BAB II

LANDASAN TEORI

2.1.1 Berita

Berita merupakan laporan mengenai peristiwa terkini yang mencakup topik umum. Biasanya, berita disajikan secara singkat dan jelas untuk memberikan informasi yang objektif dan tepat waktu kepada pembaca atau penonton. Sumber berita dapat berasal dari berbagai media, baik cetak maupun elektronik. Kecepatan penyebaran berita sangat penting karena informasi dalam berita sangat terkait dengan waktu peristiwa. Kualitas suatu berita, yang mencakup fakta dan ketertarikan, ditentukan oleh beberapa faktor. Salah satu faktor kualitasnya adalah kemampuan berita tersebut dalam mempengaruhi audiens [10]. Esensi dari berita adalah menyampaikan peristiwa atau kejadian yang memiliki nilai informasi yang dianggap relevan atau penting bagi audiens. Meskipun suatu peristiwa mengandung fakta, jika dianggap tidak relevan, kurang aktual, atau kurang menarik, kemungkinan besar tidak akan diangkat sebagai bahan berita.

Konsep pentingnya berita terletak pada nilai informasinya bagi masyarakat. Berita yang disampaikan bertujuan untuk memberikan tambahan informasi kepada audiens, membantu mereka untuk memahami apa yang terjadi di sekitar mereka, baik secara lokal maupun global. Ketika suatu peristiwa dianggap penting, itu berarti peristiwa tersebut memiliki dampak atau relevansi yang signifikan bagi kehidupan sehari-hari, keamanan, politik, sosial, atau ekonomi.

Berita berperan penting dalam meningkatkan pengetahuan dan pemahaman orang-orang terhadap dunia di sekitar mereka. Dengan menyediakan informasi yang terpercaya, terkini, dan relevan, berita membantu pembaca atau penonton untuk membuat keputusan yang lebih baik, memahami dampak suatu peristiwa, atau bahkan meningkatkan kesadaran

mereka terhadap isu-isu penting yang mungkin memengaruhi kehidupan mereka.

Dalam era informasi saat ini, akses terhadap berita yang akurat dan bermanfaat menjadi krusial bagi individu dalam memahami realitas kompleks yang ada di sekitar mereka. Oleh karena itu, berita memiliki nilai signifikan sebagai sumber informasi yang dapat membantu meningkatkan pengetahuan, pemahaman, dan kesadaran masyarakat akan isu-isu yang terjadi di dunia.

2.1.2 Berita Hoaks

Hoaks adalah usaha untuk memutarbalikkan fakta dengan menggunakan informasi yang terlihat seperti kebenaran, namun kebenarannya tidak dapat dipastikan. Dalam Kamus Besar Bahasa Indonesia (KBBI), hoaks diartikan sebagai berita bohong atau berita yang tidak memiliki sumber yang jelas. Keberadaan berita hoaks dapat merugikan baik individu maupun kelompok, namun juga bisa memberikan keuntungan pada pihak tertentu. Dampak dari berita hoaks dapat memecah belah pandangan masyarakat, membuat kesulitan dalam membedakan antara berita yang benar dan palsu, serta memicu emosi pada pendengarnya karena disampaikannya informasi yang tidak selalu valid. Berita hoaks seringkali menyerupai berita asli, tetapi dapat dibedakan dengan kecermatan dan kemampuan untuk menganalisis dengan cermat [11].

Membedakan antara berita hoaks dengan informasi yang faktual bisa menjadi sebuah tantangan, terutama bagi mereka yang kurang berpengalaman. Hoaks seringkali digambarkan sebagai informasi palsu yang menyesatkan, dimana para penipu dunia maya menggunakan taktik tertentu untuk membingungkan masyarakat dalam membedakan antara kebenaran dan kebohongan. Menurut David Harley selaku Direktur Intelijen Malware, menguraikan beberapa metode identifikasi pesan hoaks dalam bukunya, "*Common Hoaxes and Chain Letters*". Salah satunya adalah melihat adanya pesan yang tersebar secara berantai, seperti ancaman yang mendesak untuk menyebarluaskan pesan tersebut kepada semua orang. Pesan juga seringkali mencantumkan sumber yang tidak dikenal atau berisi tautan dan gambar yang

tidak relevan dengan informasi yang disajikan. Lebih lanjut, pesan hoaks cenderung menciptakan rasa cemas atau panik pada pembacanya, dan seringkali pengirimnya tidak mengungkapkan identitasnya [12].

2.1.3 *Data Mining*

Data mining adalah proses ekstraksi pola atau pengetahuan berharga dari kumpulan data besar menggunakan teknik matematika, statistik, kecerdasan buatan, dan metode lainnya. Tujuannya adalah mengidentifikasi pola tersembunyi, tren, dan informasi penting yang mendukung pengambilan keputusan yang lebih akurat, memprediksi tren masa depan, serta meningkatkan pemahaman atas data yang ada.

Proses *data mining* tidak dilakukan secara manual, melainkan secara semi-otomatis dengan memanfaatkan berbagai teknik seperti statistik, matematika, kecerdasan buatan, dan machine learning. Teknik-teknik ini digunakan untuk mengekstraksi dan mengidentifikasi informasi yang berharga dari kumpulan data yang besar dan kompleks.

Melalui *data mining*, dapat menemukan pola yang tidak terlihat secara langsung dalam data, mengidentifikasi keterkaitan antara variabel, dan memahami perilaku atau tren yang mungkin tersembunyi. Proses ini memungkinkan para profesional, peneliti, dan pengambil keputusan untuk mengoptimalkan penggunaan data dalam pengambilan keputusan yang lebih baik.

