KLASTERISASI PORTAL BERITA ONLINE MENGGUNAKAN ALGORITMA *K-MEANS* DENGAN TEKNIK *KNOWLEDGE DISCOVERY IN DATABASE*

Novita Ramadhini, Salsabila Alfath Annisaa', Adiaz Salsabilah Putri, Ken Ditha Tania, Ahmad Rifai
Sistem Informasi, Universitas Sriwijaya
Jl. Raya Palembang – Prabumulih No.KM. 32, Indralaya Indah,
Kec. Indaralaya, Kabupaten Ogan Ilir, Sumatera Selatan, Indonesia
novitaramadhini0@gmail.com

ABSTRAK

Lonjakan volume informasi pada portal berita online menciptakan tantangan besar dalam pengorganisasian data yang tidak terstruktur dan tidak berlabel. Ketidakmampuan sistem dalam mengelompokkan berita secara otomatis menyebabkan redundansi informasi serta penurunan efisiensi pencarian bagi pengguna. Penelitian ini bertujuan untuk mengimplementasikan klasterisasi berita secara otomatis dengan mengintegrasikan algoritma *K-Means* dan kerangka kerja *Knowledge Discovery in Databases* (KDD). Metode KDD digunakan untuk memastikan proses pengolahan data dilakukan secara sistematis, meliputi tahapan *data selection*, *preprocessing* (*case folding*, *normalization*, *tokenizing*, *stopword removal*, dan *stemming*), *data transformation*, hingga tahap *data mining* menggunakan *K-Means*. Penentuan jumlah klaster optimal dilakukan dengan metode *Elbow* dan divalidasi menggunakan *Silhouette Coefficient*. Hasil penelitian menunjukkan bahwa nilai k **= 7** merupakan jumlah klaster optimal dalam memisahkan topik-topik berita. Evaluasi menggunakan *Silhouette Score* menghasilkan nilai rata-rata positif yang menunjukkan kualitas klaster yang baik. Penelitian ini memberikan kontribusi dalam pengorganisasian berita secara otomatis sehingga mempermudah pengguna dalam menemukan informasi yang relevan dengan lebih cepat dan akurat.

Kata kunci : *Klasterisasi Berita, K-Means, Knowledge Discovery in Databases, Text Mining, Portal Berita Online.*

1. PENDAHULUAN

Akses informasi melalui portal berita online telah mengalami pertumbuhan yang sangat masif, di mana volume dokumen digital diperkirakan telah melampaui angka 550 triliun [1].

Fenomena ini menyebabkan data berita tersaji dalam bentuk yang tidak terstruktur dan seringkali tidak disertai dengan label kategori yang konsisten [2].

Masalah nyata yang muncul adalah tingginya dimensi data yang menghambat proses pengolahan informasi secara cepat, serta adanya redundansi informasi di mana satu topik berita yang sama dipublikasikan oleh berbagai sumber media dengan sudut pandang yang berbeda-beda [3].

Hal ini menyulitkan sistem dalam menemukan pola informasi yang bermakna dari sekumpulan data yang sangat besar.

Dampak dari permasalahan tersebut sangat krusial, terutama pada penurunan efisiensi pencarian informasi oleh pengguna. Rendahnya kualitas pengelompokan berita mengakibatkan terjadinya information *overload*, di mana pembaca justru mengalami kesulitan dalam menyaring berita yang relevan [3].

Secara teknis, pengolahan teks tanpa metode yang tepat menyebabkan peningkatan waktu komputasi dan ketidakakuratan dalam pemetaan topik, sebagaimana terlihat pada penggunaan metode ekstraksi fitur konvensional yang belum mampu menangkap kedalaman semantik antar dokumen.

Beberapa penelitian sebelumnya telah mencoba mengatasi permasalahan ini dengan menerapkan berbagai metode analisis data. Salah satunya adalah penggunaan algoritma *Support Vector Machine* (SVM) untuk klasifikasi berita yang mampu mencapai akurasi hingga 99,90% [4].

Namun demikian, metode klasifikasi tersebut bersifat *supervised* sehingga memerlukan dataset berlabel dalam jumlah besar, yang sering kali sulit diperoleh pada data berita yang bersifat dinamis [5].

Di sisi lain, algoritma *K-Means* telah banyak digunakan untuk proses klasterisasi data, namun penerapannya sering kali belum optimal karena tidak didukung oleh tahapan pengolahan data yang komprehensif [6].

Oleh karena itu, masih terdapat research gap di mana integrasi tahapan *Knowledge Discovery in Databases* (KDD) secara utuh belum banyak diimplementasikan pada portal berita online dalam meningkatkan kualitas pengelompokan berita secara otomatis.

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan penerapan algoritma *K-Means* dengan pendekatan *Knowledge Discovery in Databases* (KDD) untuk melakukan klasterisasi portal berita secara otomatis. Pendekatan *Knowledge Discovery in Databases* (KDD) meliputi beberapa tahapan utama, yaitu *data selection*, *preprocessing*, *data transformation*, *data mining*, hingga *evaluation* diharapkan dapat menangani kompleksitas data berita secara lebih terstruktur [1].

Melalui pendekatan tersebut, tujuan dari penelitian ini adalah menghasilkan sistem pengelompokan berita yang efisien dan akurat.

Kontribusi yang diharapkan adalah mempermudah organisasi informasi berita berskala besar dan meningkatkan efisiensi waktu bagi pengguna dalam memperoleh informasi yang mereka butuhkan.

