JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 13 No. 3S1, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v13i3S1.8041

IMPLEMENTASI ETL DAN TOPIC MODELING MENGGUNAKAN ALGORITMA LATENT DIRICHLET ALLOCATION UNTUK IDENTIFIKASI TOPIK KRIMINAL PADA BERITA ONLINE

Ahista Tasya Kamila[1], Bintang Ary Pradana[2], Ruth Tika Sarwanti[3], Ririn Medistarani[4], Chaerur Rozikin[5].

[1,2]Universitas Singaperbangsa Karawang; Jl. HS. Ronggowaluyo, Telukjambe Timur, Karawang - 41363; 0812 - 1866 - 9229
[3,4,5]Universitas Singaperbangsa Karawang; Jl. HS. Ronggowaluyo, Telukjambe Timur, Karawang - 41363; 0812 - 1866 - 9229

Keywords:
Extract, Tranform, Load (ETL), Latent Dirichlet Allocation (LDA), Berita Kriminal, Web Scraping.

Corespondent Email:
2210631170043@student.unsika.ac.id

Abstrak. Peningkatan angka kriminalitas di Indonesia berdampak pada semakin masifnya pemberitaan kejahatan di media daring. Data berita yang tidak terstruktur memerlukan pendekatan sistematis agar dapat dianalisis secara efektif. Penelitian ini bertujuan untuk mengimplementasikan proses Extract, Transform, Load (ETL) dan metode topic modeling menggunakan algoritma Latent Dirichlet Allocation (LDA) untuk mengidentifikasi topik kriminal pada artikel berita online, khususnya dari portal Detik.com. Data dikumpulkan melalui teknik web scraping, kemudian diproses melalui tahapan transformasi untuk pembersihan dan standarisasi, serta dimuat ke dalam basis data agar lebih terorganisasi. Selanjutnya, dilakukan text preprocessing dan representasi teks menggunakan Bag of Words sebelum dimodelkan dengan LDA. Hasil penelitian menunjukkan bahwa dari 4.105 artikel kriminal, diperoleh 7 topik utama, yaitu pencurian dan kekerasan fisik, kejahatan seksual dan kekerasan anak, kejahatan politik dan pelanggaran HAM, kekerasan bersenjata dan separatisme, kejahatan finansial dan narkotika, pembunuhan dan penganiayaan, serta pelanggaran hukum dan penyalahgunaan wewenang. Analisis tren menunjukkan dua topik dominan sepanjang periode penelitian, yakni pencurian serta kejahatan seksual. Temuan ini menegaskan pentingnya integrasi ETL dan LDA untuk memahami pola kriminalitas secara sistematis, serta dapat menjadi dasar pengambilan kebijakan berbasis data.

Abstract. *he increasing crime rate in Indonesia has led to a surge in online news reports on criminal cases. Since news articles are unstructured, a systematic approach is required to enable effective analysis. This study aims to implement the Extract, Transform, Load (ETL) process combined with topic modeling using the Latent Dirichlet Allocation (LDA) algorithm to identify dominant criminal topics in online news articles, specifically from Detik.com. Data were collected through web scraping, then transformed for cleaning and standardization before being loaded into a database to ensure better organization. The texts were preprocessed through normalization, stopword removal, tokenization, and stemming, and then represented using the Bag of Words model prior to topic modeling with LDA. The results from 4,105 crime-related articles revealed seven main topics: theft and physical violence, sexual crimes and child abuse, political crimes and human rights violations, armed violence and separatism, financial crimes and narcotics, murder and assault,*

as well as legal violations and abuse of power. Trend analysis indicated that theft and sexual crimes consistently dominated the news throughout the study period. These findings highlight the significance of integrating ETL with LDA to systematically identify patterns in criminal news reporting and provide valuable insights for evidence-based policy making.

1. PENDAHULUAN

Dokumen Kriminalitas merupakan salah satu isu serius yang memiliki dampak langsung terhadap rasa aman masyarakat. Fenomena ini tidak hanya sebatas pelanggaran hukum, tetapi juga merupakan konsekuensi dari persoalan yang bersifat kompleks dan melibatkan berbagai dimensi. Beberapa faktor diantaranya yaitu tingginya tingkat pengangguran, rendahnya rata-rata lama sekolah, banyaknya jumlah penduduk miskin, hingga kondisi perekonomian daerah yang tercermin dari Produk Domestik Regional Bruto yang terbukti berkorelasi dengan tingginya angka kriminalitas di Indonesia [1] . Oleh sebab itu, permasalahan ini tidak dapat dipandang sederhana, karena melibatkan interaksi antara sosial, ekonomi, dan hukum secara bersamaan. Berdasarkan data Badan Pusat Statistik, jumlah kejadian kejahatan di Indonesia meningkat dari 372.965 kasus pada tahun 2022 menjadi 584.991 kasus pada tahun 2023. Peningkatan ini juga tercermin dari naiknya tingkat risiko masyarakat menjadi korban kejahatan (*crime rate*) dari 137 menjadi 214 per 100.000 penduduk, atau setara dengan satu kejadian kejahatan setiap 53 detik (BPS, 2024). Sejalan dengan itu, laporan Global Organized Crime Index 2023 yang dirilis melalui GoodStats menempatkan Indonesia sebagai negara dengan tingkat kriminalitas tertinggi kedua di ASEAN dengan skor 6,85, jauh di atas rata-rata global sebesar 5,03 Peningkatan angka kriminalitas membuat pemberitaan kejahatan di portal media daring akan semakin masif . Portal berita online memudahkan masyarakat mengakses informasi dari berbagai lokasi dan tanpa batasan waktu [2], sehingga berita kriminal tersebar dengan cepat dan dalam jumlah besar. Karakteristik data berita yang tidak terstruktur dan terus berubah membuat pengolahan manual menjadi sulit, sehingga diperlukan pendekatan sistematis untuk menyusun informasi secara lebih terstruktur sebelum dilakukan analisis lebih lanjut [3].

