JEMSI: Jurnal Ekonomi Manajemen Sistem Informasi
E-ISSN: 2686-5238
P-ISSN: 2686-4916
https://dinastirev.org/JEMSI dinasti.info@gmail.com +62 811 7404 455
DOI: https://doi.org/10.38035/jemsi.v6i3
https://creativecommons.org/licenses/by/4.0/
Received: 14 Januari 2025 | **Revised:** 30 Januari 2025 | **Accepted:** 09 Februari 2025

Analisis Pengelompokan Data Kriminalitas dan Kejahatan di Indonesia dari Portal Berita Online menggunakan Metode Latent Dirichlet Allocation

Adi Widyanto[1], Yan Puspitarani[2], Adi Purnama[3]
[1]Universitas Widyatama, Bandung, Indonesia, widyanto.adi07@gmail.com
[2]Universitas Widyatama, Bandung, Indonesia, yan.puspitarani@widyatama.ac.id
[3]Universitas Widyatama, Bandung, Indonesia, adi.purnama@widyatama.ac.id

Corresponding Author: widyanto.adi07@gmail.com[1]

Abstract: *Criminality and crime are social phenomena that violate legal norms and have a serious impact on the stability of the state as well as social and economic structures. In the digital era, online news portals make it easier to access crime-related information. However, the high volume and diversity of data often make it difficult for people to understand crime trends. The purpose of this study is to identify the main patterns or themes and categorize crime data that occurred in Indonesia from the news sites detik.com and cnnindonesia.com in the last five years. This research uses the Latent Dirichlet Allocation (LDA) method to group crime news into categories. Data collection uses web scraping from online news portals to get a representative dataset. The collected data will be processed through text preprocessing and topic modelling using the LDA method. From the 20 topics tested, the best model was found in 9 topics with a coherence score of 0.538163893830327 and perplexity of -7.85722881473597, indicating interpretative topics and good data distribution. The main topics include social issues, violence, police investigations, legal cases, justice, fraud, and theft. The dominant themes are violence against children and women (21.59%) with a total of 5,060 documents and theft (23.26%) with a total of 5,241 documents. The results of this study provide insight into crime trends in Indonesia in a social and legal context.*

Keyword: *Criminality, Online News Portal, Web Scraping, Topic Modelling, LDA*

Abstrak: Kriminalitas dan kejahatan merupakan fenomena sosial yang melanggar norma hukum dan berdampak serius pada stabilitas negara serta struktur sosial dan ekonomi. Dalam era digital, portal berita *online* mempermudah akses informasi terkait kejahatan. Namun, tingginya volume dan keberagaman data sering menyulitkan masyarakat memahami tren kejahatan. Tujuan dari penelitian ini adalah untuk mengidentifikasi pola atau tema utama dan mengelompokkan data kejahatan yang terjadi di Indonesia dari situs berita detik.com dan cnnindonesia.com pada lima tahun terakhir. Penelitian ini menggunakan metode *Latent Dirichlet Allocation* (LDA) untuk mengelompokkan berita kriminal ke dalam beberapa kategori. Pengumpulan data menggunakan *web scraping* dari portal berita *online* untuk

mendapatkan *dataset* yang representatif. Data yang telah dikumpulkan akan diolah melalui tahap *text preprocessing* dan *topic modelling* menggunakan metode LDA. Dari 20 topik yang dilakukan pengujian, model terbaik ditemukan pada 9 topik dengan *coherence score* sebesar 0,538163893830327 dan *perplexity* sebesar -7,85722881473597, menunjukkan topik yang interpretatif dan distribusi data yang baik. Topik utama meliputi isu sosial, kekerasan, investigasi polisi, kasus hukum, peradilan, penipuan, hingga pencurian. Tema dominan adalah kekerasan terhadap anak dan perempuan (21,59%) dengan jumlah sebanyak 5.060 dokumen serta pencurian (23,26%) dengan jumlah sebanyak 5.241 dokumen. Hasil penelitian ini memberikan wawasan mengenai tren kriminalitas di Indonesia dalam konteks sosial dan hukum.

Kata Kunci: Kriminalitas, Portal Berita *Online*, *Web Scraping, Topic Modelling*, LDA

PENDAHULUAN

Kriminalitas dan kejahatan merupakan fenomena sosial yang terlibat dalam pelanggaran terhadap norma-norma hukum yang berlaku dalam suatu masyarakat. Dua konsep tersebut memiliki dampak serius terhadap stabilitas dan keamanan suatu negara, serta dapat merusak struktur sosial dan ekonomi. Saat ini, perhatian masyarakat Indonesia terhadap isu-isu dan peristiwa terkait kriminalitas dan kejahatan yang terjadi di sekitarnya semakin meningkat. Hal ini sejalan dengan berkembangnya era digital, ketersediaan akan situs media berita *online* memudahkan masyarakat dalam mendapatkan informasi secara cepat dan mudah.

Meskipun situs berita *online* telah menjadi sumber informasi utama bagi masyarakat, banyak orang masih kesulitan dalam memperoleh ringkasan mengenai jenis-jenis kejahatan yang sering terjadi di Indonesia. Kesulitan ini disebabkan oleh volume berita kejahatan yang terus meningkat setiap harinya, serta keberagaman data yang disajikan dengan berbagai tingkat kualitas dan kedalaman informasi. Hal ini acapkali menyebabkan kebingungan di kalangan masyarakat, yang mungkin merasa kesulitan dalam menavigasi informasi yang ada secara efektif. Oleh karena itu, sangat penting bagi masyarakat untuk memiliki akses terhadap ringkasan yang lebih terstruktur dan mudah dipahami mengenai tren kriminalitas dan kejahatan.

Ringkasan mengenai jenis-jenis kriminalitas dan kejahatan yang umum di Indonesia dapat diperoleh dengan menganalisis topik-topik yang relevan dari situs media berita *online*. Data dikumpulkan melalui metode *web scraping*, yaitu teknik untuk mendapatkan informasi dari situs web secara otomatis tanpa harus menyalinnya secara manual (Ayani, Pratiwi, & Muhardi, 2019). Berita yang dianalisis dibatasi pada lima tahun terakhir untuk menjaga relevansi dan aktualitas. Penelitian ini menggunakan situs cnnindonesia.com dan detik.com yang merupakan salah satu situs berita paling banyak diakses dan termasuk ke dalam situs web dengan *traffic* tertinggi di Indonesia (Satriajati, Panuntun, & Pramana, 2020).

Data yang dikumpulkan melalui *web scraping* akan diproses menggunakan teknik *text preprocessing*, yang mencakup pembersihan dan standarisasi data karena data mentah yang diperoleh tidak terstruktur. Dalam penelitian ini, terdapat beberapa langkah dalam teknik *text preprocessing* yang digunakan, yaitu *cleansing, case folding, tokenizing*, *stopword removal*, dan *stemming*. *Text preprocessing* melibatkan transformasi teks sebelum dianalisis dengan mengidentifikasi unit mana yang akan digunakan, menghapus konten yang tidak relevan, menggabungkan istilah-istilah yang terkait secara semantik untuk mengurangi kerenggangan data dan meningkatkan prediksi, serta meningkatkan jumlah informasi semantik yang ditangkap (Hickman, Thapa, Tay, Cao, & Srinivasan, 2020).

