukungan sistem otomatis. Kondisi ini menyebabkan proses identifikasi topik yang sedang *trending* membutuhkan waktu lebih lama dan rentan terhadap subjektivitas, sehingga respons redaksi terhadap isu yang sedang berkembang menjadi kurang optimal. Belum tersedianya sistem berbasis data yang dapat menganalisis dan memvisualisasikan *trending* topik secara otomatis menjadi kendala tersendiri bagi redaksi dalam mengambil keputusan editorial secara cepat dan tepat.

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 959−971
DOI 10.47065/bits.v8i2.10951
https://ejurnal.seminar-id.com/index.php/bits

Berdasarkan permasalahan tersebut, penelitian ini bertujuan mengimplementasikan algoritma BERTopic berbasis IndoBERT untuk mengekstraksi dan mengelompokkan topik berita *online* secara otomatis, serta mengembangkan sistem *dashboard* berbasis *website* yang menyajikan visualisasi *trending* topik, *word cloud*, dan rekomendasi judul berita untuk membantu tim redaksi Sumatera Ekspres dalam pemantauan isu populer dan pengambilan keputusan editorial berbasis data. Implementasi diuji pada empat periode data (1, 7, 14, dan 30 hari) untuk melihat konsistensi kualitas topik yang dihasilkan, dievaluasi menggunakan metrik *coherence score* dan *topic diversity*, serta dilengkapi metrik tambahan *trend_score* untuk menentukan topik yang benar-benar sedang meningkat intensitas pemberitaannya, bukan sekadar topik dengan volume berita terbanyak.

Kontribusi utama penelitian ini terletak pada tiga hal: (1) integrasi data multi-sumber dari delapan portal berita nasional dan media sosial YouTube yang lebih heterogen dari segi gaya penulisan dan istilah dibandingkan sumber data tunggal pada penelitian-penelitian terdahulu; (2) pengusulan metrik *trend_score* berbasis *split-period comparison* sebagai kontribusi metodologis untuk membedakan topik yang benar-benar sedang meningkat intensitas pemberitaannya dari topik yang hanya bervolume besar; dan (3) implementasi seluruh alur kerja tersebut ke dalam *dashboard* berbasis *website* yang siap digunakan oleh tim redaksi Sumatera Ekspres sebagai alat bantu pengambilan keputusan editorial berbasis data.

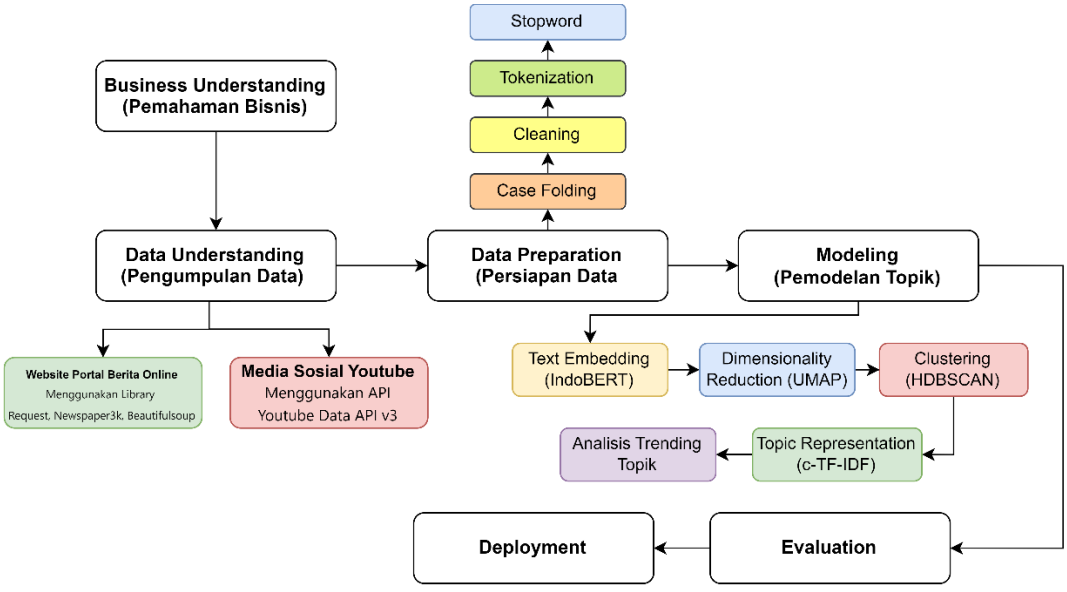
2. METODOLOGI PENELITIAN

2.1 Kajian Pustaka Algoritma

BERTopic dipilih pada penelitian ini karena memanfaatkan representasi *embedding* berbasis *transformer* sehingga menghasilkan topik yang lebih kontekstual dibandingkan pendekatan *bag-of-words* seperti LDA [4]. *IndoBERT* digunakan sebagai model *embedding* karena dilatih khusus pada korpus bahasa Indonesia melalui pendekatan *masked language modeling* dan *next sentence prediction*, sehingga menghasilkan representasi semantik yang lebih akurat untuk teks berbahasa Indonesia dibandingkan model multibahasa seperti mBERT [12], [13]. *UMAP* dipilih sebagai metode reduksi dimensi karena mampu menyederhanakan representasi *embedding* berdimensi tinggi tanpa banyak menghilangkan informasi semantik, sekaligus mempercepat proses komputasi *clustering* [14]. *HDBSCAN* digunakan pada tahap *clustering* karena secara otomatis menentukan jumlah klaster optimal berdasarkan kepadatan data serta mampu mengidentifikasi data *noise*, sehingga lebih fleksibel dibandingkan algoritma *clustering* berbasis partisi seperti K-Means yang mengharuskan jumlah klaster ditentukan di awal [15], [16], [17]. Terakhir, *c-TF-IDF* digunakan pada tahap *topic representation* untuk mengekstraksi kata kunci yang paling merepresentasikan setiap klaster berdasarkan bobot kata pada tingkat klaster, bukan pada tingkat dokumen tunggal seperti *TF-IDF* konvensional [18], [19]. Kajian pustaka terhadap kelima komponen ini menjadi dasar pemilihan kombinasi algoritma yang diterapkan secara rinci pada sub-bab Modeling.

2.2 Tahapan Penelitian

Penelitian ini menggunakan metodologi CRISP-DM (*Cross-Industry Standard Process for Data Mining*) sebagai kerangka kerja utama, karena metodologi ini bersifat fleksibel, independen terhadap jenis industri, serta memungkinkan peneliti kembali ke tahap sebelumnya apabila ditemukan permasalahan atau diperlukan penyempurnaan model selama proses pengembangan [20]. Tahapan penelitian mengikuti enam fase CRISP-DM sebagaimana ditunjukkan pada Gambar 1.



Gambar 1. Kerangka Kerja CRISP-DM

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 959−971
DOI 10.47065/bits.v8i2.10951
https://ejurnal.seminar-id.com/index.php/bits

2.3 Business Understanding

Pada tahap ini dilakukan identifikasi masalah, perumusan tujuan, dan penentuan manfaat penelitian. Penelitian ini bertujuan membantu redaksi Sumatera Ekspres mengidentifikasi judul atau topik berita populer secara otomatis menggunakan algoritma BERTopic berbasis data, yang selanjutnya diimplementasikan ke dalam aplikasi berbasis *website* untuk mendukung pengambilan keputusan editorial.

2.4 Data Understanding

Data dikumpulkan dari delapan portal berita *online* nasional, yaitu CNN Indonesia, Detik, Antaranews, tvOneNews, Suara, Kompas, Okezone, dan Liputan6, serta media sosial YouTube, Proses *web scraping* pada portal berita dilakukan menggunakan *library Newspaper3k*, *BeautifulSoup*, dan *Requests* [21], [22], sedangkan data dari YouTube diperoleh melalui *YouTube Data API v3*. Data yang dikumpulkan meliputi judul berita, isi berita, serta tanggal publikasi/*scraping*. Dalam praktiknya, proses *scraping* menghadapi tantangan teknis seperti mekanisme deteksi *bot* dan perubahan struktur halaman secara berkala, sehingga diperlukan pendekatan *ethical scraping* yang memperhatikan kebijakan penggunaan data pada situs target [23].

