BAB II

TINJAUAN PUSTAKA

2.1. *Artificial Intelligence*

Artificial Intelligence (AI) adalah istilah yang digunakan untuk menyebut yang menyiratkan penggunaan komputer untuk memodelkan perilaku cerdas, yang terintegrasi dengan *big data*, yang menyerupai kecerdasan manusia (Pabubung, 2021). Istilah AI diciptakan oleh McCarthy pada 1950-an dan mengacu pada cabang ilmu komputer dimana algoritma dikembangkan untuk meniru fungsi kognitif manusia, seperti belajar, penalaran, dan pemecahan masalah. AI juga adalah istilah yang mencakup secara luas, tetapi tidak terbatas pada, *Machine Learning* (ML), *Deep Learning* (DL), *Natural Language Processing* (NLP), dan *Computer Vision* (CV) (Yin, Ngiam, & Teo, 2021).

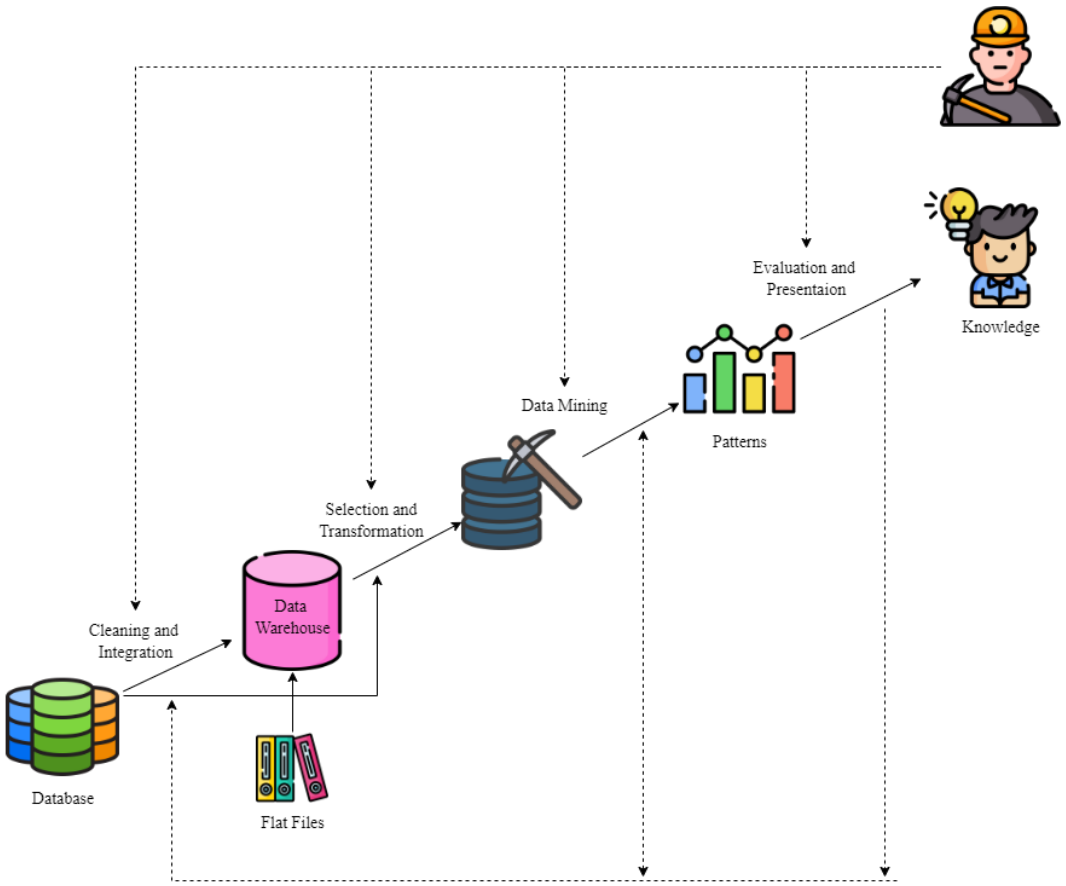
AI merupakan sebuah perkembangan dari teknologi informasi yang terus berkembang dalam sepuluh tahun terakhir. Pemanfaatan AI sangat luas dan tidak hanya terbatas di sektor industri telekomunikasi, namun juga pada berbagai sektor seperti perbankan, manufaktur, jasa, analisis, bahkan di sektor pemerintah. Di beberapa negara, implementasi AI sudah mencapai hampir 56%, terutama pada sektor industri (Ririh, Laili, Wicaksono, & Tsurayya, 2020).

2.2. *Data mining*

Data mining adalah bagaimana mencari data yang tersedia untuk menciptakan sebuah model, lalu memanfaatkan model tersebut untuk mengenali pola data lain yang tidak tersedia di dalam basis data yang tersimpan. Dalam *data mining* pengelompokan data dapat juga dilakukan untuk mengetahui pola secara

universal dari data yang tersedia agar dilakukan langkah tindak lanjut lainnya yang berguna sebagai pendukung kegiatan dan tujuan akhir tertentu (Utomo & Purba, 2019). Penjelasan lain mengatakan *Data mining* merupakan proses iteratif dan interaktif untuk menemukan pola atau model baru yang sempurna, bermanfaat dan dapat dimengerti dalam suatu basis data yang sangat besar (*massive database*) (Sikumbang, 2018). Dapat disimpulkan bahwa *data mining* adalah sebuah metode yang digunakan untuk menemukan pola dari kumpulan data, kemudian pola tersebut dimanfaatkan untuk proses analisa dari kumpulan data yang bersifat besar yang bertujuan mendapatkan suatu pengetahuan.

