JAMBURA JOURNAL OF INFORMATICS
Volume 8, No. 2 | October 2026 : 117-131
ISSN 2685-4244, doi: 10.37905/jji.v812.40448

Text summarization berbasis klasterisasi K-Means dan *TextRank* pada berita berbahasa Indonesia

Regita Putri Permata, Aqilah Jihan Nabila, Jehan Rizky Abdillah, Nadline Virgita
Program Studi Sains Data, Telkom University, Ketintang 156, Surabaya, Indonesia

Riwayat Artikel:
Diterima 16 Juli 2026
Revisi 17 September 2026
Disetujui 10 Oktober 2026

Kata Kunci:
Text Summarization
Extractive
K-Means
TextRank
Topik

ABSTRACT. The growth in the volume of digital news in Indonesia drives the need for automated methods to efficiently group and summarize information. This study developed a text mining pipeline that integrated clustering and extractive summarization to summarize Indonesian-language news by topic. From an initial pool of 80,479 articles, 67,370 remained after cleaning; 5,000 were randomly sampled for computational efficiency, and 3,331 documents remained after filtering and were used for the final analysis. Documents were represented using unigram-bigram TF-IDF, dimensionality-reduced via truncated SVD, and clustered using K-Means; the optimal number of clusters was determined by combining the Elbow Method, Silhouette Score, and Davies-Bouldin Index. Each cluster was then summarized using TextRank and evaluated with ROUGE. The analysis identified 11 clusters as optimal, with a Silhouette Score of 0.0675, producing thematically interpretable groupings such as disasters, politics, law, and crime, although the Silhouette Score itself indicates weak cluster separation. The average ROUGE-1, ROUGE-2, and ROUGE-L scores obtained were 0.1938, 0.0743, and 0.1299, respectively, with the best performance observed in clusters with specific, cohesive topics and the lowest in large, heterogeneous clusters. These results indicate an observed association between topic homogeneity and extractive-summary quality; the findings suggest the potential of the proposed pipeline for topic-based extractive summarization of Indonesian news, though this has not yet been benchmarked against a non-clustered TextRank baseline.

ABSTRAK. Pertumbuhan volume berita digital di Indonesia mendorong kebutuhan metode otomatis untuk mengelompokkan dan meringkas informasi secara efisien. Penelitian ini mengembangkan pipeline text mining yang mengintegrasikan klasterisasi dan extractive summarization untuk merangkum berita berbahasa Indonesia berdasarkan topik. Dari 80.479 artikel awal, tersisa 67.370 artikel setelah pembersihan; 5.000 artikel diambil secara acak untuk efisiensi komputasi, dan 3.331 dokumen tersisa setelah penyaringan serta digunakan pada analisis akhir. Dokumen direpresentasikan dengan TF-IDF kombinasi unigram-bigram dan direduksi dimensinya menggunakan Truncated SVD, lalu dikelompokkan dengan algoritma K-Means; jumlah klaster optimal ditentukan melalui kombinasi Elbow Method, Silhouette Score, dan Davies-Bouldin Index. Setiap klaster kemudian diringkas menggunakan metode TextRank dan dievaluasi dengan metrik ROUGE. Analisis menghasilkan 11 klaster optimal dengan Silhouette Score sebesar 0,0675, menghasilkan pengelompokan tematik yang cukup dapat diinterpretasikan seperti topik bencana, politik, hukum, dan kriminalitas, meskipun nilai Silhouette tersebut menunjukkan pemisahan klaster yang masih lemah. Rata-rata skor ROUGE-1, ROUGE-2, dan ROUGE-L yang diperoleh adalah 0,1938, 0,0743, dan 0,1299, dengan performa terbaik pada klaster bertopik spesifik dan kohesif serta terendah pada klaster besar yang heterogen. Hasil ini menunjukkan adanya asosiasi yang teramati antara homogenitas topik dan kualitas ringkasan ekstraktif; temuan ini mengindikasikan potensi pipeline yang diusulkan untuk extractive summarization berbasis topik pada berita Indonesia, meskipun belum dibandingkan dengan baseline TextRank tanpa klasterisasi.

This is an open-access article under the CC–BY-SA license.

Penulis Korespondensi:
Regita Putri Permata
Telkom University
Jalan Ketintang 156, Surabaya, Indonesia.
Email: regitapermata@telkomuniversity.ac.id

PENDAHULUAN

Perkembangan teknologi informasi dan internet mendorong pertumbuhan data digital yang sangat pesat, khususnya data teks dari portal berita daring, media sosial, dan blog. Ketersediaan informasi yang melimpah mempermudah masyarakat memperoleh berita, namun menimbulkan information overload, yaitu kondisi ketika jumlah informasi yang tersedia melebihi kemampuan pengguna untuk membaca dan memahaminya. Kondisi ini mendorong kebutuhan akan sistem yang mampu menyajikan informasi penting secara ringkas dan efisien (Jin, Paik, Yun, Jang, & Kim, 2024).

Salah satu pendekatan untuk mengatasi permasalahan tersebut adalah text summarization, yaitu proses menghasilkan ringkasan yang merepresentasikan informasi penting dari satu atau beberapa dokumen dalam bentuk yang lebih singkat tanpa menghilangkan makna utama. Dalam bidang Natural Language Processing (NLP), topik ini terus berkembang karena berperan penting dalam pengelolaan dokumen digital, pencarian informasi, dan analisis berita berskala besar.

Secara umum terdapat dua pendekatan utama, yaitu extractive summarization dan abstractive summarization. Pendekatan extractive memilih kalimat-kalimat penting dari dokumen asli, sedangkan abstractive menyusun kalimat baru melalui pemahaman dan penyusunan ulang informasi (Putri *et al*., 2025). Pendekatan extractive masih banyak digunakan karena lebih sederhana, tidak memerlukan data pelatihan dalam jumlah besar, serta stabil pada dokumen berita. Salah satu algoritmanya adalah TextRank, metode berbasis graf yang mengadaptasi konsep PageRank untuk menilai kepentingan kalimat berdasarkan kemiripannya dengan kalimat lain, dan bersifat unsupervised sehingga mudah diterapkan pada berbagai jenis dokumen.

Namun, peringkasan pada kumpulan dokumen dengan topik beragam sering menghasilkan ringkasan yang kurang fokus karena informasi lintas topik tercampur dalam satu proses. Karena itu diperlukan klasterisasi teks agar peringkasan dilakukan pada kelompok dokumen yang lebih homogen; K-Means banyak dipilih karena sederhana, efisien, dan mampu menangani dokumen dalam jumlah besar. Representasi dokumen pada penelitian ini menggunakan Term Frequency–Inverse Document Frequency (TF-IDF) yang membobot kata berdasarkan tingkat kepentingannya dalam dokumen dan koleksi dokumen (Hidayat, Yaqin, & Abidin, 2025).

Beberapa penelitian terdahulu telah mengombinasikan klasterisasi dengan multi-document summarization pada berita, misalnya integrasi TextRank dengan K-Means untuk memperbaiki koherensi ringkasan (Liu *et al*., 2024) dan pemanfaatan klasterisasi berbasis TF-IDF untuk mengelompokkan berita berbahasa Indonesia (Purniawan, Sasmita, & Pratama, 2022). Pada ranah peringkasan berita berbahasa Indonesia secara khusus, penelitian yang ada umumnya menerapkan TextRank atau LexRank pada dokumen tunggal atau kumpulan kecil dokumen (Lin, Li, & Jiang, 2022; Tuhpatussania *et al*., 2022), tanpa menguji secara eksplisit apakah tahap klasterisasi topik sebelum summarization benar-benar meningkatkan fokus dan kualitas ringkasan dibandingkan TextRank tanpa klasterisasi. Celah ini menjadi dasar penelitian ini: penelitian sebelumnya belum menyediakan perbandingan terkontrol antara peringkasan dengan dan tanpa klasterisasi pada korpus berita Indonesia berskala menengah, sehingga kontribusi klasterisasi terhadap kualitas ringkasan ekstraktif belum terukur secara kuantitatif.

