BAB II

TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

2.1.1 Analisis Sentimen

Analisis sentimen merupakan jenis pembelajaran mesin yang mengkaji sentimen, opini, penilaian, dan pandangan dalam kumpulan teks untuk mengidentifikasi karakteristik kumpulan kata. Proses analisis sentimen mengklasifikasikan dokumen teks ke dalam kelas positif dan negatif [18]. Dalam beberapa tahun terakhir, Analisis Sentimen telah diterapkan dalam berbagai cara untuk mengekspresikan sentimen dalam bentuk tekstual alternatif, terutama melalui informasi dari jejaring sosial. Dalam literatur terkini di wilayah tersebut, terdapat karya yang mengidentifikasi sentimen melalui suara, menggunakan emotikon dari jejaring sosial, dan gambar [19]. Selain itu, Analisis Sentimen juga digunakan di berbagai sektor masyarakat mengetahui opini publik.

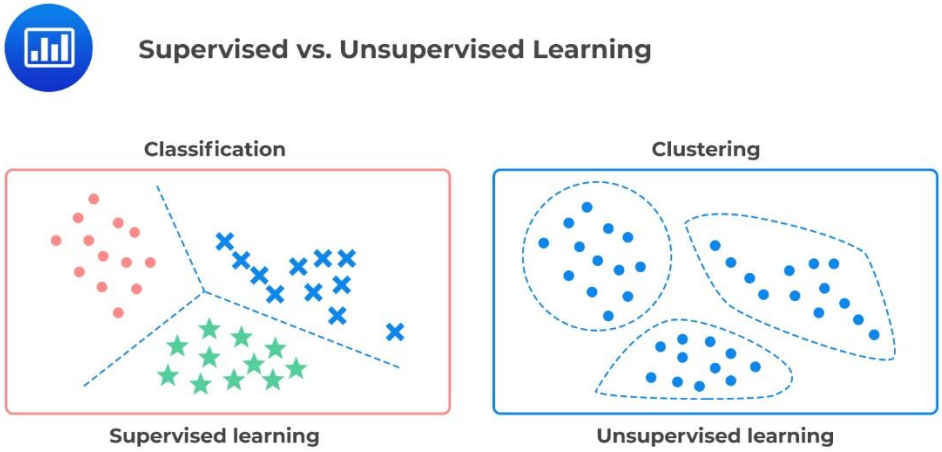
Dalam beberapa tahun terakhir, analisis sentimen telah berkembang pesat dan diterapkan dalam berbagai konteks, terutama dengan melibatkan informasi yang ditemukan di media sosial [20]. Selain itu, teknik analisis sentimen juga telah diperluas untuk mencakup sumber data lain seperti suara, emotikon, dan gambar. Misalnya, penggunaan emotikon dari jejaring sosial atau informasi audio dapat digunakan untuk memperkaya analisis sentimen, memungkinkan sistem untuk lebih memahami ekspresi emosi yang terkandung dalam teks atau pesan. Dalam konteks masyarakat, analisis sentimen digunakan untuk memahami opini publik tentang berbagai topik, mulai dari produk dan layanan hingga peristiwa politik dan sosial. Data teks yang biasanya digunakan dalam analisis sentimen termasuk artikel, *tweet*, ulasan produk, dan komentar di media sosial. Dengan menganalisis data teks ini, organisasi dan perusahaan dapat memahami reaksi dan respons publik terhadap produk, layanan, atau

kejadian tertentu, sehingga dapat mengambil tindakan yang sesuai untuk meningkatkan pengalaman atau memperbaiki masalah yang ada.

Sentimen analisis biasanya digunakan untuk mengidentifikasi sentimen yang terkandung dalam teks dengan mengukur seberapa positif atau negatif data teks. Ini dilakukan pada artikel, tweet, dan ulasan produk untuk melihat reaksi atau tanggapan publik terhadap suatu peristiwa [18].

2.1.2 Machine Learning

Machine learning merupakan cabang dari kecerdasan buatan yang membuat komputer mampu untuk belajar dari data dan pengalaman tanpa harus secara eksplisit diprogram [11]. Konsep dasar di balik machine learning adalah penggunaan algoritma matematis untuk menganalisis data yang diberikan, mengidentifikasi pola yang ada di dalamnya, dan membuat keputusan atau prediksi berdasarkan pola tersebut tanpa intervensi manusia.



Gambar 2.1 Perbedaan supervised learning dan unsupervised learning

Sumber: [43]

Dalam machine learning terdapat beberapa pendekatan, termasuk supervised learning di mana model dilatih menggunakan data yang telah dilabeli, unsupervised learning di mana model mencari pola dalam data yang tidak memiliki label, dan reinforcement learning di mana model belajar melalui trial and error berdasarkan umpan balik yang diberikan atas tindakan-

tindakannya. Konsep ini memungkinkan komputer untuk melakukan tugas-tugas kompleks seperti pengenalan wajah, analisis teks, pengenalan suara, pemrosesan bahasa alami, dan bahkan mengemudi kendaraan otonom. Dengan kemampuannya untuk belajar dari data dan pengalaman, machine learning telah menjadi alat yang sangat penting dalam berbagai bidang dan membuka banyak potensi untuk aplikasi cerdas di berbagai domain kehidupan.

Konsep machine learning telah menghadirkan kemampuan komputer untuk melakukan tugas-tugas yang kompleks dan seringkali sulit bagi manusia, seperti pengenalan wajah, analisis teks, pengenalan suara, pemrosesan bahasa alami, dan bahkan mengemudi kendaraan otonom [21]. Dengan adanya pemahaman yang kuat tentang konsep-konsep ini, dapat mengapresiasi bagaimana komputer dapat belajar dan beradaptasi secara otomatis dari data, membuka banyak potensi untuk aplikasi cerdas di berbagai domain kehidupan, mulai dari kesehatan dan keuangan hingga teknologi dan transportasi. Dengan demikian, machine learning telah menjadi alat yang sangat penting dalam revolusi teknologi modern, membantu mengatasi tantangan kompleks dan menciptakan solusi inovatif untuk masalah-masalah dunia nyata.