2.5 Data Preparation

Tahapan *data preparation* dilakukan meliputi empat proses berurutan: (1) *case folding*, mengubah seluruh teks menjadi huruf kecil untuk menghindari duplikasi kata akibat perbedaan kapitalisasi; (2) *data cleaning*, menghapus karakter HTML, tanda baca, angka, dan karakter khusus; (3) *tokenization*, memecah teks menjadi unit kata; dan (4) penghapusan *stopword* menggunakan kamus bahasa Indonesia dari *library PySastrawi* dan *NLTK*. Tahapan ini merupakan bagian penting dari *text mining* untuk mengubah data teks tidak terstruktur menjadi bentuk yang dapat dianalisis lebih lanjut [24]. Metode *topic modeling* konvensional seperti LDA dan *Non-Negative Matrix Factorization* (NMF) banyak digunakan untuk menemukan pola topik tersembunyi dari kumpulan dokumen secara otomatis [25]. namun keduanya memiliki keterbatasan karena mengandalkan pendekatan *bag-of-words* yang tidak mempertimbangkan hubungan semantik antarkata secara mendalam [26], sehingga penelitian ini memilih pendekatan berbasis *transformer* melalui BERTopic [4]

2.6 Modeling

Tahapan Pemodelan topik menggunakan BERTopic dilakukan melalui lima tahap utama tahap pertama adalah *text embedding* menggunakan model IndoBERT, secara spesifik checkpoint *firqaaa/indo-sentence-bert-base* yang merupakan varian *Sentence-BERT* hasil *fine-tuning* dari IndoBERT, yaitu model bahasa *pre-trained* berbasis arsitektur *transformer* yang dilatih khusus pada korpus bahasa Indonesia menggunakan pendekatan *masked language modeling* dan *next sentence prediction* [12], sehingga mampu menghasilkan representasi semantik yang lebih akurat dibandingkan model multibahasa untuk teks berita maupun media sosial berbahasa Indonesia [13]. Tahap kedua adalah reduksi dimensi menggunakan UMAP (*Uniform Manifold Approximation and Projection*), yang bertujuan menyederhanakan representasi *embedding* tanpa menghilangkan informasi semantik penting sekaligus mempercepat proses komputasi [14]. Tahap ketiga adalah *clustering* menggunakan HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*), yang secara otomatis menentukan jumlah klaster optimal serta mengidentifikasi data *noise* tanpa memerlukan penentuan jumlah klaster di awal seperti pada K-Means [15], [16]; dokumen yang tidak masuk ke klaster manapun dikategorikan sebagai *outlier* (topik -1) dan diabaikan dalam pembentukan topik [17]. Parameter *UMAP* yang digunakan pada penelitian ini adalah *n_neighbors* = max(2, min(15, jumlah dokumen − 1)), *n_components* = 5, *min_dist* = 0,0, *metric* = "cosine", dan *random_state* = 42; formula dinamis pada *n_neighbors* ini secara konsisten menghasilkan nilai 15 pada seluruh periode pengujian karena jumlah dokumen pada setiap periode (114–4.451 dokumen) selalu jauh melebihi ambang tersebut. Adapun *HDBSCAN* menggunakan *min_cluster_size* = max(2, min(3, jumlah dokumen // 5)), yang juga secara konsisten bernilai 3 pada seluruh periode dengan alasan yang sama, *min_samples* mengikuti nilai *default* (setara *min_cluster_size*), *metric* = "euclidean", *cluster_selection_method* = "eom" (*Excess of Mass*), dan *prediction_data* = *True*. Kombinasi parameter ini dipilih berdasarkan hasil uji coba untuk menyeimbangkan jumlah klaster yang terbentuk dengan tingkat noise pada data berita berbahasa Indonesia. Tahap keempat adalah *topic representation* menggunakan c-TF-IDF (*class-based Term Frequency–Inverse Document Frequency*), yang menghitung bobot kata berdasarkan keseluruhan dokumen dalam satu klaster dengan *ngram_range* = (1, 2) sehingga kata kunci yang dihasilkan dapat berupa kata tunggal maupun frasa dua kata, sehingga dihasilkan kata kunci yang paling representatif untuk setiap topik [18], [19]. Tahap kelima analisis *trending topic* memanfaatkan fitur *topics_over_time* pada *library* BERTopic untuk mengidentifikasi topik berdasarkan frekuensi kemunculannya dari waktu ke waktu. Karena frekuensi mentah tidak cukup untuk menentukan topik yang benar-benar sedang meningkat (mengingat total frekuensi antartopik dapat bernilai sama meskipun pola perkembangannya berbeda), penelitian ini menambahkan metrik heuristik *trend_score* yang diadaptasi dari prinsip *split-period comparison* dalam analisis deret waktu, dihitung dengan membandingkan rata-rata frekuensi kemunculan topik pada paruh akhir dikurangi paruh awal periode pengamatan. Nilai *trend_score* positif mengindikasikan topik sedang meningkat intensitas pemberitaannya, sedangkan nilai nol atau negatif mengindikasikan topik stagnan atau menurun.

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 959−971
DOI 10.47065/bits.v8i2.10951
https://ejurnal.seminar-id.com/index.php/bits

2.7 Evaluation dan Deployment

Evaluasi model dilakukan menggunakan dua metrik utama, yaitu *Coherence Score* (C_v), yang mengukur keterkaitan semantik antar kata kunci representatif dalam satu topik, dan *Topic Diversity*, yang mengukur keberagaman antartopik agar tidak redundan atau tumpang tindih. Hasil pemodelan selanjutnya diimplementasikan ke dalam sistem berbasis *website* menggunakan arsitektur *client-server*, di mana *server* menjalankan model BERTopic berbasis Python melalui FastAPI dan basis data SQLite, sementara *client* mengakses *dashboard* melalui peramban web. Rincian komponen teknologi yang digunakan disajikan pada Tabel 1. Pengujian sistem dilakukan menggunakan metode *Black-Box Testing* terhadap seluruh fitur utama untuk memastikan keluaran sistem sesuai dengan kebutuhan fungsional yang telah ditetapkan.

Tabel 1. Implementasi Sistem

Komponen	Teknologi	Fungsi
Backend	Python + FastAPI	Server-side logic dan REST API endpoint
Web Scraping	Newspaper3k, BeautifulSoup, Requests	Pengambilan konten artikel dari portal berita
Model AI	BERTopic, IndoBERT	Pemodelan dan analisis topik berita
Frontend	HTML, CSS, JavaScript	Antarmuka pengguna website
Visualisasi	Chart.js / Plotly	Grafik trending topik dan word cloud dashboard
Basis Data	SQLite	Penyimpanan data berita dan hasil analisis topik
Ekspor Data	Pandas, openpyxl	Ekspor hasil scraping berita ke format CSV
Upload Data	Pandas	Membaca dan memproses file CSV yang diunggah pengguna

3. HASIL DAN PEMBAHASAN

3.1 Data Understanding

Total data yang berhasil dikumpulkan melalui proses *scraping* dari delapan portal berita dan media sosial YouTube adalah 9.699 data, terhitung sejak 22 April 2026 hingga 12 Juli 2026. Untuk kebutuhan evaluasi pemodelan topik, data diambil berdasarkan empat rentang waktu terakhir, yaitu 1, 7, 14, dan 30 hari, sebagaimana dirangkum pada Tabel 2.

Tabel 2. Jumlah Data yang Digunakan per Periode

Periode	Jumlah Data
30 hari terakhir	4.451
14 hari terakhir	2.087
7 hari terakhir	1.049
1 hari terakhir	114

Semakin panjang rentang waktu data yang digunakan, semakin besar pula jumlah dokumen yang dianalisis, dari 114 dokumen pada periode 1 hari hingga 4.451 dokumen pada periode 30 hari.

