BAB II

LANDASAN TEORI

2.1. Penelitian Terkait

Penelitian ini tidak terlepas dari penelitian sebelumnya yang berfungsi sebagai pembanding terhadap penelitian yang akan dilaksanakan dan juga sebagai sumber inspirasi yang dapat membantu pelaksanaan penelitian. Berdasarkan hal tersebut, penelitian terkait yang peneliti gunakan sebagaimana tabel 2.1. berikut ini.

Tabel 2. 1

Penelitian Yang Relevan

Peneliti	**Judul Penelitian**	**Tahun**	**Metode**	**Hasil Penelitian**	**Kelebihan**	**Kelemahan**
Dini Hardiyanti, Faizal Mahananto	Perbandingan Algoritma *Naïve Bayes* dan *Support Vector Machine* untuk Analisis Sentimen Kurikulum Merdeka pada Komentar YouTube	2023	*Naïve Bayes* dan *Support Vector Machine*	Algoritma *Naïve Bayes* menghasilkan akurasi yang lebih tinggi (87.2%) dibandingkan SVM (84.5%) sehingga Naïve Bayes lebih efisien dalam komputasi	*Naïve Bayes* sederhana dan mudah di implementasikan serta dapat menangani data tidak berurutan	*Naïve Bayes* mengasumsikan fitur independen dan Performa SVM bergantung pada pemilihan kernel yang tepat

Peneliti	**Judul Penelitian**	**Tahun**	**Metode**	**Hasil Penelitian**	**Kelebihan**	**Kelemahan**
Achmad Fauzi, Kurniawan Teguh Martono	Analisis Sentimen Kurikulum Merdeka Menggunakan Algoritma *Logistic Regression* dan *Random Forest*	2023	*Logistic Regression* dan *Random Forest*	Algoritma *Random Forest* menghasilkan akurasi yang lebih tinggi (90.2%) dibandingkan *Logistic Regression* (86.7%) sehingga Random Forest lebih robust terhadap overfitting	*Random Forest* dapat menangani berbagai tipe data di bandingkan *Logistic Regression* sederhana dan mudah diinterpretasikan	*Logistic Regression* rentan terhadap masalah multikolinearitas dan *Random Forest* membutuhkan komputasi yang lebih tinggi
Rizky Maulana, Eliza Heliyana	Perbandingan Algoritma *Logistic Regression* dan *Decision Tree* untuk Analisis Sentimen Kurikulum Merdeka pada Komentar YouTube	2023	Logistic Regression dan *Decision Tree*	Algoritma *Decision Tree* menghasilkan akurasi yang lebih tinggi (88.4%) dibandingkan *Logistic Regression* (85.1%) sehingga Decision Tree lebih mudah untuk di interpretasikan	*Decision Tree* dapat menangani fitur kategorikal dan numerik di bandingkan *Logistic Regression* efisien dalam komputasi	*Logistic Regression* rentan terhadap masalah multikolinearitas sedangkan *Decision Tree* dapat menghasilkan model yang kompleks
Andi Suryadi, Siti Nurhaliza	Perbandingan Algoritma Lexicon-based dan Machine Learning untuk Analisis Sentimen Kurikulum Merdeka pada Dataset Komentar YouTube	2023	*Lexicon based* dan *Support Vector Machine*	Algoritma *Lexicon based* menghasilkan akurasi yang lebih tinggi (92.3%) dibandingkan SVM (87.6%) *Lexicon-based*	*Lexicon-based* tidak memerlukan pelatihan model dan SVM dapat menangani data yang kompleks	*Lexicon-based* tergantung pada kualitas leksikon yang digunakan sedangkan SVM membutuhkan optimisasi parameter yang tepat

Fitriana Destyani, Bayu Setiawan	Analisis Perbandingan Algoritma *Naive Bayes* dan *Random Forest* untuk Klasifikasi Sentimen Kurikulum Merdeka pada Komentar YouTube	2023	*Naïve Bayes* dan *Random Forest*	Algoritma *Random Forest* menghasilkan akurasi yang lebih tinggi (91.3%) dibandingkan *Naïve Bayes* (88.7%) *Random Forest* lebih robust terhadap overfitting	*Naïve Bayes* sederhana dan mudah diimplemen tasikan sedangkan *Random Forest* dapat menangani berbagai tipe fitur	*Naïve Bayes* mengasumsikan fitur independen dan *Random Forest* membutuhkan komputasi yang lebih tinggi
Rizky Armando, Siti Nurhaliza	Perbandingan Algoritma *Logistic Regression* dan *Decision Tree* untuk Analisis Sentimen Kurikulum Merdeka pada Dataset Komentar YouTube	2023	*Logistic Regression* dan *Decision Tree*	Algoritma *Decision Tree* menghasilkan akurasi yang lebih tinggi (89.2%) dibandingkan Logistic Regression (86.4%)	*Decision Tree* dapat menangani fitur kategorikal dan numerik sedangkan *Logistic Regression* efisien dalam komputasi	*Logistic Regression* rentan terhadap masalah multikolinearitas dan *Decision Tree* dapat menghasilkan model yang kompleks
Andi Surya, Dini Hardiyanti	Analisis Perbandingan Algoritma Support Vector Machine dan Naïve Bayes untuk Klasifikasi Sentimen Kurikulum Merdeka pada Komentar YouTube	2023	*Support Vector Machine* dan *Naïve Bayes*	Algoritma *Naïve Bayes* menghasilkan akurasi yang lebih tinggi (88.9%) dibandingkan SVM (86.2%) sehingga *Naïve Bayes* lebih efisien dalam komputasi	*Naïve Bayes* sederhana dan mudah diimplemen tasikan SVM dapat menangani data yang kompleks	*Naïve Bayes* mengasumsikan fitur independen dan Performa SVM bergantung pada pemilihan kernel yang tepat

Berdasarkan hasil penelitian sebagaimana diuraikan pada tabel di atas, dapat disintesakan bahwa terlepas dari banyaknya penelitian empiris tentang teknik *data*

mining dan media sosial, hanya sedikit penelitian yang membandingkan teknik penambangan data dalam hal akurasi, kinerja, dan kesesuaian. Misalnya, telah diamati bahwa keakuratan teknik pembelajaran mesin tertentu dihitung dengan cara yang berbeda, sehingga sulit untuk menemukan jawaban tentang kesesuaian teknik penambangan data, terutama terkait topik penelitian analisis sentimen terhadap Implementasi Kurikulum Merdeka Pada Satuan Pendidikan di Indonesia