Tahapan proses dalam penggunaan *data mining* merupakan salah satu dari rangkaian *Knowledge Discovery in Database* (KDD). KDD adalah sebuah kegiatan untuk mengekstrak suatu pengetahuan dari sekumpulan data, hasil *data mining* kemudian diubah menjadi informasi yang mudah dipahami (Afdal & Rosadi, 2019). Dibawah merupakan gambar untuk tahap-tahap data mining :



Gambar 2.1 Tahap-Tahap Data Mining

Sumber : (Han, Kamber, & Pei, 2012)

Gambar 2.1 menunjukkan serangkaian proses dari *data mining*. Serangkaian proses tersebut memiliki tahap sebagai berikut :

1. *Data cleaning* adalah tahap untuk menghilangkan *noise* data yang tidak konsisten dan tidak perlu.
2. *Data integration* adalah tahap mengintegrasikan data yang terpecah menjadi satu kesatuan.
3. *Data selection* adalah tahap dilakukannya penyeleksian data yang relevan dengan tugas analisis, kemudian data dikembalikan ke dalam *database.*
4. *Data transformation* adalah tahap mengubah bentuk data agar tepat dan sesuai untuk memudahkan proses penambangan.

5. *Data mining* merupakan proses esensial dimana metode yang intelijen digunakan untuk mengekstrak pola data.
6. *Pattern evaluation* adalah tahap untuk mengidentifikasi pola yang benar-benar menarik yang mewakili pengetahuan berdasarkan atas beberapa tindakan yang menarik.
7. *Knowledge presentation* adalah tahap untuk yang dimana gambaran teknik visualisasi dan pengetahuan digunakan untuk memberikan pengetahuan yang telah didapatkan kepada *user*.

Beberapa peneliti terdahulu yang telah menerapkan *data mining* sebagai media dalam penelitiannya, Rintho Rante Rerung dengan penelitiannya berjudul "Penerapan Data Mining dengan Memanfaatkan Metode *Association Rule (AR)* untuk Promosi Produk". Penelitian ini bertujuan untuk perhitungan seberapa besar kemungkinan pelanggan akan tertarik terhadap produk yang ditawarkan. Pada penelitian ini metode AR diterapkan untuk menghitung nilai asosiatif antar produk sehingga pola pelanggan bisa didapatkan. Penelitian ini menghasilkan kesimpulan bahwa metode AR bisa digunakan sebagai cara untuk menghitung persentase ketertarikan pelanggan terhadap produk yang ditawarkan. Selain kesimpulan tadi, penelitian tersebut juga menghasilkan sebuah Sistem Promosi Distro Nasional, yang di mana sistem ini menerapkan metode *association rule* (Rerung, 2018).

Kemudian peneliti Ikhsan Romli dan Rega Firana Puspita Dewi dengan judul penelitiannya "Penerapan Data Mining Menggunakan Algoritma K-Means Untuk Klasifikasi Penyakit Ispa". Tujuan penelitian ini adalah menerapkan metode K-Means untuk mengklasifikasi penyakit ISPA dan mendapatkan akurasi

yang tepat dan cepat dalam mengklasifikasi gejala penyakit ISPA menggunakan metode *K-Means*. Proses dari penelitian tersebut menghasilkan 3 klaster, yaitu *cluster* C1 (ISPA Biasa) dengan jumlah anggota 81, *cluster* C2 (ISPA sedang) dengan jumlah anggota 103, *cluster* C3 (ISPA Berat) dengan jumlah anggota 66 (Romli & Dewi, 2021).

Penelitian berjudul "*Implementation Of Data Mining Algorithms For Grouping Poverty Lines By District/City In North Sumatra*" oleh Mhd Ali Hanafiah dan Anjar Wanto. penelitian ini bertujuan untuk mengklasifikasikan garis kemiskinan menurut kabupaten/kota di Provinsi Sumatera Utara, sehingga diketahui kabupaten/kota mana yang memiliki garis kemiskinan tinggi atau rendah. Algoritma pengelompokan yang digunakan adalah *K-Means*. Dengan menggunakan algoritma ini maka data akan dikelompokkan menjadi beberapa bagian, dimana proses implementasi *K-Means* data mining menggunakan *Rapid Miner*. Data yang digunakan adalah data garis kemiskinan menurut kabupaten/kota (rupiah/kapita/bulan) di provinsi Sumatera Utara tahun 2017-2019. Data bersumber dari Badan Pusat Statistik Sumatera Utara. Nilai titik tengah pada Centroid ditentukan berdasarkan pengelompokan yang diinginkan. Di dalam penelitian ini, garis kemiskinan dibagi menjadi 3 kriteria, yaitu kelompok tinggi, kelompok menengah dan kelompok rendah. Maka berdasarkan data pada tabel 1 diambil nilai Centroid (C1) yang tinggi dari data maksimum, sedangkan untuk nilai Centroid sedang (C2) diambil dari nilai rata-rata dan nilai untuk nilai Centroid rendah (C3) diambil dari minimum nilai. Penelitian ini menghasilkan data gambaran Garis Kemiskinan Kabupaten/Kota kelompok di Provinsi Sumatera Utara adalah sebagai berikut: Kelompok garis kemiskinan yang termasuk dalam

klaster tinggi terdiri dari 5 kabupaten/kota. Kelompok garis kemiskinan yang termasuk dalam klaster sedang terdiri dari 18 kabupaten. Sedangkan klaster rendah terdiri dari 10 kabupaten (Hanafiah & Wanto, 2020).