3.2 Data Preparation

Berdasarkan tahapan *data preparation* yang telah dilakukan, diperoleh perubahan bentuk data judul berita dari data asli menjadi data yang lebih terstruktur dan siap digunakan dalam proses pemodelan topik. Proses tersebut diterapkan secara berurutan melalui *case folding*, *data cleaning*, *tokenization*, dan penghapusan *stopword*. Hasil penerapan tahapan tersebut dapat dilihat pada Tabel 3.

Tabel 3. Hasil Proses *Data Preparation*

No	Data Asli (Sebelum)	Data Bersih (Sesudah)
1	Viral Video Ojek Pangkalan di Lampung Dibacok, Pelaku Ditangkap di Sumsel	viral ojek pangkalan lampung dibacok pelaku ditangkap sumsel
2	Heboh Dentuman Misterius di Cirebon, Ini Penjelasan BRIN	heboh dentuman misterius cirebon penjelasan brin
3	Kondisi Wali Kota Bandung Mulai Stabil Usai Dilarikan ke Rumah Sakit	kondisi wali kota bandung stabil dilarikan rumah sakit
4	🔴2 Rumah Rahasia Berisi Harta Bupati Sukoharjo Diendus KPK, Diduga Timbun Hasil Pemerasan Rp21,2 M	rumah rahasia berisi harta bupati sukoharjo diendus kpk diduga timbun hasil pemerasan
5	🔴BREAKING NEWS - Presiden Prabowo Hadiri Acara Hari Koperasi (12/07)	presiden prabowo hadiri acara koperasi

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 959−971
DOI 10.47065/bits.v8i2.10951
https://ejurnal.seminar-id.com/index.php/bits

3.3 Pemodelan Topik dan Evaluasi Kualitas

Hasil pemodelan topik pada keempat periode menunjukkan pola yang konsisten: semakin panjang rentang waktu data, semakin banyak pula jumlah topik yang terbentuk, sebagaimana dirangkum pada Tabel 4.

Tabel 4. Ringkasan Hasil Pemodelan Topik pada Empat Periode

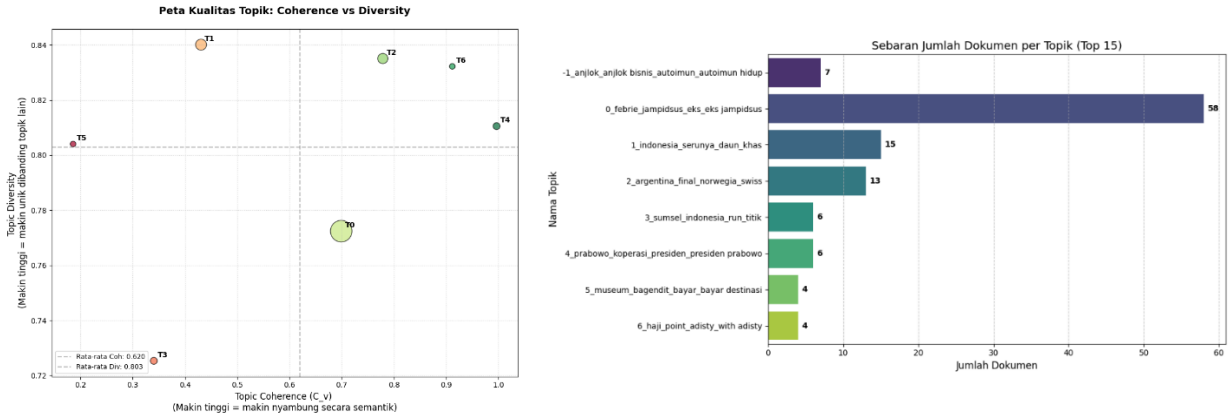
Periode	Jumlah Data	Jumlah Topik	Topik Frekuensi Terbesar	Rata-rata Coherence	Rata-rata Diversity
1 Hari	114	7	Kasus eks Jampidsus Febrie Adriansyah (58 dokumen.)	0,6204	0,8028
7 Hari	1.049	93	Praperadilan Roy Suryo (27 dokumen.)	0,654	0,757
14 Hari	2.087	156	Roy Suryo & ijazah Jokowi (116 dok.)	0,653	0,774
30 Hari	4.451	297	Demo mahasiswa tuntut perbaikan ekonomi (76 dokumen.)	0,654	0,766

Pada periode 1 hari, BERTopic mengelompokkan 114 dokumen ke dalam 7 topik dengan sebaran yang sangat timpang; topik dominan (febrie, jampidsus, eks, korupsi) mencakup 58 dari 114 dokumen, mencerminkan pemberitaan intensif mengenai kasus hukum eks Jampidsus Febrie Topik 4 (prabowo, koperasi, presiden) mencatat koherensi tertinggi (0,9965), hasil dapat dilihat pada Tabel 5.

Tabel 5. Contoh Evaluasi Coherence dan Diversity Periode 1 Hari

ID Topik	Jumlah	Kata Kunci	Koherensi	Diversitas
0	58	febrie, jampidsus, eks, eks jampidsus, korupsi	0,6991	0,7724
1	15	indonesia, serunya, daun, khas, kuliner	0,4308	0,8401
2	13	argentina, final, norwegia, swiss, world	0,7787	0,8351
3	6	sumsel, indonesia, run, titik, juli	0,3404	0,7253
4	6	prabowo, koperasi, presiden, presiden prabowo, hadiri	0,9965	0,8105
5	4	museum, bagendit, bayar, bayar destinasi, curhatan	0,1859	0,8040

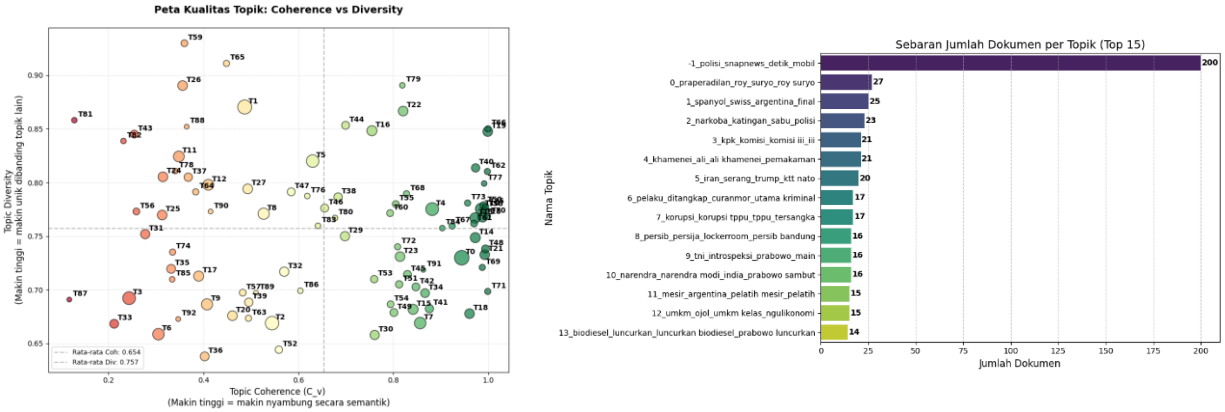
Pada periode 1 hari, jumlah data meningkat menjadi 114 dokumen dan BERTopic berhasil membentuk 7 topik, Klaster *noise* (-1) berisi 7 berita yang tidak cukup homogen, sedangkan topik non-*noise* terbesar adalah kasus praperadilan Kasus eks Jampidsus Febrie Adriansyah (58 dokumen). Pada Gambar 2 memperlihatkan peta sebar (*scatter plot*) yang memetakan nilai *coherence* terhadap *diversity* untuk seluruh 7 topik pada periode ini, dengan ukuran *bubble* merepresentasikan jumlah dokumen tiap topik. Nilai *coherence* pada topik dominan periode ini (0,6991) tergolong lebih tinggi dibandingkan skor koherensi rata-rata yang dilaporkan Ramadhan & Fernando pada ulasan pelanggan restoran (0,5541) [7] , menunjukkan bahwa BERTopic mampu menghasilkan koherensi yang baik meskipun volume data masih sangat terbatas.



