BAB II

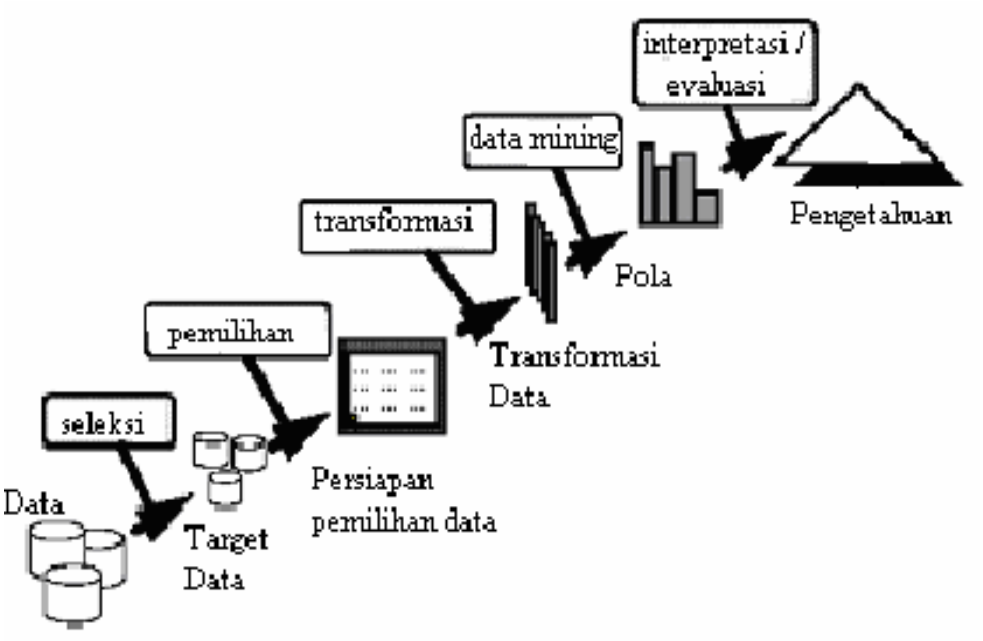
LANDASAN TEORI

2.1. Konsep Teoritis

2.1.1. Data Mining

Mining adalah proses mencari, menemukan, dan menganlisis pola, hubungan, tren, atau informasi yang tersembunyi dalam kumpulan data berukuran besar dengan menggunakan teknik statistik, matemtimati, dan kecerdasan buatan. Tujuan utama data mining adalah menggubah data mentah menjadi informasi yang berupa untuk mendukung pengambilan keputusan.

Data mining biasanya dilakukan melslui beberapa tahapan, seperti pengumpulan data, pembersihan data yang relevan, proses analisis menggunakan algoritma tertentu, serta interpretasi hasil. Metode yang sering digunakan meliputi klasifikasi, klasterisasi (*clustering*), aosiasi, prediksi, dan deteksi anomali.



Gambar 2.1 Fase-fase dalam *data mining*
Sumber : (Arfan, 2020)

Tahapan proses dalam Data *Mining* dapat dijelaskan sebagai berikut :

1. Seleksi Data

   Pemilihan (seleksi) data daru sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses *data mining*, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. *Pre-processing*/*Cleaning* (pemilihan data)

   Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Juga dilakukan proses enrichment, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

3. Transformasi

   Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. *Data Mining*

   Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau

algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. Interpretasi / Evaluasi

   Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut dengan interpretation. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

2.1.2. Naive Bayes

Naive Bayes adalah algoritma klasifikasi dalam machine learning, yang menggunakan teorema Bayes untuk menghitung probilitas suatu data termasuk kedalam kelas tertentu. Theorema Bayes dinyatakan dalam persamaan sebagai berikut :

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

Keterangan :

X : Data dengan kelas yang belum diketahui (tupel/fitur).

H : Hipotesis bahwa X merupakan data dengan kelas spesifik C.

$P(H|X) : Probabilitas posterior (probabilitas hipotesis H terjadi berdasarkan kondisi X).