Artikel dengan judul "*Educational data mining using cluster analysis and decision tree technique: A case study*" oleh Snježana Križanić. Artikel ini menjelaskan penerapan teknik *data mining* pada data pendidikan sebuah perguruan tinggi di Kroasia. Teknik data mining yang digunakan dalam penelitian ini adalah analisis *cluster* dan *Decision Tree* (DT). Tujuan utama penggunaan teknik *data mining* di bidang pendidikan adalah untuk mengembangkan model dimana kinerja siswa dapat diprediksi secara keseluruhan dalam kursus yang dipilih. Model kemudian dianalisis untuk memprediksi keberhasilan siswa. Berbagai teknik penambangan data seperti klasifikasi dan pengelompokan diterapkan untuk mengungkap pengetahuan tersembunyi dari data pendidikan. Pengelompokan digunakan dengan analisis pola, pengambilan keputusan, dan *machine learning*, yang termasuk penambangan data, pengambilan dokumen, segmentasi gambar, dan klasifikasi pola. Data siswa kemudian dipisahkan ke dalam kelompok, sehingga siswa dalam kelompok yang sama bisa mencapai kemajuan yang sama (Krizanic, 2020).

Penelitian oleh Tri Wahyudi dan Titi Silfia dengan judul "*Implementation of Data Mining Using K-Means Clustering Method to Determine Sales Strategy In S&R Baby Store*". Penelitian ini berfokus pada peningkatan strategi penjualan yang baik agar dapat meningkatkan keuntungan penjualan pada toko S&R Baby Store. Penelitian ini membahas penerapan data mining, menggunakan algoritma *K-Means Clustering* dengan metode *CRISP-DM*. Implementasi menggunakan

RapidMiner 9.10 yang dilakukan dengan memasukkan data transaksi penjualan dengan total 4 atribut dan membentuk 4 *cluster* yang terdiri dari *very in demand*, *in demand*, *moderate in demand* dan *less in demand*. *Cluster* kedua dengan 944 produk, *cluster* ketiga dengan 2 produk, dan *cluster* keempat dengan 43 produk. Hasil dari *cluster* tersebut adalah produk yang dijual yang masuk kedalam kategori produk terlaris, kemudian hasil *cluster* tersebut divalidasi menggunakan *Davies-Bouldin Index (DBI)* dengan nilai DBI yang dihasilkan dari pengelompokan 0,560 (Wahyudi & Silfia, 2022).

2.3. *Machine Learning*

Machine Learning (ML) adalah aplikasi atau bagian dari AI yang membuat sistem memiliki kemampuan belajar secara otomatis dan meningkatkan kemampuannya berdasarkan pengalaman tanpa diprogram secara eksplisit. Dalam hal ini, program komputer tidak ditulis secara statis. Fokus ML terdapat pada pengembangan program komputer yang dapat mengakses data dan belajar dari data tersebut. ML juga dapat didefinisikan sebagai algoritma yang bertujuan menemukan dan mengaplikasikan pola-pola di dalam data. Algoritma pada ML menggunakan teknik-teknik statistik untuk menemukan pola-pola tersebut (Kusuma, 2020). ML bisa diartikan bahwa adalah sebuah cara untuk mesin atau komputer untuk bisa bertindak pintar dan belajar dari data yang dimasukkan dan dengan demikian mesin tersebut bisa bertindak dengan “pintar” untuk menyelesaikan suatu tugas yang diberikan (Chinnamgari, 2019).

ML memiliki beberapa tipe algoritma yang ditujukan untuk menyelesaikan masalah atau tugas tertentu. Sebagai contoh, sebuah kendala dari masalah tertentu dapat berupa ketersediaan data berlabel yang dapat diberikan sebagai masukan ke

algoritma pembelajaran. ML memiliki beberapa metode yang populer, antara lain *supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, dan transfer learning* (Chinnamgari, 2019):

1. *Supervised Learning*, adalah sebuah algoritma dari ML yang digunakan ketika tujuan yang diinginkan jelas. Pada metode ini data yang digunakan adalah data yang memiliki kelas atau label di dalamnya. Data terbagi menjadi dua yaitu data latih dan data uji. Data latih digunakan untuk membentuk model, lalu model yang telah dibentuk itu diterapkan pada data uji, salah satu tujuannya adalah untuk mengetahui hasil atau *output* dan mengetahui tingkat akurasinya.
2. *Unsupervised Learning*, adalah sebuah algoritma dari ML yang digunakan ketika data yang digunakan tidak memiliki label. Berbeda dengan metode *supervised learning* yang memiliki data latih dan data uji, metode *unsupervised learning* tidak perlu menggunakan data untuk "berlatih", akan tetapi model akan dibentuk dari cara mengidentifikasi pola, mengenali karakteristik data, dan lain-lain. Salah satu contoh penggunaan algoritma ini adalah *clustering*.
3. *Semi-Supervised Learning*, adalah sebuah algoritma kolaborasi dari dua algoritma sebelumnya yaitu *Supervised Learning* dan *Unsupervised Learning*. ML membutuhkan data latihan yang sangat banyak untuk mesin. Algoritma ini digunakan ketika data yang ada sebagian memiliki label, namun sebagian tidak memilikinya. Algoritma ini bekerja dengan cara membuat model dengan data yang berlabel, kemudian melakukan *clustering* atau pengelompokan data yang tidak berlabel yang kemudian

akan diletakan dari model yang sudah terbentuk berdasarkan kesamaan fitur atau ciri-ciri dari data.