Gambar 2. Visualisasi Sebaran dokumen, Peta Coherence dan Diversity Periode 1 Hari

Pada Gambar 3, periode 7 hari jumlah data meningkat menjadi 1.049 dokumen dan BERTopic berhasil membentuk 93 topik, jauh lebih beragam dibandingkan periode 1 hari. Klaster *noise* (-1) menjadi kelompok terbesar dengan 200 dokumen berisi campuran berita yang tidak cukup homogen, sedangkan topik non-*noise* terbesar adalah kasus praperadilan Roy Suryo terkait ijazah Jokowi (27 dokumen). Gambar 3 memperlihatkan peta sebar (*scatter plot*) yang memetakan nilai *coherence* terhadap *diversity* untuk seluruh 93 topik pada periode ini, dengan ukuran *bubble* merepresentasikan jumlah dokumen tiap topik. Nilai *coherence* topik dominan pada periode ini (0,9439) juga melampaui rata-rata koherensi yang dilaporkan pada penelitian berbasis komentar YouTube [5] , [6], mengindikasikan bahwa data berita berbahasa formal cenderung menghasilkan koherensi topik yang lebih stabil dibandingkan data komentar media sosial yang sifatnya lebih informal.

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 959−971
DOI 10.47065/bits.v8i2.10951
https://ejurnal.seminar-id.com/index.php/bits



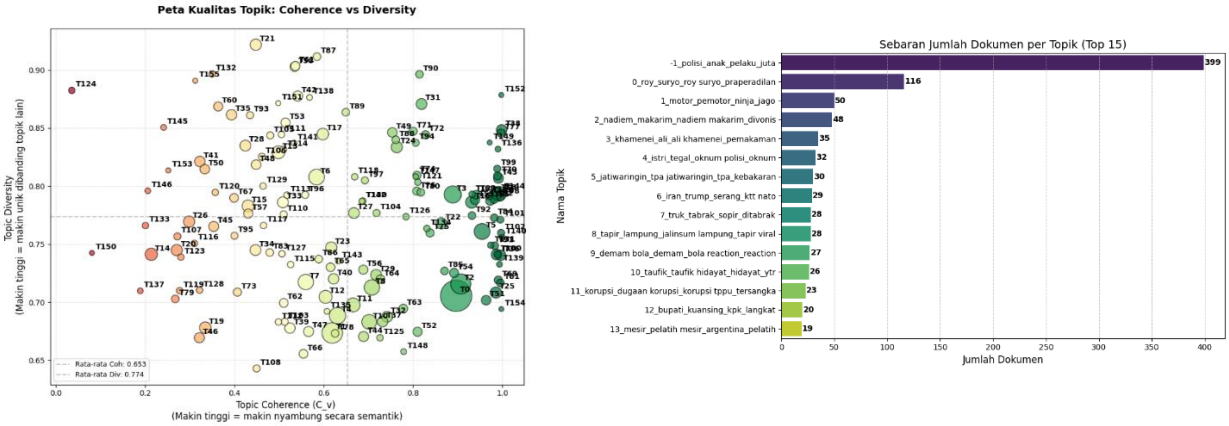
Gambar 3. Visualisasi Sebaran dokumen, Peta Coherence dan Diversity Periode 7 Hari

Tabel 6 menyajikan contoh hasil evaluasi coherence dan diversity pada periode 7 hari, difokuskan pada topik-topik dengan frekuensi dokumen teratas. Topik yang mendominasi pada periode ini adalah praperadilan Roy Suryo (27 dokumen), diikuti oleh isu Piala Dunia FIFA 2026 (25 dokumen) dan kasus narkoba di Katingan (23 dokumen), sebagaimana ditunjukkan pada tabel berikut.

Tabel 6. Contoh Evaluasi Coherence dan Diversity Periode 7 Hari

ID Topik	Jumlah	Kata Kunci	Koherensi	Diversitas
0	27	praperadilan, roy, suryo, roy suryo, menang, menang praperadilan, praperadilan roy, suryo menang, hakim, jokowi	0,9439	0,7298
1	25	spanyol, swiss, argentina, final, piala, piala dunia, fifa world, cup score, world cup, score	0,4868	0,8703
2	23	narkoba, katingan, sabu, polisi, bandar, katingan ditangkap, bandar narkoba, polisi gugur, peredaran, penyerangan	0,5444	0,6689
3	21	kpk, komisi, komisi III, III, gratifikasi, raja, raja juli, menhut, MPR, DPR	0,2433	0,6922
4	21	khamenei, ali, ali khamenei, pemakaman, pemakaman ali, tertinggi iran, iran ali, iran, tertinggi, pemimpin	0,8816	0,7752
5	20	iran, serang, trump, KTT NATO, NATO, KTT, serang iran, rusia, hormuz, selat hormuz	0,6299	0,8200

Periode 14 hari, dari 2.087 dokumen terbentuk 156 topik. Topik non-*noise* terbesar tetap seputar kasus Roy Suryo dan ijazah Jokowi (116 dokumen), diikuti insiden lalu lintas pemotor "Ninja Jago" di Jagakarsa (50 dokumen) dan vonis kasus korupsi Chromebook yang melibatkan eks Mendikbudristek Nadiem Makarim (48 dokumen). Nilai rata-rata *diversity* pada periode ini (0,774) merupakan yang tertinggi kedua di antara seluruh periode, menandakan topik-topik pada rentang waktu ini secara umum lebih terpisah dan unik satu sama lain, dapat dilihat, sebagaimana dapat dilihat pada Gambar 4. Memperlihatkan peta sebar (*scatter plot*) yang memetakan nilai *coherence* terhadap *diversity* untuk seluruh 156 topik pada periode ini, dengan ukuran *bubble* merepresentasikan jumlah dokumen tiap topik. Nilai *coherence* topik dominan pada periode ini (0,8963) jauh melampaui rata-rata koherensi yang dilaporkan Muhammad Rayhan Nur dkk. pada unggahan platform X (0,5437) [9], menegaskan keunggulan pendekatan BERTopic berbasis IndoBERT pada data multi-sumber berita nasional dibandingkan data media sosial tunggal.



Gambar 4. Visualisasi Sebaran dokumen, Peta Coherence dan Diversity Periode 14 Hari

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 959−971
DOI 10.47065/bits.v8i2.10951
https://ejurnal.seminar-id.com/index.php/bits

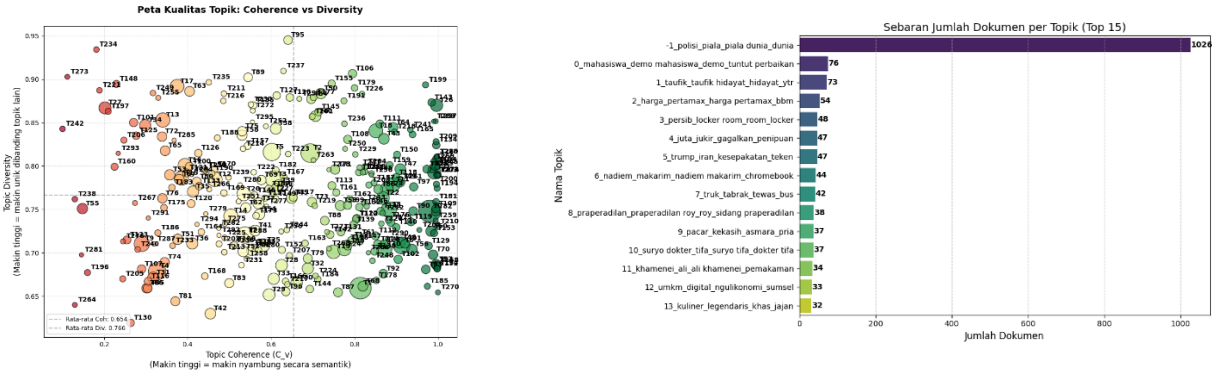
Sejalan dengan periode sebelumnya, Tabel 7 memuat contoh evaluasi coherence dan diversity untuk periode 14 hari yang disusun berdasarkan frekuensi dokumen terbanyak pada tiap topik. Pada periode ini, topik kasus Roy Suryo dan ijazah Jokowi masih mendominasi dengan 116 dokumen.