2.1.3 Supervised Learning

2.1.3.1 Decision Tree

Decision Tree merupakan sebuah diagram dengan satu node dan cabang untuk setiap pilihannya, yang setiap cabang akan membuat cabang baru. [12]. Untuk melakukan klasifikasi suatu tujuan, DT digunakan untuk membuat pohon keputusan dari data pelatihan. Algoritma DT, yang pertama kali dikembangkan oleh Ross Quinlan J. pada tahun 1979 dan dikenal sebagai pembentuk pohon keputusan ID3 [20]. Pohon keputusan ini terdiri dari simpul dan tepi. Ini dimulai dari simpul akar dan berakhir pada simpul daun. Fitur data diwakili oleh setiap simpul pohon keputusan, dan nilai fitur tersebut diwakili oleh keputusan yang diambil berdasarkan fitur tersebut.Decision tree

menemukan fitur yang paling efektif untuk membagi data menjadi subset yang lebih kecil dengan varian yang lebih sedikit. Saat semua data yang diuji telah diklasifikasikan, proses ini berhenti [22]. Terdapat dua tahapan yang perlu dilakukan pada Decision Tree yaitu *learning* dan *classification*. Pada tahap belajar, algoritma DT membuat pohon keputusan dengan data dan hasil klasifikasinya. Pohon keputusan ini kemudian menjadi model untuk digunakan dalam klasifikasi data yang belum diklasifikasikan. Langkah kerja DT mudah dipahami dan cepat dilaksanakan [23]. Salah satu keunggulan utama dari Decision Tree adalah kemampuannya dalam mengidentifikasi fitur terbaik yang membagi data menjadi subset yang lebih kecil dengan varian minimum. Proses ini disebut sebagai proses pemilihan atribut yang memaksimalkan kehomogenan atau mengurangi ketidakpastian dalam setiap subset. Dengan cara ini, DT dapat memisahkan kelas-kelas target dengan efisien, sehingga memungkinkan untuk pembuatan keputusan yang akurat.

Tahapan utama dalam penggunaan DT adalah learning (pembelajaran) dan classification (klasifikasi). Pada tahap pembelajaran, algoritma DT menggunakan data latih yang telah disertai dengan label kelasnya untuk membangun pohon keputusan [23]. Proses pembangunan pohon keputusan ini melibatkan pemilihan atribut terbaik sebagai pemisah pada setiap simpul, sehingga menghasilkan cabang-cabang yang merepresentasikan keputusan yang berbeda. Setelah model pohon keputusan dibangun, tahap klasifikasi dilakukan untuk mengklasifikasikan data baru yang belum diketahui label kelasnya berdasarkan aturan yang telah dipelajari dari data latih. Keuntungan lain dari DT adalah kemudahan dalam interpretasi dan visualisasi hasil. Karena pohon keputusan dapat direpresentasikan dalam bentuk grafis yang intuitif, maka proses pengambilan keputusan dapat dengan mudah dipahami oleh pengguna atau pemangku kepentingan. Selain itu, DT juga dapat menangani baik data kategorikal maupun numerikal tanpa perlu

transformasi data tambahan, sehingga cocok digunakan dalam berbagai kasus di berbagai bidang aplikasi.

$$1 - \sum_{i=1}^{k} p_i^2 \qquad (2.1)$$

Deskripsi:

k : jumlah kelas

p_i: proporsi sampel yang termasuk dalam kelas i pada node tersebut.

2.1.3.2 KNN (K-Nearest Neighbor)

Algoritma K-Nearest Neighbors (KNN) merupakan salah satu metode dalam machine learning yang digunakan untuk melakukan klasifikasi atau regresi berdasarkan data yang telah diberikan. Algoritma ini termasuk dalam kategori supervised learning, di mana data pelatihan (*training data*) yang memiliki label digunakan untuk mempelajari hubungan antara fitur (*features*) dan label (target) [12]. KNN memiliki beberapa keunggulan dalam penggunaannya, seperti:

1. Termasuk algoritma yang sederhana sehingga mudah untuk dipahami dan diimplementasikan.
2. Berbasis instansi sehingga tidak memerlukan pelatihan yang kompleks.
3. Dapat menangani batas keputusan yang kompleks dan non-linear sehingga lebih fleksibel.

Selain keunggulan-keunggulan yang telah disebutkan, Algoritma K-Nearest Neighbors (KNN) juga memiliki karakteristik yang unik dalam proses pengambilan keputusan. Prinsip utama dari KNN adalah dengan mengklasifikasikan data baru berdasarkan mayoritas label dari k-nearest neighbors atau tetangga terdekatnya [24]. Artinya, jika suatu data baru akan diklasifikasikan, algoritma KNN akan mencari k-nearest neighbors

dari data tersebut dalam ruang fitur, kemudian menentukan label mayoritas dari tetangga-tetangga tersebut sebagai label prediksi untuk data baru. Keunikan dari pendekatan ini adalah bahwa KNN tidak melakukan proses pembelajaran secara eksplisit seperti halnya algoritma supervised learning lainnya. Sebagai gantinya, KNN mengandalkan informasi yang terdapat pada data latih secara langsung saat proses klasifikasi. Dalam konteks ini, KNN dikenal sebagai algoritma berbasis instansi atau lazy learner, karena tidak melakukan proses pembelajaran untuk menghasilkan model internal [25]. Hal ini membuat KNN lebih fleksibel dan adaptif terhadap perubahan pada data latih, karena tidak terikat pada model tertentu.