4. *Reinforcement Learning*, merupakan metode ML yang bertujuan untk meningkatkan strategi yang bisa digunakan untuk menyelesaikan masalah secara berkelanjutan dengan mempelajari masukan yang diterima. Tujuan dari algoritma ini adalah untuk untuk mendapatkan hasil terbaik dari tindakan yang dilakukan oleh mesin untuk menyelesaikan suatu masalah. Atau dengan kata lain algoritma ini berfokus untuk mendapatkan langkah yang optimal untuk menyelesaikan sebuah masalah.
5. *Transfer Learning*, adalah sebuah algoritma yang mencoba untuk menyelesaikan permasalahan dengan menggunakan data seminimal mungkin dengan cara menggunakan pengetahuan yang sudah diketahui yang didapatkan dari berbagai model yang berkesinambungan.

Beberapa peneliti yang telah menerapkan ML sebagai media dalam penelitiannya seperti Mokhamad Ramdhani Raharjo dan Agus Perdana Windarto dengan penelitiannya berjudul "Penerapan *Machine Learning* (Prediksi Tingkat Pemahaman Mahasiswa terhadap Mata kuliah) dengan Konsep Data Mining *Rough Set (RF)*". Tujuan dari penelitian adalah untuk menggali informasi dari RF dengan menggunakan aplikasi Rosetta pada kasus prediksi tingkat pemahaman mahasiswa terhadap mata kuliah. Penelitian ini menguji data yang didapatkan dari pemberian angket kepada 165 mahasiswa. Dari penelitian ini didapatkan sebuah kesimpulan bahwa metode RF bisa diterapkan untuk memprediksi tingkat pemahaman mahasiswa terhadap mata kuliah. Penelitian ini mendapatkan hasil

90 rules berupa aturan pola dengan atribut komunikasi dan media pembelajaran menjadi atribut yang dominan (Raharjo & Windarto, 2021).

Kartika Maulida Hindrayani, Amalia Anjani, dan Afina Lina Nurlaili dengan penelitiannya yang berjudul "Penerapan *Machine Learning* pada Penjualan Produk Usaha Mikro Kecil dan Menengah (UMKM) : Studi Literatur". Tujuan penelitian ini agar dapat mengetahui pengaplikasian algoritma ML dalam mendukung aktivitas penjualan produk UMKM. Metode yang digunakan pada penelitian yaitu identifikasi permasalahan, pengumpulan artikel dengan kata kunci yang ditentukan, penyortiran dari judul artikel, penyortiran dari isi artikel yang tidak berhubungan dengan algoritma ML, dan penarikan kesimpulan (Hindrayani, Anjani, & Nurlaili, 2021).

Peneliti Tongke Fan dan Jing Xu dengan judul penelitian "*Image classification of crop diseases and pests based on deep learning and fuzzy system*". Tujuan penelitian adalah untuk menciptakan teknologi segmentasi citra berdasarkan teori graf dan penerapannya, kerangka dasar teknologi segmentasi citra International Journal of Data Warehousing and Mining. Teknologi ini nantinya dapat melakukan klasifikasi otomatis citra penyakit tanaman dipelajari dengan metode *deep learning*. Teori sistem fuzzy diterapkan pada pretreatment gambar buram. Penelitian ini juga menerapkan *data mining* dalam prosesnya untuk memproses data berupa gambar yang digunakan dalam penelitiannya (Fan & Xu, 2020).

Penelitian dengan judul "*The Implementation of Deep Reinforcement Learning in E-Learning and Distance Learning: Remote Practical Work*" oleh Abdelali El Gourari, Mustapha Raoufi, Mohammed Skouri, dan Fahd Ouatik.

Tujuan dari pekerjaan ini adalah untuk mengusulkan sistem rekomendasi berdasarkan *Deep Quality-Learning Networks* (DQNs) untuk merekomendasikan dan mengarahkan siswa terlebih dahulu melakukan kerja *Remote Practical Work* (RPW) sesuai dengan keterampilan mereka masing-masing klik mouse atau keyboard per siswa. Berbagai informasi dari siswa dan guru serta interaksi mereka dengan konten pembelajaran akan digunakan sebagai *input* kedalam sistem baru untuk mendapatkan *output* (melakukan RPW). Algoritma atau metode yang digunakan dalam penelitian tersebut adalah *deep reinforcement learning* karena efisiensinya yang tinggi (Gourari, Raoufi, Skouri, & Ouatik, 2021).

Penelitian berjudul "*Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset*)" oleh Ayushi Mitra. Penelitian ini mengadopsi pendekatan berbasis aturan yang mendefinisikan seperangkat aturan dan input seperti *Classic Natural Language Processing techniques, Stemming, tokenization* wilayah penandaan ucapan dan penguraian *machine learning* untuk analisis sentimen yang akan diimplementasikan oleh bahasa *Python* paling canggih. Tujuan penelitian ini adalah untuk mengetahui akurasi dari analisis sentimen dengan pendekatan berbasis *Lexicon* yang berisi kamus kata-kata positif dan negatif yang digunakan untuk menentukan polaritas sentimen berdasarkan kecenderungan pesan dari konten sumber *dataset*. Kombinasi pendekatan berbasis *Machine learning* dan *leksikon* kemudian digunakan oleh pendekatan berbasis *Hybrid* untuk klasifikasi. Dalam penelitian tersebut kemudian diterapkanlah beberapa algoritma untuk menguji keakurasiannya, mulai dari algoritma *NBC, SKlearnBernouliieNB, Sklearn Support Vector Classification* (SVM), DT, *Random Fores, dan K-Nearest*

Neighbor (K-NN). Dari algoritma yang telah diterapkan, diperoleh keakuratannya hasilnya yang tidak efisien. Ini terjadi karena Kekuatan analisis sentimen bergantung pada skala *lexicon* (kamus) karena ukuran *lexicon* akan meningkatkan pendekatan ini dan menjadi semakin memakan waktu (Mitra, 2020).