2.2. Landasan Teori

2.2.1. *Machine Learning*

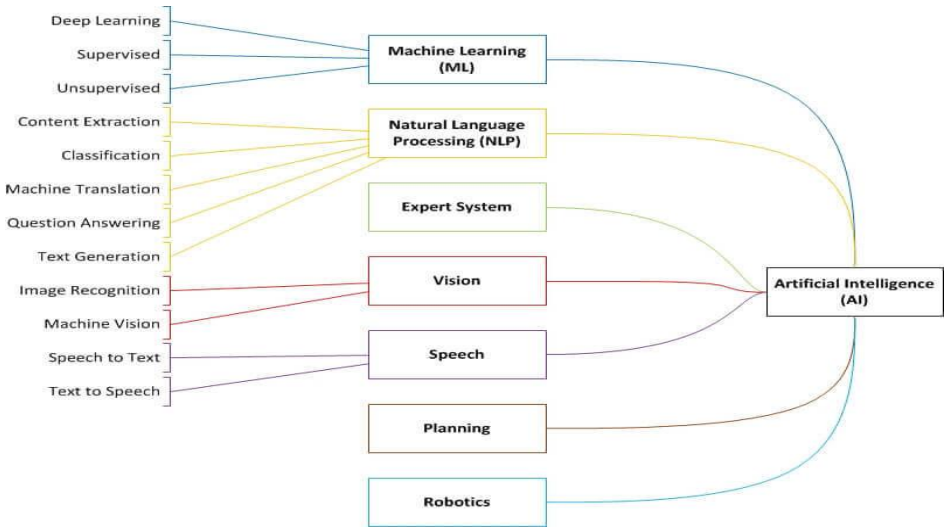
Perkembangan *artificial intelligence* (AI) atau teknologi kecerdasan buatan semakin berkembang pesat. Kecerdasan buatan terdiri dari beberapa bidang, salah satunya adalah pembelajaran mesin (*machine learning*). Pembelajaran mesin ialah cabang dari ilmu komputer dan *artificial intelligence* (AI) yang memiliki fungsi sebagai algoritma untuk meniru cara manusia belajar secara bertahap meningkatkan akurasinya dan penggunaan data [13]. Algoritma pembelajaran mesin memungkinkan sistem dapat secara otomatis belajar dan meningkatkan diri tanpa pemrograman eksplisit dan berfokus pada pengembangan program komputer sehingga dapat mengakses data dan digunakan sebagai belajar mandiri.

Proses pembelajaran mesin dimulai dari pengamatan data misalnya instruksi dan pengalaman langsung. Tujuannya adalah agar menemukan dan mempelajari pola data untuk menginformasikan keputusan masa depan yang lebih

baik. Komputer/mesin secara otomatis mempelajari dan mengadaptasi tindakan yang tepat berdasarkan itu tanpa campur tangan/bantuan manusia [14]. Dengan kata lain, pembelajaran mesin memiliki kemampuan untuk mengumpulkan data sendiri, bukan perintah manusia. Data yang dihasilkan diperiksa oleh pembelajaran mesin untuk melakukan tugas tertentu. Tugas yang dapat dilakukan pembelajaran mesin sangat bervariasi tergantung pada apa yang dipelajarinya.

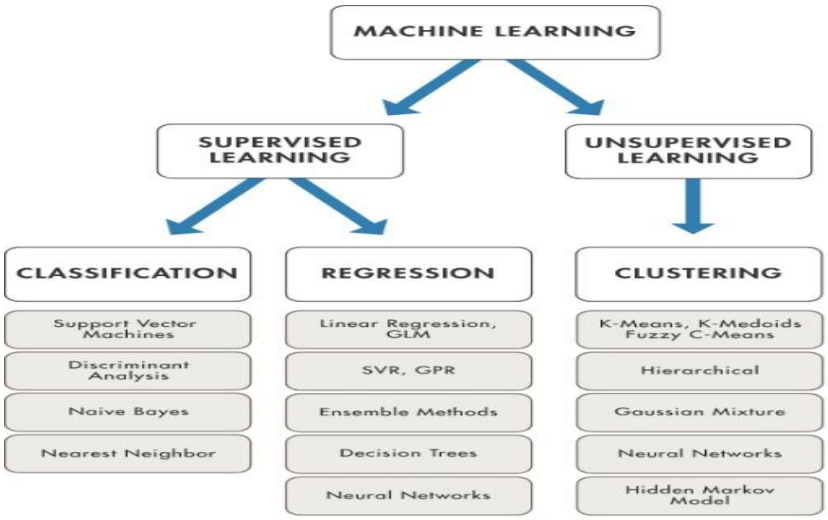
Pembelajaran mesin juga berguna dalam menggabungkan skala waktu dan panjang yang kecil dan besar. Misalnya, simulasi dinamika molekul sering dilakukan dengan model pengganti seperti medan gaya. Model pembelajaran mesin dapat memberikan akurasi tinggi sambil mempertahankan kecepatan tinggi [15].

Kecerdasan buatan terapan dapat secara luas dibagi menjadi tujuh bidang: pembelajaran mesin robotika (*robotics*), bahasa (*speech)*, sistem pakar (*expert system)*, pemrosesan bahasa alami (*natural language processing)*, perencanaan (*planning)*, visi (*vision)*, dan (*machine learning)*, Tujuan dari percabangan ini adalah untuk mengurangi cakupan aplikasi ketika mengembangkan atau mempelajari AI, karena artificial intelligence pada prinsipnya memiliki cakupan yang sangat luas. [16]. Lebih jelasnya sebagaimana gambar 2.3 berikut.