Meskipun sederhana, KNN memiliki beberapa kelemahan, terutama terkait dengan kompleksitas komputasi saat melakukan klasifikasi pada data berukuran besar [25]. Karena KNN memerlukan perhitungan jarak antara data baru dengan setiap data latih untuk menentukan tetangga terdekat, maka waktu komputasi dapat menjadi masalah terutama saat dimensi data atau jumlah data yang besar. Selain itu, KNN juga sensitif terhadap pemilihan parameter k, yaitu jumlah tetangga terdekat yang digunakan dalam proses klasifikasi. Pemilihan nilai k yang tidak tepat dapat mengakibatkan model yang tidak optimal, baik terlalu kompleks maupun terlalu sederhana. Berikut merupakan rumus dari algoritma KNN:

$$(X_{test}, X_i) = \sqrt{\sum_{j=1}^{d} (x_{test,j} - x_{i,j})^2} \qquad (2.2)$$

Deskripsi:

d : jumlah fitur (dimensi)

x_{test} : data uji

x_i : data pelatihan

2.1.3.3 Naïve Bayes

Algoritma Naïve Bayes ditemukan oleh Reverend Thomas Bayes pada pertengahan abad ke-18 [26]. Naïve bayes populer disebut sebagai teknik untuk mengkategorikan dan mengkategorikan teks berdasarkan frekuensi kata-kata. Pada penelitian ini, algoritma Naive Bayes akan digunakan untuk mengklasifikasikan data teks yang diambil dari YouTube menjadi kelas positif atau kelas negatif. Naive Bayes adalah metode klasifikasi statistik yang dapat memprediksi keanggotaan kelas di mana sampel yang ada akan termasuk ke dalam kelas tertentu. Metode ini didasarkan pada teorema Bayes, yang menunjukkan bahwa ia dapat memprediksi kemungkinan bahwa sampel yang ada akan termasuk ke dalam kelas tertentu di masa mendatang [15].

Algoritma pengklasifikasian probabilitas Naïve Bayes menghitung sekumpulan kemungkinan dengan menggabungkan nilai dari kumpulan data dan kemudian mengjumlahkan frekuensinya [23]. Teorema Bayes menganggap bahwa semua atribut yang ada adalah atribut independen atau tidak bergantung pada nilai variabel kelas. Salah satu keuntungan algoritma Naive Bayes adalah bahwa mereka memiliki tingkat akurasi yang cukup tinggi saat menangani data besar, seperti sentimen analisis [27].

Salah satu asumsi utama dari algoritma Naïve Bayes adalah bahwa semua atribut dalam dataset adalah independen satu sama lain, artinya nilai dari setiap atribut tidak bergantung pada nilai atribut lainnya ketika kelas target diketahui. Meskipun asumsi ini sering kali tidak realistis dalam konteks dunia nyata, Naïve Bayes tetap sering digunakan karena sifatnya yang cepat dan mudah diimplementasikan, serta kinerjanya yang cukup baik dalam banyak kasus, terutama ketika data cukup besar [28]. Rumus dasar Naïve Bayes menggabungkan probabilitas kelas yang diketahui dengan probabilitas atribut-atribut yang terkait dengan kelas

tersebut. Dengan menggunakan teorema Bayes, algoritma ini menghitung probabilitas posterior dari setiap kelas berdasarkan atribut-atribut yang diamati. Kemudian, kelas dengan probabilitas posterior tertinggi dipilih sebagai prediksi untuk sampel tertentu. Meskipun sederhana, Naïve Bayes sering kali memberikan hasil klasifikasi yang cukup baik, terutama dalam konteks analisis sentimen di mana fokus utamanya adalah pada pemahaman sentimen positif atau negatif dari teks yang diberikan.Berikut adalah rumus Naïve Bayes [14].

$$P(C \mid X) = \ (P(C \mid X).P(c))/(P(X)) \quad (2.3)$$

Deskripsi:

X : Data dengan kelas yang tidak diketahui

C : Data yang dihipotesiskan X adalah kelas tertentu

P(C|X) : Probabilitas hipotesis C berdasarkan kondisi X (probabilitas poterior)

P(C) : Probabilitas hipotesis C (probabilitas sebelumnya)

P(X|C) : Probabilitas X berdasarkan kondisi dalam hipotesis C (kemungkinan)

P(X) : Probabilitas X (probabilitas prediktor sebelumnya)

2.1.3.4 Support Vector Machine

Support Vector Machine (SVM) adalah algoritma klasifikasi yang populer dalam penelitian pembelajaran mesin [28]. Sebagai landasan teori statistik, SVM sangat baik dalam menyelesaikan masalah klasifikasi biner dan multikelas. Bahkan dalam kasus di mana data memiliki banyak fitur kompleks, SVM tetap menjadi algoritma pembelajaran mesin yang paling efisien untuk mengklasifikasikan sejumlah besar data [12]. Fungsi hyperplane atau pemisah yang paling baik memisahkan kelas data dapat berbentuk garis atau bidang, bergantung pada dimensi ruang yang

digunakan. SVM akan melakukan pencarian hyperplane ini.Untuk mengoptimalkan pemisahan antara kelas data.

Algoritma Support Vector Machine (SVM) memiliki beberapa fitur yang digunakan untuk mengoptimalkan pemisahan antar kelas data [25]. Fungsi tujuan SVM (minimisasi) menjumlahkan kontribusi setiap sampel data menggunakan fungsi kernel yang mengukur kedekatan antar sampel data, meminimalkan kesalahan klasifikasi, dan memaksimalkan margin. Memilih fungsi partisi, juga dikenal sebagai hyperplane, membagi ruang fitur menjadi dua kelas dengan margin terbesar. Hyperplane ditentukan oleh vektor bobot dan konstanta bias, yang dihitung menggunakan metode optimasi seperti metode gradien atau metode minimum sekuensial (SMO) [29].