Dengan menggali informasi dan pengetahuan yang tersembunyi dalam data, data mining membantu meningkatkan efisiensi dalam pengambilan keputusan bisnis, memprediksi tren pasar, mengidentifikasi pola perilaku konsumen, dan banyak lagi. Dengan memanfaatkan teknik-teknik canggih, data mining menjadi instrumen yang kuat dalam memahami dan memanfaatkan potensi besar dari kumpulan data yang ada. *Data mining* sendiri merupakan bagian dari proses KDD *(Knowledge Discovery in Databases)* yang melibatkan serangkaian tahapan, mulai dari pemilihan data,

pra-pengolahan, transformasi, proses *data mining*, hingga evaluasi hasil yang diperoleh [13]. Karakteristik dari *data mining* adalah sebagai berikut:

1. Menemukan pola tersembunyi dan informasi yang sebelumnya tidak dikenal dalam data.
2. Menggunakan data dalam skala besar, yang cenderung meningkatkan keandalan hasil.
3. Penting dalam pengambilan keputusan kritis, terutama dalam strategi [14].

2.1.4 *Clustering*

Clustering merupakan salah satu teknik dalam kategori pembelajaran tak terawasi (*unsupervised learning*) yang digunakan untuk mengelompokkan objek-objek dalam data. Tujuannya adalah untuk membentuk kelompok atau klaster di mana objek-objek dalam satu kelompok memiliki kemiripan atau kesamaan satu sama lain, sedangkan berbeda secara signifikan dengan objek-objek dalam kelompok lainnya.

Proses *clustering* bertujuan untuk mengidentifikasi struktur internal dalam data sehingga elemen-elemen atau titik data dalam satu kelompok memiliki kesamaan atau kemiripan yang tinggi, sementara kelompok lain memiliki perbedaan yang lebih besar. Teknik ini memungkinkan kita untuk mengelompokkan data berdasarkan kesamaan karakteristik atau fitur tertentu, tanpa adanya label atau klasifikasi sebelumnya.

Pengelompokan yang dihasilkan dari proses *clustering* sering digunakan dalam berbagai bidang seperti analisis data, pengenalan pola, pemrosesan gambar, biologi, dan lainnya. Misalnya, dalam analisis data bisnis, clustering dapat membantu dalam segmentasi pasar atau identifikasi profil konsumen berdasarkan perilaku pembelian yang mirip. Dalam bidang medis, *clustering* dapat digunakan untuk mengelompokkan pasien berdasarkan karakteristik klinis yang serupa.

Dengan menggunakan teknik *clustering*, kita dapat memahami karakteristik kompleks dari kumpulan data besar dan heterogen. Hal ini memungkinkan kita untuk mengidentifikasi struktur yang tersembunyi, pola yang terkandung dalam data, serta memahami relasi atau kesamaan antar-objek dalam kelompok yang telah terbentuk. *Clustering* menjadi salah satu alat penting dalam analisis data yang membantu mengungkap informasi berharga dari data yang kompleks dan tidak terlabel sebelumnya. Dalam *clustering* ada dua pendekatan utama dalam pengembangan metodenya, yaitu pendekatan partisi dan pendekatan hirarki. Pendekatan partisi atau sering disebut *partition-based clustering*, merupakan proses pengelompokkan data dengan memisahkan data yang dianalisis ke dalam kelompok atau *cluster* yang berbeda [15].

2.1.5 Analisis Jaringan

Analisis jaringan atau disebut juga *network analysis* adalah investigasi yang mencakup hubungan di antara berbagai entitas dalam suatu sistem yang terstruktur. Fokus utamanya adalah pada pemahaman mengenai interaksi, koneksi, dan hubungan antara entitas-entitas tersebut. *Network analysis* merupakan alat untuk memetakan koneksi antar individu dengan pendekatan yang mengidentifikasi aliran informasi baik horizontal maupun vertikal. Hal ini memungkinkan identifikasi sumber dan tujuan penyebaran narasi serta memahami peran *node* yang memengaruhi akses terhadap sumber daya informasi yang ada. Dengan mengidentifikasi arus informasi, dapat membantu dalam perencanaan strategi berbagi informasi daripada mengembangkan strategi baru [16]. Penggunaan metode *network analysis* untuk menemukan node antara akun-akun twitter yang terindikasi mempunyai pengaruh dalam jaringan penyebaran informasi berita hoaks [17]. *Node* dan *edge* adalah elemen krusial dalam analisis jaringan, di mana jaringan itu sendiri adalah struktur yang terdiri dari *node* yang saling terhubung oleh *edge*. *Node* menggambarkan individu atau entitas dalam jaringan, sedangkan *edge* merepresentasikan koneksi atau hubungan antara *node*. Dalam analisis jaringan, ukuran atau kepentingan *node*, yang disebut sebagai sentralitas aktor,

memainkan peran penting dalam menunjukkan seberapa besar pengaruh atau kepentingan seorang aktor dalam jaringan [18].

2.1.6 *Principal Component Analysis* (PCA)

Principal Component Analysis (PCA) adalah metode statistik yang berguna untuk mereduksi dimensi dari dataset yang kompleks. Tujuan utamanya adalah mengidentifikasi pola dalam data yang rumit dengan mengubah variabel yang saling terkait menjadi sekelompok variabel yang tidak memiliki korelasi, dikenal sebagai komponen utama. PCA mampu mengubah dataset yang memiliki banyak variabel menjadi bentuk yang lebih sederhana, memungkinkan interpretasi yang lebih mudah terhadap struktur data dan mengurangi dimensi tanpa kehilangan informasi penting.

Menurut Suyanto (2018), PCA adalah metode matematis untuk mengubah data ke dalam dimensi baru dengan menghasilkan sejumlah komponen utama yang paling signifikan. Namun, dalam proses penambangan data berdimensi tinggi, terkadang terjadi penyederhanaan berlebihan yang hanya fokus pada penghapusan fitur tertentu. Hal ini dapat membuat model tidak mampu memahami kompleksitas permasalahan. Alih-alih menghapus fitur, pendekatan lain yang lebih baik untuk mengurangi kompleksitas komputasi adalah dengan mentransformasikan data ke dalam dimensi yang lebih kecil [19].