Berdasarkan celah tersebut, penelitian ini mengusulkan dan menguji sebuah pipeline yang mengintegrasikan klasterisasi K-Means dengan TextRank untuk extractive summarization pada dataset berita berbahasa Indonesia. Kontribusi penelitian ini adalah: (1) menyusun pipeline text mining yang menggabungkan representasi TF-IDF, reduksi dimensi SVD, klasterisasi K-Means, dan TextRank untuk peringkasan berita Indonesia berdasarkan topik; dan (2) mengamati hubungan deskriptif antara homogenitas topik dalam klaster dengan kualitas ringkasan yang dihasilkan, diukur dengan ROUGE. Pertanyaan penelitian yang diajukan adalah: apakah pengelompokan dokumen berdasarkan topik sebelum tahap summarization berasosiasi dengan ringkasan ekstraktif yang lebih fokus dibandingkan tanpa pengelompokan tersebut? Dokumen berita terlebih dahulu

direpresentasikan menggunakan TF-IDF dan dikelompokkan menggunakan algoritma K-Means berdasarkan kesamaan topik. Selanjutnya, setiap klaster diringkas menggunakan algoritma TextRank, dan kualitas ringkasan yang dihasilkan dievaluasi menggunakan metrik ROUGE.

METODE

Dataset

Dataset yang digunakan dalam penelitian ini merupakan kumpulan artikel berita berbahasa Indonesia yang dikumpulkan dari berbagai portal berita daring nasional, antara lain Kompas, Detik, Tempo, CNN Indonesia, Republika, dan Tribunnews. Dataset mentah awalnya terdiri dari 80.479 artikel dengan 60 kolom atribut. Kolom-kolom utama yang relevan meliputi Judul, Content, Waktu, Link, kolom tag kategori (tag1 hingga tag59), serta kolom source yang menyimpan informasi portal asal artikel. Berikut raw dataset yang digunakan pada penelitian ini disajikan pada Gambar 1.

(80479, 60)

	Judul	Waktu	Link	Content	tag1	tag2	tag3	tag4	tag5	source	...	Unnamed: 50	Unnamed: 51	Unnamed: 52	Unnamed: 53	Unnamed: 54	Unnamed: 55	Unnamed: 56	Unnamed: 57	Unnamed: 58	Unnamed: 59
0	Viral Isu PHK Buruh Gudang Garam, Said Iqbal: ...	6-Sep-25	https://nasional.kompas.com/read/2025/09/06/14...	JAKARTA, KOMPAS.com – Presiden Konfederasi Se...	Said Iqbal	industri rokok	PT Gudang Garam	PHK massal	phk massal 2025 terbaru	kompas	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Gempa M 5,3 Guncang Pulau Doi Maluku Utara	Senin, 12 Agu 2024 21:58 WIB	https://news.detik.com/berita/d-7486691/gempa-...	Gempa bumi berkekuatan magnitudo (M) 5,3 mengg...	pulau doi	gempa	NaN	NaN	NaN	detik	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Toko Emas Palsu di Riau Dibongkar Polisi, Peri...	Rabu, 30 Jul 2025 22:22 WIB	https://news.detik.com/melindungi-tuah-manvah/...	Satreskrim Polres Bengkalis membongkar praktik...	pemalsuan emas	emas palsu	polres bengkalis	polda riau	melindungi tuah manvah	detik	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Minyakita Tak Sesuai Ukuran juga Ditemukan di ...	Senin, 10 Mar 2025 23:15 WIB	https://news.detik.com/berita/d-7816829/minyak...	Polisi mendatangi salah satu gudang Minyakita ...	minyakita	kudus	NaN	NaN	NaN	detik	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	Pimpin LDP, Sanae Takaichi Calon Kuat PM Perem...	4 Oktober 2025 | 14.00 WIB	https://www.tempo.co/internasional/pimpin-ldp-...	Baca berita dengan sedikit iklan, klik di sin...	jepang	perdana-menteri	sanae-takaichi	perempuan	ldp	tempo	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
80474	KPK Sebut SK Pencabutan IUP 4 Perusahaan Tamba...	21 Oktober 2025	https://nasional.kompas.com/read/2025/10/21/22...	JAKARTA, KOMPAS.com - Komisi Pemberantasan Ko...	KPK	Dian Patria	Raja Ampat	IUP Tambang	NaN	kompas	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
80475	10 Museum yang Paling Banyak Dikunjungi di Dunia	8 September 2025 | 16.00 WIB	https://www.tempo.co/hiburan/10-museum-yang-pa...	Baca berita dengan sedikit iklan, klik di sin...	museum	museum-louvre	museum-vatikan	metropolitan-museum-of-art	NaN	tempo	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
80476	Ongkos Transportasi di Bekasi Paling Mahal Se-...	1 Agustus 2025	https://megapolitan.kompas.com/read/2025/08/01...	KOMPAS.com – Kota Bekasi tercatat sebagai kot...	BPS	Bekasi	Biaya Transportasi	ongkos	Transportasi	kompas	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
80477	Cerita Siswa Sekolah Rakyat Dapat Secarik Pesa...	Kamis, 11 Sep 2025 22:04 WIB	https://news.detik.com/berita/d-8107916/cerita...	Presiden Prabowo Subianto menulis secarik pesa...	kemensos	prabowo subianto	sekolah rakyat	NaN	NaN	detik	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
80478	Protes Meletus di Seluruh Eropa Usai Serangan ...	2 Oktober 2025 | 12.15 WIB	https://www.tempo.co/internasional/protes-mele...	Baca berita dengan sedikit iklan, klik di sin...	israel	gaza	global-sumud-flotilla	kapal	flotilla	tempo	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

80479 rows × 60 columns

Gambar 1. Raw dataset

Artikel yang memiliki nilai kosong (*null*) pada kolom Content dihapus, begitu pula duplikasi konten, sehingga jumlah data berkurang menjadi 67.370 artikel. Karena pertimbangan efisiensi komputasi, dilakukan pengambilan sampel acak sebanyak 5.000 artikel menggunakan random_state = 42 untuk menjaga reproduksibilitas hasil. Kolom yang dipertahankan untuk proses selanjutnya hanya Judul dan Content.

Text Mining

Text mining merupakan proses ekstraksi informasi dan pengetahuan dari data teks tidak terstruktur secara otomatis menggunakan teknik kecerdasan buatan, statistik, dan linguistik komputasional. Karena sifatnya yang tidak terstruktur, data teks memerlukan tahapan pengolahan khusus yang umumnya mencakup pengumpulan data, *preprocessing* teks, representasi fitur, dan penerapan algoritma analisis seperti klasifikasi, klasterisasi, atau peringkasan teks (Purniawan *et al*., 2022).

Text Summarization

Text summarization atau peringkasan teks adalah proses menghasilkan versi ringkas dari satu atau beberapa dokumen yang tetap mempertahankan informasi penting dan makna utama dari teks aslinya. Tujuan utama dari *text summarization* adalah membantu pengguna memperoleh inti informasi secara cepat tanpa harus membaca keseluruhan dokumen, yang sangat relevan dalam kondisi ledakan informasi digital seperti saat ini (Jin *et al*., 2024). Dalam bidang Natural Language Processing (NLP), *text summarization* telah menjadi salah satu topik penelitian yang aktif dikembangkan karena memiliki aplikasi luas, mulai dari peringkasan berita, abstraksi dokumen ilmiah, hingga pembuatan laporan otomatis (Widyassari *et al*., 2022).