SVM merupakan algoritma yang sangat fleksibel yang dapat digunakan untuk menyelesaikan banyak masalah klasifikasi, termasuk masalah dengan data nonlinier dan banyak atribut [30]. SVM juga berguna untuk menangani unbalanced data atau data tidak seimbang dimana jumlah sampel pada suatu kelas tertentu sedikit. SVM juga memiliki kelemahan. Perbedaan pada kumpulan data yang besar memerlukan waktu pelatihan yang lama [31]. Kumpulan data yang berisik juga dapat menyebabkan model yang dihasilkan menjadi kurang optimal dan menurunkan performa pada data pengujian.

Selain itu, SVM bisa gagal jika jumlah fitur di setiap data lebih besar daripada jumlah sampel di data pelatihan. Secara keseluruhan, algoritma klasifikasi SVM sangat populer di kalangan peneliti pembelajaran mesin. SVM ideal untuk menangani data nonlinier dan tidak seimbang, serta masalah klasifikasi biner dan multikelas [32]. Namun, seperti algoritma pembelajaran mesin lainnya, SVM memiliki kekurangan yang perlu diperhatikan, yaitu performa buruk pada kumpulan data yang berisik dan berdurasi panjang. Berikut rumus yang dapat digunakan:

$$w^T x + b = 0 \quad (2.4)$$

Deskripsi:

w : vektor bobot (weight)

x : vektor fitur input

b : bias

2.1.3.5 Artificial Neural Network

Model penalaran yang didasarkan pada otak manusia disebut juga sebagai Algoritma Neural Artificial (ANN) [33]. ANN terdiri dari sejumlah prosesor yang sangat sederhana dan saling berhubungan yang disebut neuron. Sinyal mengalir dari neuron satu ke neuron lain melalui neuron yang terhubung dengan pembobotan. ANN dapat meniru jaringan neural biologis [34]. Jaringan syaraf tiruan dapat mengatur dirinya untuk menghasilkan respons yang konsisten terhadap rangkaian masukkan melalui proses belajar. Jaringan saraf imitasi dibangun dan dilatih untuk memiliki kemampuan yang sebanding dengan manusia [35]. Neuron mungkin memiliki banyak masukan dan satu keluaran. Hasil keluaran neutron dapat berupa hasil akhir atau bahan masuk untuk neutron berikutnya, sedangkan jalur masukan neuron dapat berisi data mentah atau hasil olahan neuron sebelumnya.

Model ANN ini mencerminkan cara kerja jaringan neural biologis, meskipun secara konseptual lebih sederhana daripada otak manusia. Dengan melalui proses pelatihan (training), jaringan syaraf tiruan dapat belajar dan menyesuaikan diri terhadap data yang diberikan. Hal ini dilakukan dengan mengoptimalkan bobot-bobot antar neuron agar jaringan dapat menghasilkan respons yang diinginkan terhadap rangkaian input yang diberikan. Dengan kata lain, ANN dirancang untuk mempelajari pola-pola yang tersembunyi dalam data dan menggunakan informasi tersebut untuk membuat prediksi atau keputusan. Salah satu keunggulan utama dari ANN adalah kemampuannya untuk menangani kompleksitas dan ketidaklinieran dalam data [19]. Dengan jumlah neuron

dan layer yang tepat, ANN dapat memodelkan hubungan yang sangat rumit antara variabel-variabel input dan output, bahkan dalam kasus data yang sangat besar. Hal ini membuat ANN menjadi salah satu alat yang sangat berguna dalam berbagai aplikasi machine learning dan analisis data, termasuk dalam konteks analisis sentimen di mana model perlu memahami dan menggeneralisasi pola-pola dalam data teks untuk membuat prediksi sentimen yang akurat.Di bawah ini merupakan rumus *forward propagation* dari algoritma ANN:

$$z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)} \quad (2.5)$$
$$a^{(l)} = \sigma(z^{(l)})$$

Deskripsi:

$W^{(l)}$: matriks bobot untuk lapisan *l*.

$b^{(l)}$: vektor bias untuk lapisan *l*.

σ : fungsi aktivasi yang diterapkan pada $z^{(l)}$.

2.1.3.6 Adaboost

Adaboost (*Adaptive Boosting*) adalah sebuah algoritma pembelajaran ensemble yang termasuk ke dalam kategori algoritma Machine Learning, khususnya dalam pembelajaran terawasi (*supervised learning*) [36]. Lebih khusus lagi, Adaboost adalah booster yang digunakan untuk meningkatkan kinerja model prediktif dengan menggabungkan hasil dari beberapa model lemah (*weak learner*) menjadi satu model yang lebih kuat (*strong learner*).

Adaboost merupakan algoritma yang efektif untuk masalah klasifikasi biner dan sering digunakan dalam praktik karena kemampuannya menghasilkan model yang kuat dengan memanfaatkan model-model lemah yang relatif sederhana [36]. Namun, Adaboost rentan terhadap *noise* atau *outlier* dalam data pelatihan dan bisa *overfit*

jika tidak diatur dengan benar. Berikut merupakan rumus adaboost untuk bobot sampel:

$$w_i^{(t)} = w_i^{(t-1)} e^{-\alpha_t y_i h_t(x_i)} \quad (2.6)$$

Dekripsi:

$w_i^{(t-1)}$: bobot sampel pada iterasi sebelumnya

α_t : koefisien pembelajaran (*learning rate*) dari model lemah h_t pada iterasi *t*.

y_i : label sebenarnya dari sampel x_i (-1 atau 1 dalam klasifikasi biner).

$h_t(x_i)$: prediksi model lemah h_t terhadap sampel x_i.