Cara kerja metode ini adalah dengan melakukan ekstraksi atribut untuk menyisihkan atribut-atribut tertentu sehingga hasil yang diperoleh menjadi lebih optimal. Metode ini terdiri dari empat tahapan yang meliputi:

1. Menemukan sejumlah data dengan dimensi $m \times n$, di mana m adalah jumlah sampel data dan n adalah jumlah atribut.

$$X^*_{i,j} = X_{i,j} - \bar{x} \quad (2.1)$$

Keterangan:

$X_{i,j}$ = elemen matrik X

$X^*_{i,j}$ = elemen matrik X*

$\bar{x}$ = nilai rata-rata matrik X

2. Mencari nilai kovarian (C_x) dari kumpulan data menggunakan persamaan:

$$C_x = \frac{1}{m-1} \cdot X_{i,j}^{*T} \quad (2.2)$$

3. Menghitung nilai eigen (λ) dengan menggunakan persamaan 3, di mana I adalah matriks identitas v adalah vektor eigen:

$$|C_x - \lambda I| = 0 \text{ dan } (C_x - \lambda I).\ v = 0 \quad (2.3)$$

4. Menghitung persentase kontribusi varian kumulatif (V_r), di mana d adalah jumlah awal atribut dan r adalah jumlah komponen yang dipilih.

$$V_r = \frac{\sum_j^r \lambda j}{\sum_j^d \lambda j} \cdot 100\% \text{ [20]} \quad (2.4)$$

Langkah awal dalam metode PCA melibatkan evaluasi variabel yang pantas dimasukkan ke dalam analisis selanjutnya. Proses ini mencakup inklusi semua variabel yang ada, diikuti dengan serangkaian pengujian. Ketika sebuah variabel menunjukkan kecenderungan untuk berkelompok dan membentuk faktor, itu menandakan korelasi yang signifikan dengan variabel lain. Sebaliknya, variabel yang memiliki korelasi rendah cenderung tidak membentuk faktor spesifik [21].

2.1.7 *Degree Centrality*

Degree centrality adalah suatu ukuran dalam analisis jaringan yang mengestimasi jumlah koneksi atau relasi yang dimiliki oleh sebuah *node* dalam suatu jaringan. Metrik ini merupakan salah satu dari metode sentralitas yang paling mendasar yang menandai seberapa banyak hubungan dari sebuah *node* terhadap *node* lainnya dalam jaringan. Penghitungan nilai *Degree Centrality* dapat dilakukan dengan rumus berikut:

$$CD(Ni)=d(Ni) \quad (2.5)$$

Keterangan:

CD: adalah singkatan dari *Degree Centrality*.

Ni: adalah jumlah *node* i dalam jaringan.

$d(Ni)$: merujuk pada banyakanya interaksi *node* Ni dengan *node* lainnya dalam jaringan [22].

2.1.8 *Betweenness Centrality*

Betweenness Centrality adalah ukuran dalam analisis jaringan yang mengevaluasi seberapa sering suatu node menjadi bagian dari jalur terpendek antara dua node lain di jaringan. Metrik ini menilai seberapa besar pengaruh sebuah node dalam mengontrol aliran informasi dalam jaringan, bertindak sebagai penghubung kunci atau jembatan antara bagian-bagian yang berbeda dalam jaringan. Ketika nilai *betweenness centrality* sebuah node semakin tinggi, semakin penting peran node tersebut dalam menghubungkan bagian-bagian yang berbeda dalam jaringan. Penghitungan nilai *Betweenness Centrality* dari sebuah node dapat menggunakan rumus berikut:

$$CB(ni) = \sum \frac{Gjk\ (ni)}{gjk} \qquad (2.6)$$

Dalam rumus tersebut, $\sum gjk(ni)$ menggambarkan total jalur terpendek dari node j ke node k yang melewati node i, sedangkan gjk melambangkan jumlah jalur terpendek antara dua node dalam jaringan [23].

2.1.9 *Closeness Centrality*

Closeness centrality adalah metrik analisis jaringan yang mengevaluasi seberapa cepat sebuah node dapat dijangkau oleh node lain dalam jaringan. Metrik ini mengukur jarak rata-rata antara node tertentu ke semua node lainnya dalam jaringan. Semakin kecil nilai rata-rata jarak, semakin tinggi tingkat *closeness centrality* dari suatu node, menandakan bahwa node tersebut lebih dekat secara keseluruhan dengan node lain dalam jaringan. Node dengan *closeness centrality* yang lebih tinggi memiliki dampak yang lebih besar dalam mengatur aliran informasi dalam jaringan karena mampu dengan cepat mengakses informasi dari atau menuju node lainnya. Penghitungan *closeness centrality* dapat didefinisikan ke dalam rumus berikut:

$$Cc(\mathrm{n}i) = \frac{\mathrm{N}-1}{\sum \mathrm{d}(\mathrm{n}i,\mathrm{n}j)} \quad (2.7)$$

Dalam konteks ini, N merupakan total jumlah node dalam jaringan dan $\sum d(\mathrm{n}i, \mathrm{n}j)$ adalah kumulatif dari jalur terpendek yang menghubungkan node ni dan nj [23]. Koefisien closeness centrality berada dalam rentang 0 hingga 1. Semakin mendekati nilai 1, semakin kecil jarak antara node dalam jaringan (lebih rapat). Dalam konteks ini, jarak yang lebih dekat menunjukkan kecepatan yang lebih tinggi dalam penyebaran informasi antara node [24].