Text summarization terdiri dari dua pendekatan utama, yaitu *extractive* dan *abstractive summarization*. *Extractive summarization* menghasilkan ringkasan dengan memilih kalimat-kalimat penting langsung dari dokumen asli, sedangkan *abstractive summarization* menyusun kalimat baru yang merepresentasikan isi dokumen secara keseluruhan (Putri *et al*., 2025).

Meskipun abstractive menghasilkan ringkasan yang lebih alami, pendekatan ini memerlukan model bahasa yang kompleks dan data pelatihan yang besar. Sebaliknya, *extractive summarization* lebih sederhana, tidak memerlukan data pelatihan, serta terbukti efektif dan stabil untuk berbagai jenis dokumen, termasuk dokumen berita berbahasa Indonesia (Girsang & Amadeus, 2023). Oleh karena itu, penelitian ini menggunakan pendekatan *extractive summarization* dengan algoritma *TextRank* yang memanfaatkan struktur graf dan prinsip *PageRank* untuk menentukan tingkat kepentingan setiap kalimat dalam dokumen.

Text Preprocessing

Text preprocessing merupakan tahap awal untuk membersihkan dan menormalisasi data teks sebelum dianalisis. Tahap ini penting karena teks mentah umumnya mengandung berbagai noise yang dapat memengaruhi kualitas hasil text mining. Menurut Juna & Hayaty (2023); Rifaldi *et al*., (2023), tahapan preprocessing meliputi:

Boilerplate Removal

Boilerplate removal menghapus elemen non-konten seperti iklan, tautan, teks navigasi, nama portal berita, caption foto, dan informasi hak cipta. Proses ini menggunakan fungsi clean_boilerplate() yang menerapkan pola *regular expression* untuk menghapus frasa promosi, nama domain portal berita (kompas.com, tempo.co, detik.com, cnnindonesia.com, tribunnews.com), caption foto, dan teks hak cipta, lalu menormalisasi spasi berlebih.

Case Folding

Case folding mengubah seluruh huruf menjadi lowercase agar kata dengan kapitalisasi berbeda dianggap sebagai token yang sama. Implementasi dilakukan menggunakan fungsi .lower() pada Python sehingga kata seperti "Presiden" dan "presiden" memiliki representasi yang identik.

Cleaning

Cleaning bertujuan menghapus URL, angka, tanda baca, simbol khusus, dan spasi berlebih sehingga diperoleh teks yang bersih untuk proses analisis berikutnya. Pada penelitian ini digunakan beberapa pola regular expression, yaitu r"http\S+|www\S+" untuk menghapus URL, r"[^\w\s]" untuk menghapus tanda baca dan simbol, r"\d+" untuk menghapus angka, serta r"\s+" untuk menormalkan spasi.

Tokenization

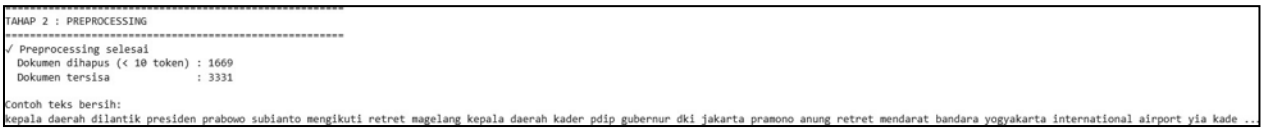
Tokenisasi merupakan proses memecah teks menjadi token atau kata-kata individual sebagai dasar representasi fitur. Pada penelitian ini tokenisasi dilakukan menggunakan fungsi word_tokenize() dari pustaka NLTK.

Stopword Removal

Stopword removal bertujuan menghilangkan kata-kata umum yang kurang informatif sehingga dapat mengurangi dimensi fitur dan meningkatkan relevansi kata yang digunakan dalam analisis (Purniawan *et al*., 2022). Penelitian ini menggunakan gabungan stopword bahasa Indonesia dari NLTK dan daftar stopword tambahan yang disusun secara manual, meliputi nama bulan, nama portal berita, serta kata-kata umum seperti foto, video, iklan, baca, dan klik. Selain itu, token dengan panjang kurang dari atau sama dengan dua karakter juga dihapus menggunakan kondisi len(t) > 2.

Filtering Dokumen Pendek

Tahap terakhir adalah penyaringan dokumen yang memiliki jumlah token hasil preprocessing kurang dari 10 menggunakan kondisi df["teks_bersih"].str.split().str.len() >= 10. Dokumen dengan jumlah token yang terlalu sedikit dianggap tidak memiliki informasi yang cukup representatif untuk proses klasterisasi dan peringkasan. Dari 5.000 artikel yang digunakan, sebanyak **1.669 dokumen** dieliminasi sehingga **3.331 dokumen** digunakan pada tahap analisis selanjutnya.



Gambar 2. Cuplikan tahap 2 *preprocessing*

Pada tahap preprocessing ini (Gambar 2), sebanyak 1.669 dokumen tereliminasi, sehingga jumlah dokumen yang lolos ke tahap berikutnya adalah 3.331 dokumen. Berikut contoh perbandingan teks sebelum dan sesudah preprocessing.

Tabel 1. Cuplikan sebelum dan sesudah preprocessing

	Teks
Sebelum	"Sejumlah kepala daerah yang sudah dilantik Presiden Prabowo Subianto mengikuti retret di Magelang. Namun, kepala daerah kader PDIP, termasuk Gubernur DKI Jakarta Pramono Anung belum ikut retret meski sudah mendarat di Bandara Yogyakarta International Airport (YIA)..."
Sesudah	"kepala daerah dilantik presiden prabowo subianto mengikuti retret magelang kepala daerah kader pdip gubernur dki jakarta pramono anung retret mendarat bandara yogyakarta international airport yia..."

Terlihat bahwa setelah preprocessing, kata-kata stopword seperti "yang", "sudah", "di", "namun", "termasuk", "meski", serta tanda baca berhasil dihilangkan. Teks yang dihasilkan lebih ringkas dan padat secara informasi topik, sehingga lebih siap untuk diproses pada tahap vektorisasi.

Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) adalah metode pembobotan kata yang digunakan untuk merepresentasikan dokumen teks dalam bentuk vektor numerik. Kata yang sering muncul dalam satu dokumen tetapi jarang muncul di dokumen lain akan mendapatkan bobot TF-IDF yang tinggi, yang mengindikasikan bahwa kata tersebut bersifat diskriminatif dan representatif bagi dokumen tersebut (Purniawan *et al*., 2022).

Secara matematis, *Term Frequency–Inverse Document Frequency* (TF-IDF) merupakan metode pembobotan kata yang menggabungkan frekuensi kemunculan kata dalam suatu dokumen dengan tingkat kelangkaannya di seluruh korpus (Salton & Buckley, 1988; Manning *et al*., 2008). *Term Frequency* (TF) mengukur frekuensi kemunculan suatu kata $t$dalam dokumen d. Dalam penelitian ini digunakan sublinear TF, yaitu transformasi logaritmik untuk mengurangi pengaruh kata yang muncul sangat sering (Manning *et al*., 2008):

$$TF(t,d) = \begin{cases} 1+\log(f_{t,d}), & f_{t,d} > 0 \\ 0, & f_{t,d} = 0 \end{cases} \quad (1)$$

dimana $f_{t,d}$ adalah jumlah kemunculan kata t dalam dokumen d. Penggunaan sublinear TF membantu mencegah kata dengan frekuensi sangat tinggi mendominasi bobot secara tidak proporsional (Manning *et al*., 2008). Inverse Document Frequency (IDF) mengukur tingkat kelangkaan suatu kata di seluruh korpus:

$$IDF(t,D) = log(\frac{|D|}{|\{d \in D : t \in d\}|}) \quad (2)$$

di mana $|D|$ adalah jumlah dokumen dalam korpus dan $|\{d \in D : t \in d\}|$ adalah jumlah dokumen yang mengandung kata t (Salton & Buckley, 1988). Nilai TF-IDF diperoleh dari perkalian kedua komponen tersebut:

$$TFIDF(t,d,D) = TF(t,d) \times IDF(t,D) \quad (3)$$

Semakin tinggi nilai TF-IDF, semakin besar tingkat kepentingan suatu kata dalam dokumen relatif terhadap keseluruhan korpus (Manning *et al.*, 2008). Pada penelitian ini TF-IDF dibangun menggunakan parameter:

- max_features = 5000
- ngram_range = (1,2)
- min_df = 2
- max_df = 0.95
- sublinear_tf = True

```
============================================================
HASIL REPRESENTASI DOKUMEN
============================================================

[1] INFORMASI MATRIX TF-IDF
  Shape matrix TF-IDF        : (3331, 5000)
  Jumlah dokumen             : 3331
  Jumlah fitur (term)        : 5000
  Sparsity                   : 98.24%
```

Gambar 3. Informasi matriks TF-IDF

Representasi dokumen menggunakan TF-IDF (Gambar 3) dengan parameter max_features = 5.000, ngram_range = (1,2), min_df = 2, max_df = 0,95, dan sublinear_tf = True. Vektorisasi menghasilkan matriks berukuran 3.331 dokumen × 5.000 fitur dengan tingkat sparsity 98,24%. Sparsity setinggi ini lazim pada representasi TF-IDF karena setiap dokumen hanya memuat sebagian kecil kosakata korpus. Dimensi yang tinggi diatasi menggunakan Truncated SVD.

N-Gram

N-Gram merupakan teknik representasi teks yang memperluas TF-IDF dengan mempertimbangkan urutan kata, tidak hanya kata tunggal (*unigram*) tetapi juga pasangan kata (*bigram*) atau lebih. Pendekatan ini memungkinkan sistem menangkap konteks dan makna frasa yang tidak dapat direpresentasikan oleh unigram saja. Sebagai contoh, frasa *"ridwan kamil"* atau *"pemberantasan korupsi"* memiliki makna yang lebih spesifik dibandingkan setiap katanya secara terpisah. Pada penelitian ini digunakan kombinasi unigram dan bigram (*ngram_range* = (1,2)) sehingga representasi dokumen mencakup kata tunggal maupun pasangan kata yang berurutan (Jain *et al.*, 2024). Secara matematis, himpunan n-gram dari suatu urutan token $W = w_1, w_2, ..., w_m$ didefinisikan pada Persamaan (4):

$$G_n(W) = \{(w_i, w_{i+1}, ..., w_{i+n-1}) | \ 1 \le i \le m - n + 1\} \quad (4)$$

dengan W = urutan token dalam dokumen, m= jumlah token, n= ukuran n-gram, $G_n(W)$ = himpunan seluruh *n-gram* yang terbentuk. Pada penelitian ini digunakan $n = 1$ (unigram) dan $n = 2$ (bigram).

Normalisasi L2

Setelah pembobotan TF-IDF, dilakukan normalisasi L2 agar panjang vektor antar dokumen sebanding. Tanpa normalisasi, dokumen yang lebih panjang cenderung memiliki nilai TF-IDF lebih besar hanya karena mengandung lebih banyak kata (Purniawan *et al.*, 2022). Persamaan normalisasi L2 disajikan pada Persamaan (5)–(7):

$$\mathbf{v} = \frac{\mathbf{v}}{\| \mathbf{v} \|_2} \quad (5)$$

dengan norma Euclidean:

$$\| \mathbf{v} \|_2 = \sqrt{\sum_{i=1}^{n} v_i^2} \quad (6)$$

sehingga setiap elemen hasil normalisasi menjadi

$$\hat{v}_i = \frac{v_i}{\sqrt{\sum_{j=1}^{n} v_j^2}} \quad (7)$$

dengan $\mathbf{v}$= vektor TF-IDF sebelum normalisasi, $\hat{\mathbf{v}}$= vektor TF-IDF setelah normalisasi, v_i= nilai TF-IDF pada dimensi ke-i, $\| \mathbf{v} \|_2$= norma Euclidean (Purniawan *et al*., 2022).

Singular Value Decomposition (SVD)

Representasi TF-IDF menghasilkan matriks berdimensi tinggi dan bersifat *sparse*, sehingga dapat menurunkan kinerja klasterisasi (*curse of dimensionality*). Untuk mengatasinya digunakan Singular Value Decomposition (SVD) yang mereduksi dimensi dengan mempertahankan informasi utama.

$$X = U \Sigma V^T \quad (8)$$

dengan $\mathbf{V}$ sebagai matriks representasi dokumen, $\sum$ sebagai matriks diagonal berisi nilai singular, dan $\mathbf{V}^{\mathbf{T}}$ sebagai matriks representasi term. Dalam penelitian ini digunakan Truncated SVD, yaitu hanya mempertahankan k nilai singular terbesar sehingga komputasi lebih efisien sekaligus mampu menangkap hubungan semantik laten antar dokumen (LSA).

```
[2] INFORMASI MATRIX SVD (setelah reduksi dimensi)
  Shape matrix SVD          : (3331, 100)
  Explained variance ratio  : 0.2577 (25.77%)
```

Gambar 4. Informasi matriks SVD

Reduksi dimensi dilakukan menggunakan *Truncated SVD* ke 100 komponen sehingga matriks representasi akhir berukuran 3.331 × 100 seperti disajikan pada Gambar 4. Nilai *explained variance ratio* sebesar 25,77% menunjukkan bahwa 100 komponen tersebut menangkap sekitar seperempat variasi informasi dalam data. Angka ini wajar untuk data berita dengan topik yang sangat beragam karena tidak ada pola dominan tunggal pada dataset.

Klasterisasi Teks dengan K-Means

K-Means merupakan algoritma klasterisasi unsupervised yang mengelompokkan dokumen berdasarkan kemiripan fitur dengan meminimalkan jarak antara dokumen dan centroid klaster (Purniawan *et al*., 2022). Fungsi objektif K-Means dinyatakan sebagai

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} || x - \mu_i ||^2 \quad (9)$$

dengan C_i adalah klaster ke-i dan μ_i centroid klaster. Klaster menggunakan input TF-IDF dapat meningkatkan performansi model untuk peringkasan dokumen (Surianto & Surianto, 2025).

Klasterisasi dilakukan menggunakan K-Means dengan input berupa representasi TF-IDF yang telah direduksi dimensinya melalui *Truncated SVD*, sehingga representasi dokumen lebih ringkas dan *sparsity* berkurang. Parameter scikit-learn yang digunakan disajikan pada Tabel 2.

Tabel 2. Parameter *K-Means*

Parameter	Nilai
init	k-means++
n_klasters	hasil optimasi (pencarian k=4–12)
n_init	10
max_iter	300
random_state	42
algorithm	lloyd
tol	1.00E-04

Penentuan Jumlah Klaster

Jumlah klaster optimal ditentukan menggunakan *Elbow Method, Silhouette Score, dan Davies-Bouldin Index* (DBI) (Wicaksono *et al*., 2024, (Chicco *et al*., 2025). *Elbow Method* digunakan untuk mengidentifikasi kandidat nilai k melalui grafik WCSS, sedangkan Silhouette Score dan DBI digunakan sebagai evaluasi kuantitatif. Silhouette Score dihitung menggunakan Persamaan (10):

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{10}$$

dengan $a(i)$ sebagai rata-rata jarak dalam klaster dan $b(i)$ sebagai rata-rata jarak ke klaster terdekat. Davies-Bouldin Index dihitung pada Persamaan (11):

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \left(\frac{s_i + s_j}{d_{ij}} \right) \tag{11}$$

Nilai klaster terbaik ditunjukkan oleh Silhouette Score yang tinggi dan DBI yang rendah.