2.1.4 Unsupervised Learning

2.1.4.1 K-means

Algoritma pengelompokan yang disebut K-Means mengelompokkan data ke dalam kelompok atau cluster dengan pola serupa atau kemiripan antar titik data [37]. Metode ini memungkinkan setiap titik data dalam suatu kelompok mempunyai sifat yang serupa satu sama lain namun berbeda dengan titik dalam kelompok lainnya. K-means dikenal dengan efisiensi dan kesederhanaannya dalam proses clustering data. Namun algoritma ini mempunyai beberapa keterbatasan terutama ketika berhadapan dengan data yang mengandung cluster yang bentuk dan ukurannya tidak beraturan.

Pengelompokan K-means cenderung kurang akurat jika cluster dalam data Anda tidak konsisten dalam bentuk, ukuran, dan kepadatan.

Selanjutnya, hasil clustering yang dihasilkan dapat dipengaruhi oleh sensitivitas inisialisasi pusat cluster awal. Namun K-Means memiliki kelebihan seperti kecepatan komputasi yang tinggi, kemampuan menangani dataset yang besar, dan hasil *clustering* yang lebih mudah

dipahami [38]. Langkah-langkah algoritma penyebaran K-means adalah sebagai berikut:

1. Menentukan jumlah cluster
2. Menentukan nilai centroid: Nilai awal centroid digunakan secara acak untuk menentukan nilai awal iterasi, tetapi untuk menentukan nilai centroid untuk tahap iterasi,

Salah satu keunggulan utama dari K-means adalah kesederhanaan dan efisiensinya dalam melakukan klasterisasi data. Algoritma ini relatif mudah dipahami dan diimplementasikan, serta memiliki kecepatan komputasi yang tinggi sehingga cocok digunakan dengan data dalam jumlah besar. Hasil clustering yang dihasilkan juga intuitif untuk diinterpretasikan, sehingga memungkinkan pengguna memahami struktur data yang ada dengan mudah. Namun, K-means juga memiliki keterbatasan [38]. Algoritme ini cenderung memiliki akurasi pengelompokan yang buruk jika cluster dalam data Anda tidak konsisten dalam bentuk, ukuran, dan kepadatan. Selain itu, sensitivitas algoritma terhadap inisialisasi pusat cluster awal dapat mempengaruhi hasil clustering yang dihasilkan. Namun demikian, mengingat keterbatasan ini, K-means tetap menjadi pilihan populer untuk analisis klaster, terutama dalam kasus-kasus di mana interpretasi intuitif dari hasil klasterisasi lebih diutamakan daripada akurasi mutlak, maka digunakan rumus sebagai berikut:

$$\overline{V_{ij}} = \frac{1}{N_i} \sum_{k=0}^{N_i} X_{kj} \qquad (2.7)$$

Deskripsi:

vij : centroid/ rata-rata cluster ke-I untuk variable ke-j

Ni : jumlah data yang menjadi anggota cluster ke-i

i,k : indeks dari cluster

j : indeks dari variabel

xkj : nilai data ke-k yang ada di dalam cluster tersebut untuk variable ke-j

2.1.4.2 Hierarchical Clustering

Hierarchical Clustering (pengelompokan hirarkis) adalah salah satu metode dalam analisis data yang digunakan untuk mengelompokkan objek atau sampel ke dalam kelompok yang beririsan berdasarkan kesamaan atau jarak antara mereka [39]. Metode ini dapat digunakan dalam berbagai bidang seperti ilmu komputer, statistik, bioinformatika, dan lain-lain untuk mengelompokkan data secara hierarkis. Metode ini tidak memerlukan label atau informasi kelas sebelumnya tentang sampel. Metode ini hanya menggunakan informasi jarak atau kesamaan antara sampel. Proses Hierarchical Clustering tidak dapat diubah (*irreversible*), artinya setelah objek atau cluster digabungkan, penggabungan tersebut tidak dapat dibatalkan pada tahap berikutnya. Hierarchical Clustering merupakan salah satu teknik yang populer dalam analisis data unsupervised untuk eksplorasi dan pemahaman struktur dalam data, terutama jika tidak diketahui sebelumnya berapa jumlah cluster yang diinginkan.

Secara umum, tujuan clustering baik hierarchical maupun partitional adalah untuk membuat kelompok yang memiliki fitur yang sama dalam satu anggota kelompok dan fitur yang berbeda di antara kelompok tersebut [40]. Konsep inilah yang mengharuskan proses pembuatan cluster untuk mempertimbangkan jarak/(dis)similarity/ukuran ketidakmiripan antar data. Untuk menghitung kemiripan antar data, metode penghitungan (dis)similarity yang dipilih harus dipilih [5]. Metode penghitungan *(dis)similarity* yang sering digunakan adalah *euclidean distance* dan *manhattan distance*. Berikut ini merupakan formula dalam perhitungan (dis)similarity dari kedua metode tersebut:

Euclidean Distance

$$Euclidean(p,q) = \sqrt{\sum_{i=1}^{n} (qi - pi)^2} \quad (2.8)$$

Deskripsi:

($qi-pi$)2 adalah kuadrat dari selisih antara koordinat titik p dan q pada dimensi ke-*i.*

Manhattan Distance

$$Manhattan(p,q) = \sum_{i=1}^{n} |qi - pi| \quad (2.9)$$

Deskripsi:

$|qi-pi|$ adalah perbedaan absolut antara koordinat titik p dan q pada dimensi ke-*i.*

2.1.4.3 DBSCAN

Berdasarkan ide kepadatan data, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) merupakan algoritma clustering yang mengelompokkan titik-titik data berdasarkan kepadatan data atau kepadatan spasial [41]. Algoritma ini berhasil mengidentifikasi kelompok padat dan dapat menangani titik data yang tidak termasuk dalam kelompok tertentu (sering disebut dengan noise). Pendekatan DBSCAN didasarkan pada konsep kepadatan data, dan algoritme mampu mendeteksi kelompok data yang relatif padat dan mengisolasi wilayah atau titik yang jarang atau tidak padat yang sering dianggap sebagai data tidak terstruktur. [42].