2.1.10 *Eigenvector Centrality*

Eigenvector centrality merupakan metode evaluasi dalam analisis jaringan yang mengevaluasi signifikansi suatu node dalam jaringan berdasarkan jumlah serta signifikansi dari *node* lain yang terhubung dengannya. Metrik ini memberikan nilai yang tinggi pada node yang memiliki keterhubungan dengan node-node yang dianggap penting dalam jaringan. Semakin tinggi nilai *eigenvector centrality* suatu *node*, semakin besar dampak serta kepentingan dari node tersebut dalam jaringan. Penghitungan *Eigenvector Centrality* dapat didefinisikan ke dalam rumus berikut:

$$C_e\,(v_i) = \sum_{j=1}^{n} \mathrm{a}ij\mathrm{C}e(vj) \quad (2.8)$$

aij dalam persamaan tersebut adalah elemen matriks A yang mencerminkan keterhubungan antara simpul i dan simpul j. Kemudian, $Ce(vj)$ merujuk pada *eigenvector centrality* dari simpul j, dan penjumlahan ini diulang sebanyak n, yang merepresentasikan jumlah simpul [25].

2.1.11 *Machine Learning*

Machine learning adalah salah satu bagian dari kecerdasan buatan (AI) yang memungkinkan komputer untuk belajar tanpa instruksi yang eksplisit. Ini memungkinkan komputer untuk mengidentifikasi pola, membuat prediksi, atau mengambil keputusan tanpa perlu diprogram secara spesifik untuk melakukan tugas tertentu. Dengan menggunakan algoritma pembelajaran dalam machine learning, komputer dapat menemukan algoritma yang lebih

akurat dengan melakukan perbandingan di antara mereka. [26]. Terdapat beberapa metode pendekatan dalam *machine learning*, seperti *supervised learning* (pembelajaran terpandu), *unsupervised learning* (pembelajaran tak terpandu), dan *reinforcement learning* (pembelajaran penguatan). Pendekatan-pendekatan ini memiliki berbagai kegunaan, termasuk untuk mengklasifikasikan data, melakukan prediksi, mengenali pola, mengklasifikasikan gambar, dan masih banyak lagi. Penggunaan machine learning sangat bervariasi, mulai dari pengenalan wajah, pemrosesan bahasa alami, pengenalan suara, analisis data, hingga pengembangan otomasi. Ini memungkinkan mesin atau komputer untuk menangani tugas-tugas yang sebelumnya membutuhkan pemrograman manual secara besar-besaran, dengan memanfaatkan kemampuan belajar dari data untuk menghasilkan keputusan atau tindakan yang lebih cerdas dan efisien.

2.1.12 *Unsupervised Learning*

Unsupervised learning tidak terikat pada label apa pun pada data, hal ini mendorong algoritma untuk mengenali pola dalam input, sehingga input yang serupa dikelompokkan dalam kategori yang sama [27]. Fokus dari *unsupervised learning* adalah memungkinkan sistem komputer untuk menemukan pola, struktur, atau korelasi yang mungkin tersembunyi dalam data tanpa mengandalkan label atau petunjuk dari luar. *Unsupervised learning* melakukan analisis terhadap dataset tanpa label tanpa memerlukan intervensi manusia, melainkan proses yang didorong oleh data itu sendiri. Ini sering digunakan untuk mengekstraksi fitur yang generatif, mengidentifikasi tren dan struktur fitur yang signifikan, mengelompokkan hasil, serta eksplorasi tujuan lainnya. Metode *unsupervised learning* yang umum meliputi pengelompokan, estimasi kepadatan, pembelajaran fitur, pengurangan dimensi, menemukan aturan asosiasi, deteksi anomali, dan sebagainya [28].

2.1.12.1 Algoritma K-Means

K-means merupakan salah satu algoritma klastering yang bekerja dengan mengelompokkan data ke dalam kelompok atau klaster dengan pola atau kesamaan yang serupa di antara titik-titik data [29]. Pendekatan ini memungkinkan setiap titik data dalam suatu kelompok memiliki karakteristik yang mirip satu sama lain, namun berbeda dengan titik-titik data dalam kelompok lainnya. Berikut adalah tahapan algoritma K-Means:

a. Menentukan jumlah kluster K.
b. Mengatur nilai awal *centroid* secara acak.
c. Mencari data yang memiliki jarak paling dekat dengan *centroid* menggunakan formula *Euclidean Distance* yang dijelaskan dalam rumus (2.9)

$$D(x,y) = \sqrt{(xi - si)^2 + (yi - ti)^2} \qquad (2.9)$$

Di mana D(x, y) adalah jarak dari data ke pusat klaster x. xi dan yi adalah data *centroid*. Serta Si dan ti adalah data *record*.

d. Mengelompokkan data berdasarkan jarak terpendek dari *centroid*.
e. Jika ada perubahan pada anggota setiap klaster pada iterasi sebelumnya. Sebelum melakukan perhitungan ulang dengan rumus (2.9), hitung kembali nilai *centroid* menggunakan rumus yang dijelaskan dalam persamaan (2.10).

$$Sl = \frac{1}{Zl}(T1l + T2l + \cdots + Tnl) \qquad (2.10)$$

Di mana S_l adalah rata-rata baru dari klaster, Z_l merupakan total data dalam klaster ke-l, dan T_{nl} adalah pola ke-n yang merupakan bagian dari klaster ke-l [29].

K-means dikenal karena kesederhanaannya namun efisiensinya dalam proses klastering data. Meskipun demikian, algoritma ini memiliki keterbatasan, terutama dalam menangani data yang memiliki klaster dengan bentuk dan ukuran yang tidak teratur. K-means cenderung menghasilkan klasterisasi yang kurang akurat jika klaster dalam data memiliki bentuk yang kompleks, ukuran yang tidak seragam, atau

kepadatan yang berbeda-beda. Selain itu, sensitivitasnya terhadap inisialisasi pusat klaster awal juga dapat memengaruhi hasil klasterisasi yang dihasilkan. Walaupun demikian, K-Means memiliki keunggulan, seperti kecepatan komputasi yang tinggi, kemampuan untuk menangani dataset yang besar, dan hasil klasterisasi yang dapat diinterpretasikan secara intuitif [17].