Untuk menangani besarnya volume data yang bersifat tidak terstruktur, salah satu pendekatan yang dapat dilakukan adalah dengan Extract, Transform, Load (ETL). Dalam tahap ekstraksi, dapat digunakan teknik *Web scraping*, yang berfungsi sebagai metode otomatisasi untuk mengumpulkan data dan menyesuaikan format agar lebih rapi dan konsisten. Data yang telah dikumpulkan kemudian dibersihkan dan diubah formatnya sehingga siap dimasukkan ke basis data untuk analisis lebih lanjut [4]. Setelah data berita tersusun secara terstruktur, pola dan tema dapat diekstrak menggunakan *topic modeling*, yang mengelompokkan topik secara otomatis sehingga tren dan isu utama terlihat jelas tanpa perlu memeriksa setiap artikel satu per satu. Salah satu algoritma yang populer dan efektif digunakan adalah *Latent Dirichlet Allocation* (LDA), yang merepresentasikan setiap dokumen sebagai campuran topik tersembunyi dengan distribusi kata-kata khas dan menggunakan Gibbs sampling untuk menghitung probabilitas topik dalam dokumen [5]. Penerapan LDA pada studi sebelumnya menghasilkan nilai koherensi yang baik, yaitu 0,7586 pada penelitian Matira et al., (2022) dan 0,53 pada penelitian Fahlevvi & Azhari, (2022).

Berdasarkan permasalahan tersebut, penelitian ini berfokus pada penerapan proses ETL dan *topic modeling* untuk mengelola dan menganalisis artikel berita kriminal di Indonesia, khususnya dari portal Detik.com. Data akan diproses secara terstruktur melalui tahapan ETL untuk menjamin kualitas dan konsistensi, kemudian dianalisis menggunakan algoritma LDA guna mengidentifikasi topik-topik dominan. Pendekatan ini akan memberikan pemahaman yang jelas mengenai tren dan pola kriminalitas yang diberitakan media daring, sekaligus menjadi dasar yang kuat untuk pengambilan keputusan dan perumusan kebijakan keamanan berbasis data.

2. TINJAUAN PUSTAKA

2.1 Berita Online

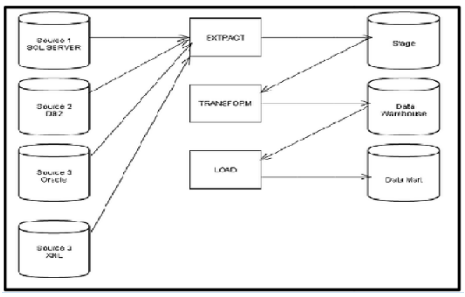
Berita online merupakan laporan kejadian yang disebarkan lewat internet dalam bentuk teks dan gambar. Berita online menyajikan informasi yang sama seperti berita cetak, tetapi dikemas ulang secara digital tanpa mengubah makna inti [6]. Berita online umumnya tersimpan dalam bentuk data semi-terstruktur, yaitu memiliki elemen yang terorganisir seperti judul, penulis, dan tanggal, serta bagian isi yang cenderung tidak terstruktur [7]. Karakteristik ini menjadikan data berita dapat diolah lebih lanjut dengan berbagai pendekatan analisis teks.

2.2 Kriminalitas

Kriminalitas merupakan perbuatan yang bertentangan dengan hukum negara dan mendapatkan penolakan dari masyarakat. Tindakan ini mencakup berbagai bentuk, seperti pembunuhan, perampokan, pemerkosaan, dan penipuan. Suatu perbuatan dapat dikategorikan sebagai kejahatan apabila terdapat niat jahat (*mens rea*) dan tindakan nyata (*actus reus*) yang menimbulkan kerugian bagi pihak lain. Pada dasarnya, tidak ada individu yang terlahir sebagai penjahat, melainkan kondisi sosial, ekonomi, biologis, dan psikologis yang mendorong seseorang melakukan tindakan kriminal [8]. Faktor-faktor penyebab kriminalitas ini meliputi ketimpangan ekonomi, pengangguran, kepadatan penduduk, kurangnya pendidikan, tekanan mental, dan trauma psikologis. Urbanisasi yang cepat tanpa dukungan infrastruktur memadai serta efektivitas penegakan hukum juga memengaruhi kelangsungan perilaku kriminal dalam masyarakat [9].

2.3 Extract, Transform, Load (ETL)

ETL (Extract, Transform, Load) merupakan rangkaian aktivitas yang digunakan dalam data warehousing untuk mengintegrasikan data dari berbagai sistem sumber ke dalam data warehouse. Proses ini mencakup extraction data dari sumber, transformation melalui pembersihan, pemfilteran, serta validasi, dan loading ke dalam repositori data. Dengan ETL, proses integrasi data dapat berlangsung lebih terstruktur sehingga mendukung analisis dan pelaporan secara optimal [10].