Tabel 7. Contoh Evaluasi Coherence dan Diversity Periode 14 Hari

ID Topik	Jumlah	Kata Kunci	Koherensi	Diversitas
0	116	roy, suryo, praperadilan, jokowi, ijazah	0,8963	0,7054
1	50	motor, pemotor, ninja, bang jago, viral	0,6187	0,6734
2	48	nadiem, makarim, chromebook, divonis, korupsi	0,9067	0,7156
3	35	khamenei, iran, pemakaman, pemimpin	0,8886	0,7928
4	32	istri, oknum polisi, siri, tegal	0,6305	0,6884
5	30	jatiwaringin, TPA, kebakaran, api, Tangerang	0,9544	0,7609

Pada Gambar 5 periode 30 hari, dari 4.451 dokumen terbentuk 297 topik, jumlah terbanyak di antara seluruh periode. Klaster noise meningkat signifikan menjadi 1.026 dokumen (23% dari total data), menunjukkan bahwa semakin panjang rentang waktu, semakin tinggi keberagaman isu sehingga proporsi berita tidak terklasifikasi turut meningkat. Topik non-noise terbesar adalah demonstrasi mahasiswa menuntut perbaikan ekonomi (76 dokumen), diikuti kasus penyekapan oleh Taufik Hidayat (73 dokumen). Nilai rata-rata diversity pada periode ini juga tertinggi (0,766), memperkuat pola bahwa semakin banyak topik terbentuk, topik-topik tersebut semakin unik dan terpisah satu sama lain, Gambar 5 memperlihatkan peta sebar (*scatter plot*) yang memetakan nilai *coherence* terhadap *diversity* untuk seluruh 297 topik pada periode ini, dengan ukuran *bubble* merepresentasikan jumlah dokumen tiap topik. Peningkatan proporsi klaster *noise* serta nilai *diversity* tertinggi pada periode ini sejalan dengan pola evolusi opini yang diamati Simanjuntak dkk. melalui *Dynamic Topic Modeling* [6], namun berbeda dengan penelitian tersebut, penelitian ini melengkapi temuan tersebut dengan metrik *trend_score* yang memungkinkan status kenaikan suatu topik diukur secara kuantitatif, bukan hanya dipetakan secara kronologis.

Semakin panjang rentang waktu data, semakin banyak pula peristiwa berdiri sendiri (isolated events) seperti berita hukum, olahraga, atau kejadian lokal yang hanya diberitakan dalam jumlah dokumen kecil dan kurang memiliki kemiripan semantik dengan topik lain, sehingga titik-titiknya tersebar renggang pada ruang embedding dan gagal memenuhi syarat kepadatan minimum untuk membentuk klaster tersendiri. Hal ini menjelaskan lonjakan proporsi klaster noise pada periode 30 hari, yang lebih disebabkan oleh akumulasi topik-topik minoritas tersebut, bukan karena topiknya tidak relevan. Nilai diversity pada periode ini yang justru tertinggi (0,766) turut mengindikasikan bahwa topik-topik yang berhasil terbentuk tetap memiliki kualitas pemisahan yang baik meskipun proporsi noise meningkat.



Gambar 5. Visualisasi Sebaran dokumen, Peta Coherence dan Diversity Periode 30 Hari

Sebagai periode dengan rentang waktu terpanjang, Tabel 8 menampilkan contoh evaluasi coherence dan diversity pada periode 30 hari yang juga diurutkan berdasarkan jumlah dokumen tiap topik. Topik dengan frekuensi tertinggi pada periode ini adalah demonstrasi mahasiswa menuntut perbaikan ekonomi (76 dokumen).

Tabel 8. Contoh Evaluasi Coherence dan Diversity Periode 30 Hari

ID Topik	Jumlah	Kata Kunci	Koherensi	Diversitas
0	76	mahasiswa, demo mahasiswa, demo, tuntut perbaikan, gedung, perbaikan ekonomi, tuntut, perbaikan, gedung dpr, ekonomi	0.9805	0.7480
1	73	taufik, taufik hidayat, hidayat, ytr, penyekapan, hidayat ditangkap, wanita bandung, hidayat pelaku, penyekap, bandung	0.8140	0.6592
2	54	harga, pertamax, harga pertamax, bbm, minyakita, minyak, pertamina, turun, solar, minyak dunia	0.6940	0.8149
3	48	persib, locker room, room, locker, persija, persib bandung, pemain, sandy walsh, walsh, lockerroom	0.6083	0.7847

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 959−971
DOI 10.47065/bits.v8i2.10951
https://ejurnal.seminar-id.com/index.php/bits

Secara keseluruhan, hasil pengujian pada keempat periode menunjukkan bahwa BERTopic berbasis IndoBERT konsisten menghasilkan nilai *coherence* pada kisaran 0,62–0,65 dan *diversity* pada kisaran 0,75–0,80 pada setiap rentang waktu, yang tergolong cukup baik dan sejalan dengan temuan penelitian sebelumnya yang melaporkan *coherence* BERTopic pada kisaran 0,46–0,55 dan *diversity* 0,71–0,76 pada berbagai domain data lain [5], [7], [9], mengindikasikan bahwa pendekatan yang diterapkan pada data berita berbahasa Indonesia menghasilkan kualitas topik yang kompetitif.

3.4 Analisis Trending Topik

Metrik *trend_score* terbukti mampu membedakan antara topik dengan volume pemberitaan besar namun sudah melewati puncaknya, dengan topik bervolume kecil yang justru sedang naik daun. Pada periode 1 hari misalnya, Topik 0 (febrie, jampidsus) memiliki volume dokumen terbanyak (58 dokumen) dapat dilihat pada Tabel 5 sebelumnya, namun mencatat skor tren terendah dan negatif (-1,25), sedangkan Topik 4 (prabowo, koperasi) dengan volume jauh lebih kecil (6 dokumen) justru mencatat skor tren tertinggi (1,00). Pola serupa terlihat pada periode 7, 14, dan 30 hari, sebagaimana dirangkum pada Tabel 9, di mana topik-topik dengan skor tren tertinggi pada umumnya bukan topik dengan volume dokumen terbanyak, melainkan topik yang mengalami lonjakan intensitas pemberitaan paling tajam pada rentang waktu tertentu.

Tabel 9. Contoh Topik dengan Skor Trend Tertinggi pada Setiap Periode

Periode	ID Topik	Kata Kunci Utama	Skor Tren
1 Hari	4	prabowo, koperasi, presiden	1,00
7 Hari	20	emas, brankas, batangan	4,50
14 Hari	22	adriansyah, febrie, jampidsus	4,50
30 Hari	86	adriansyah, febrie, jampidsus	11,00

Pada periode 30 hari, topik dengan skor tren tertinggi (11,0) adalah isu pengunduran diri eks Jampidsus Febrie Adriansyah, yang juga muncul kembali pada topik lain dengan skor tren tinggi (4,50) pada periode 14 hari, menunjukkan bahwa isu tersebut mengalami eskalasi pemberitaan yang berkelanjutan menjelang akhir periode pengamatan. Temuan ini menegaskan bahwa metrik *trend_score* berhasil menangkap dinamika laju peningkatan pemberitaan yang tidak dapat diketahui hanya dari frekuensi total kemunculan topik semata.