2.4. *Web Scraping*

Web scraping adalah suatu proses yang melibatkan pengambilan dokumen semi-terstruktur dari internet, umumnya halaman web berupa halaman dalam bahasa markup seperti *HyperText Markup Language (HTML)* atau *Extensible HyperText Markup Language (XHTML)*, dan kemudian dokumen tersebut dianalisis untuk mengekstraksi data tertentu yang dapat digunakan dalam konteks lain. *Web scraping* juga dikenal sebagai *screen scraping* (Mufidah & Siahaan, 2021).

Pada dasarnya teknik *web scraping* adalah suatu proses *copy-paste* dari suatu web namun dilakukan secara otomatis dan terorganisir. Teknik *web scraping* hanya mengambil data yang sudah tersaji dalam suatu web, bukan secara ilegal dengan langsung masuk kedalam *database* pemilik web. Dengan demikian, baik pemilik *web* masih tetap mendapatkan *traffics* yang dapat meningkatkan valuasinya, sementara pihak *scraper* mendapatkan data yang dibutuhkan (Saurkar, Pathare, & Gode, 2018).

Dari beberapa pengertian diatas, *web scraping* adalah suatu aktivitas yang dilakukan untuk mengumpulkan data dari suatu web yang dilakukan secara otomatis guna mendapatkan data yang diinginkan untuk kepentingan penelitian maupun kepentingan lain.

2.5. *Twitter*

Twitter adalah layanan jejaring sosial yang memungkinkan penggunanya untuk memposting teks, gambar dan video yang dikenal dengan sebutan kicauan (tweet). *Twitter* didirikan oleh Jack Dorsey pada bulan maret 2006, dan situs jejaring sosialnya diluncurkan pada bulan juli. Sejak diluncurkan, *Twitter* telah menjadi salah satu dari sepuluh situs yang paling sering dikunjungi di Internet (D'Monte, 2009). Pengguna *Twitter* di tahun 2023 berdasarkan laporan dari *We Are Social* dan *Hootsuite*, terdapat 556 juta pengguna *Twitter* di seluruh dunia. Jumlah tersebut meningkat 27,4% dibandingkan pada periode yang sama tahun sebelumnya. Indonesia menempati urutan ke-5 untuk pengguna *Twitter* terbanyak, dengan jumlah pengguna aktif sekitar 24 juta (Santika, 2023).

Dengan adanya *Twitter* ini, masyarakat dapat menyampaikan opini atau pendapat mereka tentang suatu kontroversi yang sedang terjadi secara langsung. Di *Twitter* pengguna bisa membuat sebuah kicauan, merupakan sebuah kegiatan untuk menuliskan sebuah status atau pesan yang nantinya bisa dibaca oleh seluruh pengguna *Twitter* lain. *Twitter* menyediakan *API (Application Programming Interface)* yang mempermudah setiap orang untuk mengambil data dari *Twitter*.

Twitter dalam penelitian ilmiah sering digunakan untuk sumber dari data, terutama penelitian untuk mengukur kepuasan masyarakat dalam satu topik tertentu. Penelitian terdahulu yang sudah menggunakan *Twitter* antara lain penelitian *"Online learning sentiment analysis during the covid-19 Indonesia pandemic using Twitter data"* (Sahir, Ramadhana, Marpaung, Munthe, & Watrianthos, 2021), penelitian *"Sentiment Analysis of Shared Tweets on Global Warming on Twitter with Data Mining Methods: A Case Study on Language"*

(Kirelli & Arslankaya, 2020), penelitian "Analisis Sentimen Pembelajaran Daring Pada *Twitter* di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes" (Samsir, Ambiyar, Verawardina, Edo, & Watrianthos, 2021).

2.6. *Term Frequency-Inverse Document Frequency* (TF-IDF)

TF-IDF merupakan algoritma yang digunakan untuk melakukan pembobotan sebuah kata dalam *dataset*. Penggunaan algoritma ini bertujuan untuk membuat sebuah vektor dengan banyak *term* sehingga tiap kata yang ada dalam *dataset* bisa dikenali dan dihitung sebagai satu fitur. Perhitungan TF-IDF terbagi menjadi dua bagian, yaitu *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). TF digunakan untuk menghitung jumlah kemunculan kata dalam sebuah *dataset*, sedangkan IDF merupakan jumlah data yang berisi *term* yang dicari (Darwis, Pratiwi, & Pasaribu, 2020). Rumus yang digunakan untuk perhitungan TF-IDF terdapat pada Persamaan 1 berikut ini (Fairuz, Ramadhani, & Tanjung, 2021):

$$TF - IDF = TF \times IDF \quad (1)$$

$$IDF = Log(D/DF) \quad (2)$$

TF = frekuensi istilah muncul di dokumen

IDF = jumlah dokumen yang memiliki istilah

DF = total istilah yang muncul di tiap dokumen

D = total dokumen

TF-IDF sering digunakan dalam penelitian didalamnya melakukan pengolahan kata, seperti penelitian oleh Susanti, Rianto, dan Acep yang berjudul *"Sentiment Analysis Provider by.U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method".* Penelitian tersebut bertujuan untuk

melakukan analisis sentimen untuk aplikasi By.U yang dimana data sentimen diambil dari sesi *review* di Google Play Store. Penelitian ini menerapkan TF-IDF sebagai algoritma untuk pembobotan kata dan juga SVM sebagai algoritma untuk analisis sentimen. Hasil dari penelitian tersebut adalah 54.6% pengguna memberikan *review* positif dan 45.4% negatif dengan tingkat akurasi 83.3% (Fransiska, Rianto, & Gufroni, 2020).