Gambar 2. 1 Alur ETL

(Sumber: Mali & Bojewar, 2015)

2.4 Web Scraping

Web scraping merupakan teknik dalam pengumpulan data dari internet melalui program otomatis yang mengakses server web, mengambil data dalam bentuk HTML atau file penyusun halaman, kemudian mengolahnya untuk mengekstrak informasi yang dibutuhkan. Aktivitas ini sebelumnya dikenal dengan istilah *screen scraping*, *data mining*, atau *web harvesting*, namun saat ini lebih umum disebut *web scraping*, sedangkan program yang menelusuri banyak halaman disebut *web crawlers* atau *bots*. Dengan cakupan yang luas, web scraping menjadi salah satu fondasi penting dalam pengolahan data [11].

2.5 Topic Modeling

Topic modeling merupakan metode dalam *text mining* yang digunakan untuk menemukan dan mengelompokkan topik dalam dokumen berdasarkan pola kata dan kesamaan temanya. Teknik ini menggunakan pendekatan statistik dan pembelajaran mesin untuk mengekstrak informasi dari kumpulan data teks tanpa memerlukan anotasi semantik, sehingga memungkinkan pengorganisasian, pemahaman, dan peringkasan dokumen secara efisien. Setiap dokumen dianggap terdiri dari campuran beberapa topik, sehingga *topic modeling* membantu mengidentifikasi tema utama dan pola tersembunyi dalam teks [12].

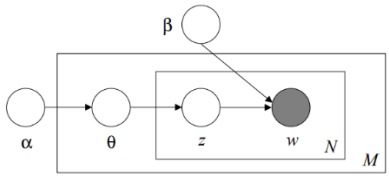
2.6 Bag of Words

Bag of words atau juga dikenal dengan BoW adalah representasi teks yang menyajikan setiap dokumen sebagai vektor berisi frekuensi kemunculan kata dalam kosakata, tanpa memperhatikan tata bahasa maupun urutan kata. Model ini hanya

menghitung jumlah kemunculan tiap kata, sehingga informasi yang digunakan terbatas pada kata itu sendiri. Pendekatan BoW ini terbukti efektif dalam berbagai tugas klasifikasi teks, karena kata-kata tertentu dapat menjadi prediktor kuat terhadap label tertentu [13].

2.7 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) adalah model probabilistik generatif yang digunakan untuk merepresentasikan kumpulan teks atau *corpus*. Setiap dokumen dipandang sebagai campuran acak dari sejumlah topik laten, sementara setiap topik dijelaskan oleh distribusi kata tertentu. Berbeda dengan metode pengelompokan tradisional yang menghubungkan dokumen pada satu topik saja, LDA memungkinkan satu dokumen terkait dengan beberapa topik sekaligus [14].



Gambar 2. 2 Representasi model grafis LDA

(Sumber: Blei, Ng, & Jordan, 2003)

$$\frac{jumlah\ topik\ \boldsymbol{K}\ \ pada\ dokumen\ \boldsymbol{D} + alpha}{panjang\ dokumen\ \boldsymbol{D} + jumlah\ topik * alpha}$$

Munculan topik tersebut pada dokumen yang bersangkutan. Nilai probabilitas tersebut dapat diperoleh melalui rumus tertentu dengan parameter:

K : Indeks Topik
D : Indeks Dokumen
α : Parameter Dirichlet (0,1)

Parameter alpha (α) berfungsi untuk mengatur banyaknya topik yang muncul dalam sebuah dokumen. Semakin besar nilai alpha, semakin banyak pula topik yang tercakup dalam dokumen tersebut (Medea et al., 2022).

2.8 Topic Coherence

Topic coherence merupakan metrik untuk menilai kualitas suatu topik berdasarkan kesamaan makna antar kata yang terkandung di dalamnya. Ukuran ini berguna untuk membedakan topik yang bermakna secara semantik dengan topik yang hanya terbentuk secara statistik. Semakin tinggi *coherence score* yang diperoleh, semakin baik kualitas model topik yang dihasilkan. Rumus perhitungan *coherence score* dituliskan sebagai berikut [15].

Rumus perhitungan *coherence score* dituliskan sebagai berikut:

$$score(v_i, v_j + \epsilon) = \log \frac{D\,(v_i, v_j) + \epsilon}{D\,(v_j)}$$

Keterangan:

v_i, v_j = kata dalam topik.
$D\,(v_i, v_j)$ = jumlah dokumen yang memuat kata v_i dan v_j.
$D\,(v_j)$ = jumlah dokumen yang memuat kata v_j.
ϵ = konstanta untuk memastikan hasil bernilai positif.