Data yang telah melalui proses *text preprocessing* akan dikelompokkan. Proses pembuatan *topic modelling* dilakukan dengan menggunakan metode *Latent Dirichlet*

Allocation (LDA). LDA adalah sebuah metode *topic modelling* yang digunakan untuk menentukan pola pada sebuah dokumen yang dapat menghasilkan topik (Tong & Zhang, 2016). Tujuan dari *topic modelling* adalah (Sari & Purnomo, 2022) untuk memperoleh model topik abstrak pada kumpulan dokumen. Hasil dari proses *topic modelling* yaitu *data clustering* digunakan untuk menganalisis distribusi kelompok data yang mirip berdasarkan topik, serta *coherence value* untuk menentukan jumlah topik yang muncul. Evaluasi akhir dilakukan untuk merangkum temuan dari pemodelan data, serta visualisasi yang menunjukkan hubungan antara topik-topik tersebut, guna memperoleh ringkasan mengenai jenis-jenis kriminalitas dan kejahatan yang umum di Indonesia.

Penelitian yang menerapkan metode *topic modelling* dengan LDA pernah dilakukan oleh Ahmad Fathan Hidayatullah dkk. (Hidayatullah, Ma'arif, Habibie, & Khomsah, 2020) pada tahun 2021 untuk mengelompokkan topik pembangunan infrastruktur di Indonesia dari situs media berita *online.* Selain itu penelitian sebelumnya mengenai *topic modelling* dengan LDA juga pernah dilakukan oleh Gede Herdian Setiawan dkk. (Setiawan, Adnyana, Sugiartha, & Budiarta, 2023) pada tahun 2023 bersumber dari laporan aduan mahasiswa di salah satu Perguruan Tinggi Swasta (PTS) untuk mengekstraksi topik dari data tersebut. Penelitian serupa terkait pengumpulan data dari situs media berita *online* yaitu situs berita detik.com dengan teknik *web scraping* pernah dilakukan oleh Salim Satriajati dkk. (Satriajati, Panuntun, & Pramana, 2020) pada tahun 2020 untuk mengelompokkan jenis kasus kriminal pada masa pandemi COVID-19.

Tujuan dari penelitian ini adalah untuk mengidentifikasi pola atau tema utama dan mengelompokkan data kejahatan yang terjadi di Indonesia dari situs berita detik.com dan cnnindonesia.com pada lima tahun terakhir menggunakan *topic modelling* dengan metode LDA. Kinerja LDA lebih baik dibandingkan dengan metode pemodelan topik yang lainnya dan dapat diimplementasikan untuk identifikasi topik, klasifikasi, dan *clustering* (Astuti & Cahyono, 2023).

METODE

Tahap Pengumpulan Data

Pada tahap ini, proses pengumpulan data dilakukan dengan fokus pada artikel-artikel yang diterbitkan di situs berita *online* detik.com dan cnnindonesia.com dalam rentang waktu lima tahun terakhir, yaitu dari 1 Januari 2020 hingga 1 Oktober 2024. Sebelum memulai pengambilan data, ditentukan beberapa kata kunci yang berkaitan dengan topik kriminalitas dan kejahatan. Kata kunci tersebut meliputi pembegalan, pembunuhan, penculikan, pemerkosaan, dan perampokan. Kata kunci ini berfungsi sebagai alat pencarian untuk menemukan artikel-artikel yang relevan dengan fokus penelitian.

Metode yang digunakan untuk pengumpulan data adalah *web scraping*. *Web scraping* adalah proses pengambilan sebuah dokumen semi-terstruktur dari internet, umumnya berupa halaman-halaman *web* dalam bahasa *markup* seperti HTML atau XHTML, dan menganalisis dokumen tersebut untuk diambil data tertentu dari halaman tersebut untuk digunakan bagi kepentingan lain (Turland, 2010). Proses ini dimulai dengan mengidentifikasi tag HTML di situs detik.com dan cnnindonesia.com yang memuat berita dan informasi penyertanya. Tahap ini mengubah data mentah menjadi teks yang terstruktur, sekaligus menandai bagian-bagian dalam teks seperti paragraf, kolom, atau tabel. Pada beberapa kasus, tahap ini juga dapat mengekstrak informasi tingkat dokumen, seperti <Penulis> atau <Judul>, jika posisi visual elemen-elemen tersebut memungkinkan diidentifikasi (Puspitarani & Zulpratita, 2020). Setelah proses identifikasi selesai, langkah berikutnya adalah membuat *web scraper* menggunakan bahasa pemrograman Python. *Library* Scrapy dipilih karena merupakan perangkat lunak *open source*, bersifat fleksibel, dan memungkinkan pengguna untuk melakukan sedikit penyesuaian saat menjelajahi berbagai situs dengan mekanisme yang berbeda dalam mengekstrak data dan

informasi dari dokumen HTML dengan memilih elemen tertentu (Rohman, Santoso, Saraswati, & Winarsih, 2019). Untuk mendapatkan konten artikel, URL artikel disalin sebagai parameter masukan ke *web scraper*. Hasil dari *web scraping* adalah kumpulan informasi tentang berbagai kasus kriminal dan kejahatan di Indonesia. Semua data yang terkumpul disimpan dalam format CSV.

Tahap *Text Preprocessing*

Pada tahap ini, data yang dikumpulkan melalui *web scraping* akan diproses menggunakan teknik *text preprocessing*. *Text preprocessing* adalah teknik yang mencakup pembersihan dan standarisasi data mentah menjadi format yang mudah dimengerti (Wardhani, Astuti, & Saputra, 2024). Tujuan dari *text preprocessing* adalah untuk menghilangkan data yang tidak konsisten, data duplikat, dan data yang tidak mempengaruhi polaritas dokumen yang ada (Julianto, Kurniadi, & Jr, 2023). *Text preprocessing* melibatkan transformasi teks sebelum dianalisis dengan mengidentifikasi unit mana yang akan digunakan, menghapus konten yang tidak relevan, menggabungkan istilah-istilah yang terkait secara semantik untuk mengurangi kerenggangan data dan meningkatkan prediksi, serta meningkatkan jumlah informasi semantik yang ditangkap (Hickman, Thapa, Tay, Cao, & Srinivasan, 2020). Dalam penelitian ini, terdapat beberapa langkah dalam teknik *text preprocessing*, yaitu (Ridwansyah, 2022):

1. *Cleansing,* yaitu pembersihan data dengan melakukan penghapusan untuk karakter atau elemen yang tidak relevan, seperti simbol kusus, angka, tanda baca, angka, *emai,* dan URL.
2. *Case folding,* yaitu proses dimana berbagai jenis huruf yang sama disatukan dengan mengubah semua karakter menjadi huruf kecil *(lowercase)*.
3. *Tokenizing,* yaitu proses yang membagi kalimat menjadi kata-kata yang terpisah. Dalam tahap ini, teks dipecah menjadi token dengan memisahkan setiap kata menggunakan spasi. Token dapat terdiri dari berbagai elemen, seperti kata, simbol, dan angka.
4. *Stopword removal,* yaitu proses menghilangkan kata-kata yang tidak memberikan informasi penting, biasanya kata-kata yang sering muncul dalam teks dan tidak menambah makna pada data, sehingga dianggap tidak relevan untuk proses klasifikasi. Kata-kata ini bisa berupa istilah apa pun yang tidak memiliki relevansi yang jelas.
5. *Stemming,* yaitu proses mencari kata dasar atau akar katanya dengan menghilangkan kata imbuhan.