TextRank

TextRank merupakan algoritma *extractive summarization* berbasis graf yang mengadaptasi prinsip PageRank (Gulati *et al*., 2023). Setiap kalimat direpresentasikan sebagai simpul, sedangkan bobot sisi ditentukan menggunakan *cosine similarity* antar vektor TF-IDF (Yulianti *et al*., 2023). Cosine similarity dipercaya sangat sesuai untuk mengitung sebuah kemiripan kalimat (Permata *et al*., 2025). Kemiripan dua kalimat dihitung dengan Persamaan (12):

$$CosineSimilarity(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\| \mathbf{a} \| \| \mathbf{b} \|} \tag{12}$$

Skor setiap kalimat dihitung secara iteratif menggunakan Persamaan (13):

$$WS(V_i) = (1-d) + d \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \tag{13}$$

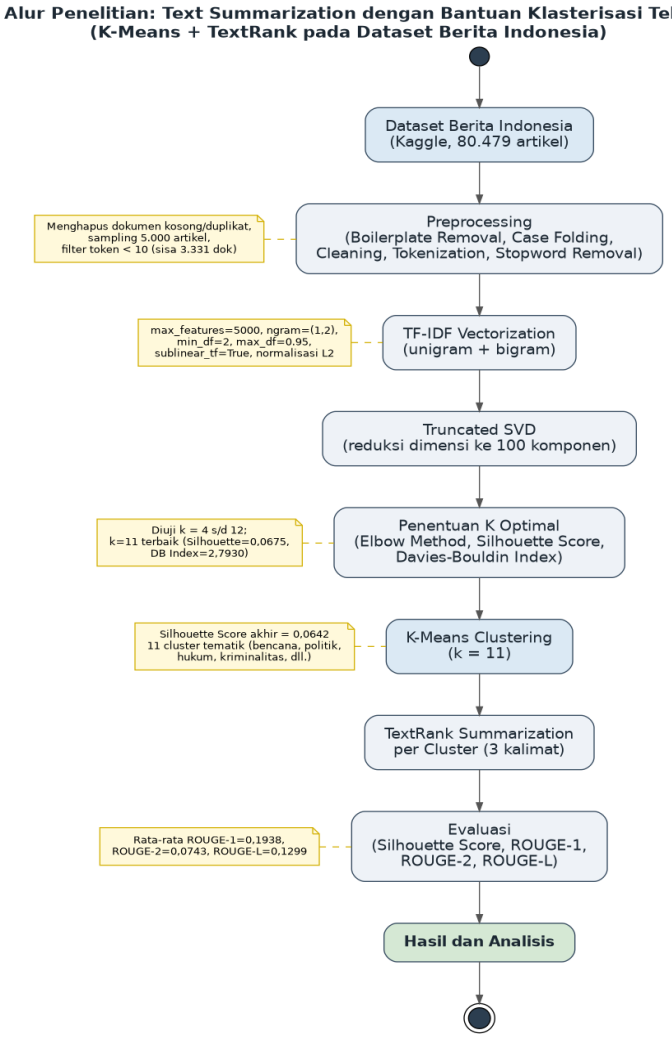
dengan $d = 0.85$. Kalimat dengan skor tertinggi dipilih sebagai ringkasan akhir.

Evaluasi Menggunakan ROUGE

ROUGE merupakan metrik evaluasi otomatis yang membandingkan ringkasan sistem dengan ringkasan referensi berdasarkan tingkat *overlap* (Gulati *et al*., 2023). Penelitian ini menggunakan ROUGE-1, ROUGE-2, dan ROUGE-L. ROUGE-1 dihitung sebagai:

$$ROUGE-1 = \frac{\sum Count_{match}(gram_1)}{\sum Count(gram_1)} \tag{14}$$

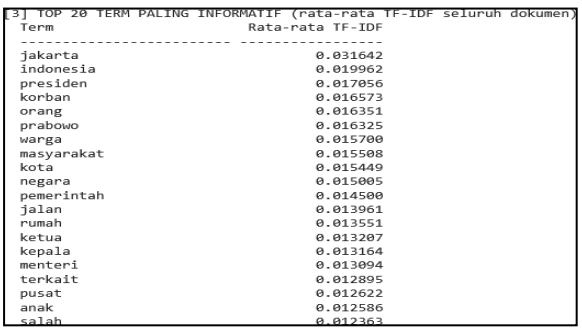
sedangkan ROUGE-2 menggunakan *bigram* dengan persamaan serupa. ROUGE-L mengukur Longest Common Subsequence (LCS) sehingga mampu mengevaluasi kesamaan struktur kalimat. Nilai evaluasi yang digunakan adalah F1-score, karena menyeimbangkan nilai presisi dan recall. Urutan langkah penelitian disajikan pada Gambar 5.



Gambar 5. Flowchart alur penelitian: peringkasan berita indonesia menggunakan metode K-Means dan textrank

HASIL DAN DISKUSI

Hasil Representasi Dokumen TF-IDF



[3] TOP 20 TERM PALING INFORMATIF (rata-rata TF-IDF seluruh dokumen)

Term	Rata-rata TF-IDF
jakarta	0.031642
indonesia	0.019962
presiden	0.017056
korban	0.016573
orang	0.016351
prabowo	0.016325
warga	0.015700
masyarakat	0.015508
kota	0.015449
negara	0.015005
pemerintah	0.014500
jalan	0.013961
rumah	0.013551
ketua	0.013207
kepala	0.013164
menteri	0.013094
terkait	0.012895
pusat	0.012622
anak	0.012586
salah	0.012363

Gambar 6. 20 Term informatif dari TF-IDF

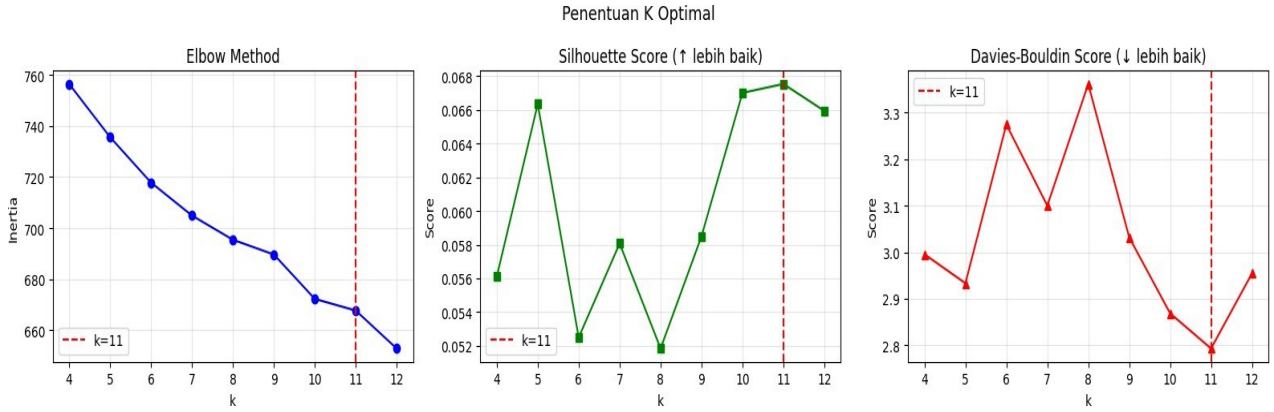
Gambar 6 menampilkan 20 term dengan nilai rata-rata TF-IDF tertinggi di seluruh dokumen. Term "jakarta" mendominasi dengan nilai 0,0316, mencerminkan bahwa sebagian besar artikel berpusat pada peristiwa di wilayah Jakarta. Kehadiran term "presiden", "prabowo", "menteri", "ketua", dan "negara" menegaskan dominasi berita politik dan pemerintahan, sedangkan "korban", "warga", dan "jalan" menunjukkan banyaknya artikel bertema kejadian atau bencana perkotaan. Distribusi ini selaras dengan klaster yang terbentuk pada tahap berikutnya.