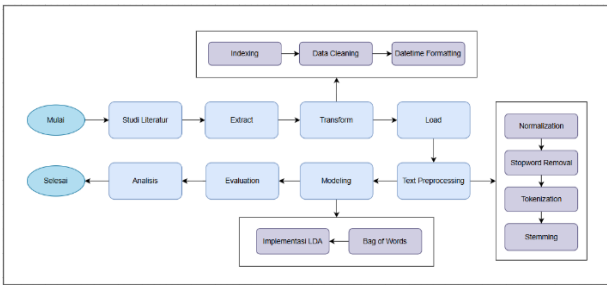
2.9 Penelitian Terdahulu Mengenai Berita Online Kriminal

Beberapa penelitian sebelumnya telah memanfaatkan teknik pengolahan teks dalam konteks yang berbeda. Satriajati dkk. (2020) meneliti pemberitaan kriminal pada masa pandemi COVID-19 dengan menggunakan *web scraping* untuk mengumpulkan data daring. Hasil penelitian menunjukkan adanya perubahan frekuensi dan jenis kasus kriminal yang diberitakan. Akan tetapi, penelitian ini masih terbatas pada tahap ekstraksi data dan belum melibatkan transformasi teks maupun pemodelan topik.

Berbeda dengan penelitian tersebut, Medea dkk. (2024) menerapkan Latent Dirichlet Allocation (LDA) pada data berita online yang juga diperoleh melalui *web scraping*, dengan analisis yang difokuskan pada *headline* berita. Hasilnya menunjukkan bahwa LDA mampu mengidentifikasi pola tema dominan dengan baik. Namun, penggunaan headline sebagai dasar analisis memiliki keterbatasan karena judul berita cenderung ringkas, berpotensi *clickbait*, dan tidak sepenuhnya merepresentasikan isi berita yang lebih informatif.

Sementara itu, penelitian lain oleh Tanasescu dkk. (2021) menyoroti peran penting proses *Extract, Transform, Load* (ETL) dalam analisis *text mining*. ETL berfungsi mengekstraksi data dari berbagai sumber, merapikan data tidak terstruktur, dan memuatnya dalam format siap dianalisis. Hasilnya menunjukkan bahwa kualitas pra-pengolahan data sangat berpengaruh terhadap kualitas analisis teks. Dengan demikian, kajian yang mengintegrasikan ETL dengan pemodelan topik pada pemberitaan kriminal masih jarang dilakukan dan perlu dikembangkan lebih lanjut.

3. METODE PENELITIAN



Gambar 3. 1 Rancangan Penelitian

3.1. Studi Literatur

Tahap ini bertujuan untuk memperdalam pemahaman mengenai penelitian-penelitian sebelumnya yang membahas penerapan ETL dan *text mining* pada artikel berita, khususnya yang berkaitan dengan isu kriminal. Selain itu, studi literatur juga mencakup kajian mengenai *text preprocessing*, metode evaluasi seperti *coherence score*, serta algoritma *Latent Dirichlet Allocation* (LDA). Berbagai sumber digunakan sebagai acuan, baik dari penelitian terdahulu maupun dokumentasi pustaka Python yang relevan.

3.2. Extract

Pada tahap ini, data dikumpulkan dari sumber berita yang relevan yaitu Detik.com. Pengambilan data dilakukan menggunakan teknik *web scraping* sehingga diperoleh informasi penting seperti judul, tanggal publikasi, isi berita, serta tautan artikel. Data yang terkumpul kemudian menjadi dasar untuk proses pengolahan lebih lanjut.

3.3. Transform

Tahap ini bertujuan untuk menyiapkan data hasil ekstraksi agar lebih bersih dan konsisten. Proses ini meliputi penghapusan elemen yang tidak relevan, penyeragaman format, serta pengorganisasian data ke dalam struktur yang lebih teratur. Dengan demikian, data menjadi siap digunakan untuk tahap selanjutnya.

3.4. Load

Data yang telah melalui proses transformasi disimpan ke dalam sistem basis data. Penyimpanan ini dilakukan agar data dapat terkelola dengan baik, mudah diakses, serta aman untuk keperluan analisis lanjutan.

3.5. Text Prepocessing

Tahap ini berfokus pada persiapan teks agar dapat diolah dengan metode analisis topik. Proses yang dilakukan mencakup penyeragaman bentuk kata (*normalization*), penghapusan kata umum yang tidak bermakna (*stopword removal*), pemisahan kata (*tokenization*), dan pengembalian kata ke bentuk dasar (*stemming*). Hasil akhir tahap ini adalah teks bersih yang siap diubah menjadi representasi numerik.

3.6. Modeling

Tahap ini diawali dengan pembentukan representasi numerik dari teks melalui metode *Bag of Words* (BoW). Dari hasil tokenisasi, dibuat sebuah *dictionary* yang berisi seluruh kosakata unik dalam korpus, kemudian dibentuk *corpus* berupa representasi dokumen dalam bentuk frekuensi kata berdasarkan *dictionary* tersebut. Representasi inilah yang menjadi masukan utama bagi algoritma topik.

Selanjutnya, digunakan algoritma *Latent Dirichlet Allocation* (LDA) untuk memodelkan distribusi topik pada kumpulan berita. Dari proses ini dihasilkan distribusi topik pada tiap dokumen serta kata-kata yang paling merepresentasikan masing-masing topik, sehingga isu-isu dominan dapat diidentifikasi secara sistematis.

3.7. Evulation

Untuk memastikan kualitas topik yang dihasilkan, dilakukan evaluasi menggunakan *topic coherence*. Nilai ini mengukur keterkaitan antar kata dalam satu topik. Selain itu, ditentukan pula jumlah topik terbaik, termasuk visualisasi hasil topiknya. Tahap ini dilakukan untuk menjamin topik yang dihasilkan relevan dan mudah diinterpretasikan.