Setelah proses *text preprocessing* selesai, hasilnya adalah artikel yang telah bersih dan siap untuk analisis lebih lanjut. *File* yang dihasilkan dari *text preprocessing* ini kemudian disimpan dalam format CSV *(Comma-Separated Values)*, mudah digunakan karena formatnya sederhana dan jelas. Data CSV disimpan dalam teks dengan pemisah koma, sehingga mudah dibaca, ditulis, dan diproses tanpa memerlukan perangkat lunak khusus (Tapsai, 2018).

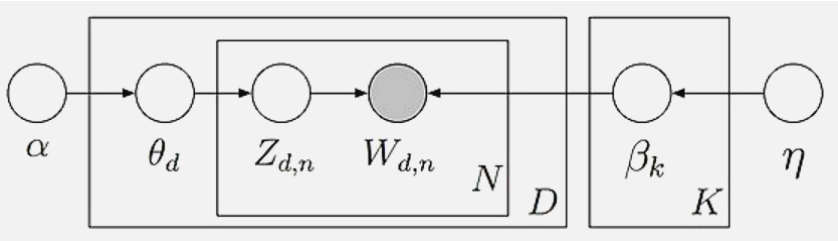
Tahap *Topic Modelling* dengan *Latent Dirichlet Allocation* (LDA)

Proses *topic modelling* dilakukan dengan menggunakan metode *Latent Dirichlet Allocation* (LDA). *Topic modelling* menyediakan cara yang baik untuk menganalisis data teks dalam jumlah besar yang belum diklasifikasikan (Alghamdi & Alfalqi, 2015). *Topic modelling* adalah pendekatan *text mining* yang dapat diandalkan dalam menemukan data teks yang tersembunyi dan menemukan hubungan antar kata dari sebuah *corpus* (Jelodar, et al., 2019). *Topic modelling* termasuk kedalam *clustering* dengan mengelompokkan dokumen berdasarkan kemiripannya (Guo, Han, Li, Zhang, & Bai, 2018). Pemahaman dasarnya adalah tiap dokumen direpresentasikan sebagai sekumpulan topik yang berbeda, dimana tiap topik memiliki karakteristik berupa distribusi dari berbagai kata (Blei, Ng, & Jordan, Latent Dirichlet Allocation, 2003). Alur general dari LDA adalah dengan memilih distribusi topik secara acak

untuk suatu dokumen dan untuk setiap kata dalam dokumen akan dipilih salah satu topik sesuai dengan distribusi (Blei, Probabilistic Topic Models, 2012).

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d \quad \text{(i)}$$

Pada rumus (i), baik α dan β mewakili *hyperparameter* untuk mendapatkan kemungkinan pemodelan. α merepresentasikan *dirichlet prior parameter* untuk distribusi topik pada *level* dokumen dan β menggambarkan distribusi probabilitas kata untuk topik tertentu. Distribusi topik dari dokumen d direpresentasikan dalam bentuk vektor θ. Notasi z mengacu pada topik tersembunyi dari dokumen d. Notasi M dan N masing-masing merepresentasikan panjang dokumen dan jumlah istilah dalam dokumen (Blei, Ng, & Jordan, Latent Dirichlet Allocation, 2003).



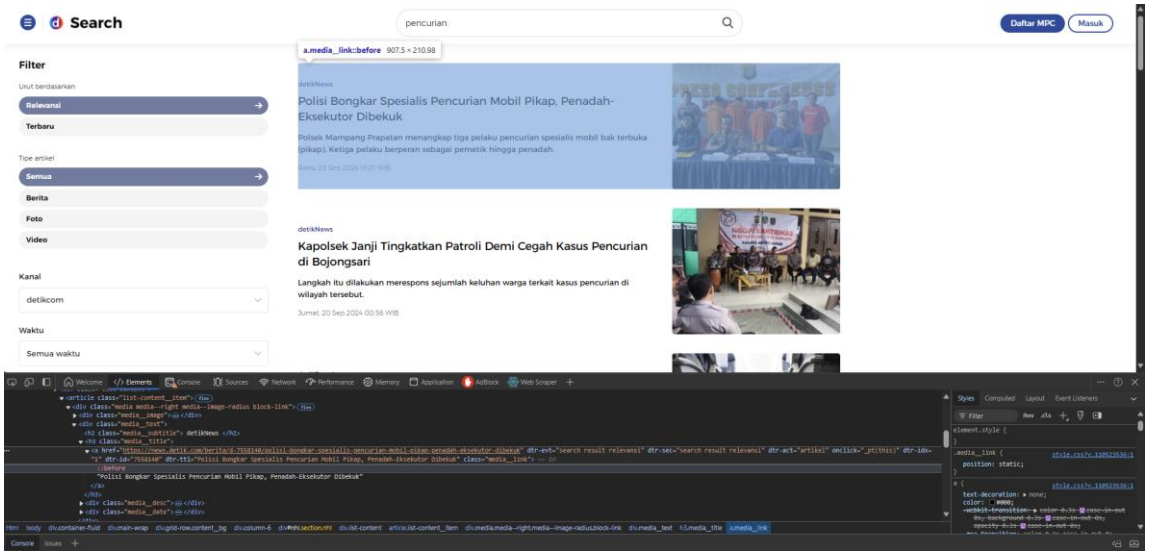
Sumber: Blei, D. (2012)
Gambar 1. Cara Kerja LDA

Hasil dari proses *topic modelling* ini adalah *data clustering*, yang sangat berguna untuk memahami penyebaran kelompok data yang serupa dan memiliki kesamaan berdasarkan topik. Dengan melakukan *clustering*, kita dapat mengidentifikasi pola dalam berita dan menentukan bagaimana berbagai topik saling berhubungan (Shevendrakumar, 2023). Selain itu, proses ini juga menghasilkan *coherence value*, yang digunakan untuk menilai jumlah topik yang teridentifikasi. Salah satu keunggulan *coherence c_v* adalah kemampuannya untuk mencerminkan relevansi semantik yang dirasakan manusia. *Coherence c_v* memiliki korelasi yang lebih tinggi dengan evaluasi manusia dibandingkan metrik koherensi lainnya, seperti *coherence u_mass* (Röder, Both, & Hinneburg, 2015).

HASIL DAN PEMBAHASAN

Metadata Portal Berita

Setiap halaman yang ada di internet memiliki *metadata* yang berbeda-beda. Untuk itu diperlukan analisis terhadap portal berita yang akan dijadikan sumber untuk pengumpulan berita. Elemen pada *metadata* ini akan menjadi penentu bagian mana saja data yang perlu diambil pada saat melakukan ekstraksi pada proses *web scraping*.