Hasil Klasterisasi Dokumen

Hasil klasterisasi dokumen akan melalui tahapan sebagai berikut:

Penentuan Jumlah Klaster Optimal

Penentuan jumlah klaster dilakukan menggunakan tiga pendekatan sekaligus, yakni Elbow Method, Silhouette Score, dan Davies-Bouldin Score, pada rentang $k = 4$ hingga $k = 12$. Hasil selengkapnya disajikan pada Gambar 7.



Gambar 7. Penentuan *K* optimal

Berdasarkan Gambar 7 di atas, terlihat bahwa $k = 11$ menghasilkan Silhouette Score tertinggi sebesar **0,0675** sekaligus Davies-Bouldin Score terendah sebesar **2,7930** di antara semua nilai *k* yang diuji. Karena kedua metrik memberikan rekomendasi yang sama, maka **k = 11** dipilih sebagai jumlah klaster optimal. Pada Elbow Method, penurunan inertia mulai melandai setelah k = 9–11, yang turut mendukung pemilihan k = 11.

Distribusi Dokumen per Klaster

Setelah model K-Means dijalankan dengan k = 11, diperoleh Silhouette Score final sebesar **0,0675** dengan distribusi dokumen ke masing-masing klaster sebagaimana ditampilkan pada Tabel 3.

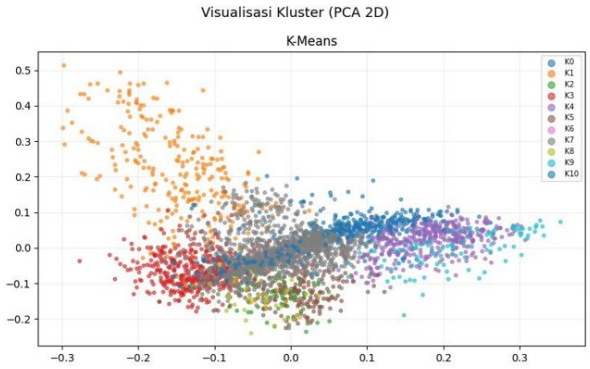
Tabel 3. Distribusi dokumen per klaster

Klaster	0	1	2	3	4	5	6	7	8	9	10	Total
Jumlah Dokumen	380	263	93	496	402	167	23	1.295	43	98	71	**3.331**
Proporsi (%)	11,41	7,90	2,79	14,89	12,07	5,01	0,69	38,88	1,29	2,94	2,13	**100**

Menurut Tabel 3, Klaster 7 mendominasi dengan 1.295 dokumen atau hampir 39% dari seluruh data. Hal ini mencerminkan banyaknya artikel yang bersifat umum dan tidak terikat pada satu topik tunggal, yang wajar mengingat data berasal dari portal berita umum dengan cakupan tema yang sangat luas. Sementara itu, Klaster 6 adalah yang paling kecil dengan hanya 23 dokumen, karena berisi konten yang sangat spesifik dan bersifat musiman (jadwal ibadah Ramadan).

Hasil Visual SVD

Visualisasi PCA 2D pada Gambar 8 memperlihatkan bahwa beberapa klaster relatif terpisah, khususnya kelompok klaster berwarna oranye (K1/Pilkada) serta K10/MBG dan K9/Polda Metro di sisi kanan. Namun, sebagian besar klaster di tengah plot masih bercampur, konsisten dengan nilai Silhouette Score yang rendah dan tingginya tumpang tindih semantik antar topik berita.



Gambar 8. Visualisasi klaster PCA 2D

Interpretasi Topik per Klaster

Interpretasi topik dilakukan berdasarkan 10 term TF-IDF tertinggi pada setiap klaster. Hasil interpretasi selengkapnya ditampilkan pada Tabel 4.

Tabel 4. Interpretasi per topik

Klaster	Top Terms	Label Topik
0	kebakaran, jalan, jakarta, warga, banjir, lokasi, air, petugas, hujan, wib	Bencana & Kejadian Kota
1	pilkada, calon, gubernur, pasangan, ridwan, pilkada jakarta, kamil, ridwan kamil, jakarta, pramono	Pilkada DKI Jakarta
2	kpk, korupsi, pemberantasan, pemberantasan korupsi, komisi pemberantasan, korupsi kpk, komisi, gedung merah, tersangka, gedung	KPK & Pemberantasan Korupsi
3	prabowo, presiden, prabowo subianto, subianto, presiden prabowo, indonesia, menteri, negara, partai, ketua	Pemerintahan Pusat & Presiden
4	korban, pelaku, polisi, polres, metro, kejadian, tersangka, rumah, berinisial, polres metro	Kriminalitas & Kepolisian
5	jaksa, korupsi, pengadilan, pidana, hakim, kejaksaan, terdakwa, agung, hukum, tindak pidana	Persidangan Korupsi
6	puasa, ibadah puasa, ibadah, pengingat, jadwal, jadwal buka, buka puasa, imsak pengingat, imsak, sahur jadwal	Jadwal Ibadah Ramadan
7	jakarta, indonesia, masyarakat, pemerintah, kota, kementerian, warga, negara, menteri, orang	Berita Umum
8	hasto, harun, harun masiku, masiku, kpk, hasto kristiyanto, kristiyanto, suap, pdi, penyidikan	Kasus Harun Masiku
9	polda metro, metro jaya, metro, polda, ade, jaya, ade ary, ary, jaya kombes, korban	Polda Metro Jaya
10	mbg, bergizi, gratis, bergizi gratis, makan bergizi, makan, gizi, program, program makan, sppg	Program Makan Bergizi Gratis

Dari sebelas klaster pada Tabel 4, model menghasilkan pengelompokan yang tampak cukup dapat dibedakan secara tematik pada level kualitatif, meskipun nilai Silhouette Score 0,0675 dan DBI 2,7930 menunjukkan pemisahan yang secara kuantitatif masih lemah. Klaster 0 merepresentasikan kejadian darurat dan bencana perkotaan di Jakarta, khususnya kebakaran dan banjir; kemunculan kata “wib” menandakan banyaknya artikel yang mencantumkan waktu kejadian. Klaster 1 berisi liputan Pilkada DKI Jakarta dan terfokus pada persaingan Ridwan Kamil dan Pramono Anung.

Klaster 2 dan Klaster 5 sama-sama menyentuh isu korupsi namun dari sudut pandang yang berbeda sehingga keduanya terpisah: Klaster 2 berfokus pada proses penyelidikan dan penangkapan tersangka oleh KPK, sedangkan Klaster 5 merepresentasikan liputan persidangan kasus korupsi di Pengadilan Tipikor. Pemisahan ini menunjukkan bahwa model mampu membedakan tahapan penegakan hukum dalam domain topik yang sama. Klaster 3 merepresentasikan liputan kebijakan dan aktivitas pemerintahan Presiden Prabowo Subianto beserta jajaran kabinetnya, menjadikannya klaster terbesar kedua dengan 496 dokumen.

Klaster 4 berisi laporan kriminal dengan ciri khas penyamaran identitas tersangka, tercermin dari kemunculan term “berinisial”. Klaster 6 adalah klaster paling homogen meski hanya beranggotakan 23 dokumen, karena seluruh term dominannya khas pada konten jadwal ibadah Ramadan dan tidak ditemukan di klaster lain. Sebaliknya, Klaster 7 merupakan klaster terbesar (1.295 dokumen) yang menampung artikel tanpa topik dominan tunggal dengan term yang generik dan lintas topik, sehingga menjadi salah satu penyebab rendahnya Silhouette Score secara keseluruhan.