Algoritma DBSCAN mendapatkan popularitas karena dapat menangani data yang tidak beraturan dan mengatasi noise pada data spasial yang besar. DBSCAN menggunakan pendekatan berbasis

kepadatan untuk memberikan hasil clustering yang dapat mengatasi kekurangan K-Means dalam menangani struktur cluster yang kompleks dan tidak beraturan pada data dengan kepadatan berbeda. DBSCAN sendiri dapat mengklasifikasikan setiap titik menjadi titik inti, titik batas, atau titik gangguan [42]. Untuk menghitung jarak dari titik *core point* dengan *point* lain pada DBSCAN dapat menggunakan rumus berikut:

$$Jarak = \sqrt{(x - xp)^2 + (y - yp)^2} \quad (2.10)$$

Deskripsi:

x: koordinat sumbu x titik tujuan

y: koordinat sumbu y titik tujuan

2.1.4.4 Fuzzy C-Means

Metode Fuzzy C-Means membagi data berdasarkan tingkat keanggotaan [43]. Data dapat dikelompokkan berdasarkan tingkat keanggotaan, yang berkisar dari 0 hingga 1, dan beberapa tipe data hanya menunjukkan keanggotaan sebagian. Fuzzy C-Means menggunakan fuzzy clustering untuk menetapkan kepemilikan data pada setiap cluster, yang masing-masing memiliki keanggotaan yang berbeda. Derajat keanggotaan mengontrol rentang antara keberadaan data dalam cluster dan 0.

Metode fuzzy C-Means sangat baik untuk mendeteksi cluster tingkat tinggi dan mengungkapkan hubungan antara berbagai model cluster. Untuk mendapatkan struktur klaster dan nilai klaster yang optimal, fungsi fitnes atau fungsi objektif dapat diminimalkan dengan menggunakan persamaan berikut [43]. Berikut merupakan rumus yang dapat digunakan dalam perhitungan matriks keanggotaannya:

$$u_{ij} = \left(\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}} \right)^{-1} \quad (2.11)$$

Deskripsi:

m adalah nilai fuzziness (biasanya >1m>1).

C adalah jumlah cluster.

$\|x_i - c_j\|$ adalah jarak antara titik data x_i dan pusat cluster c_j.

$\|x_i - c_k\|$ adalah jarak antara titik data x_i dan pusat cluster c_k.

2.1.4.5 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) merupakan sebuah teknik dalam analisis statistik dan machine learning yang digunakan untuk mengurangi dimensi dari dataset yang kompleks menjadi dimensi yang lebih rendah, sehingga mempertahankan informasi yang signifikan dari dataset asli [23]. PCA bertujuan untuk menemukan pola atau struktur yang dominan dalam data dengan cara mentransformasi data asli ke dalam sistem koordinat baru (ruang fitur) yang disebut sebagai principal components (komponen utama).

PCA dapat Mengurangi dimensi dataset dapat membantu mengurangi overfitting pada model machine learning, Dengan mengurangi dimensi, data dapat divisualisasikan lebih mudah dalam bentuk yang lebih sederhana, dan PCA membantu mengidentifikasi pola atau hubungan penting antara variabel dalam dataset. Tahapan awal dalam PCA adalah standarisasi data yang dilakukan Jika X adalah matriks data $N \times D$ (dengan N jumlah sampel dan D jumlah fitur), maka data standar X' dihitung sebagai berikut:

$$X' = \frac{X - \mu}{\sigma} \qquad (2.12)$$

Deskripsi:

μ : vektor rata-rata dari setiap fitur.

σ : vektor deviasi standar dari setiap fitur.

2.1.5 Pemilihan Umum

Demokrasi berlaku di Indonesia. Proses penyaluran pendapat rakyat melalui

pemilihan umum yang diadakan secara berkala adalah pilar utama dalam setiap sistem demokrasi [17]. Pemilihan umum, juga dikenal sebagai pemilihan umum, merupakan salah satu implementasi demokrasi di Indonesia, di mana mereka memungkinkan masyarakat untuk memilih wakil rakyat dan pejabat publik lainnya [10].

Pemilihan umum biasanya diadakan secara periodik di negara-negara demokrasi ini. Indonesia akan mengadakan pemilu serentak untuk presiden dan wakil presiden pada tahun 2024. Sudah banyak tokoh politik yang dicalonkan menjadi presiden berdasarkan opini masyarakat. Hal ini disebabkan fakta bahwa opini masyarakat yang berkaitan dengan pemilu dapat digunakan sebagai alat untuk menggambarkan gambaran opini masyarakat terhadap para calon presiden.

2.1.6 Systematic Literature Review

Systematic Literature Review (SLR) adalah metode penelitian literatur yang sistematis dan objektif yang digunakan untuk mengumpulkan, mengevaluasi, dan membuat kesimpulan tentang semua bukti yang relevan dan sah tentang masalah yang telah ditentukan [12]. Tujuan utama dari SLR adalah untuk mengidentifikasi, mengevaluasi, dan menyintesis informasi dari berbagai sumber literatur secara obyektif dan metodis. Penelitian ini diharapkan dapat menjadi sumber rujukan untuk mendukung dalam pengambilan keputusan yang bersifat teknik maunpun strategis serta mendorong pengembangan penelitian baru.