2. TINJAUAN PUSTAKA

Seiring meningkatnya volume informasi digital di portal berita online, topik tentang pengelompokan dokumen berita menjadi penting. Oleh karena itu, penelitian ini membahas tentang algoritma *K-Means* menggunakan teknik *Knowledge Discovery in Databases* (KDD) sebagai metode klasterisasi karena mampu mengelompokkan dokumen secara otomatis melalui tahapan pengolahan data yang sistematis tanpa memerlukan data berlabel. Dalam penelitian ini pemilihan metode tersebut sejalan dengan penelitian sebelumnya yang menyatakan bahwa data berita pada portal online menyediakan informasi dalam format teks dengan jumlah yang besar, beragam topik, dan pembaruan real-time yang terus-menerus, sehingga sulit dikelompokkan menggunakan pendekatan konvensional [1].

Selain itu, penelitian terdahulu menunjukkan bahwa banyaknya berita tentang topik serupa dari berbagai media menyebabkan redudansi informasi, sehingga pengguna kesulitan menemukan topik [3].

Hasilnya menunjukkan bahwa metode pengelompokan harus digunakan agar dokumen berita dapat diorganisasi secara sistematis tanpa bergantung pada data berlabel. Oleh karena itu, pendekatan klasterisasi yang menggunakan algoritma *K-Means* dikombinasikan dengan pendekatan *Knowledge Discovery in Databases* (KDD) dipilih karena efektivitasnya dalam mengorganisasi data teks tidak berlabel pada portal berita online dengan volume besar. Meskipun *K-Means* efektif dalam klasterisasi berita, beberapa penelitian menunjukkan bahwa penerapan algoritma ini tanpa tahapan pengolahan data yang terstruktur dapat menghasilkan klaster yang kurang representatif. Dalam sistem agregasi berita, algoritma *K-Means* mengorganisasi artikel berdasarkan kemiripan topik dan mengurangi redudansi informasi. Algoritma *K-Means* ini dapat membentuk klaster topik yang konsisten dalam berita. Tetapi temuan penelitian menunjukkan bahwa kualitas klaster yang dibuat sangat bergantung pada tahapan pengolahan data, terutama proses preprocessing dan representasi fitur sebelum [1].

Hal ini diperkuat oleh penelitian yang menunjukkan bahwa menggunakan tahapan pengolahan data dan ekstraksi fitur yang dilakukan secara sistematis dan terorganisir dapat meningkatkan kualitas klasterisasi *K-Means* [7].

Berdasarkan temuan tersebut, dapat disimpulkan bahwa integrasi algoritma K-Means dengan kerangka *Knowledge Discovery in Databases* (KDD) merupakan pilihan yang tepat. Hal ini karena *Knowledge Discovery in Databases* (KDD) merupakan proses yang kompleks dalam menemukan pola-pola penting pada data sehingga lebih mudah dipahami. Proses ini mencakup beberapa tahapan yang saling berkaitan, yaitu *data selection*, *data preprocessing*, *data transformation, data mining*, dan *evaluation*, yang secara bersama-sama mendukung pelaksanaan proses klasterisasi secara sistematis.

Dengan mengacu pada berbagai penelitian yang ada pendekatan klasterisasi menggunakan algoritma *K-Means* efektif digunakan dalam pengelompokan dokumen berita ketika didukung oleh tahapan pengolahan data yang terstruktur. Karena Algoritma *K-Means* merupakan metode klasterisasi berbasis unsupervised learning yang mengelompokkan data berdasarkan kedekatan terhadap pusat klaster (*centroid*) melalui proses iteratif hingga diperoleh klaster yang stabil.

Dalam konteks klasterisasi portal berita online, penelitian ini diawali dengan data berita dikumpulkan dari berbagai portal berita digital dan digunakan sebagai data mentah untuk dianalisis. Setelah itu, data diproses melalui tahap *preprocessing*, yang mencakup pembersihan dan normalisasi teks untuk mengurangi kebisingan dan mempersiapkan data untuk proses pengolahan [1].

Setelah proses pembersihan data selesai, dokumen berita kemudian direpresentasikan dalam bentuk numerik menggunakan metode TF-IDF. Langkah ini bertujuan untuk mengukur tingkat kepentingan setiap kata yang muncul dalam dokumen [7].

Selanjutnya, data yang telah direpresentasikan tersebut diproses menggunakan algoritma *K-Means* untuk mengelompokkan berita berdasarkan kemiripan topik. Untuk menilai tingkat kohesi dan pemisahan antar klaster, digunakan indeks *Calinski–Harabasz sebagai* metode evaluasi kualitas klaster [3].

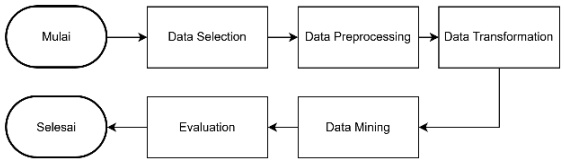
Pemanfaatan matrik evaluasi ini membantu dalam menentukan konfigurasi klaster yang paling sesuai untuk karakteristik data pada portal berita online.

3. METODE PENELITIAN

Pada Metodologi penelitian merupakan tahapan yang menjelaskan mengenai gambaran umum dari metode serta tahapan proses klasterisasi portal berita yang akan digunakan dalam penelitian [8].

Proses penelitian dilakukan dengan memanfaatkan algoritma *K-Means* sebagai metode klasterisasi dengan pendekatan *Knowledge Discovery in Database* (KDD), yaitu *data selection*, *data preprosesing*, *data transformation, data mining*, dan evaluasi hasil [9].

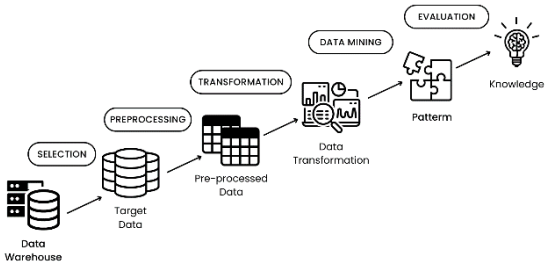
Setiap tahap dalam metodologi dirancang secara sistematis agar mampu menghasilkan klaster berita yang akurat dan relevan. Alur proses penelitian dapat terlihat pada gambar 1.