3.8. Analisis

Setelah model terbaik diperoleh analisis dilakukan dengan meninjau topik dominan yang muncul pada artikel berita. Hasil analisis divisualisasikan agar lebih mudah dipahami, baik dalam bentuk grafik distribusi maupun analisis tren dari data yang tersedia. Dengan cara ini, isu-isu kriminal yang berkembang dapat dipetakan dan dianalisis secara lebih mendalam.

4. HASIL DAN PEMBAHASAN

4.1 Studi Literatur

Berdasarkan studi literatur, penerapan *topic modeling* dengan algoritma Latent Dirichlet Allocation (LDA) telah banyak dilakukan pada dokumen teks seperti berita, artikel ilmiah, dan media sosial. Penelitian sebelumnya menunjukkan bahwa LDA mampu menemukan pola tersembunyi

dalam dokumen tanpa anotasi semantik, sehingga cocok diterapkan pada berita kriminal yang bersifat tidak terstruktur. Dengan mengintegrasikan tahapan ETL (Extract, Transform, Load), proses pengolahan data menjadi lebih terstruktur dan siap untuk analisis topik.

4.2 Extract

Data berita dikumpulkan dari portal detik.com pada periode 25 Maret - 25 September 2025 dengan metode *web scraping* untuk mengekstrak data menggunakan Python. Total data yang berhasil diperoleh sebanyak 4000++ artikel dengan atribut link, title, location, datetime, dan content.

Tabel 4. 1 Contoh Data Hasil Ekstrak

link	title	datetime	location	content
https://www.detik.com/jabar/hukum-dan-kriminal/d-8129247/...	Jejak Eksekutor Demo Ricuh di Bandung Ditelusuri Polda Jabar	Kamis, 25 Sep 2025 10:30 WIB	Bandung	Direktorat Reserse Kriminal Umum (Ditreskrimum) Polda Jabar masih mendalami...
https://www.detik.com/jabar/hukum-dan-kriminal/d-8130057/...	Kedok Tugas Rahasia Negara Pembobol Rekening Dormant Rp 204 M	Kamis, 25 Sep 2025 15:37 WIB	Jakarta	Sembilan tersangka pembobol rekening dormant senilai Rp 204 miliar pada bank....
...	....	...	...	...
https://www.detik.com/bali/hukum-dan-kriminal/d-7840146/...	Mengenang Rosalina, Guru Muda Asal NTT Korban Penyerangan OPM di Yahukimo	Selasa, 25 Mar 2025 21:24 WIB	Flores Timur	Sebanyak tiga oknum anggota TNI diperiksa terkait kasus penjualan dan...
https://www.detik.com/sulsel/hukum-dan-kriminal/d-7840066/...	6 Fakta Brutalnya KKB di Yahukimo Bunuh Guru gegara Permintaan Uang Ditolak	Selasa, 25 Mar 2025 08:00 WIB	Yahukimo	Gerombolan anggota kelompok kriminal bersenjata (KKB) membunuh seorang guru di...

4.3 Transform

Tahap transformasi dilakukan untuk menyiapkan data agar bersih, konsisten, dan memiliki identitas unik. Proses ini mencakup pembuatan kolom id menggunakan hash dari link artikel sebagai kunci utama, case folding untuk menyeragamkan teks, pembersihan karakter non-alfabet, tanda baca, dan elemen HTML, penghapusan duplikat serta baris kosong, serta standarisasi format tanggal. Dengan tahapan ini, data menjadi lebih terstruktur dan siap dimuat pada tahap berikutnya.

Tabel 4. 2 Contoh Data Hasil Transform

id	link	title	datetime	location	content
04414afaa2ae2be6c41513bb0bf52393	https://news.detik.com/berita/d-8130788/...	jejak eksekutor demo ricuh di bandung ditelusuri polda jabar	2024-09-25	bandung	direktorat reserse kriminal umum ditreskrimum polda jabar masih mendalami...
d842f29c55f7ba3c5f85ee91e1e50a49	https://www.detik.com/sumut/hukum-dan-kriminal/d-8130772/...	kedok tugas rahasia negara pembobol rekening dormant rp m	2024-09-25	jakarta	sembilan tersangka pembobol rekening dormant senilai rp miliar pada bank...
...	...	...	...	...	...
6905fbc6c1a72d7ad93bb95ef3db1782	https://www.detik.com/bali/hukum-dan-kriminal/d-7840146/...	mengenang rosalina guru muda asal ntt korban penyerangan opm di yahukimo	2024-03-25	flores timur	sebanyak tiga oknum anggota tni diperiksa terkait kasus penjualan dan...
513bb46c8c67d7af285b3e6695dd7251	https://www.detik.com/sulsel/hukum-dan-kriminal/d-7840066/...	fakta brutalnya kkb di yahukimo bunuh guru gegara permintaan uang ditolak	2024-03-25	yahukimo	gerombolan anggota kelompok kriminal bersenjata kkb membunuh seorang guru...

4.4 Load

Proses load dilakukan dengan memindahkan data hasil transformasi ke dalam *database* "artikel_berita" pada tabel "data_artikel". Data disusun ulang sesuai urutan kolom yang telah ditetapkan, kemudian dimuat ke PostgreSQL, dengan mekanisme yang hanya menambahkan baris baru berdasarkan identitas unik, sementara apabila tabel belum tersedia maka sistem secara otomatis membuat tabel baru dan mengisinya dengan seluruh data.