2.1.12.2 Algoritma DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) merupakan algoritma klastering yang bekerja dengan mengelompokkan titik-titik data berdasarkan kerapatan atau kepadatan dalam ruang data [30]. Algoritma ini berhasil mengidentifikasi kelompok dengan tingkat kerapatan yang tinggi serta dapat menangani titik-titik data yang tidak masuk ke dalam kelompok tertentu, yang sering disebut sebagai *noise*.

Pendekatan DBSCAN didasarkan pada konsep kerapatan data, di mana algoritma ini mampu mengenali kelompok data dengan kerapatan yang cukup tinggi dan memisahkan daerah atau titik-titik yang jarang atau tidak padat, yang sering kali dianggap sebagai *noise* atau data yang tidak terstruktur. Langkah-langkah yang dilakukan dalam mengelompokkan data menjadi beberapa klaster menggunakan algoritma DBSCAN dijabarkan sebagai berikut:

a. Menetapkan nilai *Minimal Points* (minPts) dan *Epsilon* (eps).
b. Menetapkan titik awal secara acak.
c. Lakukan perhitungan jarak antara setiap titik (eps) menggunakan metode *Euclidean distance* dengan memanfaatkan rumus (2.11).

$$E(x,y) = \sqrt{\sum_{i=0}^{n}(Xi - Yi)^2} \qquad (2.11)$$

Dimana E(X,Y) adalah jarak antara titik X_i dengan titik Yi, X_j merupakan nilai titik 1 pada cluster ke-j, dan Y_i merupakan nilai *centroid* I pada cluster ke-j.

d. Klaster terbentuk berdasarkan tingkat kepadatan data.

e. Apabila titik yang terletak dalam radius eps mencapai atau melebihi minPts, titik p dianggap sebagai inti dan membentuk sebuah klaster.
f. Apabila p merupakan titik batas dan tidak ada titik lain yang memiliki kepadatan yang dapat dijangkau oleh p, proses tersebut akan melanjutkan ke titik lainnya.
g. Iterasi langkah c hingga f akan diulang hingga semua titik telah diproses [31].

Algoritma DBSCAN menjadi populer karena kemampuannya dalam menghadapi data yang tidak beraturan dan kemampuannya mengatasi *noise* dalam data spasial yang besar. Dengan pendekatan berbasis kepadatan, DBSCAN memberikan hasil klasterisasi yang dapat mengatasi kekurangan K-means dalam menangani struktur klaster yang kompleks dan tidak teratur pada data yang memiliki kepadatan yang berbeda-beda. DBSCAN sendiri dapat mengklasifikasikan setiap titik sebagai titik inti, titik batas, atau titik kebisingan [32].

2.1.13 *Davies-Bouldin Index (DBI)*

Indeks *Davies-Bouldin Index* (DBI) merupakan parameter evaluasi yang diterapkan dalam analisis pengelompokan (*clustering*) untuk menilai kualitas suatu partisi data dalam proses pengelompokan. Fokusnya adalah untuk mengevaluasi efektivitas pengelompokan yang dihasilkan dengan membandingkan kepadatan antara berbagai kelompok (*clusters*) dan rata-rata jarak antara titik data dalam setiap kelompok dengan pusat kelompoknya. Tujuan penggunaan DBI pada tahapan pengujian adalah untuk memaksimalkan jarak antara satu kelompok dengan kelompok lainnya, sambil juga mencari nilai yang dapat mengurangi jarak antara data atau dokumen yang berada dalam satu kelompok yang sama [33]. Apabila terdapat perbedaan yang signifikan antar *cluster* dengan jarak yang maksimum, maka perbedaan yang lebih kecil antar kelompok akan menjadi lebih terlihat. Ketika jarak

dalam *cluster* minimal, ini menunjukkan bahwa setiap objek dalam kelompok memiliki karakteristik yang sangat serupa [34].

Semakin kecil nilai *Davies-Bouldin Index*, semakin baik hasil klasteringnya. Rumus *Davies-Bouldin Index* untuk mengevaluasi kualitas klastering antara dua klaster C*i* dan C*j* adalah sebagai berikut:

$$Rij = \frac{Si+Sj}{dij} \qquad (2.12)$$

Di mana:

1. R_{ij} adalah nilai *Davies-Bouldin Index* antara klaster C_i dan C_j.
2. S_i adalah ukuran dispersi atau penyebaran data dalam klaster C_i, yang dihitung dengan variasi atau jarak rata-rata antara titik data dalam klaster dan pusat klaster m_i.
3. S_j adalah ukuran dispersi atau penyebaran data dalam klaster C_j, yang dihitung dengan variasi atau jarak rata-rata antara titik data dalam klaster dan pusat klaster m_i.
4. d_{ij} adalah jarak antara pusat klaster m_i dan m_j [35].

2.1.14 *Silhoutte Score*

Silhouette score merupakan parameter evaluasi dalam analisis pengelompokan (*clustering*) yang menilai kualitas serta kesesuaian pembagian data. Skor ini mengevaluasi sejauh mana tiap objek dalam suatu kelompok (*cluster*) sepadan dengan kelompoknya sendiri dibandingkan dengan kelompok lainnya. Pengelompokan yang baik akan menghasilkan klaster di mana elemen-elemen yang sama terletak dekat satu sama lain dalam klaster tersebut, sementara elemen-elemen dari klaster yang berbeda berada jauh dari elemen-elemen klaster lainnya. Skor *Silhouette* mempertimbangkan kedua aspek ini. Rentang skornya dari -1,0 hingga 1,0, dengan nilai yang lebih tinggi menandakan kualitas pengelompokan yang lebih baik [36].

Koefisien siluet per-sampel digunakan untuk mengidentifikasi dokumen yang kurang terkelompok dengan baik, dokumen-dokumen ini cenderung memiliki nilai siluet yang lebih rendah, yang menunjukkan letaknya pada batas antara dua kelompok atau lebih. Ini mengakibatkan tingkat ketidakpastian yang lebih tinggi dalam tugas pengelompokan. Meskipun demikian, ini bukan satu-satunya faktor, terkadang ada sampel yang, meskipun tidak terletak di batas, memiliki hubungan yang kurang kuat dengan artikel lain dalam kelompok yang sama. Hal ini sering terjadi pada dokumen yang terletak jauh dari pusat kelompok tempat mereka berada. Karena itu, pendekatan kami dalam mengurangi *noise* menggunakan dua metrik yang berbeda untuk mengidentifikasi *outliers*, koefisien siluet per sampel dan jarak kosinus dari *centroid* [37].