Selanjutnya penelitian berjudul "Implementasi Algoritma Multiclass SVM Pada Opini Publik Berbahasa Indonesia Di *Twitter*" oleh Debby, Yusra, dan Heni. Penelitian tersebut bertujuan untuk melakukan analisa performa dari algoritma SVM untuk analisis sentimen terhadap opini publik di media sosial *Twitter*. Opini publik dalam penelitian tersebut dibagi menjadi tiga kelas yaitu netral, positif, dan negatif dengan menerapkan algoritma TF-IDF untuk melakukan pembobotan kata dan juga menggunakan algoritma SVM yang telah dioptimasi yaitu dengan SVM *One Against One* dan SVM *One Against Rest* untuk analisis sentimen. Hasil dari penelitian tersebut adalah hasil SVM *One Against One* lebih unggul untuk nilai presisi, *recall* dan *F1Score*, sedangkan untuk akurasi SVM *One Against Rest* lebih unggul dengan nilai perbedaan 0,06 (Alita, Fernando, & Sulistiani, 2020).

Penelitian berjudul "Penerapan Algoritma SVM Untuk Analisis Sentimen Pada Data *Twitter* Komisi Pemberantasan Korupsi Republik Indonesia" oleh Dedi, Eka, dan Ferico. Penelitian tersebut bertujuan untuk melakukan analisis sentimen masyarakat di media sosial *Twitter* tentang KPK RI yang akan diklasifikasikan menjadi tiga kelas yaitu netral, positif, dan negatif. Penelitian tersebut menggunakan 2000 data dengan mengimplementasikan algoritma TF-IDF untuk melakukan pembobotan kata serta menggunakan algoritma SVM untuk

proses klasifikasinya. Hasil dari penelitian tersebut adalah 25% data masuk kedalam kategori netral, 8% positif, dan 77% negatif dengan tingkat akurasi 82% (Darwis, Pratiwi, & Pasaribu, 2020).

2.7. *Synthetics Minority Over-sampling Technique* (SMOTE)

SMOTE adalah sebuah metode yang pertama kali diusulkan oleh Chawla pada tahun 2002. Metode ini bertujuan untuk melakukan *oversamping* pada kelas minoritas dan membuat data *training* sintetik. *Imbalance dataset* adalah sebuah kondisi dimana perbandingan masing-masing kelas di *dataset* yang tidak seimbang. Secara umum ada dua cara untuk mengatasi *imbalance dataset*, yaitu bisa dengan *Random Undersampling* (RUC) dan *Random Oversampling* (ROC). ROS akan menduplikasi secara acak data dari kelas yang minoritas, secara umum *oversampling* lebih baik dari *undersampling*. *Oversampling* sendiri adalah sebuah cara untuk menaikan jumlah kelas minoritas sampai setara dengan kelas mayoritas. ROC bisa menjadi salah satu cara yang cukup baik untuk membantu proses penyeimbangan *dataset*, namun kemungkinan hasil dari klasifikasi nantinya mengalami *overfitting* karena metode ini membuat salinan/duplikat yang sama persis dari data yang berasal dari kelas minoritas (Amalia & Wahyuni, 2020).

SMOTE sendiri adalah sebuah metode yang mirip dengan ROC namun data sintetik yang dibuat tidak hanya asal duplikat, melainkan dibuat dengan menggunakan konsep *nearest neighbour*. Data sintetik yang dibuat dengan metode SMOTE dihitung dari jarak kedekatan fiturnya. Jumlah tetangga terdekat dari *k* yang dipilih menyesuaikan dengan jumlah data sintetik yang dibutuhkan. Data sintetik dibuat dengan mengambil perbedaan antara vektor fitur yang

dipertimbangkan dan tetangga terdekatnya dan mengalikan perbedaan ini dengan angka acak antara 0 dan 1. Data sintetik yang dihasilkan menyebabkan wilayah keputusan kelas minoritas menjadi lebih umum, yang mana mengarah ke generalisasi yang lebih baik untuk proses klasifikasi nantinya (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

Metode SMOTE untuk mengatasi permasalahan ketidak seimbangan data telah banyak digunakan dalam penelitian. Penelitian dengan judul "Smote: Metode Penyeimbang Kelas Pada Klasifikasi Data Mining" oleh Amalia Anjani Arifiyanti dan Eka Dyar Wahyuni, yang membahas mengenai metode SMOTE untuk mengatasi ketidak seimbangan *dataset*. Sebagian besar, algoritma pengklasifikasi cenderung secara implisit menganggap bahwa data yang diproses memiliki distribusi yang seimbang, karenanya pengklasifikasi standar lebih condong kearah data yang jumlah kelasnya dominan. Pada penelitian ini, model klasifikasi yang dihasilkan oleh *Logistic Linear*, KNN, dan *Naive Bayes* menunjukkan bahwa metode SMOTE meningkatkan performa model klasifikasi, sedangkan decision tree tidak menunjukkan hasil yang berbeda baik sebelum *oversampling* maupun setelah *oversampling* (Amalia & Wahyuni, 2020).

Penelitian oleh Sulistiyono, Yoga, Sumarni, dan Gagah dengan judul "Implementasi Algoritma Synthetic Minority Over-Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada *dataset* Klasifikasi". Penelitian ini membahas tentang penanganan terhadap distribusi kelas yang tidak seimbang pada *dataset* dengan menggunakan algoritma SMOTE untuk meningkatkan nilai akurasi maupun g-mean pada algoritma NBC, SVM, K-NN, dan Decision Tree. Data pengujian yang digunakan dalam penelitian ini adalah data public yang

didapatkan dari situs KEELS data mining yang menyediakan *dataset* dengan angka Imbalance Ratio yang beragam. Hasil dari penelitian tersebut adalah sebuah kesimpulan bahwa metode SMOTE dapat meningkatkan nilai akurasi maupun g-mean pada algoritma NBC, SVM, K-NN dan Decision Tree (Sulistiyono, Pristyanto, Adi, & Gumelar, 2021).