3.5 Rekomendasi Judul Berita

Berdasarkan hasil analisis *trending topic*, sistem menghasilkan rekomendasi judul berita untuk topik dengan skor trend tertinggi pada setiap periode, dengan kesesuaian yang baik antara kata kunci dan judul, sebagaimana dirangkum pada Tabel 10. Pada periode 1 hari, kata kunci "prabowo, koperasi, presiden" menghasilkan judul "Presiden Prabowo Hadiri Acara Hari Koperasi", selaras dengan pola serupa pada periode 7 hari. Namun, topik dengan skor trend tertinggi pada periode 14 dan 30 hari menghasilkan kata kunci maupun judul yang identik ("adriansyah, febrie, jampidsus" — "Kejagung Umumkan Febrie Adriansyah Mundur dari Jampidsus"), menandakan isu tersebut tetap dominan pada kedua periode. Temuan ini menjadi catatan bahwa mekanisme pemilihan judul representatif masih dapat disempurnakan..

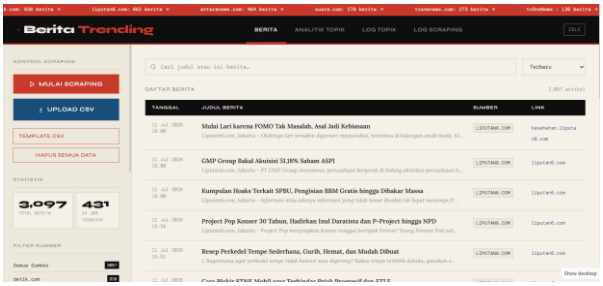
Tabel 10. Contoh Rekomendasi Judul dengan Skor Trend Tertinggi pada Setiap Periode

Periode	ID Topik	Kata Kunci	Contoh Judul Berita
1 Hari	4	prabowo, koperasi, presiden, presiden prabowo	BREAKING NEWS - Presiden Prabowo Hadiri Acara Hari Koperasi (12/07)
7 Hari	20	emas, brankas, batangan, emas batangan	Polisi Dibantu Ahli Kunci Buka Paksa Brankas saat Geledah Rumah Mewah di Sentul terkait Korupsi
14 Hari	22	adriansyah, febrie adriansyah, febrie, jampidsus	Kejagung Umumkan Febrie Adriansyah Mundur dari Jampidsus
30 Hari	86	adriansyah, febrie adriansyah, febrie, jampidsus	Kejagung Umumkan Febrie Adriansyah Mundur dari Jampidsus

3.6 Implementasi Sistem dan Pengujian

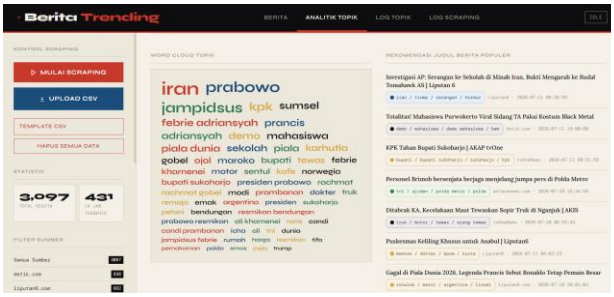
Hasil pemodelan topik diimplementasikan ke dalam sistem *dashboard* berbasis website bernama "Berita Trending" yang dibangun menggunakan *FastAPI* sebagai *backend*, *SQLite* sebagai basis data, serta *Chart.js/Plotly* untuk visualisasi. Sistem menyediakan halaman Berita untuk menampilkan daftar artikel hasil *scraping* beserta fitur pencarian dan filter sumber, halaman *Log Scraping* yang menampilkan proses pengambilan data secara *real-time*, halaman Analitik Topik yang memvisualisasikan tren pergerakan topik dari waktu ke waktu, halaman Log Topik yang mencatat tahapan pemodelan BERTopic, serta dashboard Word Cloud, Rekomendasi Judul, Evaluasi Coherence/Diversity, hasil dapat dilihat pada Gambar 6-9.

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 959−971
DOI 10.47065/bits.v8i2.10951
https://ejurnal.seminar-id.com/index.php/bits



Gambar 6. Halaman Berita

Gambar 7. Halaman Dasboard Analitik Topik



Gambar 8. WordCloud dan Rekomendasi Judul

Gambar 9. Evaluasi Coherence dan Diversity

Pengujian sistem menggunakan metode *Black-Box Testing* dilakukan terhadap 20 skenario yang mencakup seluruh alur kerja sistem, mulai dari proses *scraping* berita, pengelolaan dan penelusuran data, hingga pemodelan topik dan visualisasi hasil analitik. Tabel 11 menyajikan contoh beberapa skenario pengujian yang mewakili alur utama sistem. Seluruh 20 skenario pengujian dinyatakan berstatus "Valid", yang berarti setiap fitur menghasilkan keluaran sesuai dengan hasil yang diharapkan tanpa ditemukan kesalahan fungsional, sehingga sistem dinyatakan siap digunakan oleh tim redaksi Sumatera Ekspres untuk mendukung proses pemantauan dan pengambilan keputusan editorial berbasis data, hasil *Black-Box Testing* dapat dilihat pada Tabel 11.

Tabel 11. Skenario Pengujian Black-Box

No	Skenario Pengujian	Hasil yang Diharapkan	Hasil Pengujian	Status
1	Memulai proses scraping berita	Sistem memulai proses scraping dan mengembalikan status proses berjalan	Proses scraping berhasil dimulai	Valid
2	Memantau status & log scraping	Sistem menampilkan status terkini beserta log proses scraping	Status dan log scraping tampil sesuai proses yang berjalan	Valid
3	Menampilkan daftar berita dengan pagination, pencarian, dan filter	Sistem menampilkan daftar berita sesuai parameter yang diberikan	Daftar berita tampil sesuai pagination, pencarian, dan filter	Valid
4	Menampilkan detail satu berita	Sistem menampilkan detail lengkap berita sesuai ID	Detail berita tampil dengan benar	Valid
5	Mengunggah berita dari file CSV	Sistem membaca dan menyimpan data berita dari file CSV ke database	Data berita dari CSV berhasil tersimpan	Valid
6	Mengunduh template CSV	Sistem mengembalikan file template CSV dengan format kolom yang sesuai	File template CSV berhasil terunduh	Valid
7	Mengekspor seluruh data berita ke CSV	Sistem menghasilkan file CSV berisi seluruh data berita di database	File CSV hasil export berhasil terunduh	Valid
8	Menampilkan statistik database	Sistem menampilkan ringkasan statistik (jumlah berita, sumber, topik, dll.)	Statistik database tampil sesuai kondisi	Valid
9	Menghapus seluruh data berita	Sistem menghapus semua data berita dari database	Seluruh data berita berhasil terhapus	Valid
10	Menampilkan daftar sumber scraping	Sistem menampilkan daftar sumber berita yang digunakan untuk scraping	Daftar sumber scraping tampil dengan benar	Valid
11	Memulai pemodelan topik	Sistem memulai proses pemodelan topik terhadap data berita	Proses pemodelan topik berhasil dimulai	Valid

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 959−971
DOI 10.47065/bits.v8i2.10951
https://ejurnal.seminar-id.com/index.php/bits

No	Skenario Pengujian	Hasil yang Diharapkan	Hasil Pengujian	Status
12	Memantau status & log pemodelan topik	Sistem menampilkan status dan log proses pemodelan topik	Status dan log pemodelan topik tampil sesuai proses	Valid
13	Menampilkan daftar topik hasil pemodelan	Sistem menampilkan seluruh topik hasil pemodelan BERTopic	Daftar topik tampil dengan lengkap	Valid
14	Menampilkan detail topik dan artikel terkait	Sistem menampilkan detail topik beserta artikel yang tergabung di dalamnya	Detail topik dan artikel terkait tampil dengan benar	Valid
15	Menampilkan judul berita per topik	Sistem menampilkan daftar judul berita yang termasuk dalam topik tersebut	Daftar judul berita per topik tampil sesuai cluster	Valid
16	Menampilkan data tren topik	Sistem menampilkan data frekuensi kemunculan topik dari waktu ke waktu	Data tren topik tampil sesuai periode yang dianalisis	Valid
17	Menampilkan data word cloud	Sistem menampilkan kata kunci representatif tiap topik untuk visualisasi word cloud	Data word cloud tampil sesuai kata kunci topik	Valid
18	Menampilkan rekomendasi judul berita	Sistem menampilkan rekomendasi judul berita berdasarkan topik yang sedang trending	Rekomendasi judul berita tampil sesuai topik trending	Valid
19	Menampilkan metrik kualitas topik	Sistem menampilkan nilai coherence score dan topic diversity hasil pemodelan	Metrik kualitas topik tampil sesuai hasil evaluasi model	Valid
20	Mengunduh laporan PDF analitik topik	Sistem menghasilkan file laporan PDF berisi ringkasan analitik topik	File laporan PDF berhasil terunduh	Valid