Gambar 2.1. Bidang Ilmu AI [16]

Simulasi model *machine learning* sangat terkait dengan *Computational Statistics* yang tujuan utamanya adalah untuk fokus membuat prediksi melalui komputer. Terdapat banyak algoritma pembelajaran mesin yang diawasi (*supervised learning*) yang melatih model pada data masukan dan keluaran yang diketahui sehingga dapat memprediksi keluaran di masa mendatang, dan tidak diawasi (*unsupervised learning*) yang menemukan pola tersembunyi atau struktur intrinsik dalam data masukan, dan masing-masing menggunakan pendekatan pembelajaran yang berbeda [17]. Selain itu, tidak ada metode terbaik atau satu ukuran cocok untuk semua. Pemilihan algoritma juga bergantung pada ukuran dan jenis data yang dikerjakan. Teknik *machine learning* ini dapat dijelaskan dalam gambar 2.2. berikut ini.



Gambar 2.2 Tekhnik machine learning [18]

2.2.2. *Data Mining*

Pada bidang statistik, *data mining* dan *machine learning* semuanya berperan dalam memahami data, mendeskripsikan karakteristik kumpulan data, dan menemukan hubungan dan pola dalam data tersebut untuk membangun model [13]. *Data mining* adalah teknik penambangan yaitu proses mengidentifikasi pola dan tren data untuk mendapatkan informasi yang berguna dari kumpulan data yang sangat besar untuk membuat penilaian atau keputusan. Banyak teknik *data mining* telah dikembangkan dan digunakan dalam *data mining*. Termasuk asosiasi, klasifikasi, pengelompokan, pohon keputusan, prediksi, jaringan saraf, dan banyak lagi. Setiap teknik memiliki aturan dan metode yang menentukan jenis masalah yang dipecahkannya [7].

Data mining dapat diterjemahkan sebagai proses menemukan struktur yang menarik dalam data. Struktur dapat mengambil banyak bentuk, seperti sekumpulan aturan, grafik atau jaringan, pohon, atau satu atau lebih persamaan. Struktur ini bisa menjadi bagian dari dasbor visual yang kompleks, atau bisa sesederhana daftar kandidat pemilu dan nomor terkait yang mewakili sentimen pemilih berdasarkan *feed Twitter* [19]. Proses menemukan struktur data yang ada membutuhkan sebuah teknik yang disebut teknik *data mining*. Teknik *data mining* adalah proses identifikasi pola dan tren data untuk mendapatkan informasi yang berguna dalam kumpulan data yang sangat besar sehingga pengguna dapat menilai atau memutuskan [11].

Data mining tidak seperti metode penelitian tradisional. *Data mining* terdiri dari penggalian informasi dan penemuan pengetahuan tanpa asumsi eksplisit, yaitu tanpa penelitian dan desain sebelumnya, informasi yang diperoleh harus memiliki tiga karakteristik: tidak diketahui sebelumnya, efektif, dan dapat ditindaklanjuti [11] [19]. Setiap bidang dibagi menjadi tiga kategori yang merujuk pada jenis pembelajaran mesin atau tugas *data mining* dan mengejar tujuan: deskriptif (misalnya mengidentifikasi pola yang tidak diketahui), prediktif (misalnya perkiraan berdasarkan pengetahuan yang tersedia), dan preskriptif (misalnya optimalisasi berdasarkan *machine learning* yang dikendalikan pengambilan keputusan).

Sesi penambangan data menggunakan satu atau lebih algoritma untuk mengidentifikasi tren dan pola yang menarik dalam data. Pengetahuan yang

diperoleh dari sesi penambangan data adalah model data umum. Tujuan utamanya adalah menerapkan apa yang ditemukan pada situasi baru.

Berdasarkan pengertian di atas, dapat disarikan bahwa *data mining* adalah proses otomatis dari jumlah data yang sangat besar, bertujuan untuk mendapatkan koneksi dan pola yang membawa informasi baru yang bernilai. *Data mining* juga merupakan proses pendukung keputusan yang mencari pola informasi dalam data.

2.2.3. Pengelompokkan *Data Mining*

Banyak teknik penggalian data telah dikembangkan dan diterapkan dalam proyek penambangan data. Setiap teknik memiliki tugas dan metodenya sendiri berdasarkan jenis masalah yang dipecahkannya. Tugas-tugas yang dilakukan oleh data mining terbagi menjadi beberapa kelompok [11] [13] yaitu:

a. Asosiasi (*Association*)

   Asosiasi ialah sebuah teknik penggalian data yang paling populer, menemukan pola berdasarkan hubungan antara variabel dalam transaksi yang sama. Asosiasi juga dikenal sebagai teknik relasi karena menggunakan hubungan antara item dan menemukan seringnya kemunculan item berbeda yang muncul dengan frekuensi tertinggi dalam kumpulan data [11]. Oleh karena itu, tugas asosiasi dalam data mining ialah untuk mendeteksi atribut-atribut yang muncul secara bersamaan. Contoh: mencari jumlah murid di sekolah tertentu yang cenderung bereaksi positif pada kualitas pelayanan pendidikan yang tersedia.

b. Klasifikasi (*classification*)

Teknik klasifikasi digunakan untuk mengklasifikasikan suatu koleksi data ke dalam kelompok atau kelas yang berbeda untuk mendapatkan prediksi dan analisis yang akurat dalam kumpulan data yang sangat besar [11]. Klasifikasi memiliki target variabel kategorikal. Misalnya, kelompok pendapatan dapat dibagi menjadi tiga kategori: pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.

c. *Clustering*

Klasifikasi adalah pengelompokan catatan, disebut juga sebagaipengamatan atau observasi bertujuan guna pembentukan kelas dari objek-objek yang sama.*Cluster* adalah kumpulan *record* yang mirip satu sama lain dan berbeda dari *record* di *cluster* lain. *Clustering* adalah salah satu teknik pertama yang digunakan dalam *data mining*. Proses *clustering* melibatkan analisis satu atau lebih atribut untuk mengidentifikasi data yang mirip satu sama lain untuk memahami perbedaan dan persamaan antara kumpulan data [11].

d. Pohon keputusan (*Decision tree*)

Teknik pohon keputusan dapat diterapkan sebagai bagian kriteria seleksi. Selain itu, untuk membantu penggunaan dan pemilihan data tertentu dalam keseluruhan struktur [11] .

e. Prediksi (*Prediction*)

Prediksi adalah topik yang komprehensif dan berangkat dari memprediksi kegagalan komponen untuk memahami penipuan dan bahkan memprediksi keuntungan perusahaan. Prediksi termasuk analisis tren, klasifikasi,

pencocokan pola, dan hubungan. Prediksi dibuat dengan menganalisis peristiwa atau contoh masa lalu [11]. Prediksi mirip dengan klasifikasi dan estimasi, kecuali bahwa membuat prediksi nilai hasil masa depan. Contoh: Perkiraan curah hujan untuk wilayah tertentu.

f. Jaringan saraf Tiruan (*Neural Network*)

   Neural Network adalah teknik penting yang banyak digunakan oleh orang zaman sekarang. Teknik yang sering digunakan pada tahap awal teknologi *data mining*. Jaringan syaraf tiruan terbentuk dari komunitas kecerdasan buatan [11].