Gambar 1. Diagram alur proses klasifikasi portal berita dengan pendekatan KDD

3.1. Tahapan *Knowledge Discovery in Database* (KDD)

Tahapan metode pengolahan data yang digunakan pada penelitian ini adalah metode *Knowledge Discovery in Database* (KDD). *Knowledge Discovery in Database* (KDD) adalah proses sistematis yang dilakukan untuk mendapatkan pengetahuan baru dari dataset yang besar dan kompleks.



Gambar 2. Proses *knowledge discovery in database* (KDD)

Pada gambar 2 menggambarkan tentang tahapan-tahapan pada *Knowledge Discovery in Database* (KDD) yang mana terdiri dari tahapan *data selection, data preprocessing, data transformation, data mining dan evaluation.*

3.2. *Data Selection*

Tahap pertama dalam penelitian ini adalah proses pemilihan data yang bertujuan untuk memperoleh informasi yang relevan dan sesuai dengan kebutuhan analisis. [10].

Data yang digunakan dalam penelitian ini berasal dari sumber kumpulan dataset yang ada di situs kaggle dengan link https://www.kaggle.com/datasets/sh1zuka/indonesia-news-dataset-2024. Data diambil dari periode April 2025 hingga September 2025. Dataset yang diperoleh dengan cara mengcopy table data dan menyimpannya menjadi tabel dataset dengan jumlah 3036.

3.3. *Data Preprocessing*

Tahap ini bertujuan untuk melakukan proses pra-pemrosesan pada data berita agar dapat mengurangi kata-kata yang tidak diperlukan atau tidak relevan, sehingga mempermudah proses pengolahan pada tahap selanjutnya. Data juga diperiksa konsistensinya dan disesuaikan agar layak untuk pemrosesan lebih lanjut. Tahapan *data preprocessing* yang dilakukan yaitu, *data cleaning*, *case folding*, *normalization*, *tokenizing, filtering, stemming*.

a. *Data cleaning*
Proses pembersihan data merupakan tahap yang dilakukan untuk menghilangkan elemen-elemen yang tidak valid, duplikat, tidak konsisten, maupun data yang tidak lengkap dalam suatu dataset. Tujuan utama dari tahap ini adalah mengubah data mentah yang masih mengandung informasi yang tidak relevan menjadi data yang lebih bersih, sehingga dapat mendukung proses analisis dengan hasil yang lebih akurat dan optimal [11].

b. *Case Folding*
Case folding adalah proses konversi seluruh huruf dalam teks menjadi huruf kecil (*lowercase*) agar variasi penggunaan huruf besar dan kecil tidak dianggap berbeda saat pengolahan teks, sehingga setiap kata dapat dikenali secara konsisten dalam tahap selanjutnya. Hal ini merupakan bagian dari *data preprocessing* agar data teks lebih seragam sebelum dilakukan analisis lanjutan dalam *Natural Language Processing* [12].

c. *Normalization*
Normalization adalah menyetarakan rentang nilai dari hasil pembobotan kata, seluruh data numerik disesuaikan kembali ke dalam skala standar yang seragam, umumnya antara nilai 0 hingga 1. Tahap ini sangat krusial karena setiap artikel berita memiliki panjang teks yang berbeda-beda, sehingga dapat memicu ketimpangan rentang angka. Dengan melakukan penyeragaman skala ini, perhitungan tingkat kemiripan antar berita oleh algoritma *K-Means* menjadi lebih adil dan terhindar dari bias terhadap dokumen yang berukuran lebih panjang serta data berada dalam proporsi yang seimbang dan optimal untuk dikelompokkan.

d. *Tokenizing*
Pada tahap pemrosesan ini, teks akan diuraikan menjadi bagian-bagian yang lebih kecil. Hasil dari proses *tokenizing* berupa kata-kata yang terdapat dalam sebuah kalimat yang disebut sebagai *token*. *Tokenizing* merupakan proses membagi teks berita menjadi token yang dapat berupa kata, frasa, simbol, angka, maupun bagian penting lainnya yang terdapat di dalam teks. [13].

e. *Stopword Removal*
Stopword merupakan kata-kata yang bersifat umum, tidak mengandung informasi khusus, serta tidak memiliki unsur sentimen, namun sering muncul berulang dalam sebuah teks. Proses *stopword removal* dilakukan untuk mengurangi dimensi data, menurunkan beban komputasi, serta meningkatkan efisiensi dan kinerja analisis. Oleh karena itu, tahap stopword menjadi langkah penting untuk menghapus kata-kata yang tidak diperlukan berdasarkan daftar stopwords yang telah ditentukan dalam database. [12].

f. *Stemming*
Stemming merupakan proses yang digunakan untuk mengubah kata menjadi bentuk dasarnya dengan cara menghapus berbagai imbuhan atau

akhiran, sehingga kata-kata yang memiliki akar kata yang sama dapat direpresentasikan secara lebih konsisten. Tahapan ini membantu menyiapkan data teks agar lebih bersih dan terstruktur, sehingga dapat digunakan dalam model analisis sentimen maupun aplikasi NLP lainnya dengan tingkat akurasi yang lebih baik [14].

3.4. *Data Transformation*

Data Transformation merupakan tahap yang dilakukan untuk mengubah data ke dalam format yang lebih sesuai sehingga dapat diproses pada tahap pengolahan selanjutnya. [15].