2.1.7 Evaluasi

Pada tahap terakhir analisis sentimen, evaluasi dilakukan untuk mengukur kinerja model atau sistem analisis sentimen yang telah dikembangkan. Berikut adalah beberapa jenis evaluasi yang umum dilakukan pada tahap terakhir analisis sentimen:

1. *Accuracy* (Akurasi)

   Akurasi merupakan ukuran umum yang menggambarkan seberapa baik model memprediksi kelas dengan benar secara keseluruhan [28]. Akurasi dihitung dengan membagi jumlah prediksi yang benar (positif dan negatif) dengan total jumlah prediksi yang dilakukan oleh model. Akurasi bermanfaat saat kelas-kelas dalam dataset seimbang (*balance*). Namun, pada dataset yang tidak seimbang, akurasi dapat menjadi misleading karena model mungkin cenderung memprediksi mayoritas kelas, tanpa memperhatikan kinerja pada kelas minoritas.Ini dihitung sebagai jumlah prediksi yang benar dibagi dengan jumlah total prediksi seperti rumus dibawah ini:

$$\frac{TP + TN}{TP + TN + FP + FN} \quad (2.13)$$

2. *Precision* (Presisi)

   Presisi mengukur seberapa banyak prediksi positif yang benar dari semua prediksi positif yang dilakukan oleh model [28]. Presisi dihitung dengan membagi jumlah prediksi positif yang benar (*true positive*) dengan total jumlah prediksi positif yang dilakukan oleh model (*true positive* dan *false positive*). Presisi berguna untuk memahami seberapa akurat model dalam mengidentifikasi kelas positif, dan penting ketika ingin meminimalkan jumlah *false positive*. Ini dihitung sebagai jumlah prediksi positif yang benar dibagi dengan total prediksi positif seperti rumus dibawah ini:

$$\frac{TP}{TP + FP} \quad (2.14)$$

3. Recall

   Recall mengukur seberapa baik model dalam mendeteksi semua instance kelas positif yang sebenarnya dalam dataset [42]. Recall dihitung dengan membagi jumlah prediksi positif yang benar (*true*

positive) dengan total jumlah instance kelas positif yang sebenarnya. Recall penting ketika ingin meminimalkan *false negative*, yaitu ketika model salah mengklasifikasikan instance kelas positif sebagai negatif. Ini dihitung sebagai jumlah prediksi positif yang benar dibagi dengan total instance kelas positif yang sebenarnya.

$$\frac{TP}{TP + FN} \quad (2.15)$$

4. F1-score

   F1-score adalah metrik evaluasi yang menggabungkan presisi dan *recall* menjadi satu nilai tunggal [42]. F1-score memberikan keseimbangan antara presisi dan *recall*. Nilai F1-score yang tinggi menunjukkan bahwa model memiliki presisi dan *recall* yang baik secara bersamaan. F1-score dihitung sebagai dua kali perkalian presisi dengan *recall*, dibagi oleh jumlah presisi dan *recall*.

$$2\,x\,\frac{PrecisionxRecall}{Precision + Recall} \quad (2.16)$$

2.2 Penelitian Terdahulu

Berikut adalah beberapa penelitian yang melakukan analisis sentimen dengan algortima yang berbeda-beda sebagai pengklasifikasian pada review online yang digunakan peneliti sebagai perbandingan tinjauan studi terdahulu, sebagai berikut:

Tabel 2.1 Penelitian Terdahulu

1.	Nama Penulis	Macrohon Julio Jerison E; Villavicencio Charlyn Nayve; Inbaraj X. Alphonse; Jeng Jyh-Horng
	Nama Jurnal, Volume dan Nomor, Tahun	Information,, Vol. 13 No.10 (2022)
	Judul Artikel	*A Semi-Supervised Approach to Sentiment Analysis of Tweets during the 2022 Philippine Presidential Election*
	Metode	Naïve bayes
	Hasilnya	Melalui penelitian ini, didapatkan hasil 83,90% tweet yang dikirimkan bersifat negatif, dan tidak peduli apakah tweet tersebut mendukung atau menentang kandidat tertentu. Diikuti oleh hanya 13,49% dan 2,60% tweet positif dan netral.

		Model Multinomial Naïve Bayes, yang kemudian digunakan sebagai parameter model Self-Training untuk pendekatan pembelajaran semi-supervised dengan 30% data tidak berlabel, akurasi tingkat akurasi 84,83% lebih tinggi dibandingkan tingkat akurasi penelitian peneliti sebelumnya.
2.	Nama Penulis	Olusola Olabanjo, Ashiribo Wusu, Oseni Afisi, Mauton Asokere, Rebecca Padonu, Olufemi Olabanjo, Oluwafolake Ojo, Olusegun Folorunso, Benjamin Aribisala, Manuel Mazzara
	Nama Jurnal, Volume dan Nomor, Tahun	Heliyon, vol. 9, No.5 (2023)
	Judul Artikel	From Twitter to Aso-Rock: A sentiment analysis framework for understanding Nigeria 2023 presidential election
	Metode	Long Short-Term Memory (LSTM) Recurrent Neural Network, Bidirectional Encoder Representations from Transformers (BERT) and Linear Support Vector Classifier (LSVC)
	Hasilnya	Model sentimen memberikan akurasi, presisi, perolehan, AUC, dan f-measure masing-masing sebesar 88%, 82,7%, 87,2%, 87,6% dan 82,9% untuk LSTM; 94%, 88.5%, 92.5%, 94.7% dan 91.7% masing-masing untuk BERT dan 73%, 81.4%, 76.4%, 81.2% dan 79.2% masing-masing untuk LSVC.
3.	Nama Penulis	Ghulam Asrofi Buntoro, Rizal Arifin1, Gus Nanang Syaifuddiin1, Ali Selamat, Ondrej Krejcar, Hamido Fujita
	Nama Jurnal, Volume dan Nomor, Tahun	IIUM Engineering Journal, Vol. 22, No.1 (2021)
	Judul Artikel	Implementation of a Machine Learning Algorithm for Sentiment Analysis of Indonesia‘s 2019 Presidential Election
	Metode	Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM)
	Hasilnya	Hasil dari gabungan algoritma SVM dan tokenisasi abjad menghasilkan nilai ketepatan tertinggi yaitu 79.02%. Sedangkan nilai ketepatan terendah diperoleh dari kombinasi algoritma NBC dan tokenisasi N-gram dengan nilai ketepatan 44.94%
4.	Nama Penulis	April Lia Hananto, Aprilia Putri Nardilasari, Ahmad Fauzi, Agustia Hananto, Bayu Priyatna, Aviv Yuniar Rahman
	Nama Jurnal, Volume dan Nomor, Tahun	International Journal of Intelligent Systems and Applications in Engineering, Vol.11 No.6 (2023)