2.2.4. Tahapan Proses *Data Mining*

Data Mining adalah bagian dari seri Knowledge Discovery in Databases (KDD), yang berkaitan pada teknik integrasi dan temuan ilmiah, visualisasi dan interpretasi terhadap sebuah data.Ungkapan KDD diciptakan pada tahun 1989 untuk menekankan bahwa pengetahuan dapat diturunkan dari data-*driven discovery* dan sering digunakan secara bergantian dengan *data mining* [19]. Proses penemuan pengetahuan meliputi langkah-langkah berikut [20]:

a. Penemuan Data (*data discovery*): tahap ini adalah tahap pengumpulan data dan mencakup deteksi, identifikasi, dan karakterisasi data yang tersedia.
b. Pembersihan dan Pembersihan Data (*data cleaning and cleaning*): Kebisingan dan data yang tidak penting dihapus pada tahap ini, dan data yang bertentangan dan data yang tidak konsisten dihapus.

c. Integrasi data (*data integration*): pada tahap ini, data serupa dan terkait dikumpulkan dari beberapa sumber data dan digabungkan menjadi satu.
d. Pemilihan data (*data selection*): pada tahap ini, data yang sesuai diidentifikasi dan diambil dari kumpulan data.
e. Transformasi data (*data transformation*): pada tahap ini, data diubah menjadi bentuk khusus yang sesuai untuk prosedur pencarian dan pengambilan melalui umpan pencapaian atau operasi pengelompokan.
f. Penambangan data (*data mining*): pada tahap ini, penggunaan metode cerdas yang diterapkan untuk mengekstraksi pola data dan ekstraksi model yang berguna mungkin.
g. Evaluasi pola (*pattern evaluation*): pada tahap inilah pola yang sangat penting yang mewakili basis pengetahuan untuk penggunaan beberapa metrik penting diidentifikasi.
h. Presentasi pengetahuan (*knowledge presentation*): tahap ini adalah tahap terakhir dari penemuan pengetahuan dalam basis data, dan ini adalah tahap yang dilihat pengguna. Tahap dasar ini menggunakan metode visual untuk membantu pengguna memahami dan menginterpretasikan hasil ekstraksi data.

2.2.5. *Text Mining*

Pengolahan data yang banyak biasanya disebut *data mining* dan dalam ranah pengolahan data mentah teks disebut penambangan teks *(Text mining)* [21]. *Text mining* juga dikenal sebagai penambangan data teks, dirancang untuk memperoleh

pengetahuan implisit yang tersembunyi di dalam yang tidak terstruktur teks [22]. *Text mining* adalah metode penggalian informasi dari data yang tidak terstruktur. Data yang tidak terstruktur diproses menggunakan teknik tertentu untuk menghasilkan informasi yang berguna bagi pengguna [23].

Text mining adalah eksplorasi menggunakan komputer untuk menghasilkan informasi baru yang sebelumnya belum pernah diketahui, dari informasi yang diekstrak secara otomatis dari berbagai sumber data tekstual. Proses *text mining* memiliki kemiripan dengan penelitian *data mining*, dengan perbedaan bahwa pada *text mining*, model yang digunakan berasal dari sekumpulan bahasa alami yang tidak terstruktur, sedangkan pada *data mining*, model berasal dari bahasa terstruktur. Proses penambangan teks mirip dengan data klasik pengolahan. Pengambilan informasi, dimaksudkan untuk memperoleh teks yang diinginkan, mirip dengan pengumpulan data. Ekstraksi informasi digunakan untuk mengekstrak informasi yang telah ditentukan sebelumnya, yaitu pemrosesan awal dari data yang dikumpulkan [22].

Tujuan dari tahap pemilihan fitur adalah untuk mengurangi dimensi dari koleksi teks. Dengan kata lain, bertujuan agar proses klasifikasi menjadi lebih efektif dan akurat dengan menghilangkan kata-kata yang dianggap tidak penting atau tidak menggambarkan isi dokumen yang diambil dari data twitter. Tindakan yang dilakukan peneliti pada tahap ini adalah menghilangkan kata hendti/*stopword* (*remove stopwords*) dan menghilangkan kata (*stemming*) yang berimbuhan.

2.2.6. Pra-Pemrosesan Data (*Preprocessing*)

Preprocessing merupakan proses normalisasi istilah yang berasal dari kalimat. Tahapan ini bertujuan mendapatkan data penelitian yang baik dan fitur yang diekstraksi tersinkronisasi dengan fitur yang diinginkan sehingga mempermudah pengolahan data. Pengumpulan data opini dari media sosial Twitter tidak boleh identik dengan kata baku, kata kamus atau bahasa daerah yang digunakan atau dihilangkan [23]. Teks-teks dikembalikan ke teks alami dengan melakukan eliminasi ekspresi tipikal agar dapat meminimalisir noise pada tahap selanjutnya sehingga perlu dilakukan *pre-processing* atau normalisasi untuk mengatasi hal tersebut.