Pada tahap ini, data melalui proses transformasi untuk menyesuaikan format atribut agar sesuai dengan kebutuhan algoritma yang akan digunakan. Pada penelitian ini proses klasterisasi dilakukan dengan memanfaatkan representasi fitur TF-IDF. Metode TF-IDF digunakan untuk mengubah setiap dokumen teks menjadi vektor numerik berdasarkan tingkat kepentingan kata dalam dokumen dan keseluruhan korpus. Vektor TF-IDF tersebut kemudian dijadikan input pada algoritma *K-Means* untuk membentuk kelompok dokumen yang memiliki kemiripan konten [12].

3.5. *Data Mining*

Data mining adalah tahap untuk mengekstraksi informasi yang bernilai dari kumpulan data berukuran besar, baik yang bersifat terstruktur maupun tidak terstruktur, dengan memanfaatkan berbagai metode atau algoritma tertentu [16].

Pada penelitian ini, proses *data mining* difokuskan pada penerapan algoritma *K-Means* sebagai metode klasterisasi. *K-Means Clustering* merupakan algoritma pembelajaran tanpa pengawasan (*unsupervised learning*) yang efektif dalam mengelompokkan data berdasarkan tingkat kemiripan karakteristiknya, sehingga cocok digunakan untuk menganalisis data teks yang tidak memiliki label [17].

Algoritma ini bekerja dengan cara menentukan Pada penelitian ini, proses *data mining* difokuskan pada penerapan algoritma *K-Means* sebagai metode klasterisasi. *K-Means Clustering* merupakan algoritma pembelajaran tanpa pengawasan (*unsupervised*) yang efisien dalam mengelompokkan data berdasarkan kesamaan karakteristik, cocok untuk menganalisis data teks yang tidak berlabel [18].

Dalam penentuan jumlah k cluster yang akan dibentuk terdapat banyak metode. Dalam penelitian ini akan menggunakan metode *Elbow*. Metode *Elbow* adalah salah satu metode yang sering digunakan untuk menentukan jumlah cluster yang optimal dengan melihat persentase varians dalam setiap klaster, di mana titik optimalnya ditandai oleh bentuk grafik yang menyerupai siku [19].

3.6. *Evaluation*

Evaluasi model *K-Means* dilakukan untuk menentukan jumlah klaster yang paling optimal dengan menggunakan metrik evaluasi internal, yaitu *Silhouette Score. Silhouette Score* merupakan salah satu metode evaluasi yang digunakan untuk mengukur seberapa baik suatu data berada dalam klasternya dibandingkan dengan klaster lainnya. Nilai skor ini berada pada rentang -1 hingga 1. Nilai yang mendekati 1 menunjukkan bahwa data telah dikelompokkan dengan baik pada klaster yang tepat, nilai yang mendekati 0 menandakan bahwa data berada di batas antara dua klaster, sedangkan nilai negatif menunjukkan bahwa proses pengelompokan kurang tepat [20].

4. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan dataset yang berjumlah 3036 data portal berita yang diambil dari website dataset Kaggle.

4.1. *Data Collection*

Tahap pertama adalah proses pemilihan data yang digunakan dalam proses pengelompokan data yang bertujuan untuk memilih atau menyeleksi data yang diperoleh [21].

Data atau atribut diambil dari dataset indonesia news di situs kaggle yang diambil dari periode April 2025 hingga September 2025 yang berjumlah 3036 data. Data yang diambil berisi atribut Judul dan Source terlihat pada tabel 1, dibawah ini.

Tabel 1. *Data collection*

No	Judul	Source
1	Viral Isu PHK Buruh Gudang Garam, Said Iqbal: Suplier hingga Pemilik Kontrakan Juga Akan Terdampak	Kompas
2	Remaja Ditindak Polisi di Serpong karena Bawa Mobil Berpelat Asing	Kompas

4.2. *Data Preprocessing*

Tahap Data yang telah dipilih sebelumnya dilakukan proses pengecekan terhadap keberadaan data yang hilang *(missing value).*

a. *Data Cleaning*

Dalam proses ini, sistem akan mencari dan menghapus informasi yang tidak utuh, data yang muncul dua kali (duplikat), serta karakter yang tidak perlu atau simbol aneh yang ikut terbawa saat pengambilan berita dari internet. Dengan membersihkan data dari gangguan tersebut, hasil pengelompokan berita nantinya akan menjadi jauh lebih akurat karena sistem hanya fokus pada isi berita yang asli. Tabel 2 menunjukkan perbedaan antara data yang masih berantakan dengan data yang sudah rapi setelah melalui proses pembersihan. Setelah melalui tahap data cleaning, jumlah data yang semula berjumlah 3036 berkurang menjadi sebanyak 3.023 data berita yang siap digunakan pada tahap pengolahan selanjutnya. Data yang telah dibersihkan ini kemudian digunakan sebagai dataset utama dalam proses analisis dan pengelompokan berita.

Tabel 2. Hasil *data cleaning*

Sebelum *Cleaning*	**Setelah *Cleaning***
Remaja Ditindak Polisi di Serpong karena Bawa Mobil Berpelat Asing	Remaja Ditindak Polisi di Serpong karena Bawa Mobil Berpelat Asing
Viral Isu PHK Buruh Gudang Garam, Said Iqbal: Suplier hingga Pemilik Kontrakan Juga Akan Terdampak	Viral Isu PHK Buruh Gudang Garam Said Iqbal Suplier hingga Pemilik Kontrakan Juga Akan Terdampak

b. *Case Folding*

Untuk menjaga konsistensi dan keseragaman dalam teks, seluruh huruf pada dokumen diubah menjadi huruf kecil (*lowercase*). Pada tahap ini, elemen yang tidak relevan seperti tanda baca, tautan, tagar, mention, dan spasi berlebih juga dihapus untuk mempersiapkan data secara optimal. Tabel 3 menampilkan hasil proses *case folding*, dimana teks menjadi lebih seragam dan bersih dari karakter yang tidak diperlukan.