Klaster 8 paling spesifik secara tematis karena seluruh term dominannya merujuk pada satu kasus hukum, yaitu penyidikan KPK terkait suap penetapan anggota DPR yang menyeret Sekjen PDIP Hasto Kristiyanto; kekhususan inilah yang membuatnya memperoleh nilai ROUGE tertinggi. Klaster 9 berbeda dari Klaster 4 meski sama-sama berkaitan dengan kepolisian, sebab lebih terfokus pada pernyataan dan konferensi pers resmi Polda Metro Jaya. Klaster 10 terbentuk dengan bersih karena Program Makan Bergizi Gratis memiliki kosakata khas seperti “mbg” dan “sppg” yang belum bercampur dengan topik lain.

Hasil *Text Summarization*

Setelah proses klasterisasi selesai, dilakukan *text summarization* menggunakan metode *TextRank* yang bersifat *extractive*. Untuk setiap klaster, diambil maksimal 15 dokumen yang digabung, lalu diringkas menjadi 3 kalimat representatif. Tabel 5 merupakan contoh ringkasan yang dihasilkan untuk beberapa klaster.

Tabel 5. Ringkasan teks berdasarkan klaster

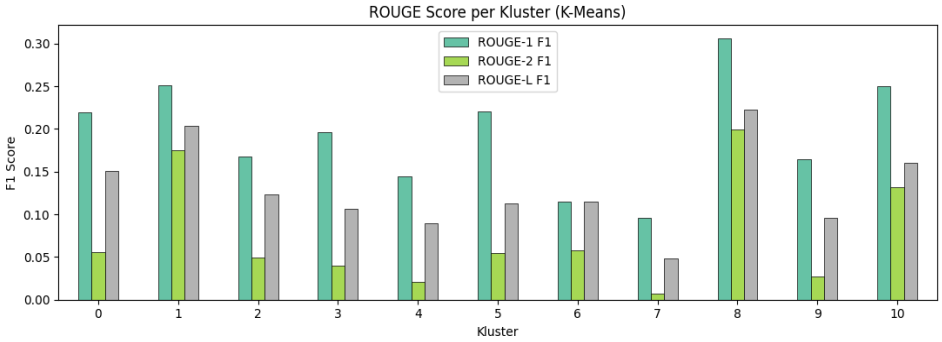
Klaster	Topik	Ringkasan
0	Bencana & Kejadian Kota	"JAKARTA, - Hujan yang mengguyur Jakarta pada Rabu (22/10/2025) malam menyebabkan kepadatan lalu lintas di sepanjang Jalan Sudirman. Bantuan tersebut disalurkan ke tiga wilayah yang paling terdampak banjir, yakni di Jakarta..."
1	Pilkada DKI Jakarta	“JAKARTA, - Calon wakil gubernur (cawagub) Jakarta Rano Karno sempat menyinggung program calon gubernur (cagub) Ridwan Kamil (RK) yang ingin memberi anggaran Rp 200 juta per RW. Bakal calon gubernur DKI Jakarta..."
5	Persidangan Korupsi	"Ketua Majelis Hakim Pengadilan Tipikor Jakarta Pusat, Tony Irfan, menyatakan bahwa Albani, Hasan, dan Buyung terbukti bersalah melakukan tindak pidana korupsi secara bersama-sama, sebagaimana diatur dalam Pasal 2 Ayat (1)..."
8	Kasus Harun Masiku	"Sekretaris Jenderal PDIP Hasto Kristiyanto menyampaikan pembelaan terakhirnya sebelum mendengarkan vonis. Mantan komisioner KPU Wahyu Setiawan tiba di Gedung KPK, Jakarta, 6 Januari 2025..."
10	Program MBG	"JAKARTA, - Pemerintah Provinsi (Pemprov) Jakarta sedang menggodok anggaran untuk program makan bergizi gratis (MBG) dan pendidikan gratis di sekolah swasta. Presiden Prabowo Subianto mempersilakan pemerintah daerah untuk turut berkontribusi..."

Secara umum, ringkasan yang dihasilkan *TextRank* mampu menangkap inti topik dari masing-masing klaster, khususnya pada klaster dengan topik yang kohesif dan jumlah dokumen yang tidak terlalu besar. Sebagai contoh, Klaster 8 menghasilkan ringkasan yang terfokus pada konteks penyidikan KPK terkait kasus Harun Masiku, sementara Klaster 10 berhasil merangkum inti kebijakan Program Makan Bergizi Gratis dengan baik. Adapun Klaster 7 yang sangat besar dan

heterogen menghasilkan ringkasan yang kurang representatif karena dokumen-dokumennya tidak kohesif secara topik.

Evaluasi *Text Summarization*

Evaluasi kualitas ringkasan dilakukan menggunakan ROUGE Score (*Recall-Oriented Understudy for Gisting Evaluation*), yang mengukur tingkat tumpang tindih n-gram antara ringkasan yang dihasilkan model (*hypothesis*) dengan teks referensi. Teks referensi dibentuk dari penggabungan kalimat pertama (*lead sentence*) dari maksimal 10 dokumen sampel acak per klaster, yang diasumsikan merepresentasikan inti topik masing-masing klaster. Tiga varian ROUGE yang digunakan adalah ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), dan ROUGE-L (longest common subsequence).



Gambar 9. Visualisasi Skor ROUGE per Klaster

Berdasarkan hasil evaluasi Gambar 9, Klaster 8 (Kasus Harun Masiku) memperoleh skor ROUGE tertinggi (ROUGE-1 = 0,3063; ROUGE-2 = 0,1994; ROUGE-L = 0,2222), menunjukkan bahwa klaster dengan topik yang homogen menghasilkan ringkasan yang lebih representatif. Klaster 1 (Pilkada Jakarta) juga menunjukkan performa yang baik karena memiliki konsistensi penggunaan istilah dan frasa. Sebaliknya, Klaster 7 memperoleh skor ROUGE terendah (ROUGE-1 = 0,0959; ROUGE-2 = 0,0069; ROUGE-L = 0,0479) akibat tingginya keragaman topik sehingga sulit direpresentasikan oleh satu ringkasan. Secara keseluruhan, rata-rata ROUGE-1 sebesar 0,1938, ROUGE-2 sebesar 0,0743, dan ROUGE-L sebesar 0,1299 menunjukkan performa peringkasan yang cukup baik. Meskipun evaluasi menggunakan lead sentence sebagai referensi sehingga memiliki keterbatasan, hasil yang diperoleh menunjukkan bahwa *TextRank* lebih efektif menghasilkan ringkasan pada klaster yang tematik dibandingkan klaster yang heterogen.