Penelitian berjudul ”Implementasi XGBoost Pada Keseimbangan Patient *Dataset* dengan SMOTE dan *Hyperparameter Tuning Bayesian Search*” oleh Rahmad dkk. Penelitian ini melakukan klasifikasi menggunakan metode XGBoost yang didukung oleh SMOTE untuk membantu menyeimbangakan *dataset*. Dari penelitian hasil yang didapatkan dari model XGBoost memperoleh nilai AUC sebesar 0.618, untuk model XGBoost dengan Bayesian search memperoleh nilai AUC sebesar 0.658, kemudian untuk model XGBoost SMOTE memperoleh nilai AUC sebesar 0.716, lalu untuk model XGBoost SMOTE dengan Bayesian search memperoleh nilai AUC sebesar 0.767. Kesimpulan dari penelitian ini adalah dengan bantuan SMOTE dan hyperparameter tuning Bayesian search hasil kinerja model klasifikasi menggunakan metode XGBoost pada prediksi penderita penyakit liver dapat ditingkatkan (Ubaidillah, Muliadi, Nugrahadi, Faisal, & Herteno, 2022).

2.8. *Naïve Bayes Classifier* (NBC)

Algoritma *Naïve Bayes Classifier* (NBC) adalah algoritma klasifikasi berdasarkan probabilitas dalam statistik yang diusulkan oleh Thomas Bayes yang digunakan untuk memprediksi peluang masa depan berdasarkan peluang masa lalu (Teorema Bayes). Algoritma tersebut kemudian dikombinasikan dengan *“naiv”* dimana kondisi antara atribut satu sama lain tidak terikat satu sama lain

(Pratmanto, Rousyati, Wati, Widodo, & Suleman, 2020). NBC memiliki beberapa jenis, salah satunya *Multinomial Naïve Bayes* (*Multinomial* NB) yang merupakan jenis model klasifikasi NBC yang sering digunakan untuk melakukan penelitian di bidang NLP. Teorema dasar *Multinomial* NB mengikuti aturan yang sama dengan pengklasifikasian NBC yang dapat dilihat pada Persamaan (3) (Mustakim & Priyanta, 2022):

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (3)$$

X = data dengan kelas tidak dikenal

H = hipotesis data X adalah kelas khusus

P(H|X) = probabilitas hipotesis H didasarkan pada kondisi X

P(H) = probabilitas hipotesis H

P(X|H) = probabilitas hipotesis X didasarkan pada kondisi H

P(X) = probabilitas X

NBC sering digunakan oleh banyak peneliti sebagai algoritma yang digunakan untuk menyelesaikan masalahnya. Berikut adalah beberapa penelitian yang menggunakan algoritma NBC. Penelitian yang dilakukan oleh Samsir, Ambiyar, Unung Verawardina, Firman Edi, Ronal Watrianthos dengan judul penelitiannya "Analisis Sentimen Pembelajaran Daring Pada *Twitter* di Masa Pandemi COVID-19 Menggunakan Metode *Naïve Bayes*". Penelitian ini berfokus untuk menganalisis sentimen masyarakat mengenai pembelajaran daring yang dilakukan semasa pandemi Covid-19. Penelitian ini dilakukan dengan mengambil data opini atau pendapat masyarakat dari media sosial *Twitter* dengan kata kunci "pembelajaran daring", "kuliah", "belajar", "online", "daring", dan tagar

#BelajarDariRumah yang difilter dengan kata kunci “online” dan “rumah” pada cuitan-cuitan dalam bahasa Indonesia pada minggu pertama November 2020 dan kemudian di analisa dengan metode NBC. Hasil dari penelitian ini adalah 1% netral, 30% data positif, 69% data negatif dan dengan tingkat akurasi mencapai 97.15% (Samsir, Ambiyar, Verawardina, Edo, & Watrianthos, 2021).

Penelitian berjudul “Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma Naïve Bayes Classifier” oleh Winda Yulita, Eko Dwi Nugroho, Muhammad Habib Algifari yang berfokus untuk melakukan analisis sentimen masyarakat tentang vaksin Covid-19. Data dari penelitian ini didapatkan dari media sosial *Twitter* dengan jumlah data 3780 cuitan yang berkaitan dengan vaksinasi Covid-19. Data tersebut kemudian di analisa dengan menggunakan NBC dan mendapatkan hasil bahwa jumlah cuitan yang netral (34.4%), sebagian besar cuitan memiliki sikap positif (60.3%), melebihi jumlah cuitan yang menentang (5.4 %). Nilai akurasi yang dihasilkan sebesar 0.93 (93 %) (Yulita, Nugroho, & Algifari, 2021).

Penelitian berjudul “*Skin disease detection and segmentation using dynamic graph cut algorithm and classification through Naive Bayes classifier*” yang dilakukan oleh V.R. Balaji, S.T. Suganthi, R. Rajadevi, V. Krishna Kumar, B. Saravana Balaji, dan Sanjeevi Pandiyan. Tujuan penelitian tersebut adalah untuk menggunakan algoritma pemotongan grafik dinamis baru untuk segmentasi lesi kulit diikuti oleh pengklasifikasi probabilistik yang disebut pengklasifikasi NBC untuk tujuan klasifikasi penyakit kulit. Metode pemotongan grafik digunakan karena membawa keunggulan akurasi dan kinerja dibandingkan dengan metode segmentasi gambar lainnya. NBC dalam penelitian tersebut digunakan