Tabel 3. Hasil *case folding*

Sebelum C*ase Folding*	**Setelah C*ase Folding***
Remaja Ditindak Polisi di Serpong karena Bawa Mobil Berpelat Asing	remaja ditindak polisi di serpong karena bawa mobil berpelat asing
Viral Isu PHK Buruh Gudang Garam Said Iqbal Suplier hingga Pemilik Kontrakan Juga Akan Terdampak	viral isu phk buruh gudang garam said iqbal suplier hingga pemilik kontrakan juga akan terdampak

c. *Normalization*

Kata tidak baku diubah menjadi baku pada tahap ini. Tabel 4 menampilkan hasil proses *normalization*, dimana teks menjadi lebih baku.

Tabel 4. Hasil *normalization*

Sebelum *Normalization*	**Setelah *Normalization***
remaja ditindak polisi di serpong karena bawa mobil berpelat asing	remaja ditindak polisi di serpong karena bawa mobil berpelat asing
viral isu phk buruh gudang garam said iqbal suplier hingga pemilik kontrakan juga akan terdampak	viral isu phk buruh gudang garam said iqbal suplier hingga pemilik kontrakan juga akan terdampak

d. *Tokenizing*

Kalimat-kalimat panjang dalam berita dipotong-potong menjadi satuan kata tunggal. Selain memisahkan kata berdasarkan spasi, tahap ini juga membersihkan tanda baca atau angka yang tidak diperlukan agar struktur data menjadi lebih rapi. Gambaran mengenai teks yang sudah terbagi menjadi potongan kata ini disajikan pada tabel 5.

Tabel 5. Hasil *tokenizing*

Sebelum *Tokenizing*	**Setelah *Tokenizing***
remaja ditindak polisi di serpong karena bawa mobil berpelat asing	['remaja', 'ditindak', 'polisi', 'di', 'serpong', 'karena', 'bawa', 'mobil', 'berpelat', 'asing']
viral isu phk buruh gudang garam said iqbal suplier hingga pemilik kontrakan juga akan terdampak	['viral', 'isu', 'phk', 'buruh', 'gudang', 'garam', 'said', 'iqbal', 'suplier', 'hingga', 'pemilik', 'kontrakan', 'juga', 'akan', 'terdampak']

e. *Stopword Removal*

Tahapan ini bertujuan untuk menyaring dan membuang kata-kata yang terlalu umum sehingga tidak memiliki makna penting dalam menentukan topik berita. Kata-kata seperti "di", "ke", "dari", dan "dan" dihapus agar sistem hanya fokus mengolah kata kunci yang benar-benar mewakili isi berita. Dengan menghilangkan kata-kata tersebut, proses pengelompokan menjadi lebih ringan dan cepat. Perubahan data setelah penyaringan ini dapat dilihat pada tabel 6, di mana teks menjadi lebih padat dan hanya menyisakan kata-kata inti.

Tabel 6. Hasil *stopword removal*

Sebelum *Stopword Removal*	**Setelah *Stopword Removal***
['remaja', 'ditindak', 'polisi', 'di', 'serpong', 'karena', 'bawa', 'mobil', 'berpelat', 'asing']	['remaja', 'ditindak', 'polisi', 'serpong', 'bawa', 'mobil', 'berpelat', 'asing']
['viral', 'isu', 'phk', 'buruh', 'gudang', 'garam', 'said', 'iqbal', 'suplier', 'hingga', 'pemilik', 'kontrakan', 'juga', 'akan', 'terdampak']	['viral', 'isu', 'phk', 'buruh', 'gudang', 'garam', 'said', 'iqbal', 'suplier', 'pemilik', 'kontrakan', 'terdampak']

f. *Stemming*

Untuk menyatukan berbagai variasi kata, semua kata yang memiliki imbuhan diubah kembali menjadi kata dasarnya. Proses ini sangat penting dalam bahasa Indonesia agar kata-kata yang asalnya sama (seperti "mengerjakan" dan "dikerjakan") tidak dianggap sebagai dua hal yang berbeda oleh sistem. Pada tahap akhir ini, kata-kata berimbuhan diubah menjadi bentuk dasar untuk menyeragamkan variasi kata dalam proses analisis teks.

Tabel 7. Hasil *stemming*

Sebelum *Stemming*	**Setelah *Stemming***
['remaja', 'ditindak', 'polisi', 'serpong', 'bawa', 'mobil', 'berpelat', 'asing']	remaja tindak polisi serpong bawa mobil pelat asing
['viral', 'isu', 'phk', 'buruh', 'gudang', 'garam', 'said', 'iqbal', 'suplier', 'pemilik', 'kontrakan', 'terdampak']	viral isu phk buruh gudang garam said iqbal suplier milik kontra dampak

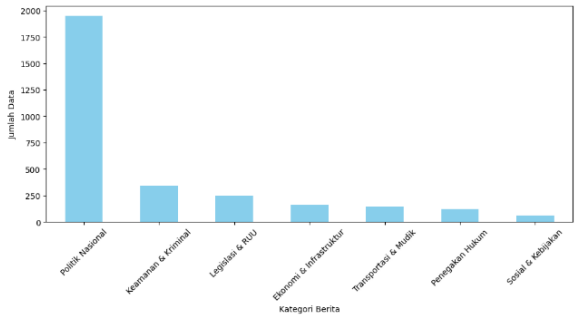
4.3. *Data Transformation*

Tahap transformasi data berfokus pada konversi representasi teks yang telah melalui proses pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF) menjadi format matriks numerik. Transformasi ini mutlak diperlukan karena algoritma K-Means tidak dapat memproses data dalam bentuk alfabetis, melainkan beroperasi secara matematis dengan mengkalkulasi titik koordinat data dalam ruang vektor multidimensi. Peneliti memberikan label pada seluruh data dari tahap preprocessing untuk menentukan topik yang ada menjadi 7 kelas. Setiap kelas yang diberi label akan diubah menjadi angka agar dapat diproses oleh model machine learning. Tabel di bawah menunjukkan output label dari kelas-kelas yang ada saat ini.