Teks yang diproses dalam proses *text mining* biasanya memiliki karakteristik seperti ukuran yang besar, *noise* pada data dan struktur teks yang kurang baik. Caranya dengan terlebih dahulu menentukan fitur yang mewakili setiap kata untuk setiap fitur dalam dokumen. Sebelum menentukan fitur yang representatif, diperlukan langkah *pre-processing*, yang biasanya dilakukan selama *text mining* dokumen [21] [24].

a. *Case folding*

Teks yang diperoleh dari media sosial Twitter memiliki huruf besar atau kecil yang beragam. Supaya tidak mengganggu proses analisis data perlu diubah menjadi huruf kapital atau kecil semua. Proses ini disebut *case folding*, yaitu metode untuk mengubah semua huruf dalam *dataset* menjadi kapital atau kecil

semua. Proses ini bertujuan untuk memudahkan proses analisis *dataset* dan mengurangi jumlah penggunaan memori [21].

b. *Data cleaning*

Data cleaning merupakan proses membersihkan data yang dikombinasikan dengan *regex* untuk mendeteksi karakter yang tidak berguna dan langsung dihapus dari data utama untuk meningkatkan kualitas *dataset* [21]. Setiap data tweet dari media sosial Twitter biasanya mengandung banyak kata dan karakter yang tidak berguna untuk proses analisis data.

c. *Lemmatization*

Pada tahap ini, kata-kata tidak baku sering digunakan dalam berkomunikasi dan berinteraksi dengan orang lain diolah agar tidak mempengaruhi hasil perhitungan karena terkadang jauh dari kaidah baku bahasa aslinya. Pada analisis sentimen kata tidak baku mempengaruhi perhitungan analisis data [21].

d. *Automate*

Penandaan data menggunakan Vader, sebuah metode yang pelatihannya didasarkan pada pendekatan yang berpusat pada manusia, yang menggabungkan analisis kualitatif dan validasi empiris dengan menggunakan kebijaksanaan dan penilaian manusia [21]. Penilaian polaritas menggabungkan fitur kamus leksikal dengan skor sentimen dari 5 kriteria tambahan, yaitu tanda seru, huruf besar, tingkat urutan kata, perubahan polaritas karena kata 'tetapi' dan menggunakan fitur trigram untuk memeriksa keberadaan negasi.

e. *Remove Stopwords*

Hapus *stopwords* sangat berguna seperti preposisi, konjungsi, kata sifat, kata slank, kata ganti dan sebagainya. Kata-kata seperti ini biasanya muncul bersamaan dengan kata utama sehingga tidak unik, tidak memiliki arti tertentu, dan tidak berkontribusi. Daftar kata yang tidak berkontribusi terlalu banyak pada teks analitik disebut *stopword* atau *stoplist* [21].

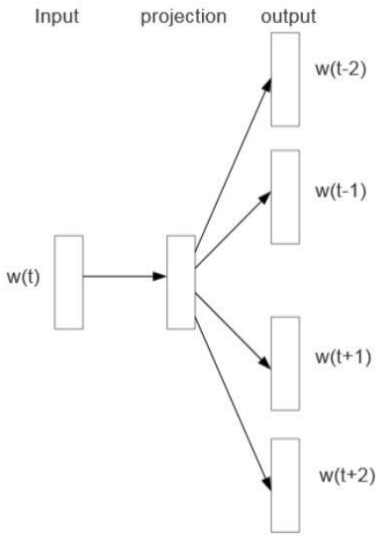
f. *Stemming*

Stemming adalah proses menghilangkan atribut tambahan dari kata seperti menghapus “mem” dan “-kan” dari “making” menjadi “making” [21]. *Stemming* adalah membuat kata mempengaruhi kata dasar dan digunakan untuk sistem pencarian informasi seperti mesin pencari dan analisis teks lainnya.

2.2.7. Pembobotan kata

Word2vec adalah representasi vektor dari kata-kata yang dibuat oleh Google dan baru-baru ini dikembangkan oleh Thomas Mikolov. Word2vec juga merupakan kumpulan dari beberapa model yang digunakan untuk membuat jangkar kata. Word Embeddings adalah nama dari sekumpulan model bahasa dan menyajikan teknik pembelajaran pemrosesan bahasa alami, di mana setiap kata dalam kosakata memiliki vektor yang merepresentasikan arti dari kata tersebut, dan kata-kata tersebut dipetakan ke sebuah vektor bilangan real. Word2vec menggunakan kumpulan teks (corpus) yang besar sebagai data pelatihan untuk

membangun kosakata dan menghasilkan ruang vektor yang dapat berisi beberapa ratus dimensi, dan setiap kata unik dalam corpus adalah sebuah vektor, di mana pelatihan vektor valid. Model Skip-Gram dan model CBOW (Continuous Bag-of-Words). Metode Word2vec terdiri dari dua algoritma integrasi kata utama, yaitu Continuous Bag of Words (CBOW) dan Skip Gramm. Algoritma CBOW digunakan untuk menentukan panjang spesifik dari sebuah kata dalam dokumen input. Algoritma Skip-Gram digunakan untuk memprediksi konteks sebuah kata dengan melihat kedekatan kata tersebut dengan kata lain yang mendahului atau mengikutinya. Model yang digunakan dalam penelitian ini adalah arsitektur Skip-Gram. gram Model ini bekerja dengan mengambil sebuah korpus teks sebagai input dan membuat representasi vektor dari setiap kata di dalam korpus tersebut sebagai output. File vektor yang dihasilkan dapat digunakan dalam studi pemrosesan bahasa alami dan aplikasi pembelajaran mesin. Vektor kata ini juga dapat digunakan untuk mengukur jarak antara vektor kata lainnya. Word2vec memiliki dua arsitektur pemodelan yang dapat digunakan untuk merepresentasikan vektor kata. Arsitektur tersebut adalah Continuous Bag of Words (CBOW) dan Skip-Gram.[25]



Gambar 2. 3 Arsitektur Skip-Gram. (Sumber: (Choi & Lee, 2020))

Pada Gambar 2.1, berdasarkan Rahutomo dkk, apabila sebuah dokumen disimbolkan dengan D dan setiap kata pada dokumen tersebut dinotasikan dengan wi, dimana i merupakan indeks kata pada dokumen, dan T merupakan jumlah kata pada dokumen tersebut, maka sebuah dokumen dapat dinyatakan sebagai D={w1,w2,w3,...,wT}.

2.2.8. Analisis Sentimen

Analisis sentimen ialah salah satu proses menganalisis kumpulan teks korpus yang bertujuan untuk menganalisis polaritas emosi baik emosi negatif, emosi netral maupun emosi positif [21]. Analisis sentimen adalah teknik yang secara otomatis mengenali, mengekstrak, dan mengeksekusi informasi tekstual untuk menemukan

informasi emosional dari ekspresi pikiran [23]. Analisis sentimen adalah jenis pemrosesan bahasa alami untuk melacak sentimen orang terhadap produk atau topik tertentu [26].