untuk membantu pengelompokan penyakit kulit setelah wilayah yang terpengaruh tersegmentasi dan fitur-fiturnya diambil. Atribut diagnostik dari kumpulan data dapat diklasifikasikan ke dalam tipe jinak atau ganas, atribut klinis meliputi usia, lokasi, diameter, tipe diagnosis, riwayat keluarga, kelas melanoma, mitosis indeks, ketebalan, tipe, ulserasi, melanositik, tipe nevus, dan data terkait riwayat pribadi. Total ada sekitar 23.906 gambar untuk pelatihan dan pengujian algoritma NBC. Kumpulan fitur dimulai dengan data tekstur yang diikuti oleh warna dan fitur asimetri. Data tekstur meliputi nilai kontras, korelasi, dan energi. Perangkat yang digunakan untuk *training* dan *testing* adalah perangkat dengan The Intel i7-10510 8 M cache, 4.8 GHz. Hasil dari penelitian tersebut adalah terbentuknya algoritma baru untuk melakukan segmentasi gambar kulit yang digunakan sebagai diagnosa awal apakah kejadian abnormal yang terjadi pada kulit tersebut merupakan penyakit atau bukan. Segmentasi potongan graf yang diusulkan dalam penelitian tersebut memiliki keuntungan memberikan efisiensi praktis, pelabelan yang optimal secara global mengintegrasikan banyak isyarat dan kendala, ketahanan numerik dan sifat topologi daerah yang tidak terbatas (Balaji, et al., 2020).

Beberapa penelitian terdahulu yang telah melakukan penelitian terkait analisis sentimen seperti yang ditampilkan pada Tabel 2.1 berikut:

Tabel 2.1 Penelitian Yang Relevan

No	Judul Penelitian	Metode penelitian	Peneliti	Hasil Penelitian
1	Analisis Sentimen Masyarakat Terhadap COVID-19 Pada Media Sosial *Twitter*	NBC & K-NN	(Fairuz, Ramadhani, & Tanjung, Analisis Sentimen Masyarakat Terhadap COVID-19 Pada	Hasil akurasi dari algoritma NBC 85% dan algoritma K-NN adalah 72%

No	Judul Penelitian	Metode penelitian	Peneliti	Hasil Penelitian
			Media Sosial Twitter, 2021)	
2	Analisis Sentimen Pembelajaran Daring Pada *Twitter* di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes	NBC	(Samsir, Ambiyar, Verawardina, Edo, & Watrianthos, 2021)	Dari 12,9 ribu data yang dianalisis, diperoleh 1% netral, 30% positif, dan 69% negatif. Dengan akurasi 97.15%.
3	Sentiment Analysis of Shared Tweets on Global Warming on *Twitter* with Data Mining Methods: A Case Study on Turkish Language	NBC, SVM, & K-NN	(Kirelli & Arslankaya, 2020)	Akurasi algoritma NBC 65.43%. K-NN 74.63%. SVM 73.51%. Penelitian tersebut menghasilkan bahwa algoritma K-NN adalah yang terbaik pada studi kasus Bahasa Turki
4	Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma Naïve Bayes Classifier	NBC	(Yulita, Nugroho, & Algifari, 2021)	Penelitian tersebut menganalisa data sebanyak 3780, yang menghasilkan respon 34,4% netral, positif 60,3%, dan 5,4% negatif dengan akurasi 93%.
5	Penerapan Algoritma Naive Bayes untuk Analisis Sentimen Review Data *Twitter* BMKG Nasional	NBC	(Darwis, Siskawati, & Abidin, Penerapan Algoritma Naive Bayes untuk Analisis Sentimen Review Data Twitter BMKG Nasional, 2021)	Penerapan metode NBC pada penelitian tersebut menghasilkan akurasi 68.97%.
6	Aspect-Based Sentiment Analysis of KAI Access Reviews Using NBC and SVM	NBC & SVM	(Mustakim & Priyanta, 2022)	Penerapan metode SVM lebih baik untuk mendapatkan tingkat akurasi yang lebih tinggi. Tingkat akurasi metode SVM adalah 91,63%
7	Sentiment Analysis of Customer Feedback Reviews Towards Hotel"s Products and Services in Labuan Bajo	SVM, NBC, & K-NN	(Christanto & Singgalen)	Dari ketiga metode yang digunakan, K-NN dengan akurasi 85,24%, NBC 78,29%, dan SVM 78,30%. Dengan

No	Judul Penelitian	Metode penelitian	Peneliti	Hasil Penelitian
				menghasilkan kelas positif dan negatif.
8	Online learning sentiment analysis during the covid-19 Indonesia pandemic using *Twitter* data	NBC	(Sahir, Ramadhana, Marpaung, Munthe, & Watrianthos)	Penelitian tersebut meneliti 40.438 data yang menghasilkan kelas netral 1%, positif 25%, negatif 74%.
9	Sentiment Analysis of Online Food Reviews using Big Data Analytics	*Linear Support Vector Classification* (SVC), NBC, *Logistic Regression*	(Ahmed, Awan, Khan, Yasin, & Shehzad)	SVC menampilkan performa terbaik untuk penelitian yang dilakukan dengan akurasi 88,38%.
10	App Review Sentiment Analysis Shopee Application In Google Play Store Using Naive Bayes Algorithm	NBC	(Pratmanto, et al.)	Penelitian menggunakan metode NBC dengan akurasi 96,667% dan menghasilkan kelas positif dan negatif.

Penelitian-penelitian yang telah dilakukan yang menerapkan algoritma NBC untuk melakukan analisis sentimen dan kemudian validasi menggunakan akurasi seperti yang ada di Tabel 2.1 kemudian akan digunakan sebagai acuan untuk metodologi penelitian dalam penelitian ini.