Tabel 8. Label

Label	Label Encoding
Ekonomi dan Infrastruktur	6
Penegakan Hukum	5
Legislasi dan RUU	4
Keamanan dan Kriminal	3
Politik Nasional	2
Tansportasi dan Mudik	1
Sosial dan Kebijakan	0

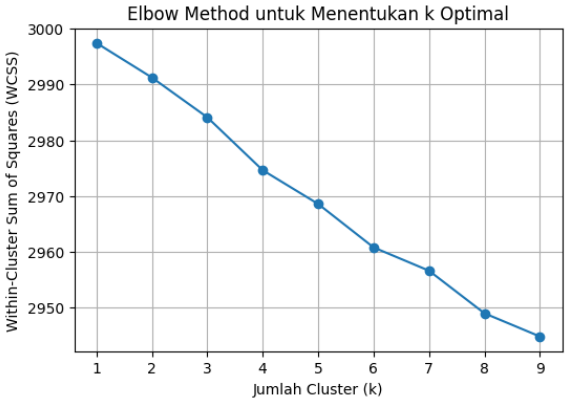
Proses *label* menghasilkan jumlah topik politik nasional sebanyak 1946 data, topik keamanan dan kriminal sebanyak 340 data, topik legislasi dan RUU sebanyak 249 data, topik ekonomi dan infrastruktur sebanyak 158 data, topik transportasi dan mudik sebanyak 144 data, topik penegakan hukum sebanyak 123 data, serta topik sosial dan kebijakan sebanyak 63 data. Grafik visualisasi distribusi topik berita dari hasil proses *label* ditunjukkan oleh gambar 3.



Gambar 3. Grafik Distribusi Topik

4.4. *Data Mining*

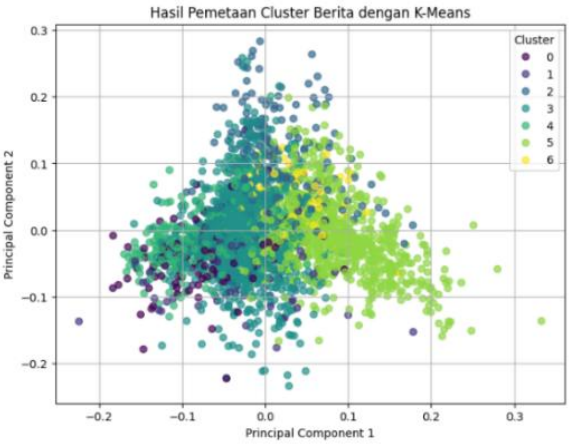
Data berita yang telah melalui proses pembersihan selanjutnya dimasukkan ke dalam tahap penemuan pola (*data mining*). Proses pengelompokan data ini dilakukan secara menyeluruh menggunakan algoritma *K-Means* tanpa memerlukan pembagian data latih (*training*) dan data uji (*testing*), karena algoritma ini mampu mengenali pola kesamaan teks secara mandiri (*unsupervised learning*). Sebelum menjalankan proses clustering akhir, dilakukan analisis terlebih dahulu untuk menentukan jumlah kelompok optimal menggunakan *Elbow Method*.



Gambar 4. Hasil *elbow method*

Berdasarkan grafik *Elbow* pada gambar 4, terlihat hubungan antara jumlah klaster pada sumbu X dan nilai *Inertia* (Within-Cluster Sum of Squares) pada sumbu Y. Grafik menunjukkan penurunan nilai *Inertia* yang sangat drastis dari k = 1 hingga k = 6, namun setelah itu penurunan mulai melandai secara konsisten. Melalui deteksi otomatis menggunakan algoritma *KneeLocator*, titik siku (*elbow point*) ditemukan berada pada k = 7. Oleh karena itu, jumlah klaster optimal yang dipilih untuk mengelompokkan berita ini berada pada k = 7.

Melanjutkan tahap penentuan jumlah cluster yang optimal, dilakukan proses clustering pada data portal berita berdasarkan kemiripan isi judul berita. Proses pengelompokan ini menggunakan algoritma *K-Means Clustering* dengan jumlah klaster (k) sebanyak 7. Tujuan dari proses ini adalah untuk mengelompokkan berita berdasarkan kesamaan topik atau kata-kata yang sering muncul pada judul berita. Hasil dari proses *clustering* tersebut kemudian disimpan dalam atribut *cluster* pada dataset. Visualisasi hasil *clustering* pada data portal berita dapat dilihat pada gambar 5.



Gambar 5. Hasil pemetaan dengan *k-means*

Hasil pemetaan clustering pada portal data berita pada gambar 5 menggunakan algoritma *K-Means* yang menghasilkan 7 cluster, yaitu:

a. Cluster 0 (Ungu Tua): Mencakup berita yang berkaitan dengan ekonomi dan kebijakan pemerintah, seperti pembahasan mengenai kebijakan tarif, impor, harga bahan bakar, serta berbagai keputusan ekonomi yang dipengaruhi oleh kebijakan pemerintah pusat.
b. Cluster 1 (Ungu Muda): Berisi berita mengenai politik dan hubungan internasional, yang mencakup aktivitas tokoh politik, pernyataan pejabat negara, serta hubungan diplomatik Indonesia dengan negara lain atau organisasi internasional.
c. Cluster 2 (Biru): Mengelompokkan berita peristiwa sosial atau kriminal, seperti kejadian yang terjadi di masyarakat, konflik sosial, kecelakaan, maupun berbagai peristiwa yang melibatkan warga dalam kehidupan sehari-hari.
d. Cluster 3 (Biru Kehijauan): Mencakup berita terkait keamanan dan penegakan hukum, seperti tindakan aparat kepolisian, proses penanganan kasus hukum, serta upaya menjaga keamanan dan ketertiban masyarakat.
e. Cluster 4 (Hijau Tua): Berisi berita yang berkaitan dengan transportasi dan infrastruktur, termasuk isu kemacetan, kondisi jalan, pembangunan fasilitas transportasi, serta pengelolaan sarana prasarana publik.
f. Cluster 5 (Hijau Muda): Mengelompokkan berita mengenai politik dan pemerintahan, seperti aktivitas lembaga legislatif, pembahasan kebijakan di parlemen, serta dinamika politik dalam pemerintahan.
g. Cluster 6 (Kuning): Berisi berita yang membahas isu sosial dan demonstrasi, termasuk aksi mahasiswa, gerakan masyarakat sipil, serta berbagai bentuk protes atau aspirasi publik terhadap kebijakan pemerintah.