Analisis sentimen, juga dikenal sebagai *opinion mining,* adalah bidang yang menganalisis pendapat, perasaan, penilaian, evaluasi, sikap, dan sentimen orang tentang entitas seperti produk, layanan, organisasi, individu, peristiwa, isu, dan atribut mereka [27]. Dalam penelitian ini, analisis sentimen digunakan untuk menemukan informasi yang berguna dari data yang tidak terstruktur. Tujuannya supaya dapat diketahui sentimen pengguna twitter terhadap Implementasi Kurikulum Merdeka Pada Satuan Pendidikan di Indonesia. Berdasarkan uraian tersebut dapat disintesakan bahwa analisis sentimen adalah proses menentukan opini atau sentimen seseorang yang diungkapkan dalam bentuk tekstual dan dapat diklasifikasikan sebagai positif atau negatif.

2.2.9. Teknik Klasifikasi

Teknik klasifikasi dijelaskan sebagai sebuah model dalam *data mining* dimana *classifier* diatur untuk memprediksi *categorical label*. Adapun algoritma klasifikasi yang akan digunakan dalam penelitian ini sebagai berikut.

a. *K-Nearest Neighbors (KNN)*

   K-Nearest Neighbor yang disingkat KNN merupakan algoritma klasifikasi yang prinsip dasarnya adalah perhitungan jarak terdekat [21] [12]. KNN adalah salah satu algoritma paling sederhana yang digunakan untuk menyelesaikan masalah klasifikasi. Pandangan lain menjelaskan bahwa KNN

merupakan algoritma yang mengklasifikasikan objek baru berdasarkan atribut dan sampel pelatihan [28].

Selain itu, J. Alzubi menegaskan bahwa KNN ialah metode non-parametrik yang digunakan untuk klasifikasi dan regresi. Diberikan N vektor pelatihan, algoritma KNN mengidentifikasi k-tetangga terdekat dari vektor fitur yang tidak diketahui yang kelasnya akan diidentifikasi [17]. Algoritma KNN adalah penentu klasifikasi berdasarkan contoh yang tidak membangun representasi kategori secara eksplisit dan deklaratif, tetapi bergantung pada label kategori yang dilampirkan pada dokumen pelatihan yang mirip dengan dokumen uji [26].

Berdasarkan uraian di atas, dapat disimpulkan bahwa KNN adalah salah satu algoritma yang paling populer untuk pengenalan pola dengan prinsip dasarnya adalah perhitungan jarak terdekat yang digunakan untuk klasifikasi dan regresi.

b. *Naïve Bayes*

Naive Bayes adalah algoritma machine learning. Dalam pengembangan basis data, Naïve Bayes melibatkan supervised learning, yaitu pembelajaran mesin yang memerlukan sampel berlabel sebagai data pelatihan. Supervised learning dikelompokkan menjadi dua bagian yaitu klasifikasi 19 dan regresi. Klasifikasi, ketika variabel menjadi kategori, seperti merah atau kuning, sakit atau tidak sakit, dll. Regresi bila variabelnya berupa nilai riil seperti bobot, nilai moneter, dll. Naive Bayes memasukkan klasifikasi pembelajaran yang diawasi seperti contoh lain yaitu Support Vector Machine (SVM), K-Nearest

Neighbor (KNN), Artificial Neural Network (ANN), Tree Gradient Boosted (TGB) dan Random Trees (RT) sedangkan Regresi seperti Decision Tree , Logistic Regresion dan Kernel Regresion [29]. Metode ini merupakan bagian dari metode bayes yang digunakan dalam klasifikasi teks didasarkan pada model penyederhanaan bahwa nilai atribut secara kondisional bersifat bebas jika nilai output diberikan. Metode Naïve Bayes banyak digunakan dalam teknik klasifikasi pada twitter. Metode ini juga digunakan dalam penambangan teks dalam analisis sentimen, memprediksi probabilitas berdasarkan data masa lalu. Naive Bayes tidak bisa mengenali gambar, hanya teks dan angka. Metode ini menggunakan metode teorema Bayes untuk menghitung probabilitas[30]. Teorema bayes ditemukan oleh Thomas Bayes yaitu seorang pendeta dari Inggris pada tahun 1763 dan disempurnakan oleh Laplace. Teorema bayes adalah pengenalan pola melalui pendekatan statistik yang fundamental. Teorema bayes dapat dideskripsikan seperti probabilitas terjadinya hubungan A dengan syarat hubungan B sudah terjadi, begitupun sebaliknya. Dalam bidang kedokteran modern, teorema bayes sering digunakan. Teorema bayes berperan dalam memperbaiki hitungan probabilitas dengan memanfaatkan data informasi tambahan [31]. Berikut persamaan teori bayes[30]:

$$\mathrm{P}\,(Y|X) = \frac{\mathrm{P}\,(x|y)(Y)}{P(X)}$$

Persamaan di atas menunjukan bahwa Y sebagai kelas spesifik, sedangkan X sebagai kelas yang belum diketahui. ($Y|X$) merupakan probabilitas dari kelas berdasarkan hipotesa sebelumnya, sedangkan P(X) merupakan probabilitas dari Y. 20 P (Y|X) adalah hasil perkalian antara likelihood dan prior dibagi edivence. Likelihood adalah probabilitas atribut data X pada kelas Y, prior adalah probabilitas kelas Y dari total data set, dan evidence adalah probabilitas atribut data X dari total dataset.

2.2.10. Evaluasi dan Validasi Model

Model yang digunakan dievaluasi menggunakan metode *confusion matrix* sebagai indikasi aturan sifat klasifikasi (diskriminan). *Confusion matrix* adalah metode evaluasi yang menggunakan tabel *matrix*. *Confusion matrix* berisikan sejumlah elemen yang telah diklasifikasikan dengan benar atau salah untuk setiap kelas.