Gambar 2. Contoh Mengakses HTML DOM Daftar Berita

Keluaran dari proses ini adalah *metadata* halaman hasil pencarian dan halaman detail berita dari portal berita. Selain elemen HTML, pada *metadata* juga terdapat nama portal berita dan URL utama dari portal berita. Data *metadata* digunakan sebagai masukan untuk proses *web scraping*.

Tabel 1. Rincian Tag HTML Proses Web Scraping

Website	Halaman	Data	Tag/Elemen HTML	Class/ Keterangan
Detik	Pencarian	URL	`<a href="...">`	*Link* di elemen judul
		Pagination	`<a>`	`pagination__item`
	Detail Berita	Judul Berita	`<h1>`	`detail__title`
		Waktu Terbit	`<h1>`	`date`
		Konten Berita	`<div>` atau `<span>`	`detail__body-text`
CNN Indonesia	Pencarian	URL	`<a href="...">`	*Link* di elemen judul
		Pagination	`<a>` atau `<button>`	`page-link`
	Detail Berita	Judul Berita	`<h1>`	`title`
		Waktu Terbit	`<span>` atau `<div>`	`date`
		Konten Berita	`<p>` di dalam `<div>`	`content`

Scraping Data

Proses *web scraping* terdapat dua sub proses utama yaitu ekstraksi halaman hasil pencarian dan ekstraksi halaman detail berita. Ekstraksi halaman hasil pencarian bertujuan untuk mendapatkan daftar URL yang harus dikunjungi berdasarkan hasil pencarian. Ekstraksi halaman detail berita bertujuan untuk mendapatkan detail berita. Pada implementasinya proses ini memanfaatkan *library* Beautiful Soup pada bahasa pemrograman Python.

Data *metadata* diperlukan sebagai penentu elemen yang akan diambil informasinya dari struktur HTML halaman pencarian dan halaman detail berita dan data *keyword* diperlukan sebagai kata kunci pencarian pada portal berita. Setelah kedua data ini didapatkan dilanjutkan ke proses mengunjungi semua portal berita yang ada pada *metadata*. Pada setiap porta berita, *keyword* yang telah didefinisikan digunakan sebagai kata kunci pencarian.

Tabel 2. Data Field dalam Proses Web Scraping

Field	Deskripsi
title	Judul berita
category	Situs asal berita
publish_date	Waktu publikasi berita

article_url	URL dari berita
content	Konten berita

Hasil dari pengumpulan berita dengan *web scraping* adalah file CSV yang berisi daftar detail berita yang selanjutnya akan digunakan untuk proses pengelompokan berita. Berikut merupakan jumlah data yang diperoleh serta contoh data yang didapatkan dari proses *scraping*:

Tabel 3. Rincian Data Hasil Web Scraping

No	Sumber Data	*Keyword*	Jumlah Data
1	Detik	Pencurian	4.580
		Pembunuhan	1.592
		Penculikan	1.553
		Penipuan	993
		Pembegalan	606
		Kekerasan Seksual	4.121
		Sub Total	**13.445**
2	CNN Indonesia	Pencurian	1.083
		Pembunuhan	5.920
		Penculikan	401
		Penipuan	1.580
		Pembegalan	138
		Kekerasan Seksual	871
		Sub Total	**9.993**
		Total Data	**23.438**

Text Preprocessing

Hasil dari proses *text preprocessing* ini akan menjadi data teks yang lebih terstruktur dan siap untuk dieksplorasi lebih lanjut dalam *topic modelling*. Dengan teks yang sudah bersih dan seragam, algoritma *topic modelling* akan lebih efektif dalam mengidentifikasi topik-topik utama dari kumpulan data yang dianalisis. Data yang diolah oleh proses *text preprocessing* hanya konten berita saja, karena data yang akan digunakan untuk proses selanjutnya hanya data dari konten berita.

Tabel 4. Contoh Data Hasil Tahapan Text Preprocessing

Data mentah	Seorang pria berinisial II (55) meninggal dunia setelah menjadi korban begal di Ciampea, Kabupaten Bogor, Jawa Barat. Berdasarkan keterangan keluarga, korban hendak menjemput putrinya. "Menurut keterangan keluarga, korban keluar rumah dengan maksud menjemput putrinya menggunakan sepeda motor," kata Kapolsek Ciampea Kompol Suminto, Senin (30/9/2024). Korban sendiri berangkat dari rumahnya sekitar pukul 01.00 WIB dini hari tadi. Namun nahas, korban dibegal dan motornya diduga dirampas pelaku. "Namun korban justru ditemukan dalam keadaan tak bernyawa dengan sepeda motornya hilang diduga dibawa oleh pelaku," tuturnya. Sebelumnya, video yang memperlihatkan seorang pria tergeletak di Ciampea, Kabupaten Bogor, Jawa Barat, viral di media sosial (medsos). Dalam video itu dinarasikan pria tersebut korban begal. Suminto membenarkan kejadian tersebut. Korban kemudian dilarikan ke rumah sakit seusai kejadian. "Siap benar, korban ada di RSUD Leuwiliang," kata Suminto. Suminto mengatakan korban mengalami luka pada kepala bagian belakang. Korban diketahui meninggal dunia setelah sampai di rumah sakit. "(Korban) Meninggal dunia setelah sampai rumah sakit," ungkapnya. Korban diketahui berinisial II (55). Terkait kejadian tersebut, pihak kepolisian masih melakukan penyelidikan untuk mengungkapnya. "Masih lidik (penyelidikan)," pungkasnya.
Data setelah *text preprocessing*	["pria", "inisial", "ii", "tinggal", "dunia", "korban", "begal", "ciampea", "kabupaten", "bogor", "jawa", "barat", "dasar", "terang", "keluarga", "korban", "jemput", "putri", "terang", "keluarga", "korban", "rumah", "maksud", "jemput", "putri", "sepeda", "motor", "kapolsek", "ciampea", "kompol", "suminto", "senin", "korban", "berangkat", "rumah", "wib", "nahas", "korban", "begal", "motor", "duga", "rampas", "laku", "korban", "temu", "nyawa", "sepeda", "motor", "hilang", "duga", "bawa", "laku", "video", "pria", "geletak",

	"ciampea", "kabupaten", "bogor", "jawa", "barat", "viral", "media", "sosial", "medsos", "video", "narasi", "pria", "korban", "begal", "suminto", "benar", "jadi", "korban", "lari", "rumah", "sakit", "jadi", "korban", "rsud", "leuwiliang", "suminto", "suminto", "korban", "alami", "luka", "kepala", "korban", "tinggal", "dunia", "rumah", "sakit", "korban", "tinggal", "dunia", "rumah", "sakit", "korban", "inisial", "ii", "kait", "jadi", "polisi", "lidi", "ungkap", "lidik", "lidi", "pungkas"]

Pembentukan *Bigram*, *Dictionary*, dan *Corpus*

Bigram, atau pasangan dua kata yang muncul berurutan dalam teks, membantu menangkap konteks kata-kata yang sering muncul bersama. Ini penting dilakukan karena beberapa konsep dalam teks hanya bermakna ketika kata-katanya digabungkan. Penggunaan *bigram* dapat meningkatkan akurasi hasil *topic modelling* karena topik akan mencerminkan frasa atau istilah yang lebih bermakna.