4.5. *Data Evaluation*

Gambar 6. Visualisasi *silhouette score*

Pada tahap ini, proses *clustering* dilakukan menggunakan algoritma *K-Means* dengan beberapa variasi jumlah klaster. Untuk menilai kualitas hasil pengelompokan, digunakan metode *Silhouette Score* yang memberikan gambaran mengenai seberapa baik data ditempatkan pada klaster yang sesuai.

Berdasarkan visualisasi *Silhouette Score* pada gambar 6, dapat diketahui bahwa proses pengelompokan dengan 7 klaster menghasilkan nilai rata-rata *silhouette coefficient* yang berada pada nilai positif, yang ditunjukkan oleh garis vertikal putus-putus berwarna merah. Meskipun nilai koefisiennya relatif kecil (mendekati 0.01), hasil ini menunjukkan bahwa sebagian besar data dalam setiap klaster memiliki nilai positif, yang mengindikasikan bahwa data tersebut telah berada pada kelompok yang cenderung tepat. Setiap bilah warna mewakili satu klaster topik berita, di mana lebar bilah menunjukkan jumlah anggota dalam klaster tersebut. Secara keseluruhan, visualisasi ini membuktikan bahwa metode *K-Means* berhasil memisahkan topik-topik berita seperti politik, kriminalitas, dan ekonomi dengan batas yang cukup jelas, sehingga homogenitas anggota di dalam setiap klaster terjaga dengan baik. Dapat diamati bahwa terdapat variasi ukuran klaster, di mana beberapa klaster memiliki anggota yang sangat banyak (bilah merah dan cokelat), sementara klaster lainnya lebih kecil.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian dan pembahasan mengenai klasterisasi portal berita online menggunakan algoritma *K-Means* dan kerangka kerja *Knowledge Discovery in Databases* (KDD), dapat disimpulkan bahwa tahapan pra-pemrosesan data terbukti sangat penting dalam mengoptimalkan data mentah. Melalui serangkaian proses terstruktur yang meliputi *data cleaning*, *case folding*, *normalization*, *tokenizing*, *stopword removal*, dan *stemming* , dataset awal yang berjumlah 3.036 berita berhasil diekstraksi menjadi 3.023 data teks bersih yang siap diolah. Kemudian dilanjutkan pada *data transformation*, setelah data dipersiapkan, pengujian penentuan klaster menggunakan metode *Elbow* menunjukkan penurunan nilai inertia yang signifikan dan membentuk titik siku, sehingga diputuskan bahwa jumlah klaster optimal (k) untuk dataset berita ini adalah sebanyak 7 klaster. Penerapan algoritma *K-Means* dengan k=7 tersebut telah berhasil memetakan dokumen berita ke dalam kategori topik yang spesifik dan terstruktur. Pengelompokan tersebut menghasilkan Klaster 0 untuk berita ekonomi dan kebijakan pemerintah, klaster 1 untuk politik dan hubungan internasional, klaster 2 untuk peristiwa sosial atau kriminal, klaster 3 untuk keamanan dan penegakan hukum, klaster 4 untuk transportasi dan infrastruktur, klaster 5 untuk politik dan pemerintahan, serta klaster 6 untuk isu sosial dan demonstrasi. Keberhasilan pemetaan ini divalidasi oleh hasil evaluasi menggunakan *Silhouette Score* yang menunjukkan nilai rata-rata koefisien positif. Nilai positif ini membuktikan bahwa sebagian besar data telah berada pada kelompok yang tepat, di

mana metode *K-Means* mampu menjaga homogenitas anggota di dalam setiap klaster serta memisahkan topik-topik berita yang kompleks dengan batas pemisah yang cukup jelas.

Meskipun penelitian ini telah berhasil melakukan klasterisasi dengan baik, terdapat beberapa peluang pengembangan untuk penelitian selanjutnya. Dari sisi pengolahan teks, penggunaan teknik ekstraksi fitur lanjutan atau *word embedding* dapat dipertimbangkan untuk menangkap kedalaman semantik antar dokumen secara lebih optimal. Selain itu, hasil pengelompokan 7 topik spesifik ini dapat dimanfaatkan sebagai pelabelan awal untuk membangun sistem klasifikasi otomatis menggunakan algoritma *supervised learning* seperti *Support Vector Machine* (SVM) atau *Decision Tree*. Integrasi dengan analisis sentimen juga disarankan untuk mengevaluasi kecenderungan opini publik pada masing-masing klaster isu, sehingga penerapan *Artificial Intelligence* dalam pengelolaan informasi portal berita menjadi semakin komprehensif.

DAFTAR PUSTAKA

[1] H. T. A. Simanjuntak, P. E. P. Silaban, J. K. S. Manurung, and V. H. Sormin, "Klasterisasi Berita Bahasa Indonesia Dengan Menggunakan K-Means Dan Word Embedding," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 3, pp. 641–652, 2023.