Tabel 4. 3 Hasil Load data Ke Database

4.6 Text Preprocessing

Tahap *text preprocessing* dilakukan untuk mereduksi kompleksitas teks sehingga informasi lebih konsisten, relevan, dan siap digunakan. Penelitian ini menggunakan variabel *location*, *date*, dan *content* untuk analisis lanjutan, dengan variabel *content* yang perlu melalui tahapan *text preprocessing* terlebih dahulu untuk mengekstraksi topik. Ringkasan tahapan preprocessing dan contoh hasilnya ditunjukkan pada Tabel 4.4 sebagai berikut.

Tabel 4. 4 Contoh Hasil *Text Preprocessing*

Tahap	Hasil
Teks Asli	sopir yang membawa kabur mobil berisi uang rp miliar milik bank tempat ia bekerja menghabiskan sekitar rp juta dalam sepekan polisi pun mengungkap…
Normalisasi	sopir yang membawa kabur mobil berisi uang rupiah miliar milik bank tempat ia bekerja menghabiskan sekitar rupiah juta dalam sepekan polisi pun mengungkap…
Stopword Removal	sopir membawa kabur mobil berisi uang rupiah miliar bank bekerja menghabiskan rupiah juta sepekan polisi mengungkap…
Tokenisasi	['sopir', 'membawa', 'kabur', 'mobil', 'berisi', 'uang', 'rupiah', 'miliar', 'bank', 'bekerja', 'menghabiskan', 'rupiah', 'juta', 'sepekan', 'polisi', …]
Stemming	['sopir', 'bawa', 'kabur', 'mobil', 'isi', 'uang', 'rupiah', 'miliar', 'bank', 'kerja', 'habis', 'rupiah', 'juta', 'pekan', 'polisi', 'ungkap', …]

4.5 Modeling

Pemodelan topik dilakukan menggunakan algoritma Latent Dirichlet Allocation (LDA) pada kumpulan artikel berita kriminal. Berbagai model diuji dengan jumlah topik berbeda (2–20) untuk mengevaluasi distribusi kata terhadap topik dan distribusi topik terhadap dokumen. Parameter iterasi (passes) ditetapkan sebanyak 20 dan alpha otomatis digunakan untuk meningkatkan kualitas model.

Tabel 4.5 menyajikan 10 kata paling dominan dari tiap topik sebagai representasi awal hasil pemodelan. Kata-kata ini memberikan gambaran tentang pola kata yang membentuk setiap topik, namun belum disertai interpretasi atau label topik.

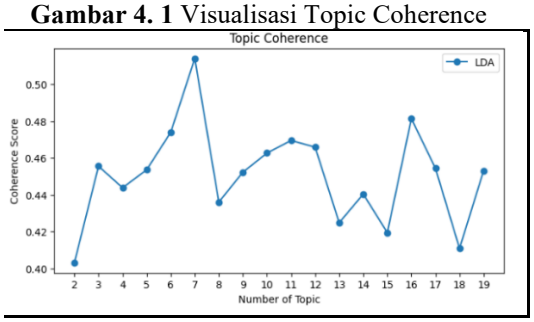
Tabel 4. 5 Contoh 10 Kata Dominan dari Model LDA (Output Awal)

Topik	Kata Dominan
1	motor, aksi, curi, pukul, mobil, tugas, bakar, tajam, pria, anggota
2	sidang, oknum, bukti, gedung, adil, indonesia, putus, anestesi, terima, laksana
3	anak, periksa, lapor, minta, perkosa, video, keluarga, keras, jelas, sakit
4	negara, tindak, amerika, serikat, perintah, layan, kerja, tegas, presiden, video
5	tni, kkb, serang, papua, senjata, damai, yahukimo, evakuasi, kelompok, jenazah
6	rupiah, uang, pasal, barang, obat, jual, juta, undang, penjara, bukti
7	bunuh, luka, tewas, aniaya, pukul, kamar, saksi, tusuk, teman, keluarga

4.7 Evaluasi Model

Setelah tahap pemodelan, setiap model dianalisis menggunakan *topic coherence* untuk menilai keterkaitan semantik kata-kata dalam masing-masing topik. Grafik coherence score ditampilkan pada Gambar 4.1, yang digunakan untuk menentukan jumlah topik optimal. Hasil evaluasi menunjukkan bahwa model dengan nilai coherence tertinggi 0,5140 memberikan struktur topik yang paling konsisten.Visualisasi Topic Coherence.

Gambar 4. 1 Visualisasi Topic Coherence



Berdasarkan model terbaik ini, teridentifikasi 7 topik utama dalam pemberitaan kriminal selama periode penelitian. Topik-topik tersebut disajikan dalam Tabel 3.5 beserta kata dominan, interpretasi ringkas, dan label topik, yang menjadi acuan untuk tahap analisis selanjutnya.