Tabel 5. Contoh Teks Bigram

Teks *Bigram*
kabupaten_bogor
jawa_barat
sepeda_motor
rumah_sakit
tinggal_dunia

Pada tahap ini juga telah dilakukan pembuatan *dictionary* menggunakan *library* Gensim sebagai bagian dari proses *topic modelling*. *Dictionary* berfungsi sebagai peta dari semua kata unik yang muncul dalam kumpulan dokumen setelah melalui proses *text preprocessing*. Setiap kata dalam *dictionary* diberi ID numerik yang unik, memungkinkan kata-kata tersebut diakses dengan mudah oleh model untuk analisis selanjutnya.

Tabel 6. Contoh Dictionary

ID Kata	Kata
3	begal
24	korban
34	sepeda_motor
44	rumah_sakit
38	polisi
42	rampas
26	pria

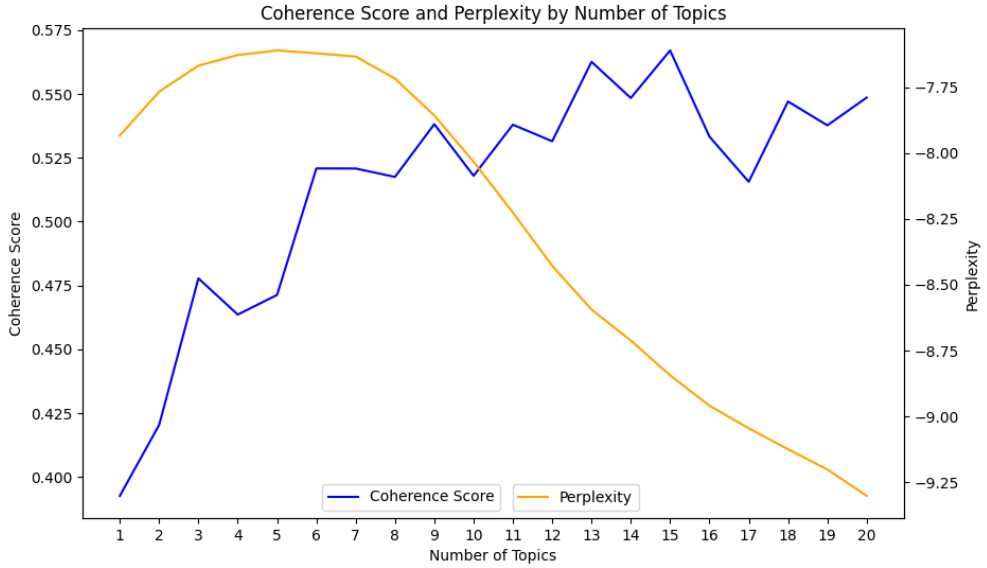
Proses pembuatan *corpus* menggunakan *library* Gensim telah berhasil dilakukan setelah tahap pembuatan *dictionary*. Dalam Gensim, *corpus* biasanya direpresentasikan dalam bentuk *Bag-of-Words* (BoW), di mana setiap dokumen diubah menjadi representasi numerik berdasarkan ID kata yang terdapat pada *dictionary* yang telah dibuat sebelumnya. Dengan *corpus* yang sudah terstruktur ini, model LDA dapat mengidentifikasi pola distribusi topik dalam dokumen berdasarkan frekuensi kata yang muncul dalam setiap topik.

Tabel 7. Contoh Corpus

ID Kata	Frekuensi Kata
3	3
24	13
34	2
44	3
38	1
42	3
26	3

Membangun Model Latent Dirichlet Allocation

Dalam evaluasi model LDA, *coherence score* dan *perplexity* adalah dua metrik penting yang bisa digunakan untuk menilai kualitas model. *Coherence score* mengukur seberapa koheren atau konsisten topik yang dihasilkan, yaitu apakah kata-kata yang ada dalam satu topik saling berkaitan dan relevan. *Coherence score* yang lebih tinggi menunjukkan bahwa topik lebih mudah diinterpretasi, karena kata-kata di dalamnya cenderung memiliki keterkaitan semantik yang kuat. Sementara itu, *perplexity* adalah metrik yang mengukur seberapa baik model tersebut mampu memprediksi dokumen-dokumen baru berdasarkan distribusi kata yang dipelajari. Nilai *perplexity* yang lebih rendah biasanya menunjukkan model yang lebih sesuai dengan data, karena menunjukkan bahwa model memiliki keyakinan yang tinggi terhadap distribusi kata di setiap topik.



Gambar 3. Grafik Coherence Score dan Perplexity Distribusi Topik

Dalam penelitian ini, untuk mengevaluasi kualitas model LDA dengan 20 topik, menggunakan metrik *coherence value* dengan metode *coherence c_v* dan juga *perplexity*. *Coherence c_v* lebih sesuai dengan penilaian manusia dalam hal kemiripan semantik. *coherence c_v* menghitung kesamaan kata berdasarkan *Normalized Pointwise Mutual Information* (NPMI) dan *cosine similarity*, yang membuatnya lebih mampu menangkap kualitas tematik dari topik secara konsisten. Tidak seperti *coherence u_mass* yang lebih bergantung pada *co-occurrence* kata dalam dokumen tertentu dan dapat kurang akurat pada *corpus* baru, *Coherence c_v* menggunakan pendekatan berbasis distribusi kata yang lebih fleksibel. Hal ini memungkinkan *coherence c_v* bekerja lebih baik dalam berbagai jenis data atau domain teks (Röder, Both, & Hinneburg, 2015). Selain itu, model LDA dilatih dengan parameter *passes* sebanyak 15 kali, yang mengatur jumlah iterasi di mana model mengunjungi setiap dokumen untuk mempelajari distribusi kata. Oleh karena itu, untuk menentukan topik terbaik, dalam percobaan ini yaitu dengan menggunakan *trade-off* (persimpangan) antara *coherence score* dan *perplexity* (Tresnasari, Adji, & Permanasari, 2020).