2.1.15 *Sum of Squared Error (SSE)*

Sum of Squared Error (SSE) merupakan parameter yang diterapkan dalam analisis klastering (*clustering*) untuk menilai seberapa jauh titik-titik data dalam suatu klaster (*cluster*) berada dari pusat klaster (*centroid*) mereka. SSE dihitung dengan menjumlahkan kuadrat dari jarak antara setiap titik data dalam klaster dengan pusat klasternya. Semakin kecil nilai SSE menunjukkan semakin konsisten atau seragamnya data di dalam setiap kelompoknya, yang mengindikasikan kualitas yang lebih baik dari kelompok yang dihasilkan. Dalam menghitung nilai *Sum of Square Error* (SSE), rumus yang digunakan adalah sebagai berikut:

$$SSE = \sum_{K=1}^{K} \sum_{x1 \epsilon Sk} ||Xi - Ck||_2^2 \qquad (2.13)$$

Keterangan:

K: jumlah *cluster*

Xi: atribut data

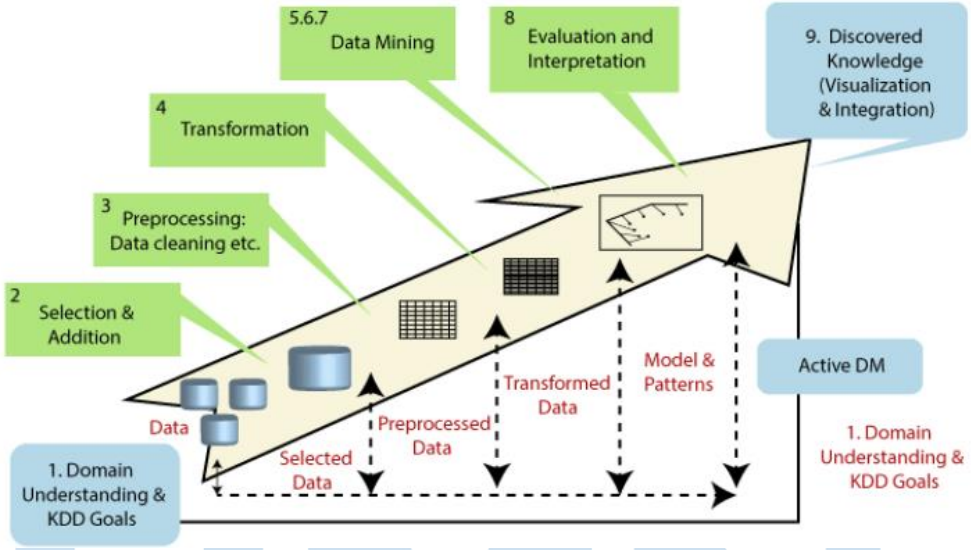
Xci: atribut *centroid*

Ci: *centroid* [38]

2.1 Algoritma dan *Framework* yang digunakan

2.2.1 KDD

KDD (*Knowledge Discovery in Databases*) merupakan suatu proses yang terorganisir untuk menemukan pola yang bermanfaat, informasi baru, atau pengetahuan yang mendalam dari data yang disimpan dalam basis data. Proses KDD melibatkan serangkaian langkah dari pemilihan data hingga evaluasi hasil analisis data, bertujuan untuk meningkatkan pengambilan keputusan, membuat prediksi, atau memperdalam pemahaman tentang data yang ada.



Gambar 2.1 Tahap Knowledge Discovery in Databases | Sumber: Javatpoint

Menurut Gustientiedina, bahwa dalam proses *Knowledge Discovery in Database* (KDD), langkah seleksi variabel diperlukan untuk menghindari kemungkinan kesamaan atau pengulangan dalam proses *Data Mining* [39]. Secara umum, proses KDD (*Knowledge Discovery in Databases*) dapat diuraikan sebagai berikut:

1. Pemilihan Data (*Data Selection*), merujuk pada proses seleksi data dari kumpulan operasional untuk digunakan dalam KDD. Data yang terpilih kemudian disimpan dalam berkas terpisah.

2. Pre-processing atau Pembersihan (*Cleaning*), tahap awal di mana data yang dipilih melalui proses pembersihan. Ini mencakup eliminasi duplikasi, pengecekan konsistensi, perbaikan kesalahan data, dan penghapusan inkonsistensi.

3. Transformasi (*Transformation*), proses mengubah skala atau bentuk distribusi data sehingga sesuai dengan distribusi yang diharapkan atau yang lebih cocok untuk analisis tertentu [40]. Kemudian, melibatkan *encoding* atau transformasi data yang telah dipilih agar sesuai untuk analisis data. Proses ini bergantung pada jenis informasi yang diinginkan dari basis data.

4. *Data Mining*, fase kunci dalam KDD di mana teknik, algoritma, atau metode tertentu digunakan untuk menemukan pola menarik, korelasi, atau informasi tersembunyi dalam data yang dipilih. Pada proses data mining, informasi diekstraksi melalui analisis pola dari kumpulan data yang berukuran besar. Penggunaan data yang besar umumnya menghasilkan hasil yang lebih dapat dipercaya [40].

5. Interpretasi atau Evaluasi (*Interpretation or Evaluation*), pola atau informasi yang ditemukan dari analisis dievaluasi dan ditafsirkan. Tahap ini melibatkan penafsiran hasil analisis, penilaian keberhasilan teknik data mining, dan pembandingan informasi baru dengan pengetahuan sebelumnya atau kebutuhan pengguna.