Manfaat menggunakan *confusion matrix* salah satunya mudah untuk melihat sistem *confusion* dua kelas. Untuk setiap contoh di *test set*, akan membandingkan kelas yang sebenarnya dengan kelas *classifier*. Contoh positif (negatif) yang diklasifikasikan dengan benar oleh *classifier* disebut *True Positive (true negative)*, contoh positif (negatif) yang salah diklasifikasikan disebut *False Negative (false positive)* [32]. Tabel *matrix* yang digunakan dalam *mining data* disajikan dalam tabel 2.2. berikut ini.

Tabel 2.1 Confusion Matrix

Correct classification	*Classified as*	
	+	-
+	*True positive*	*False negative*
-	*False positive*	*True negative*

Sumber : Bramer dalam Ibrahim (2017) dalam [32]

Witten menjelaskan bahwa model validasi umum menggunakan 10 *fold cross validation* untuk data *learning* dan pengujian. Dengan kata lain, data *training* dibagi menjadi 10 bagian yang sama dan dilakukan proses *learning* sebanyak 10 kali. Bagian dari *dataset* untuk menguji 9 bagian yang tersisa dan digunakan dalam proses *learning*. Kemudian menghitung *mean* dan deviasi dari 10 hasil uji beda. Validasi silang 10 *fold cross validation* telah menjadi metode standar dan metode praktis validasi *state-of-the-art* [32].

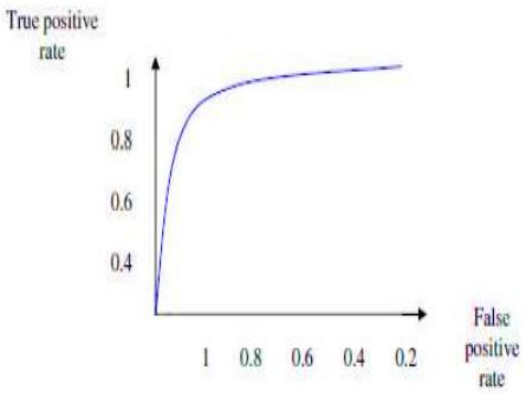
Performa *classifier* dievaluasi menggunakan *Area under curve (AUC).* AUC adalah area di bawah kurva *Receiver Operation Charactheristic* (ROC). AUC memisahkan kinerja prediktif dari kondisi operasi dan karena merupakan ukuran prediktif umum dapat secara signifikan meningkatkan konvergensi eksperimen empiris. AUC memiliki statistik yang jelas interpretasinya sebagaimana tabel 2.3 berikut ini.

Tabel 2.2 Nilai AUC

AUC	***Meaning***	***Symbol***
0.90 – 1.00	*Excellent classification*	
08.80 – 0.90	*Good classification*	
0.70 – 0.80	*Fair classification*	
0.60 – 0.70	*Poor classification*	
<0.60	*Failure*	

Sumber: Gorunescu dalam Wahono (2014) dalam [7]

Provost and Fawcett menjelaskan bahwa pengukuran lain dapat dilakukan menggunakan kurva *Receiver Operation Charactheristic* (ROC). Kurva ini menggambarkan *trade off* antara *true positive* terhadap *false positive*.



Gambar 2. 1. Kurva ROC

Sumber: Rokach & Maimon (2015) dalam [7]

Pada kurva ROC di atas dijelaskan bahwa sumbu X mewakili tingkat *false positive* dan Y-axis merupakan tingkat *true positive*. (0,100) adalah titik ideal

pada ROC kurva, hal ini berarti bahwa semua contoh positif dapat diklasifikasikan dengan benar dan tidak ada contoh negatif yang salah klasifikasi sebagai positif.

2.3. Tinjauan Objek Penelitian

2.3.1. Kurikulum Merdeka

Kurikulum Merdeka adalah program pembelajaran intra-kurikuler yang isinya beragam sehingga siswa memiliki waktu yang cukup untuk mengeksplorasi konsep dan memperkuat keterampilan mereka [33]. Guru dapat memilih dari berbagai alat pengajaran untuk menyesuaikan pembelajaran dengan kebutuhan dan minat siswa. Kurikulum berfokus pada mata pelajaran inti sehingga pembelajaran lebih mendalam, lebih banyak waktu yang diberikan untuk mengembangkan keterampilan dan karakter, serta lebih sederhana dan lebih dalam karena berfokus pada hal-hal yang penting.[3]

2.3.2. Youtube

Youtube merupakan website media sosial berbasis video yang muncul pada tahun 2005. Youtube merupakan salah satu produk dibawah naungan Google yang memiliki jumlah pengguna bulanan mencapai 2.6 triliun pengguna dan menjadi website yang paling banyak dikunjungi dengan jumlah pengunjung hingga 60 miliar pengunjung per bulan (Dixon, 2022) Menurut Kamus Oxford, Video adalah suatu sistem perekaman gambar bergerak dan suara dengan media penyimpanan secara digital atau fisikal sedangkan berdasarkan KBBI, video merupakan rekaman gambar hidup atau program televisi yang ditayangkan

melalui pesawat televisi Beberapa fitur yang banyak digunakan oleh pengguna Youtube yaitu Komentar, Like, Subscribe dan Share. Subscribe yaitu fitur dimana pengguna Youtube dapat berlangganan kepada satu kanal dan mendapatkan notifikasi apabila kanal tersebut mengunggah video. Share yaitu fitur yang digunakan untuk menyebarluaskan video youtube ke media sosial lainnya. Like yaitu fitur dimana pengguna dapat memberikan apresiasi terhadap video Youtube yang disukai. Dan komentar dimana pengguna Youtube dapat mengutarakan opininya terhadap video atau topik yang dibahas pada video tersebut. Komentar berdasar KBBI adalah suatu ulasan atau tanggapan atas terhadap suatu hal sedangkan menurut Oxford Dictionaries, komentar diartikan sebagai sesuatu yang dikatakan atau ditulis yang berguna untuk memberikan suatu opini atau menjelaskan sesuatu atau seseorang. [34]

2.3.3. Youtube API

Youtube API adalah antarmuka aplikasi pemrograman yang dapat digunakan untuk menyimpan metadata video, membuat playlist atau fungsifungsi lainnya dari Youtube. Youtube API dapat membantu memudahkan penampilan video dari Youtube ke website lainnya dengan mudah. Selain itu Youtube API juga memudahkan pengguna dalam pengambilan data-data tertentu seperti komentar yang ada dalam suatu video. (Google Developers,2022)[35]