KESIMPULAN

Penelitian ini berhasil menerapkan pendekatan Text Summarization dengan bantuan klasterisasi teks pada dataset berita Indonesia. Proses penelitian diawali dengan tahap preprocessing, representasi dokumen menggunakan TF-IDF, klasterisasi dokumen menggunakan algoritma K-Means, serta pembuatan ringkasan menggunakan metode *TextRank*. Hasil klasterisasi menunjukkan bahwa dokumen berita dapat diklasterkan berdasarkan kesamaan topik sehingga menghasilkan klaster dokumen yang lebih homogen. Ringkasan yang dihasilkan mampu merepresentasikan informasi utama dari dokumen dalam bentuk yang lebih singkat. Berdasarkan hasil evaluasi menggunakan metrik ROUGE, pendekatan yang digunakan dapat digunakan untuk mendukung proses peringkasan otomatis pada kumpulan dokumen berita berbahasa Indonesia. Selain itu, penerapan klasterisasi sebelum proses summarization membantu menghasilkan ringkasan yang lebih terfokus sesuai dengan topik dokumen yang dibahas. Pada tahap summarization, evaluasi ROUGE menunjukkan rata-rata ROUGE-1 sebesar 0,1938, ROUGE-2 sebesar 0,0743, dan ROUGE-L sebesar 0,1299. Performa terbaik diperoleh pada kluster 8 (topik Hasto

Kristiyanto/Harun Masiku) dengan ROUGE-1 0,3063, sedangkan performa terendah terjadi pada kluster 7 yang merupakan kluster terbesar dan paling heterogen, dengan ROUGE-1 hanya 0,0959. Hal ini mengindikasikan bahwa kualitas ringkasan *TextRank* sangat dipengaruhi oleh homogenitas topik dalam kluster; kluster dengan topik yang lebih spesifik dan dokumen lebih sedikit cenderung menghasilkan ringkasan yang lebih representatif dibandingkan kluster besar dengan topik yang beragam. Untuk penelitian selanjutnya, disarankan mengeksplorasi metode reduksi dimensi alternatif atau menambah jumlah komponen SVD untuk meningkatkan explained variance ratio yang saat ini masih tergolong rendah, sehingga representasi dokumen lebih informatif untuk proses klasterisasi. Selain itu, penyeimbangan ukuran kluster dapat dipertimbangkan, misalnya dengan teknik klasterisasi hierarkis atau pembatasan ukuran kluster maksimum, agar kluster besar seperti kluster 7 dapat dipecah menjadi sub-topik yang lebih homogen sehingga kualitas ringkasan meningkat. Penggunaan metode embedding kontekstual seperti IndoBERT atau Sentence-BERT sebagai pengganti TF-IDF juga dapat dipertimbangkan untuk menangkap makna semantik teks berita Bahasa Indonesia secara lebih baik dibandingkan representasi berbasis frekuensi kata. Terakhir, untuk meningkatkan skor ROUGE, dapat digunakan metode yang lebih modern dan dataset yang lebih besar guna meningkatkan kualitas ringkasan.

Ucapan Terima Kasih

Penulis menyampaikan terima kasih kepada Program Studi Sains Data, Telkom University Surabaya, atas dukungan akademik selama pelaksanaan penelitian ini. Apresiasi juga disampaikan kepada dosen pengampu Mata Kuliah Penambangan Teks atas bimbingan dan masukan yang diberikan dalam proses penelitian dan penyusunan artikel.

Referensi

Chicco, D., Campagner, A., Spagnolo, A., Ciucci, D., & Jurman, G. (2025). The Silhouette coefficient and the Davies-Bouldin index are more informative than Dunn index, Calinski-Harabasz index, Shannon entropy, and Gap statistic for unsupervised clustering internal evaluation of two convex clusters. *PeerJ Computer Science, 11*, e3309. https://doi.org/10.7717/peerj-cs.3309

Girsang, A. S., & Amadeus, F. J. (2023). Extractive text summarization for Indonesian news article using ant system algorithm. *Journal of Advances in Information Technology, 14*(2), 295–301. https://www.jait.us/show-227-1320-1.html

Gulati, V., Kumar, D., Popescu, D. E., & Hemanth, J. D. (2023). Extractive article summarization using integrated *TextRank* and BM25+ algorithm. *Electronics, 12*(2), 372. https://doi.org/10.3390/electronics12020372

Hidayat, M. Y., Yaqin, M. A., & Abidin, Z. (2025). Semantic-enhanced news clustering using TF-IDF and WordNet with K-Means. *Journal of Information Systems and Informatics*. https://doi.org/10.63158/journalisi.v7i4.1360

Jain, R., Singh, P., & Puri, S. (2024). Summarization of daily news using *TextRank* and TF-IDF algorithm. In S. Kumar, B. K., J. H. Kim, & J. C. Bansal (Eds.), *Fourth Congress on Intelligent Systems* (Vol. 865, pp. 313–324). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-9043-6_26

Jin, K., Paik, S., Yun, J., Jang, S., & Kim, Y. (2024). PRISM: Personalizing reporting with intelligent summarization through multiple frames. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3505612

Juna, M. F., & Hayaty, M. (2023). The observed preprocessing strategies for doing automatic text summarizing. *Computer Science and Information Technologies, 4*(2), 119–126. https://doi.org/10.11591/csit.v4i2.p119-126

Lin, N., Li, J., & Jiang, S. (2022). A simple but effective method for Indonesian automatic text summarisation. Connection Science, 34(1), 29–43. https://doi.org/10.1080/09540091.2021.1937942

Liu, W., Sun, Y., Yu, B., Wang, H., Peng, Q., Hou, M., Guo, H., Wang, H., & Liu, C. (2024). Automatic text summarization method based on improved TextRank algorithm and K-Means clustering. Knowledge-Based Systems, 287, 111447. https://doi.org/10.1016/j.knosys.2024.111447

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. https://nlp.stanford.edu/IR-book/

Permata, R., Rendika, & Julianty, L. C. (2025). Towards an Automated Essay Evaluation System NLP Based Text Embeddings and Similarity Metrics. *Digital Zone: Jurnal Teknologi Informasi Dan Komunikasi*, *16*(1), 37–46. https://doi.org/10.31849/digitalzone.v16i1.26541

Purniawan, I. M. A., Sasmita, G. M. A., & Pratama, I. P. A. E. (2022). Clustering berita menggunakan algoritma TF-IDF dan K-Means dengan memanfaatkan sumber data crawling pada situs Detik.com. *JITTER: Jurnal Ilmiah Teknologi dan Komputer, 3*(1), 821. https://doi.org/10.24843/JTRTI.2022.v03.i01.p18

Putri, I. G. A. I. U., Trisna, I. N. P., & Rusjayanthi, N. K. D. (2025). Abstractive Text Summarization to Generate Indonesian News Highlight Using Transformers Model. *Journal of Information Systems and Informatics*, *7*(2), 1248–1263. https://doi.org/10.51519/journalisi.v7i2.1082

Rifaldi, D., Fadlil, A., & Herman. (2023). Teknik preprocessing pada text mining menggunakan data tweet "Mental Health." *Decode: Jurnal Pendidikan Teknologi Informasi, 3*(2), 161–171. https://doi.org/10.51454/decode.v3i2.131

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Surianto, D. F., & Surianto, D. F. (2024). Enhancing K-Means clustering for journal articles using TF-IDF and LDA feature extraction. *Brilliance: Research of Artificial Intelligence, 4*(2), 964–972. https://doi.org/10.47709/brilliance.v4i2.5547

Tuhpatussania, S., Utami, E., Hartanto, A. D., *et al*. (2022). Comparison of LexRank algorithm and maximum marginal relevance in summary of Indonesian news texts online. Pilar Nusa Mandiri, 18(2), 187–192. https://doi.org/10.33480/pilar.v18i2.3190

Wicaksono, A. P., Widjaja, S., Nugroho, M. F., & Putri, C. P. (2024). Elbow and silhouette methods for K value analysis of ticket sales grouping on K-Means. *SISTEMASI, 13*(1), 28. https://doi.org/10.32520/stmsi.v13i1.2889

Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., Affandy, A., & Setiadi, D. R. I. M. (2022). Review of automatic text summarization techniques & methods. *Journal of King Saud University-Computer and Information Sciences, 34*(4), 1029–1046. https://doi.org/10.1016/j.jksuci.2020.05.006

Yulianti, E., Pangestu, N., & Jiwanggi, M. A. (2023). Enhanced *TextRank* using weighted word embedding for text summarization. *International Journal of Electrical and Computer Engineering (IJECE), 13*(5), 5472. https://doi.org/10.11591/ijece.v13i5.pp5472-5482