3.7 Pembahasan

Secara umum, temuan penelitian ini sejalan sekaligus melengkapi hasil penelitian terdahulu. Dari sisi kualitas topik, nilai rata-rata *coherence score* BERTopic pada penelitian ini (0,62–0,65) tergolong lebih tinggi dibandingkan hasil Ramadhan & Fernando yang melaporkan koherensi 0,5541 pada ulasan pelanggan restoran [7] dan Muhammad Rayhan Nur dkk. yang melaporkan rata-rata koherensi 0,5437 pada unggahan platform X [9], sedangkan nilai *diversity* penelitian ini (0,75–0,80) juga tergolong baik dibandingkan kisaran yang umum dilaporkan pada penelitian berbasis komentar YouTube [5]. Perbedaan ini mengindikasikan bahwa integrasi data dari delapan portal berita yang gaya penulisannya relatif lebih formal dan terstruktur dibandingkan komentar media sosial yang informal turut berkontribusi terhadap kualitas topik yang lebih baik, meskipun keragaman gaya penulisan antarportal tetap menghasilkan variasi koherensi pada topik-topik individual (Tabel 5 sampai dengan Tabel 8).

Dari sisi metodologis, temuan mengenai efektivitas metrik *trend_score* dalam membedakan topik bervolume besar namun sudah melewati puncak pemberitaannya dari topik yang benar-benar sedang naik daun (Tabel 9) memperluas temuan Simanjuntak dkk. yang menggunakan *Dynamic Topic Modeling* untuk memetakan evolusi opini secara kronologis tanpa metrik kuantitatif serupa [6]; dengan kata lain, penelitian ini menyediakan ukuran temporal yang lebih operasional untuk kebutuhan pemantauan tren harian. Selain itu, keberhasilan integrasi hasil pemodelan topik ke dalam *dashboard* berbasis *website* yang lulus 20 skenario *black-box testing* (Tabel 11) menunjukkan kesiapan praktis yang belum ditunjukkan pada penelitian [5], [7]–[9], yang umumnya berhenti pada tahap analisis akademis tanpa disertai sistem yang siap digunakan oleh pengguna akhir. Lonjakan proporsi klaster *noise* pada periode data yang lebih panjang (30 hari) turut memperkuat karakteristik BERTopic pada pengelompokan artikel bervolume besar sebagaimana dilaporkan Aryani dkk. [8], yaitu semakin heterogen dan besar volume data, semakin banyak pula berita yang bersifat *isolated events* yang sulit membentuk klaster padat.

4. KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan yang telah diuraikan, penelitian ini berkontribusi dalam mengintegrasikan data multi-sumber yang lebih heterogen dari delapan portal berita nasional dan YouTube, mengusulkan metrik trend_score berbasis split-period comparison, serta mewujudkannya ke dalam sistem dashboard siap pakai bagi redaksi; secara teknis, implementasi algoritma *BERTopic* berbasis *IndoBERT* terbukti mampu mengekstraksi dan mengelompokkan topik berita secara otomatis dari 9.699 data yang dikumpulkan melalui *web scraping* terhadap delapan portal berita nasional dan media sosial *YouTube*, dengan kombinasi *IndoBERT* untuk *text embedding*, *UMAP* untuk reduksi dimensi, *HDBSCAN* untuk *clustering*, dan *c-TF-IDF* untuk *topic representation* menghasilkan topik yang representatif secara semantik (rata-rata *coherence score* 0,62–0,65 dan *topic diversity* 0,75–0,80 secara konsisten pada empat periode pengujian), sedangkan metrik tambahan *trend_score* yang diadaptasi dari prinsip *split-period comparison* terbukti efektif membedakan topik yang benar-benar sedang meningkat intensitas

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 959−971
DOI 10.47065/bits.v8i2.10951
https://ejurnal.seminar-id.com/index.php/bits

pemberitaannya dari topik bervolume besar namun sudah melewati puncak pemberitaannya. Hasil ini diimplementasikan ke dalam *dashboard* berbasis *website* menggunakan *FastAPI* dan *SQLite* yang telah diuji melalui 20 skenario *black-box testing* dengan status valid, sehingga secara praktis sistem ini memberi manfaat manajerial bagi redaksi Sumatera Ekspres berupa pemantauan isu populer yang lebih cepat tanpa harus membaca satu per satu berita dari delapan sumber secara manual, serta rekomendasi judul dan topik *trending* yang dapat langsung dijadikan bahan rapat redaksi harian untuk menentukan prioritas liputan. Adapun keterbatasan penelitian ini terletak pada sumber data yang masih terbatas pada portal berita *online* dan *YouTube*, sehingga penelitian selanjutnya disarankan untuk memperluas sumber data ke platform media sosial lain seperti *X*, *Instagram*, atau *TikTok*, serta mempertimbangkan implementasi pembaruan model topik secara inkremental agar sistem lebih efisien menyesuaikan data baru.

5. PERNYATAAN PENULIS

5.1 Kontribusi Penulis

Konseptualisasi: M. W., A. M., W., dan A.;
Metodologi: M. W., dan A. M.;
Perangkat Lunak: M. W.;
Validasi: M. W., A. M., W., dan A.;
Analisis Formal: M. W., A. M., W., dan A.;
Investigasi: M. W., A. M., W., dan A.;
Sumber Daya: M. W., A.M., dan W.;
Kurasi Data: M. W., A. M., W., dan A.;
Penulisan Draf Awal: M. W., dan A.M.;
Penulisan, Peninjauan, dan Penyuntingan: M. W., A. M., W., dan A.;
Visualisasi: M. W., A. M., W., dan A.;
Seluruh penulis telah membaca dan menyetujui versi manuskrip yang telah diterbitkan.

5.2 Ketersediaan Data

Data yang disajikan dalam penelitian ini tersedia berdasarkan permintaan kepada penulis korespodensi.

5.3 Pendanaan

Penelitian ini tidak menerima pendanan eksternal.

5.4 Pernyataan Dewan Peninjau Institusional

Tidak Berlaku (Not applicable).

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak Berlaku (Not applicable).

5.6 Pernyataan Konflik Kepentingan

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan keuangan yang bersaing atau hubungan pribadi yang diketahui yang dapat dianggap memengaruhi pekerjaan yang dilaporkan dalam artikel ini.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan

Para penulis menggunakan AI Generatif untuk meningkatkan kejelasan penulisan dalam artikel ini. Para penulis telah meninjau dan menyunting konten yang dihasilkan dengan bantuan AI serta bertanggung jawab penuh atas versi akhir publikasi.

REFERENCES

[1] A. Khaer, N. Khoir, and Y. Arini Hidayati, "Senjakala Media Cetak: Tantangan Jurnalisme Cetak di Era Digital," *TRILOGI: Jurnal Ilmu Teknologi, Kesehatan, dan Humaniora*, vol. 2, no. 3, pp. 324–331, Nov. 2021, doi: https://doi.org/10.33650/trilogi.v2i3.3080.

[2] N. A. Rakhmawati, A. Cisatra, D. D. M. Ansori, D. N. F. A. Akmal, and S. Ramadhani, "Identifikasi Topik Hangat di Media Berita Menggunakan Latent Dirichlet Allocation," *Journal of Information Engineering and Educational Technology*, vol. 8, no. 1, pp. 14–17, Jun. 2024, doi: 10.26740/jieet.v8n1.p14-17.