	Judul Artikel	Best Algorithm in Sentiment Analysis of Presidential Election in Indonesia on Twitter
	Metode	support vector machine (SVM),K-Nearest Neighbor (K-NN) and Naïve Bayes (NB)
	Hasilnya	Hasil tingkat akurasi algoritma SVM sebesar 79,57%, Naïve Bayes memiliki tingkat akurasi sebesar 77,21%, dan algoritma KNN sebesar 55,80%
5.	Nama Penulis	Hassan Nazeer Chaudhry,Yasir Jave, Farzana Kulsoom, Zahid Mehmoo, Zafar Iqbal Khan, Umar Shoaib, and Sadaf Hussain Janjua
	Nama Jurnal, Volume dan Nomor, Tahun	Electronics (Switzerland), Vol. 10 No.17 (2021)
	Judul Artikel	Sentiment analysis of before and after elections: Twitter data of U.S. election 2020
	Metode	TF-IDF, Naive Bayes Classifier
	Kesimpulan	Pengklasifikasi sentimen menghasilkan akurasi 94,58% dan presisi 93,19%
6.	Nama Penulis	Rodrigue Rizk, Dominick Rizk, Frederic Rizk & Sonya Hsu
	Nama Jurnal, Volume dan Nomor, Tahun	Computational and Mathematical Organization Theory, Vol.29 No.4 (2023)
	Judul Artikel	280 characters to the White House: predicting 2020 U.S. presidential elections from twitter data
	Metode	Naïve Bayes Classifier (NBC)
	Hasilnya	Efektivitas model dalam memprediksi hasil pemilu menghasilkan presentase sebesar 89,9%
7.	Nama Penulis	Mohammad Nur Habibi, Sunjana
	Nama Jurnal, Volume dan Nomor, Tahun	International Journal of Modern Education and Computer Science, Vol.11 No.11 (2019)
	Judul Artikel	Analysis of Indonesia Politics Polarization before 2019 President Election Using Sentiment Analysis and Social Network Analysis
	Metode	Naïve Bayes Classifier (NBC), ROC Validation, dan Social Network Analysis
	Hasilnya	Nilai rata-rata hasil klasifikasi adalah 91,27% sentimen positif, 7,56% negatif, dan 1,17% sentimen netral. Klasifikasi ini menghasilkan akurasi rata-rata sebesar 69,2% untuk sentimen jokowi dan untuk sentimen prabowo 100%
8.	Nama Penulis	Brandon Joyce, Jing Deng
	Nama Jurnal, Volume dan Nomor, Tahun	Jurnal Ilmu Komputer dan Informatika, Vol.1 No.1 (2024)
	Judul Artikel	Sentiment analysis of tweets for the 2016 US presidential election

	Metode	Naïve Bayes and Support Vector Machine (SVM) methods with Term Frequency-Inverse Document Frequency (TF-IDF)
	Hasilnya	Persentase sentimen positif tertinggi pada bulan Maret 2023 adalah Ganjar Pranowo sebesar 77,94% dan persentase sentimen negatif tertinggi adalah Anies Baswedan sebesar 31,39%. Sedangkan untuk data yang diperoleh pada bulan November 2023, sentimen positif tertinggi diperoleh pada pasangan Ganjar - Mahfud sebesar 69,16%, dan sentimen negatif tertinggi terdapat pada Prabowo - Gibran sebesar 52,12%
9.	Nama Penulis	Marcel Afandi, Khairunnisak Nur Isnaini
	Nama Jurnal, Volume dan Nomor, Tahun	Journal of Computer Science and Engineering (JCSE), Vol. 5 No. 1 (2024)
	Judul Artikel	Analyzing Public Trust in Presidential Election Surveys: A Study Using SVM and Logistic Regression on Social Media Comments
	Metode	Support Vector Machine (SVM) dan Logistic Regression
	Hasilnya	Regresi Logistik, dengan akurasi tertinggi 89,79% dari Instagram dan 88,01% dari Twitter pada skenario yang sama. Analisis sentimen menggunakan SVM menghasilkan 195 komentar positif dan 216 komentar negatif. Skenario Regresi Logistik menunjukkan 180 sentimen positif dan 216 sentimen negatif
10.	Nama Penulis	Ramdhan Hakiki, Agung Pambudi, Asriyanik
	Nama Jurnal, Volume dan Nomor, Tahun	Journal of Artificial Intelligence and Engineering Applications, Vol. 3 No. 2 (2024)
	Judul Artikel	Classification of Public Sentiment Toward 2024 Presidential Candidates on Social Media Platform X Using Naïve Bayes Algorithm
	Metode	Naïve Bayes (NB)
	Hasilnya	Hasil dari dari segi akurasi, model naïve bayes yang dikembangkan menunjukkan tingkat keberhasilan dengan akurasi sebesar 74% untuk Anies Baswedan, untuk Ganjar 74%, dan untuk Prabowo sebesar 88%
11.	Nama Penulis	Dinar Ajeng Kristiyanti, Normah, dan Akhmad Hairul Umam
	Nama Jurnal, Volume dan Nomor, Tahun	IEEE, Vol. 1 No. 1 (2019)
	Judul Artikel	Prediction of Indonesia Presidential Election Results for the 2019-2024 Period Using Twitter Sentiment Analysis
	Metode	Support Vector Machine (SVM) dengan selection features of Particle Swarm

		Optimization (PSO) dan Genetic Algorithms (GA)
	Hasilnya	Dengan accuracy 86.20% dan AUC value 0.934, combination PSO SVM method is the best.