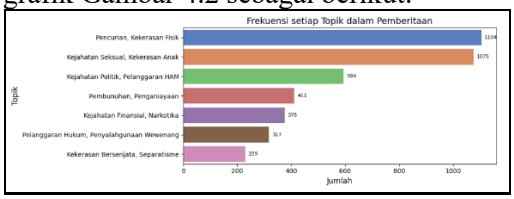
Tabel 4. 6 Interpretasi dan Label tiap Topik

Topik	Kata Dominan	Interpretasi	Label Topik
1	motor, aksi, curi, pukul, mobil, tugas, bakar, tajam, pria, anggota	Tindak kriminal berupa pencurian kendaraan yang sering disertai kekerasan fisik.	Pencurian, Kekerasan Fisik
2	sidang, oknum, bukti, gedung, adil, indonesia, putus, anestesi, terima, laksana	Kasus pelanggaran hukum yang melibatkan aparat atau individu dalam proses peradilan.	Pelanggaran Hukum, Penyalahgunaan Wewenang
3	anak, periksa, lapor, minta, perkosa, video, keluarga, keras, jelas, sakit	Kejahatan seksual dan kekerasan terhadap anak dalam lingkup keluarga atau sosial.	Kejahatan Seksual, Kekerasan Anak
4	negara, tindak, amerika, serikat, perintah, layan, kerja, tegas, presiden, video	Tindak kriminal yang terkait politik, hukum negara, dan pelanggaran hak asasi manusia.	Kejahatan Politik, Pelanggaran HAM
5	tni, kkb, serang, papua, senjata, damai, yahukimo, evakuasi, kelompok, jenazah	Kejahatan yang menggunakan senjata.	Kekerasan Bersenjata, Separatisme
6	rupiah, uang, pasal, barang, obat, jual, juta, undang, penjara, bukti	Kejahatan ekonomi berupa penipuan keuangan, penyalahgunaan obat, dan transaksi ilegal.	Kejahatan Finansial, Narkotika
7	bunuh, luka, tewas, aniaya, pukul, kamar, saksi, tusuk, teman, keluarga	Kekerasan fisik yang menimbulkan luka berat hingga kematian.	Pembunuhan, Penganiayaan

title	content	topic_label
diduga cabuli siswinya guru sd di lombok barat...	Is seorang guru sekolah dasar sd berstatus peg...	Kejahatan Seksual, Kekerasan Anak
pelajar siram air keras ke siswa smk jakut jpp...	siswa smk di koja jakarta utara jakut disiram ...	Kejahatan Seksual, Kekerasan Anak
polisi bongkar ruko kasino di bandung orang di...	polisi membongkar praktik perjudian konvension...	Kejahatan Finansial, Narkotika
penjaga kios tewas ditembak kkb papua di intan...	kelompok kriminal bersenjata kkb pimpinan apen...	Kekerasan Bersenjata, Separatisme
amarah driver ekspedisi hingga todongkan pisto...	sebuah video viral memperlihatkan aksi nekat s...	Pencurian, Kekerasan Fisik

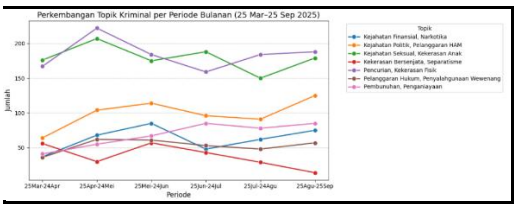
4.8 Analisis

Dari 4.105 artikel berita hasil pemrosesan ETL dan pemodelan topik, data kemudian dianalisis untuk memperoleh *insight* terkait fokus pemberitaan kriminal. Salah satu hal yang dapat ditelusuri adalah distribusi kemunculan berita berdasarkan jenis kejahatan, yang divisualisasikan dalam grafik Gambar 4.2 sebagai berikut.



Gambar 4. 2 Frekuensi setiap Topik dalam Pemberitaan

Selama periode 25 Maret - 25 September 2025, topik Pencurian & Kekerasan Fisik dan Kejahatan Seksual & Kekerasan Anak menempati jumlah artikel terbanyak, masing-masing 1.104 dan 1.075 artikel. Hasil ini menunjukkan bahwa kasus yang sering terjadi dan berdampak langsung pada masyarakat lebih banyak muncul di periode tersebut.



Gambar 4. 3 Perkembangan Topik Kriminal per Periode Bulanan

Secara tren yang terlihat pada Gambar 3.4, topik Pencurian dan Kekerasan Fisik serta topik Kejahatan Seksual dan Kekerasan Anak secara konsisten mendominasi pemberitaan, dengan puncak aktivitas pada awal periode dan fluktuasi ringan di bulan-bulan berikutnya. Topik lain tetap muncul, namun jumlahnya jauh lebih rendah sehingga terlihat selisih signifikan antara dua topik teratas dan kategori lainnya. Beberapa periode menunjukkan lonjakan kasus Kejahatan Politik dan Pelanggaran HAM, sementara kekerasan bersenjata cenderung menurun menjelang akhir periode.

5 KESIMPULAN

Berdasarkan hasil penelitian, penerapan ETL menjadi langkah krusial dalam menyiapkan data artikel berita kriminal yang diperoleh melalui web scraping, karena mampu mengubah data mentah yang tidak terstruktur menjadi format yang bersih, konsisten, dan siap dianalisis. Tahapan ETL memastikan kualitas dan integritas data sebelum dilakukan text preprocessing, termasuk normalisasi, penghapusan *stopword*, tokenisasi, dan *stemming*, sehingga teks siap dimanfaatkan dalam pemodelan topik.