2.2 *Tools* dan *Software* yang digunakan

2.3.1 Python

Python merupakan salah satu bahasa pemrograman yang menggunakan interpreter untuk mengeksekusi dan menerjemahkan kode program secara langsung. Python dapat dioperasikan pada berbagai *platform* seperti Linux, Windows, dan sistem operasi lainnya. Bahasa pemrograman Python pertama kali dikembangkan oleh Guido van Rossum di Stichting Mathematisch Centrum, Amsterdam, pada tahun 1991. Pengembangan Python terinspirasi

dari bahasa pemrograman ABC yang sedang berkembang pada masa itu. Python dapat dimanfaatkan untuk beragam keperluan, termasuk pengembangan aplikasi web, aplikasi *desktop*, *Internet of Things* (IoT), serta berbagai jenis aplikasi lainnya.

Gambar2.2 Logo Python | Sumber: PNGEgg

Dikarenakan kemampuannya dalam berintegrasi dengan sistem database serta kemampuan untuk membaca dan mengubah file, Python sering digunakan untuk *prototyping* atau pengembangan perangkat lunak dengan cepat dan efisien. [41]. Selain itu, dalam konteks bahasa pemrograman Python, terdapat beragam perpustakaan (*library*) yang bermanfaat dan dapat dimanfaatkan secara luas oleh berbagai pengguna di berbagai sistem operasi karena sifatnya yang bersifat *open source*. Beberapa contoh perpustakaan Python meliputi NumPy, Pandas, Matplotlib, dan Scikit-learn, yang masing-masing memiliki kegunaan khusus dalam analisis data, statistik, visualisasi data, dan pembelajaran mesin. Python juga memiliki kemudahan dalam integrasi dengan teknologi lain yang terkait dengan analisis data, seperti basis data, alat big data, framework web, dan sebagainya. Ini memungkinkan akses dan pengelolaan data dari berbagai sumber dengan lebih mudah. [42].

2.3.2 Google Colab

Colaboratory, atau yang disebut *'Colab'* singkatnya, merupakan hasil dari penelitian *Google*. *Colab* memungkinkan siapa pun menulis dan mengeksekusi kode Python tanpa batasan melalui peramban, sangat sesuai untuk kegiatan pembelajaran mesin, analisis data, dan keperluan Pendidikan

[43]. *Google Colab* adalah sebuah *platform cloud* dari *Google* yang memberikan kemampuan kepada pengguna untuk membuat dan menjalankan kode Python di lingkungan *notebook* yang bisa diakses secara *online*.

Gambar 2.3 Logo Google Colab | Sumber: wikipedia

Platform ini memberikan akses gratis terhadap sumber daya komputasi berbasis cloud, termasuk notebook interaktif yang berjalan di *server Google*. Dengan *Colab*, pengguna dapat menulis dan menjalankan kode, menyimpan dan berbagi notebook, serta menggunakan sumber daya komputasi seperti CPU, GPU, dan TPU (*Tensor Processing Unit*) yang disediakan oleh *Google*. Umumnya, *platform* ini digunakan untuk keperluan eksperimen, pengembangan model *machine learning*, analisis data, serta berbagai kegiatan pemrograman Python. Google Colab hadir dengan pustaka Python seperti Pandas, Matplotlib, dan Plotly yang siap digunakan untuk memanipulasi data serta membuat representasi visual dari data [44].

2.3.3 Microsoft Excel

Microsoft Excel adalah sebuah perangkat lunak spreadsheet yang dibuat oleh Microsoft. Microsoft Excel beroperasi di bawah sistem operasi *Windows*. Seperti Microsoft Word, Excel sering digunakan di berbagai bidang, terutama dalam konteks yang memerlukan perhitungan matematika yang rumit [45]. Program ini didesain untuk membantu pengguna dalam membuat, mengelola, dan menganalisis data yang disusun dalam bentuk tabel. Excel menyediakan beragam fungsi dan fitur untuk melakukan perhitungan matematika, analisis data, visualisasi data, serta pembuatan grafik dan laporan. Dengan tampilan yang intuitif, pengguna dapat mengelola data dalam sel-sel yang membentuk

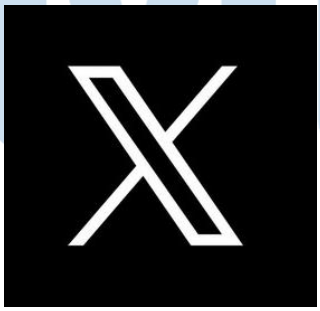
baris dan kolom, serta memanfaatkan formula, fungsi, dan alat analisis bawaan untuk mengubah data sesuai kebutuhan.

Gambar 2.4 Logo Microsoft Excel | Sumber: Logo.wine

Excel digunakan secara luas dalam berbagai industri, baik untuk keperluan bisnis, pendidikan, keuangan, dan keperluan lainnya yang melibatkan manajemen data. Microsoft excel biasanya digunakan untuk membantu menyelesaikan permasalahan administratif dengan menggunakan rumus atau formula pada lembar kerja [46].

2.3.4 X

X atau sebelumnya dikenal dengan nama Twitter adalah suatu *platform* media sosial yang memungkinkan individu untuk berbagi gagasan, informasi, atau pikiran dalam format pesan singkat yang disebut "*tweet*". Setiap pesan dibatasi hingga 280 karakter [47]. Pengguna dapat membagikan teks, foto, video, dan tautan, serta berinteraksi dengan pengguna lain melalui *like*, *retweet*, dan tanggapan pada *tweet* yang diposting oleh pengguna lainnya.