P(X|H) : Probabilitas kondisional (probabilitas X berdasarkan kondisi hipotesis H).

$P(H)$: Probabilitas prior dari hipotesis H.

P(X) : Probabilitas dari data X (sering kali dianggap sebagai konstanta pembagi).

Berikut adalah tahapan kerja algoritma *Naïve Bayes* :

1. Tahap Asumsi Independensi Fitur (*Naive Assumption*)

Karakteristik utama yang membedakan *Naive Bayes* dari algoritma klasifikasi lainnya adalah penerapan asumsi "naif" (*naive*). Algoritma ini mengasumsikan bahwa seluruh atribut atau fitur yang menjadi parameter klasifikasi bersifat independen atau saling bebas satu sama lain.

Dalam konteks ini, keberadaan atau nilai dari suatu atribut tertentu sama sekali tidak memiliki korelasi dan tidak memengaruhi nilai dari atribut lainnya di dalam kelas yang sama. Meskipun pada realitas data riil sering kali ditemukan keterkaitan antar-atribut, penyederhanaan asumsi ini secara empiris terbukti memangkas kompleksitas komputasi secara signifikan tanpa menurunkan akurasi model secara drastis.

2. Tahap Pembelajaran Data Historis (*Training Phase*)

Pada tahap awal, algoritma melakukan pemindaian terhadap seluruh dataset latih untuk membangun pengetahuan dasar. Proses yang terjadi pada tahap ini meliputi:

1. Analisis Proporsi Kelas : Algoritma menghitung bobot atau frekuensi kemunculan masing-masing kelas target dari total keseluruhan data. Langkah ini bertujuan untuk mengetahui kelas mana yang secara statistik paling dominan pada masa lalu.

2. Analisis Distribusi Karakteristik : Algoritma mengisolasi setiap atribut secara mandiri, kemudian mengukur seberapa sering nilai spesifik dari atribut tersebut muncul pada tiap-tiap kelas target.

3. Tahap Penggabungan Probabilitas Karakteristik

Ketika sistem menerima data baru yang belum diketahui kelasnya, algoritma akan membedah data tersebut berdasarkan atribut-atribut penyusunnya. Selanjutnya, pengetahuan yang telah diperoleh pada tahap pembelajaran akan diterapkan melalui langkah-langkah berikut :

1. Algoritma memanggil kembali data frekuensi kemunculan masing-masing atribut data baru tersebut pada tiap kelompok kelas.
2. Berdasarkan asumsi independensi, tingkat keyakinan untuk setiap kelas dihitung dengan cara mengombinasikan (mengalikan) nilai probabilitas dasar kelompok kelas dengan seluruh nilai probabilitas kemunculan atribut secara simultan.
3. Proses ini dilakukan secara terpisah untuk setiap kelas target yang ada, menghasilkan sebuah skor akumulatif atau bobot keyakinan untuk masing-masing kelas.

4. Tahap Pengambilan Keputusan Akhir (*Classification*)

Tahap akhir dari mekanisme *Naive Bayes* adalah proses komparasi nilai. Algoritma akan membandingkan skor akumulatif yang dihasilkan oleh masing-masing kelas target. Keputusan klasifikasi diambil berdasarkan prinsip *Maximum A Posteriori* (MAP), di mana kelas yang memiliki nilai akumulatif tertinggi atau paling maksimum akan dipilih sebagai hasil prediksi akhir untuk data baru tersebut. Penanganan Kendala Nilai Nol (*Laplace Smoothing*)

Dalam implementasi riil, algoritma *Naive Bayes* rentan terhadap kendala *Zero Probability* (Probabilitas Nol). Masalah ini muncul apabila pada data baru terdapat nilai atribut yang belum pernah terekam sama sekali pada data latih untuk kelas tertentu. Karena mekanisme penggabungan algoritma ini menggunakan operasi perkalian, keberadaan satu saja nilai nol akan menyebabkan skor akhir kelas tersebut menjadi nol secara mutlak, sehingga mengaburkan objektivitas penilaian atribut lainnya.