Tabel 8. Nilai Coherence Score dan Perplexity Setiap Topik

Number of Topics	*Coherence Score*	*Perplexity*
1	0,392650981724461	-7,93342534490154
2	0,420406649902948	-7,76712179093139
3	0,477792576675946	-7,66787180304251

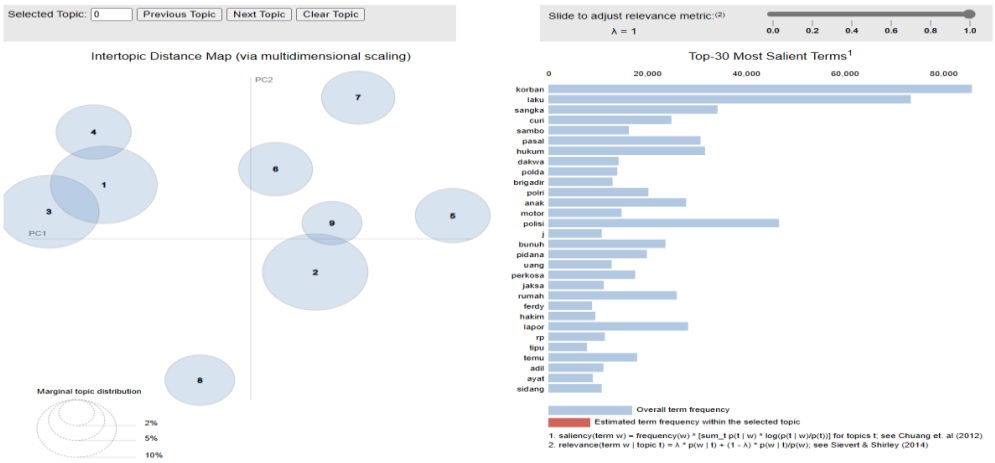
4	0,463615659021384	-7,62784314250634
5	0,471306951014255	-7,61036247097731
6	0,520842028825675	-7,62114938690989
7	0,52080088675129	-7,63358851375903
8	0,51751367146982	-7,71711212883877
9	0,538163893830327	-7,85722881473597
10	0,517958475094319	-8,03311604321893
11	0,537959534100008	-8,22586163056313
12	0,531477794434485	-8,42807243263483
13	0,562582507106429	-8,59441544005829
14	0,548423355069479	-8,71196683074083
15	0,567030902169049	-8,84438145224184
16	0,533300034410521	-8,9593612280847
17	0,515600256902523	-9,04533652742862
18	0,547044122751619	-9,12478418798714
19	0,537709327652947	-9,20151289937224
20	0,548575419659374	-9,30166228596539

Setelah dilakukan pengujian, terlihat bahwa *coherence score* mengalami kenaikan yang sangat signifikan pada jumlah topik 1 hingga -9. Namun setelah topik ke-9, *coherence score* tidak mengalami kenaikan yang signifikan dan cenderung menjadi tidak stabil. *Coherence score* tertinggi terdapat pada jumlah topik ke-15. Sedangkan untuk *perplexity* terlihat bahwa nilai mengalami penurunan dari jumlah topik ke-7 hingga ke-20. Terlihat pada Gambar 3 menunjukan bahwa grafik *coherence score* dan *perplexity* mengalami perpotongan yang mendekati di antara topik ke-9 dan ke-10. Pada topik ke-9, *coherence score* mendapatkan nilai sebesar 0,538163893830327 dan *perplexity* sebesar -7,85722881473597. *Coherence score* memiliki angka cukup tinggi dan *perplexity* memiliki angka yang relatif rendah.

Coherence score sekitar 0,5 atau lebih dianggap baik untuk model interpretatif (Röder, Both, & Hinneburg, 2015). Kualitas model LDA sebaiknya dievaluasi menggunakan kedua metrik, dengan fokus lebih besar pada *coherence score* jika interpretasi topik menjadi tujuan utama. Sehingga fokus utama pertimbangannya yaitu mengambil *coherence score* yang tinggi tetapi juga dengan nilai *perplexity* yang ideal. Angka ini cukup optimal untuk menghasilkan pengelompokan topik yang memiliki makna yang berkaitan erat dan konsisten, sekaligus baik dalam memprediksi distribusi kata di dalam setiap topik. Dengan demikian maka jumlah topik yang digunakan pada langkah selanjutnya yaitu 9 topik.

Evaluasi Hasil

Setelah dilakukan *topic modelling* menggunakan *Latent Dirichlet Allocation* dengan jumlah topik sebanyak 9 dari data yang sudah dilakukan *text-preprocessing*, didapatkan hasil berupa *dominant topic, number of documents, topic contribution,* dan *topic keywords.*



Gambar 4.Visualisasi Topic Modelling pyLDAvis

Visualisasi hasil LDA menggunakan pyLDAvis menampilkan *Intertopic Distance Map*, di mana setiap lingkaran mewakili topik yang teridentifikasi. Topik 1 memiliki kontribusi terbesar (18% dari total *corpus*) dengan cakupan tema yang luas, sedangkan Topik 9 memiliki kontribusi terkecil (5,7%) dan mencerminkan tema yang lebih spesifik. Kedekatan antar *cluster*, seperti pada Topik 1, 3, dan 4, serta Topik 2 dan 9, menunjukkan kesamaan distribusi *keyword*, yang mengindikasikan hubungan tema di antara topik tersebut. Sebaliknya, *cluster* yang terpisah, seperti Topik 5 hingga 8, menunjukkan tema yang lebih unik dan spesifik.

Pada *Top-30 Most Salient Terms*, visualisasi ini menggambarkan kata-kata dengan frekuensi tinggi dan relevansi kuat dalam *corpus*. Kata seperti "korban", "laku", dan "sangka" muncul dengan frekuensi tertinggi, menyoroti pentingnya kata-kata ini dalam menggambarkan data secara keseluruhan. Kata-kata lain seperti "sambo", "polisi", dan "pidana" juga sering muncul, memberikan konteks terkait hukum dan kriminalitas. *Slider* relevansi (λ) yang diatur pada nilai 1 menekankan frekuensi kata tanpa mempertimbangkan eksklusivitasnya terhadap topik tertentu. Visualisasi ini memberikan wawasan penting tentang distribusi kata dan topik dalam *corpus*.

Tabel 9. Topic Volume Distribution

Dominant Topic	*Num of Documents*	*Topic Contribution*	*Topic Keywords*
1	3572	0,152402082088915	indonesia, masyarakat, orang, negara, perintah, hukum, tindak, keras, kait, anggota, kerja, langgar, tingkat, menteri, atur
2	1313	0,056020138237051	bunuh, temu, rumah, korban, tinggal, polisi, keluarga, tewas, luka, saksi, nggak, duga, hilang, mati, bawa
3	5060	0,215888727707142	korban, laku, anak, perkosa, polisi, sangka, lapor, orang, rumah, aku, duga, tangkap, polres, perempuan, inisial
4	1362	0,0581107603037802	sangka, polisi, polda, duga, metro_jaya, periksa, polri, lapor, kombes, sidik, tangkap, tahan, kait, jakarta, tetap
5	938	0,0400204795631027	lapor, surat, hukum, bandung, aku, kait, duga, bukti, nama, terima, acara, video, kpk, jabar, proses
6	2547	0,108669681713457	sambo, brigadir, polri, j, ferdy, bunuh, putri, tembak, bharada_e, yosua, irjen, rencana, pasal, jakarta, putri_candrawathi
7	1667	0,0711238160252581	pasal, hukum, dakwa, pidana, jaksa, hakim, adil, ayat, sidang, penjara, tuntut, kuhp, undang, putus, tindak

8	1738	0,0741530847341923	uang, rp, tipu, juta, lapor, indonesia, korban, milik, usaha, rugi, orang, data, miliar, nomor, bayar
9	5241	0,223611229627101	laku, curi, motor, warga, polisi, tangkap, aksi, korban, aman, mobil, hasil, orang, barang, milik, jalan

Tabel 9 menunjukkan hasil analisis terhadap kumpulan dokumen yang telah dikelompokkan menjadi 9 topik utama. Setiap topik diidentifikasi berdasarkan kata kunci dominan *(topic keywords)*, jumlah dokumen yang berkontribusi pada topik tersebut *(num of documents)*, dan proporsi kontribusi topik terhadap keseluruhan *corpus (topic contribution)*. Informasi ini memberikan gambaran menyeluruh tentang tema-tema yang muncul dalam *corpus* serta relevansi masing-masing topik terhadap keseluruhan data.