Hasil pemodelan topik menggunakan algoritma LDA mengidentifikasi 7 topik utama dalam pemberitaan kriminal, yakni Pencurian dan Kekerasan Fisik, Kejahatan Seksual dan Kekerasan Anak, Kejahatan Politik dan Pelanggaran HAM, Kekerasan Bersenjata dan Separatisme, Kejahatan Finansial dan Narkotika, Pembunuhan dan Penganiayaan, serta Pelanggaran Hukum dan Penyalahgunaan Wewenang. Dari hasil analisis distribusi dan tren bulanan menunjukkan bahwa Pencurian dan Kekerasan Fisik serta Kejahatan Seksual dan Kekerasan Anak mendominasi, sementara topik lain muncul lebih jarang. Beberapa lonjakan terjadi pada Kejahatan Politik dan Pelanggaran HAM di akhir periode, sedangkan Kekerasan Bersenjata justru menurun. Temuan ini menegaskan fokus pemberitaan pada kasus dengan frekuensi tinggi dan dampak nyata bagi masyarakat, serta memberikan pemahaman sistematis mengenai pola kriminalitas yang diberitakan.

6 SARAN

Untuk penelitian selanjutnya, disarankan untuk menerapkan *framework* lain seperti ELT (Extract, Load, Transform) agar proses menjadi fleksibilitas pada data berskala besar. Selain itu, pendekatan pemodelan topik berbasis neural seperti BERTopic atau Top2Vec dapat dicoba untuk meningkatkan akurasi identifikasi topik serta menangkap pola semantik yang lebih kompleks. Perluasan sumber data dari berbagai portal berita dan media sosial juga dianjurkan agar analisis mencakup variasi kasus yang lebih luas, sehingga memberikan pemahaman yang lebih komprehensif mengenai tren kejahatan yang sedang terjadi.

DAFTAR PUSTAKA

[1] D. M. Hulu, "Faktor-Faktor Yang Memengaruhi Jumlah Kriminalitas di Indonesia Dengan Regresi Data Panel Pada Tahun 2016-2020," *Indones. Counc. Prem. Stat. Sci.*, vol. 3, no. 2, p. 37, 2024, doi: 10.24014/icopss.v3i2.32237.

[2] I. G. B. Premana Putra, M. Sudarma, and I. B. G. Manuaba, "Penerapan Metode Extreme Programming pada Rancang Bangun Sistem Analisis Sentimen Portal Berita," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 6, pp. 1369–1378, 2023, doi: 10.25126/jtiik.2023106904.

[3] G. N. Zamroji, R. A. Syahputra, S. Z. Rohman, Y. P. Astuti, and I. F. Kurniawan, "Pipeline ETL Terdistribusi untuk Klasifikasi Berita Clickbait dan Topik Berita," vol. 2025, no. Senada, pp. 165–174, 2025.

[4] Fatmasari, Y. N. Kunang, and S. D. Purnamasari, "Web Scraping Techniques to Collect Weather Data in South Sumatera," *Proc. 2018 Int. Conf. Electr. Eng. Comput. Sci. ICECOS 2018*, no. December, pp. 385–390, 2019, doi: 10.1109/ICECOS.2018.8605202.

[5] M. J. Medea, V. P. Rantung, and O. Kembuan, "Metode Latent Dirichlet Allocation dalam Pemodelan Topik Headline Berita Online tentang Hukum dan Kriminal," *JOINTER J. Informatics Eng.*, vol. 5, no. 02, pp. 1–7, 2024, doi: 10.53682/jointer.v5i02.63.

[6] E. L. Cohen, "Online Journalism as," vol. 1, 2014.

[7] F. Muiz, "98-Article Text-338-1-10-20220605," vol. 3, no. 3, pp. 56–58, 2021.

[8] T. S. T. SOWMYYA, "Crime: A Conceptual Understanding," *Indian J. Appl. Res.*, vol. 4, no. 3, pp. 196–198, 2011, doi: 10.15373/2249555x/mar2014/58.

[9] A. A. Munajat and H. Yusuf, "Dinamika

Kriminalitas Urban:Studi Tentang Faktor-Faktor Yang Mempengaruhi Tingkat Kejahatan Di Kota Besar Dynamics of Urban Criminality: a Study of the Factors Affecting Crime Rates in Large Cities," *JICN J. Intelek dan Cendikiawan Nusant.*, vol. 1, no. 2, pp. 1330–1339, 2024, [Online]. Available: https://jicnusantara.com/index.php/jicn

[10] N. Mali, "A Survey of ETL Tools," *Int. J. Comput. Tech.* -–, vol. 2, no. 5, pp. 20–27, 2015, [Online]. Available: http://www.ijctjournal.org

[11] R. Mitchell, *Ryan Mitchell Web Scraping with Python*. 2018. [Online]. Available: www.allitebooks.com

[12] A. Dwiyoga Widiantoro Mustafid Ridwan Sanjaya, *Pengantar Nlp Dan Topik Model Lda Sampul Dalam*. 2024.

[13] S. Eisenstein, "Introduction," *Give Us Bread but Give Us Roses*, pp. 9–17, 2020, doi: 10.4324/9780203103517-5.

[14] D. M. Blei and A. Y. Ng, "Latent Dirichlet Allocation," no. January 2001, 2014.

[15] D. L. C. Pardede and M. A. I. Waskita, "Analisis Pemodelan Topik Untuk Ulasan Tentang Peduli Lindungi," *J. Ilm. Inform. Komput.*, vol. 28, no. 1, pp. 17–26, 2023, doi: 10.35760/ik.2023.v28i1.7925.