Gambar 2.5 Logo X | Sumber: twitter.com

X juga berfungsi sebagai sumber berita, tempat diskusi tentang topik tertentu, *platform* untuk kampanye, dan kegiatan jaringan sosial lainnya. *Platform* ini sangat populer di Indonesia dan diminati oleh berbagai kelompok

usia dan latar belakang. Penggunaannya meliputi remaja hingga orang dewasa dari berbagai segmen masyarakat. Keberhasilannya disebabkan oleh reputasinya sebagai *platform* sosial yang memberikan informasi cepat dan akurat dari pengguna lain. Foto dan video yang dilampirkan pada postingan membuat informasi di Twitter menjadi lebih menarik. *Platform* ini menjadi tempat mencari berita terbaru, mengikuti kehidupan selebritas, memperoleh edukasi dan tips, membangun hubungan, dan mempromosikan serta mendiskusikan tren terkini [48].

2.3.5 NodeXL

NodeXL adalah perangkat lunak analisis jaringan yang terintegrasi dalam *Microsoft Excel*. Dengan NodeXL, pengguna dapat mengimpor data jaringan sosial dari berbagai *platform* seperti X, Facebook, dan LinkedIn, serta menganalisis dan memvisualisasikan data tersebut dalam bentuk grafik dan tabel di Excel. Perangkat ini berguna dalam memahami hubungan dan pola dalam jaringan sosial, serta mengukur sentralitas, peran, dan interaksi antar entitas di dalam jaringan. NodeXL sering digunakan untuk riset akademis, analisis jaringan sosial, dan pemahaman perilaku komunitas online. Penelitian ini memanfaatkan NodeXL, sebuah aplikasi yang dikembangkan oleh smrfoundation. NodeXL merupakan ekstensi tambahan untuk Microsoft Excel yang dirancang khusus untuk penelitian yang berfokus pada Analisis Jaringan Sosial.

Gambar 2.6 Logo NodeXL | Sumber: nodexl.com

Ada beberapa konsep penting di dalam NodeXL yang perlu dipahami, di antaranya *Vertices* yang mewakili organisasi, kelompok, atau individu dalam jaringan, *Edges* yang merupakan relasi atau hubungan yang menghubungkan

Vertices, serta *cluster* yang merupakan fitur yang memungkinkan pemisahan dan pembentukan kelompok-kelompok dengan kesamaan di antara *Vertices*. NodeXL juga menyediakan fitur perhitungan metrik yang membantu dalam menganalisis kekuatan pengaruh aktor terhadap publik. Dikarenakan berbasis Excel, data hasil perhitungan dapat disajikan dalam format tabel yang mudah dipahami [49].

2.3 Penelitian Terdahulu

Tabel 2.1 Penelitian Terdahulu

Nama Jurnal	Judul Artikel	Penulis	Metode	Hasil
International Journal of Electronics and Communication Engineering [50]	*Improving Fake News Detection Using K-Means and Support Vector Machine Approaches*	Kasra Majbouri Yazdi, Adell Majbouri Yazdi, Saeid Khodayi, Jingyu Hou, Wanlei Zhou, Saeed Sae dy	• *K-Means* • *Support Vector Machine* • *Decision Tree* • *Naïve Bayes*	Hasil yang didapat pada penelitian tersebut adalah SVM Classifier yang memiliki hasil presisi yang lebih tinggi dibandingkan dengan *Decision Tree* dan *Naïve Bayes*.
Social and Information Networks [51]	*Fakeswarm: Improving Fake News Detection With Swarming Characteristics*	Jun Wu, Xuesong Ye	*Principal Component Analysis Metric Representation Position Encoding DBSCAN*	Hasil yang didapat pada penelitian ini adalah penggabungan ketiga fitur swarm mencapai f1-score dan akurasi lebih dari 97 %.
Journal of Computer System and Informatics [9]	*Clustering Content Types and User Motivation using DBSCAN on Twitter*	Made Mita Wikantari, Yuliant Sibaroni, Aditya Firman Ihsan	• *DBSCAN*	Hasil yang didapat pada penelitian ini adalah penerapan metode DBSCAN Clustering terbukti optimal dilihat dari nilai sillhoutte score, yaitu 0.29 yang menghasilkan total 3 cluster.
TEM Journal [52]	*Application of Density Based Clustering of*	Mochammad Haldi Widianto, Ivan Diryana	• *DBSCAN*	Hasil yang didapat dalam penelitian ini adalah metode

Nama Jurnal	Judul Artikel	Penulis	Metode	Hasil
	Disaster Location in Realtime Social Media	Sudirman, Muhammad Hanif Awaluddin		ini bermanfaat dalam mengklasifikasikan data berdasarkan Tingkat kemiripannya dan NER rule-based dapat mendeteksi setiap tweet yang sudah dikelompokkan.
Intelligent Systems and Applications in Engineering [8]	*Unsupervised Misinformation Detection Model using Incremental K-Means Algorithm*	Yashoda Barve, Jatinderkumar R. Saini	• *Latent Dirichlet Allocation Method* • *K-Means*	Hasil yang didapat dalam penelitian ini adalah pembentukan cluster rata-rata dengan sillhoute score sebesar 0.57. Dapat dilihat fitur khusus pengguna yang baru dibentuk seperti reputasi, negara dan fitur tekstual seperti ISPA sangat berkontribusi dalam pembentukan cluster.

Pada tabel 2.1, terdapat beberapa metode yang digunakan dalam mendeteksi berita hoaks di media sosial dengan menggunakan teknik algoritma K-Means dan DBSCAN. Dari hasil pemahaman penelitian terdahulu terlihat bahwa algoritma K-Means dan DBSCAN mampu memberikan hasil akurasi yang terbaik dalam konteks pendeteksian berita palsu. Setelah diteliti kembali, penelitian terdahulu yang didapatkan, masih belum ada yang membahas perbandingan antara kedua algoritma DBSCAN dan K-Means, sehingga hal ini menjadi salah satu pembaruan dalam penentuan penelitian mengenai komparasi algoritma dalam mendeteksi berita hoaks.

Secara keseluruhan, penelitian ini memberikan gambaran mendalam tentang keberhasilan berbagai teknik dan metode dalam mengklasifikasikan data, membentuk *cluster*, serta mengidentifikasi fitur yang berperan dalam proses tersebut.