Tabel 10. Hasil Interpretasi Topik

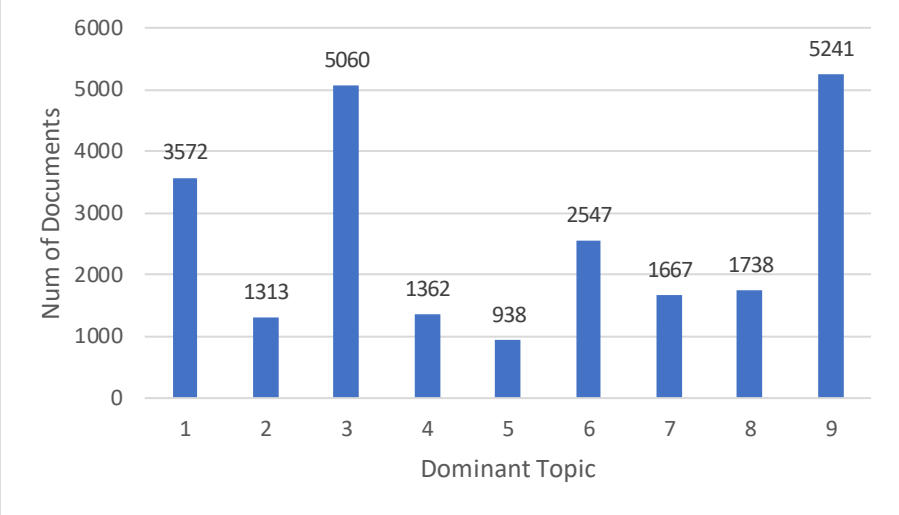
Topic	*Keywords*	*Label*
1	indonesia, masyarakat, orang, negara, perintah, hukum, tindak, keras, kait, anggota, kerja, langgar, tingkat, menteri, atur	Isu Sosial, Masyarakat, dan Hukum
2	bunuh, temu, rumah, korban, tinggal, polisi, keluarga, tewas, luka, saksi, nggak, duga, hilang, mati, bawa	Kasus Pembunuhan dan Kekerasan Berat
3	korban, laku, anak, perkosa, polisi, sangka, lapor, orang, rumah, aku, duga, tangkap, polres, perempuan, inisial	Kekerasan terhadap Anak dan Perempuan
4	sangka, polisi, polda, duga, metro_jaya, periksa, polri, lapor, kombes, sidik, tangkap, tahan, kait, jakarta, tetap	Investigasi dan Penegakan Hukum oleh Polisi
5	lapor, surat, hukum, bandung, aku, kait, duga, bukti, nama, terima, acara, video, kpk, jabar, proses	Kasus Hukum Spesifik dan Regional
6	sambo, brigadir, polri, j, ferdy, bunuh, putri, tembak, bharada_e, yosua, irjen, rencana, pasal, jakarta, putri_candrawathi	Kasus Hukum Besar dan Tokoh Publik (Kasus Ferdy Sambo)
7	pasal, hukum, dakwa, pidana, jaksa, hakim, adil, ayat, sidang, penjara, tuntut, kuhp, undang, putus, tindak	Proses Peradilan dan Putusan Hukum
8	uang, rp, tipu, juta, lapor, indonesia, korban, milik, usaha, rugi, orang, data, miliar, nomor, bayar	Kasus Penipuan, Utang, dan Kerugian Materi
9	laku, curi, motor, warga, polisi, tangkap, aksi, korban, aman, mobil, hasil, orang, barang, milik, jalan	Pencurian dan Pelanggaran Hak Milik

Dari 9 topik yang sudah dihasilkan disimpulkan dengan mengamati korelasi antara 15 kata yang paling sering muncul di setiap topik. *Human judgement* dilakukan untuk membantu memeriksa kata kunci dari hasil topik dan kemudian memberikan label topik secara manual (Chang, Gerrish, Wang, Blei, & Boyd-Graber, 2009).

Tabel 11. Contoh Berita dari Topik 3

Topik	3
Judul Berita	Kata Keluarga Korban soal Viral Pemotor Culik dan Buka-buka Rok Bocah
Tanggal Publikasi	Jumat, 13 Mar 2020 16:04 WIB
URL	https://news.detik.com/berita-jawa-tengah/d-4937894/kata-keluarga-korban-soal-viral-pemotor-culik-dan-buka-buka-rok-bocah
Isi Berita	Video pemotor yang diduga berusaha menculik dan melecehkan seorang bocah di Kota Yogyakarta viral di media sosial (medsos). Korban seorang anak usia 5 tahun disebut mau diajak pelaku karena diiming-imingi jajanan dan diajak ke JEC.

	Hal itu diungkapkan oleh salah seorang kerabat korban, berinisial S (50). Saat ditemui, S mengaku mengetahui korban diantar seorang ibu-ibu ke rumahnya dengan kondisi masih menangis, Kamis (12/3) sekitar pukul 16.00 WIB kemarin. "Jadi (korban) diantar ibu dari teman sekolah anak itu (korban), sampai sini (rumah korban) masih nangis," katanya saat ditemui di rumah korban, Kecamatan Kotagede, Kota Yogyakarta, Jumat (13/3/2020)…



Gambar 5. Grafik Jumlah Dokumen Setiap Topik

Gambar 5 menggambarkan jumlah dokumen yang mendominasi setiap topik dari hasil analisis topik menggunakan model *topic modelling*. Berdasarkan grafik, terlihat bahwa Topik 9 adalah topik dengan jumlah dokumen terbanyak, mencapai 5.241 dokumen, menunjukkan dominasi yang signifikan dibandingkan dengan topik lainnya. Topik 3 berada di posisi kedua, dengan jumlah dokumen sebanyak 5.060, yang juga memberikan kontribusi yang besar terhadap keseluruhan dataset. Kedua topik ini bersama-sama mencakup lebih dari sepertiga dari total dokumen, menegaskan bahwa data cenderung terfokus pada dua topik utama ini.

Secara keseluruhan, grafik ini menunjukkan adanya variasi yang signifikan dalam distribusi jumlah dokumen per topik. Topik 9 dan Topik 3 mendominasi data, sementara topik lainnya memiliki jumlah dokumen yang lebih kecil dan distribusinya lebih merata.