2.1.3. Sistem Analisis Sentimen

Analisis sentimen (atau sering disebut *opinion mining*) adalah sebuah cabang penelitian di bidang *Natural Language Processing* (NLP) dan *Computational Linguistics* yang berfokus pada proses ekstraksi, identifikasi, dan pengolahan data tekstual secara otomatis untuk mengetahui kecenderungan opini, emosi, atau sikap yang terkandung di dalamnya.

Tujuan utama dari sistem analisis sentimen dalam penelitian adalah untuk mengklasifikasikan polaritas teks ke dalam kategori-kategori tertentu, umumnya berupa sentimen positif, negatif, atau netral. Sistem ini memungkinkan peneliti atau organisasi untuk mengubah data tekstual yang tidak terstruktur (seperti ulasan produk, komentar media sosial, atau artikel berita) menjadi informasi kuantitatif yang dapat dianalisis. tertentu, seperti kata kunci pencarian atau *ID* saluran. Setelah daftar video diperoleh, permintaan tambahan dapat dikirimkan untuk mengambil detail lebih lanjut tentang setiap video, termasuk deskripsi, tag, dan komentar. Data yang diperoleh kemudian dapat disimpan dalam database untuk analisis lebih lanjut atau digunakan langsung dalam aplikasi tertentu.

Salah satu aplikasi utama dari *crawling* data *youtube* adalah analisis sentimen. dengan mengumpulkan dan menganalisis komentar pada video *youtube*, peneliti dapat memahami bagaimana perasaan dan pendapat public tentang topik tertentu. Teknik ini dapat digunakan dalam berbagai konteks, mulai dari analisis sentiman masyarakat terhadap kualitas pelayanan (Agarwal & Sureka, 2015).

Selain analisis sentimen, data yang diperoleh dari *crawling Youtube* juga dapat digunakan untuk pemantauan trend. Dengan menganalisis pola tayangan, suka dan komentar dari waktu ke waktu peneliti dapat mengidentifikasi tren populer dan memahami dinamika yang mendorong popularitas konten tertentu. Informasi ini dapat sangat berharga bagi pembuat konten yang ingin mengoptimalkan strategi mereka untuk menarik lebih banyak penonton dan meningkatkan keterlibatan. Secara umum, sistem analisis sentimen yang berbasis *machine learning* (seperti Naive Bayes, SVM, atau K-Nearest Neighbor) bekerja melalui alur proses yang terstruktur. Proses ini terdiri dari beberapa tahapan utama sebagai berikut :

1. Pengumpulan Data
2. Prapecossing Data
3. Transformasi Data
4. Proses Klasifikasi Niave Bayes
5. Evaluasi Model
6. Analisa Hasil

2.1.4. Naive Bayes Classifier

Naive bayes classifier merupakan suatu metode klasifikasi yang menggunakan teorema bayes dengan asumsi naïve atau sederhana, yaitu

menganggap bahwa setiap fitur yang digunakan untuk klasifikasi adalah independent satu sama lain. Keunggulan utama dari *naive bayes classifier* adalah adanya asumsi yang sangat kuat untuk setiap kondisi atau kejadian. *Naive Bayes Classifier* adalah algoritma pembelajaran mesin berbasis probabilitas yang memanfaatkan Teorema Bayes. Algoritma ini banyak digunakan untuk klasifikasi teks seperti analisis sentimen atau penyaringan spam. Berikut adalah rincian kelebihan dan kekurangannya.