2.3.4. Google Colab

Google Colab atau Colaboratory merupakan suatu produk google yang menyediakan tempat untuk membuat dan mengeksekusi kode python. Python adalah suatu bahasa pemrograman berorientasi objek yang banyak digunakan dalam machine learning. Google Colab sendiri menyediakan layanan gratis terkait pemrograman python yang diproses dalam cloud atau internet. (Google, 2022.) Google Colab banyak digunakan karena mudah digunakan tanpa perlu menginstal atau setup dan hanya membutuhkan akun Google. Limitasi Google Colab terdapat pada penggunaan sumber daya yang sedikit (perlu membayar untuk menggunakan sumber daya lebih banyak).[36]

2.3.5 Lexicon Based

Lexicon Based merupakan metode berbasis kamus yang bertujuan untuk memperoleh bobot suatu kalimat pada suatu dataset untuk mengetahui opini kelas terhadap dataset tersebut. [37]. Metode lexicon dapat dilakukan secara manual atau otomatis dari seed of word yang diperluas. metode lexicon based dianggap sesuai, dalam analisis data survei, data komentar YouTube, data Twitter atau media sosial lainnya yang memuat pendapat pengguna dan masyarakat luas.[38]

Lexicon Based mempunyai banyak kelebihannya, salah satunya adalah metode ini dapat memberi label pada kalimat secara otomatis sehingga menghemat waktu saat memproses dataset yang besar atau besar [38]. Kemudian dengan menggunakan metode ini, sentimen didalam dataset digunakan untuk menghindari kalimat bias dari opini pribadi seseorang.

Komponen terpenting didalam metode lexicon based adalah kamus. Kamus digunakan untuk menormalisasikan suatu kalimat dan mengekstrak kata kunci. Dengan contoh sebagai berikut :

a. Kata kunci positif : senang, bagus, keren, baik, cerdas.

b. Kata kunci negatif : bodoh, jelek, jahat, gagal, sulit, lemah.

c. Kata kunci negasi : tetapi, tidak, bukan, sebaiknya.

Orientasi sentimen suatu kalimat ditentukan dengan menambahkan nilai orientasi seluruh kata pada sentimen dalam kalimat tersebut. Kata positif bernilai +1, negatif -1. Kata-kata negatif dan antonim juga dipertimbangkan. Ada empat langkah dalam memutuskan sesuatu berdasarkan pendekatan lexicon, yaitu:

1. Identifikasi sentimen: Untuk setiap kalimat yang mengandung satu atau lebih kata-kata sentimen, langkah ini memilih semua kata dan frasa sentimen dalam kalimat sentimen. Setiap kata positif diberi skor +1 dan kata negatif -1. Contoh “Kualitas produk ini kurang bagus [+1], tapi daya tahannya lama [+1]”. Berdasarkan contoh ini, kata baik bernilai +1 dan tahan lama +1 karena merupakan kata positif.
2. Pengubah sentimen (sentimen shifter) adalah kata dan frasa yang dapat mengubah arah sentimen. Ada beberapa jenis pengonversi kata negative (shifter negasi), seperti tidak, tidak pernah, dan tidak ada yang merupakan jenis yang paling umum. Dengan demikian, kalimatnya menjadi “Kualitas dari barang ini tidak bagus [-1], tetapi kekuatannya tahan lama [+1]” karena kata negasinya “Tidak.”.

3. Agregasi: Pada langkah ini, fungsi agregasi opini diterapkan pada skor sentimen yang dihasilkan untuk menentukan arah akhir sentimen.

2.3.6 Kamus Lexicon

Dalam pendekatan analisis sentimen menggunakan leksikon, kamus adalah elemen kunci dalam ekstraksi kata-kata sentimen. Karena sebagian besar kamus seperti WordNet berisi sinonim dan antonim untuk setiap kata, menggunakan kamus untuk mengumpulkan kata-kata sentimen adalah metode yang jelas. Singkatnya, teknik atau pendekatan ini melibatkan penggunaan kata-kata sentimen awal sebagai referensi dan mencocokkannya berdasarkan struktur sinonim dan antonim dalam kamus. Lebih khusus lagi, metode ini bekerja sebagai sekumpulan kecil kata-kata sentimen yang orientasi positif atau negatifnya diketahui dan kemudian dikumpulkan secara manual. Algoritme kemudian menghitung jumlah kata dengan mencari sinonim dan antonim di WordNet atau kamus lainnya. Kata-kata yang ditemukan ditempatkan dalam daftar positif atau negatif. Proses ini berakhir ketika tidak ada kata baru yang ditemukan. Setelah proses selesai, sebuah langkah verifikasi digunakan untuk menghitung agregat positif atau negatif. [38] Beberapa kamus akan digunakan dalam pendekatan leksikon, seperti kamus leksikon positif, negatif, dan negasi, Kamus Besar Bahasa Indonesia (KBBI), kamus kata dasar, dan kamus kata kosong.

a. Kamus Positif

   Kamus positif digunakan untuk memilih kata-kata yang termasuk dalam sentimen positif dari kalimat atau kueri yang akan ditentukan.

b. Kamus Negatif

Kamus negatif digunakan untuk memilih kata-kata yang termasuk dalam sentimen negatif dari kalimat atau kueri yang akan ditentukan.

c. Kamus negasi

Kamus negasi digunakan sebagai pendeteksi kalimat atau query yang memiliki sentimen yang telah ditentukan, baik positif maupun negatif, jika sentimen tersebut diikuti dengan kata negasi. Sentimen yang diikuti dengan kata negasi akan berubah nilai sentimennya dari nilai sebelumnya.

d. Kamus kata dasar dan KBBI

Kamus kata dasar dan Kamus Besar Bahasa Indonesia (KBBI) digunakan untuk melakukan proses pemenggalan kata pada tahap pemrosesan bahasa alami. Pemenggalan kata melibatkan pengubahan kata majemuk menjadi kata dasar. Dalam proses ini, kamus kata dasar dan KBBI diperlukan untuk memilih kata yang tepat.

e. Kamus stopwords

Kamus kata kosong digunakan untuk memilih kata-kata dari dataset yang dianggap tidak penting. Proses ini mempercepat proses klasifikasi data