[3] E. Puspita, D. F. Shiddieq, and F. F. Roji, "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, Feb. 2024, doi: 10.57152/malcom.v4i2.1204.

[4] A. Gaviota and F. Ardiani, "Pemodelan Topik Menggunakan BERTopic Pada Nama Produk Template Desain di Peterdraw Studio," *CESS (Journal of Computer Engineering, System and Science)*, vol. 11, no. 1, pp. 27–40, Jan. 2026, doi: 10.24114/cess.v11i1.70335.

[5] W. Wahyuni, T. P. Lestari, M. Apriliana, and R. Gumelta, "Implementation of BERTopic for Topic Modeling Analysis of the Free Nutritious Meal Program Based on YouTube Comments," *Journal of Applied Informatics and Computing*, vol. 9, no. 4, pp. 1964–1971, Aug. 2025, doi: 10.30871/jaic.v9i4.9754.

[6] Kristine Angelina Simanjuntak, Muhamad Koyimatu, and Yolla Putri Ervanisari, "Analisis Perubahan Opini Publik Terhadap Kendaraan Listrik di Indonesia Melalui Komentar YouTube: Pendekatan Topic Modeling BERTopic," *Jurnal Inovasi Kewirausahaan*, vol. 1, no. 3, pp. 1–9, Oct. 2024, doi: 10.37817/jurnalinovasikewirausahaan.v1i3.3789.

[7] A. Ramadhan and E. Fernando, "Analisis Sentimen Berbasis Aspek pada Ulasan Pelanggan Menggunakan Pemodelan Latent Dirichlet Allocation dan BERTopic," *Musytari : Jurnal Manajemen, Akuntansi, dan Ekonomi*, vol. 20, no. 5, Jul. 2025, doi: https://doi.org/10.2324/q2e09q23.

[8] D. Aryani, I. Lucia Kharisma, A. Sujjada, and K. Kamdan, "Topic Modeling of the 2024 Election Using the BERTopic Method on Detik.com News Articles," *Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 9, no. 2, pp. 171–180, Aug. 2024, doi: 10.25139/inform.v9i2.8429.

[9] Muhammad Rayhan Nur, Yudi Wibisono, and Rani Megasari, "Analisis Sentimen dan Pemodelan Topik pada Post tentang Merek Teknologi di X Menggunakan Fine-tuning IndoBERT dan BERTopic," *Jurnal Komputer Teknologi Informasi Sistem Informasi (JUKTISI)*, vol. 4, no. 2, pp. 743–750, Jul. 2025, doi: 10.62712/juktisi.v4i2.508.

[10] C. J. L. Tobing, IGN Lanang Wijayakusuma, and Luh Putu Ida Harini, "Perbandingan Kinerja IndoBERT dan MBERT Untuk Deteksi Berita Hoaks Politik dalam Bahasa Indonesia," *JST (Jurnal Sains dan Teknologi)*, vol. 14, no. 1, pp. 114–123, May 2025, doi: 10.23887/jstundiksha.v14i1.92126.

[11] A. Pambudi, "Penerapan CRISP-DM Menggunakan MLR K-Fold pada Data Saham PT. Telkom Indonesia (Persero) Tbk (TLKM) (Studi Kasus: Bursa Efek Indonesia Tahun 2015–2022)," *Jurnal Data Mining dan Sistem Informasi*, vol. 4, no. 1, p. 1, Mar. 2023, doi: 10.33365/jdmsi.v4i1.2462.

[12] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *Proceedings of the 28th International Conference on Computational Linguistics*, vol. 1, no. 1, pp. 757–770, Nov. 2020, doi: 10.18653/v1/2020.coling-main.66.

[13] C. Ramadhan, V. Atina, and H. Permatasari, "Analisis Perbandingan Model CNN dan IndoBERT Dalam Sentimen Berita Politik Indonesia," *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis*, vol. 1, no. 1, pp. 110–118, Jul. 2025, doi: 10.47701/v1r9ka69.

[14] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *J. Open Source Softw.*, vol. 3, no. 29, p. 861, Sep. 2020, [Online]. Available: http://arxiv.org/abs/1802.03426

[15] G. G. Ghiffary, K. Alifviansyah, A. Fitrianto, E. Erfiani, and L. M. R. D. Jumansyah, "Perbandingan Algoritma HDBSCAN dan Agglomerative Hierarchical Clustering dalam Klasterisasi pada Data yang Mengandung Pencilan," *Jurnal Riset dan Aplikasi Matematika (JRAM)*, vol. 8, no. 2, pp. 122–135, Oct. 2024, doi: 10.26740/jram.v8n2.p122-135.

[16] N. Perdana and H. Santoso, "Klasterisasi Judul Berita Online Isu Pemilu Prabowo Subianto dengan Kombinasi LLMS Embedding Dengan HDBSCAN," *Jurnal Locus Penelitian dan Pengabdian*, vol. 4, no. 10, pp. 9284–9298, Oct. 2025, doi: 10.58344/locus.v4i10.4779.

[17] F. D. Handayani and Isnaini Rosyida, "Clustering Review Pengguna Aplikasi Zenius pada Layanan Google Play Store Menggunakan Metode DBSCAN dan HDBSCAN," *Emerging Statistics and Data Science Journal*, vol. 1, no. 2, pp. 178–191, May 2023, doi: 10.20885/esds.vol1.iss.2.art19.

[18] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," Mar. 2022, doi: 10.48550/arXiv.2203.05794.

[19] R. A. Kurniawan, P. Purwantoro, and I. Maulana, "Pemodelan Topik Ulasan Pengguna Honkai Star Rail Menggunakan Bertopic Berbasis Indobert," *Blend Sains Jurnal Teknik*, vol. 4, no. 4, pp. 872–881, Apr. 2026, doi: 10.56211/blendsains.v4i4.1534.

[20] N. C. Sastya and I. Nugraha, "Penerapan Metode CRISP-DM dalam Menganalisis Data untuk Menentukan Customer Behavior di MeatSolution," *UNISTEK*, vol. 10, no. 2, pp. 103–115, Oct. 2023, doi: 10.33592/unistek.v10i2.3079.

[21] Y. A. Hafiz and E. Sudarmilah, "Implementasi Web Scraping pada Portal Berita Online," *Inisiasi*, vol. 12, no. 1, pp. 55–60, Nov. 2023, doi: 10.59344/inisiasi.v12i1.120.

[22] M. R. Fikri, R. T. Handayanto, and D. Irwan, "Web Scraping Situs Berita Menggunakan Bahasa Pemograman Python," *Journal of Students' Research in Computer Science*, vol. 3, no. 1, pp. 123–136, May 2022, doi: 10.31599/jsrcs.v3i1.1514.

[23] A. Z. Rizquina and C. I. Ratnasari, "Implementasi Web Scraping untuk Pengambilan Data Pada Website E-Commerce," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 5, no. 4, pp. 377–383, Oct. 2023, doi: 10.47233/jteksis.v5i4.913.

[24] Runimeirati, Abdul Muis, and Figur Muhammad, "Pelatihan Text Mining Menggunakan Bahasa Pemrograman Python," *Abdimas Langkanae*, vol. 3, no. 1, pp. 36–46, Jan. 2023, doi: 10.53769/abdimas.3.1.2023.83.

[25] S. Saikin, S. Fadli, J. Akbar, and H. Fahmi, "Topic Modeling Judul Penelitian Menggunakan Metode Latent Dirichlet Allocation (LDA) dan Faktorisasi Matriks Non-Negatif (NMF)," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 2, pp. 3439–3445, Apr. 2025, doi: 10.36040/jati.v9i2.12998.

[26] C. Naury, D. H. Fudholi, and A. F. Hidayatullah, "Topic Modelling pada Sentimen Terhadap Headline Berita Online Berbahasa Indonesia Menggunakan LDA dan LSTM," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 1, p. 24, Jan. 2021, doi: 10.30865/mib.v5i1.2556.