Kelebihan dari *Naive bayes classifier* :

1. Cepat dan Efisien : Algoritma ini membutuhkan waktu komputasi yang sangat singkat dan memori yang sedikit saat pelatihan maupun pengujian
2. Kebutuhan Data Kecil : Hanya memerlukan sedikit data latih (*training data*) untuk memperkirakan parameter klasifikasi
3. Tanggap Terhadap Data Baru : Cepat menyesuaikan dan bekerja dengan baik pada dataset berskala besar
4. Tahan Terhadap *Outlier* **:** Algoritma ini relatif kuat (tidak mudah terdistorsi) jika terdapat data pencilan (*outlier*) atau fitur yang tidak relevan
5. Mudah Diimplementasikan : Konsep matematikanya intuitif dan mudah dipahami

Kekurangan dari *Naive bayes classifier* :

1. Asumsi Dependensi : Mengasumsikan bahwa setiap fitur/variabel bersifat independen (tidak saling berkaitan). Padahal di dunia nyata, antar fitur seringkali memiliki korelasi
2. Potensi Probabilitas Nol : Jika terdapat kombinasi kata atau data dalam pengujian yang tidak pernah muncul di data latih, nilai probabilitasnya

menjadi nol. Hal ini memerlukan teknik khusus seperti *Laplace Smoothing* untuk mengatasinya

3. Tidak Cocok untuk Fitur Numerik yang Rumit : Sangat baik untuk data kategori, namun kurang optimal jika diterapkan pada kumpulan data numerik yang kompleks dan kontinu tanpa penyesuaian distribusi tertentu (seperti *Gaussian*)
4. Kurang Presisi pada Dataset Kompleks : Akurasinya cenderung lebih rendah dibandingkan beberapa algoritma klasifikasi modern lainnya jika dihadapkan pada dataset yang sangat kompleks

Metode ini sangat sesuai untuk digunakan dalam pengklasifikasian dalam tugas akhir, karena memiliki beberapa kelebihan, seperti kesederhanaan, kecepatan dan tingkat akurasi yang tinggi (Assidiq, 2024), secara *matematis naive bayes* memiliki persamaan sebagai berikut :

$$P\ (A|B) = \frac{(B|A)(A)}{1!\ P(B)}$$

Keterangan :

P(A|B) = Probabilitas akhir bersyarat (conditional probability) suatu hipotesis A terjadi jika diberikan bukti (evidence) B terjadi

P(B|A) = Probabilitas sebuah bukti B terjadi akan memengaruhihipotesis

A P(A) = Probabilitas awal hipotesis A terjadi tanpa memandang bukti apapun

P(B) = Probabilitas awal bukti B terjadi tanpa memandang bukti yang lain. Selanjutnya dapat disederhanakan menjadi :

$$P(A|B) \propto P(B|A)P(A)$$

Pada saat dilakukan klasifikasi, algoritma akan mencari nilai probabilitas paling tinggi dari semua kategori dokumen yang diujikan (VMAP). Adapun persamaan VMAP sebagai berikut :

$$VMAP = \arg\max (Vj|x1, x2, \dots, xk)$$

dimana xk merupakan kata kunci ke -k dan dengan menerapkan teorema *naive bayes* pada persamaan (1) maka persamaan (3) dapat ditulis menjadi :

$$VMAP = \arg max \frac{(B|A)P(A)}{1!(B)}$$

Karena nilai P(x1, x2,…. xk) untuk semua nilai Vj sama berarti nilai Vj dapat diabaikan, maka persamaan (4) dapat ditulis menjadi :

$$VMAP \propto \arg\max P(Vj)(x1, x2 \dots xk|Vj)$$

Setiap kata dalam (x1, x2, …, xk) diasumsikan bersifat independent sehingga selanjutnya dapat diberikan persamaan :

$$P(x_1, x_2 \dots x_k|V_j) = \prod_{k=1}^{n} P(x_k|V_j)$$

Maka akan didapatkan pendekatan yang dipakai dalam *naive bayes classifier* sebagai berikut :

$$V_{MAP} = \arg\max P(V_j) \prod_{k=1}^{n} P(x_k|V_j)$$

VMAP merupakan nilai output hasil klasifikasi naïve bayes. Nilai P(Vj) dapat dihitung menggunakan persamaan :

$$P(V_j) = \frac{|doc_j|}{training}$$

Dimana |docj| adalah jumlah komentar *YouTube* yang memiliki kategori j dalam training. Untuk |*training*| adalah jumlah komentar *YouTube* dalam contoh yang digunakan training. Untuk setiap probabilitas kata xk dalam setiap kategori dihitung pada saat training dengan persamaan :

$$P(x_k|V_j) = \frac{n_k + 1}{|n + kosakata|}$$

nk merupakan jumlah kemunculan pada kata xk dalam komentar Youtube berkategori Vj sedangkan n adalah banyaknya seluruh kata dalam komentar Youtube dengan kategori Vj dan |kosakata| merupakan banyaknya kata dalam training.