Tabel 12. Jumlah Topik per Tahun

Year	*Dominant Topic*									**Total**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
2020	872	249	1234	260	105	-	129	189	1128	4166
2021	962	206	1343	408	167	12	244	287	1273	4902
2022	665	275	943	332	176	2113	534	555	516	6109
2023	660	338	1015	242	122	405	531	542	1274	5129
2024	413	245	425	120	368	17	229	165	1050	3132
Total	**3572**	**1313**	**5060**	**1362**	**938**	**2547**	**1667**	**1738**	**5241**	**23438**

Secara keseluruhan, total 23.438 dokumen dianalisis selama lima tahun. Topik 3 merupakan topik dominan secara keseluruhan dengan 5.060 dokumen, menunjukkan bahwa isu-isu dalam topik ini memiliki relevansi yang konsisten dalam pemberitaan kriminalitas. Topik 9 menempati posisi kedua dengan 5.241 dokumen, diikuti oleh Topik 6 dengan 2.547 dokumen, meskipun jumlah ini sebagian besar berasal dari lonjakan di tahun 2022. Sebaliknya, Topik 5 memiliki jumlah dokumen paling sedikit secara keseluruhan (938 dokumen), menunjukkan bahwa isu ini kurang mendapat perhatian dalam berita kriminal.

KESIMPULAN

Penelitian ini berhasil mengidentifikasi topik utama dari berita kejahatan di Indonesia menggunakan metode *Latent Dirichlet Allocation* (LDA). Data yang dianalisis diperoleh melalui *web scraping* dari detik.com dan cnnindonesia.com tahun 2020–2024 sebanyak 23.248 berita, yang diolah melalui *text preprocessing*. Dari 20 topik yang dilakukan pengujian, model terbaik ditemukan pada 9 topik dengan *coherence score* sebesar 0,538163893830327 dan *perplexity* sebesar -7,85722881473597, menunjukkan topik yang interpretatif dan distribusi data yang baik. Topik utama meliputi isu sosial, kekerasan, investigasi polisi, kasus hukum, peradilan, penipuan, hingga pencurian. Tema dominan adalah kekerasan terhadap anak dan perempuan (21,59%) dengan jumlah sebanyak 5.060 dokumen serta pencurian (23,26%) dengan jumlah sebanyak 5.241 dokumen. Hasil penelitian ini memberikan wawasan mengenai tren kriminalitas di Indonesia dalam konteks sosial dan hukum. Penelitian selanjutnya dapat memperluas sumber data dengan mencakup lebih banyak situs berita atau platform media sosial untuk meningkatkan representativitas. Periode analisis juga dapat diperpanjang guna memahami tren jangka panjang. Selain itu, analisis dapat ditingkatkan dengan memasukkan aspek sosial, ekonomi, atau demografis untuk memahami faktor yang memengaruhi tren kriminalitas. Metode analisis juga dapat didiversifikasi, misalnya dengan analisis sentimen, analisis jaringan, atau evaluasi model yang lebih variatif. Fokus penelitian dapat diperluas pada jenis kejahatan tertentu, seperti kejahatan siber, terorisme, atau perdagangan manusia, untuk mengungkap tren yang lebih spesifik.

REFERENSI

Alghamdi, R., & Alfalqi, K. (2015). A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications, 6*(1), 147-153.

Astuti, A. R., & Cahyono, N. (2023). Analisis Topic Modelling Persepsi Pengguna Internet Menggunakan Metode Latent Dirichlet Allocation. *Indonesian Journal of Computer Science*, 326-334.

Ayani, D. D., Pratiwi, H. S., & Muhardi, H. (2019). Implementasi Web Scraping untuk Pengambilan Data pada Situs Marketplace. *Jurnal Sistem dan Teknologi Informasi, 7*(4), 257-262.

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM, 55*(4), 77-84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 993-1022.

Chang, J., Gerrish, S., Wang, C., Blei, D. M., & Boyd-Graber, J. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems, 32*, 288-296.

Guo, Y., Han, S., Li, Y., Zhang, C., & Bai, Y. (2018). K-Nearest Neighbor combined with guided filter for hyperspectral image classification. *Procedia Computer Science*, 159-165.

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2020). Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. *Organizational Research Methods, 25*(1), 114-146.

Hidayatullah, A. F., Ma'arif, M. R., Habibie, M., & Khomsah, S. (2020). Indonesia Infrastructure Development Topic Discovery on Online News with Latent Dirichlet Allocation. *IOP Conf. Series: Materials Science and Engineering, 1077*.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Application, 78*(11), 15169-15211.

Julianto, I. T., Kurniadi, D., & Jr, B. B. (2023). Enhancing Sentiment Analysis With Chatbots:a Comparative Study Of Text Pre-processing. *Jurnal Teknik Informatika (JUTIF)*, 1419-1430.

Puspitarani, Y., & Zulpratita, U. S. (2020). Preparatory Document Structuring Technique. *International Journal of Psychosocial Rehabilitation, 24*(2), 3293-3302.

Ridwansyah, T. (2022). Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia Di Twitter Terhadap Kota Medan Menggunakan K-Fold Cross Validation Dan Naïve Bayes Classifie. *Kajian Ilmiah Informatika dan Komputer*, 178-185.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM)*, 399-408.

Rohman, M. S., Santoso, H. A., Saraswati, G. W., & Winarsih, N. A. (2019). Pemanfaatan Topic-Focused Crawler untuk Pembangunan Corpus Berita Bencana menggunakan Teknik Scrapy CSS Selector. *Seminar Nasional APTIKOM (SEMNASTIK)*, 250-258.

Sari, W. A., & Purnomo, H. D. (2022). Topic Modeling Using The Latent Dirichlet Allocation Method On Wikipedia Pandemic Covid-19 Data In Indonesia . *Jurnal Teknik Informatika (JUTIF), 3*(5), 1223-1230 .

Satriajati, S., Panuntun, S. B., & Pramana, S. (2020). Implementasi Web Scraping Dalam Pengumpulan Berita Kriminal Pada Masa Pandemi COVID-19. *Seminar Nasional Official Statistics*, 300-308.

Setiawan, G. H., Adnyana, I. M., Sugiartha, I. G., & Budiarta, K. (2023). Ekstraksi Topik Pada Aduan Mahasiswa Dengan Pendeketan Model Latent Dirichlet Allocation (LDA). *Seminar Nasional Corisindo*, 145-150.

Shevendrakumar, D. (2023). Clustering and Retrieval of News articles using Natural Language Processing. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, 7*, 1-5.

Tapsai, C. (2018). Information Processing and Retrieval from CSV File by Natural Language. *2018 IEEE 3rd International Conference on Communication and Information Systems (ICCIS)*, 212-216.

Tong, Z., & Zhang, H. (2016). A Text Mining Research Based On LDA Topic Modelling. *Computer Science & Information Technology, 6*, 201–210.

Tresnasari, N. A., Adji, T. B., & Permanasari, A. E. (2020). Social-Child-Case Document Clustering based on Topic Modeling using Latent Dirichlet Allocation. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 14*(2), 179-188.

Turland, M. (2010). *Php-Architect's Guide to Web Scraping.* Marco Tabini & Associates.

Wardhani, D., Astuti, R., & Saputra, D. D. (2024). Optimasi Feature Selection Text Mining: Stemming dan Stopword untuk Sentimen Analisis Aplikasi SatuSehat. *Journal Of Social Science Research*.