[2] K. Amwari and T. Chmaidy, "Klasterisasi Berita Berbahasa Indonesia Menggunakan Model K-Means Dan Indobert Untuk Menentukan Berita Hoaks: Clustering Indonesian News Using K-Means and Indobert Models to Determine Hoax News," *Explor. IT J. Keilmuan dan Apl. Tek. Inform.*, vol. 17, no. 1, pp. 11–25, 2025.

[3] I. G. Zulfikar, Y. Wibisono, and A. Wahyudin, "Intelligent News Aggregation System with Automatic Classification, Clustering, and Summarization," *Brill. Res. Artif. Intell.*, vol. 5, no. 2, pp. 689–696, 2025.

[4] I. M. Tianda, M. N. Ubadah, M. F. F. Mardianto, A. Munawwarah, and E. Ana, "Clustering fake news with k-means and agglomerative clustering based on Word2Vec," *Int. J. Math. Comput. Res.*, vol. 12, no. 02, pp. 3999–4007, 2024.

[5] S. Perveen *et al.*, "Unsupervised fake news detection on social media using hybrid Gaussian Mixture Model," *PLoS One*, vol. 20, no. 8, p. e0330421, 2025.

[6] A. Ariesta and J. Annissa, "Klasterisasi Teks Berita Pemilu untuk Analisis Framing di Kompas. Com," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 12, no. 5, pp. 1185–1196, 2025.

[7] D. F. Surianto and D. F. Surianto, "Enhancing K-Means clustering for journal articles using TF-IDF and LDA feature extraction," *Brill. Res. Artif. Intell.*, vol. 4, no. 2, pp. 964–972, 2024.

[8] D. K. Wardy, I. K. G. D. Putra, and N. K. D. Rusjayanthi, "Clustering artikel pada portal berita online menggunakan metode k-means," *J. Ilm. Teknol. dan Komput.*, vol. 3, no. 1, pp. 985–993, 2022.

[9] A. Z. Musthafa, B. M. Basuki, and A. Habibi, "IMPLEMENTASI SISTEM KLASTERISASI MENGGUNAKAN ALGORITMA K-MEANS UNTUK SELEKSI PENERIMA BANTUAN SOSIAL," *Sci. ELECTRO*, vol. 19, no. 1, 2025.

[10] B. Z. Ramadhan, R. I. Adam, and I. Maulana, "Analisis sentimen ulasan pada aplikasi e-commerce dengan menggunakan algoritma naïve bayes," *J. Appl. Informatics Comput.*, vol. 6, no. 2, pp. 220–225, 2022.

[11] H. T. Ismet, T. Mustaqim, and D. Purwitasari, "Aspect based sentiment analysis of product review using memory network," *Sci. J. Informatics*, vol. 9, no. 1, pp. 73–83, 2022.

[12] K. Tsuroyya, K. Umam, W. D. Yuniarti, and M. R. Handayani, "Klasifikasi sentimen pada ulasan pengguna aplikasi Cryptocurrency di Google Play Store menggunakan algoritma Decision Tree," *AITI*, vol. 22, no. 2, pp. 279–293, 2025.

[13] M. I. Ghufron, E. Supriyati, and T. Listyorini, "Analisa Jejaring Sosial Terhadap Fenomena Cyberbullying Fandom K-Pop pada Sosial Media Twitter," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 9, no. 2, pp. 79–93, 2024.

[14] W. Nurwahyudi, M. Isnaini, S. J. Roszi, and A. W. Laily, "Analisis Sentimen Ulasan Produk Moisturizer Skintific Di Tokopedia Menggunakan Support Vector Machine," *J. Sist. Inf. dan Bisnis Cerdas*, vol. 18, no. 1, pp. 129–142, 2025.

[15] A. Supoyo and P. T. Prasetyaningrum, "Analisis Data Mining Untuk Memprediksi Lama Perawatan Pasien Covid-19 Di DIY," *Bianglala Inf.*, vol. 10, no. 1, pp. 21–29, 2022.

[16] E. R. Fitria and F. Rozci, "Penerapan Metode Regresi Least Absolute Shrinkage and Selection Operator (LASSO) dan Regresi Linier untuk Memprediksi Tingkat Kemiskinan di Indonesia," *J. Ilm. Sosio Agribis*, vol. 22, no. 2, pp. 123–132, 2022.

[17] D. Alfiyanto, A. N. Fajri, E. Liando, F. D. Prasetio, R. Fahlapi, and F. Amsury, "Analisis Sentimen Pemberitaan kompas. com Terhadap Program Merdeka Belajar Dengan Metode Clustering K-Means," *JEKIN-Jurnal Tek. Inform.*, vol. 5, no. 3, pp. 1182–1189, 2025.

[18] F. Nuraeni and M. H. M. Firdaus, "Model Klasterisasi Data Pertanian Menggunakan Algoritma K-Means," *J. Algoritm.*, vol. 22, no. 2, pp. 890–898, 2025.

[19] H. Irawan, M. I. Fawzi, and S. Hidayat, "Integration of K-Means Clustering and Elbow Method for Mapping Baccaurea spp. Distribution to Support Agroindustrial Development in West Sulawesi," *J. Tek. Inform.*, vol. 6, no. 2, pp. 1057–1068, 2025.

[20] F. Nuraeni, A. Syukur, A. Marjuni, N. Rijati, and

D. Kurniadi, "Comparison of centroid-based clustering model performance on categorical dataset," in *2024 Ninth International Conference on Informatics and Computing (ICIC)*, IEEE, 2024, pp. 1–6.

[21] P. Apriyani, A. R. Dikananda, and I. Ali, "Penerapan Algoritma K-Means dalam Klasterisasi Kasus Stunting Balita Desa Tegalwangi," *Hello World J. Ilmu Komput.*, vol. 2, no. 1, pp. 20–33, 2023.