2.2. Alat Bantu Implementasi

Dalam penelitian Analisis Sentimen Masyarakat Terhadap Kualitas Pelayanan Infrastruktur Menggunakan Metode Naive Bayes pada Dinas Pekerjaan Umum dan Tata Ruang Kabupaten Labuhanbatu Utara, digunakan beberapa perangkat lunak untuk mendukung proses pengolahan dan analisis data. Salah satu perangkat lunak utama yang digunakan adalah RapidMiner.

Gambar 2.2 RapidMiner
Sumber:linkedin.com

RapidMiner adalah platform analisis data yang memungkinkan pengguna melakukan pemodelan prediktif tanpa memerlukan keahlian pemrograman yang

mendalam. RapidMiner dikembangkan dengan pendekatan open core, yang menggabungkan elemen-elemen opensource dengan fitur-fitur tambahan yang tersedia dalam versi berbayar.

2.3. Penelitian Terdahulu

Penelitian terdahulu digunakan sebagai bahan referensi dalam penyusunan penelitian yang mendorong penulis untuk melakukan penelitian yang sama. Penelitian terdahulu yang digunakan adalah penelitian yang memiliki relevansi dengan tema yang diangkat yaitu mengenai Analisis Sentimen Masyarakat Terhadap Kualitas Pelayanan Infrastruktur Menggunakan *Naives Bayes* Pada Data di Dinas Pekerjaan Umum dan Tata Ruang, Kab. Labuhanbatu Utara.

Beberapa bahan yang penulis ambil dari penelitian terdahulu terdapat pada tabel 2.1 berikut ini :

Tabel 2.1 Penelitian Terdahulu

No	Nama Peneliti/Tahun	Judul	Hasil Penelitian
1	(Misrun et al., 2024)	Analisis Sentimen Komentar Youtube Terhadap Anies Baswedan Sebagai Bakal Calon Presiden 2024 Menggunakan Metode *Naive Bayes Classifier*	Pada penelitian ini didapatkan hasil akurasi menggunakan algoritma naive bayes classifer sebesar 78%
2	(N. Sari et al., 2024)	Analisis Sentimen Terhadap Kandidat Calon Presiden Berdasarkan Tweets Di Sosial Media Menggunakan Naive Bayes Classifier	Dari hasil proses menggunakan *naïve bayes classifier* didapatkan akurasi tertinggi 86,38%.

3	(Abdillah et al., 2024)	Analisis Sentimen Terhadap Kandidat Calon Presiden Berdasarkan Tweets Di Sosial Media Menggunakan Naive Bayes Classifier	Dari hasil penelitian terhadap empat kandidat, didapatkan hasil untuk Anies Baswedan 74% sentimen positif dan 26% sentimen negatif. Diikuti oleh Sandiaga Uno dengan 57% sentimen positif dan 43% sentimen negatif. Kemudian Ganjar Pranowo dengan 53% sentimen positif dan 47% sentimen negatif. Dan yang terakhir yaitu Prabowo Subianto yang mendapatkan 32% sentimen positif dan 68% sentimen negatif
4	(Morgan et al., 2022a)	Analisis Sentimen Pengguna Youtube Terhadap Tayangan Matanajwamenantiterawan Dengan Metode *Naïve Bayes Classifie*	Didapatkan hasil akurasi sebesar 90,36%.
5	(F. V. Sari, 2019)	Analisis Sentimen Pelanggan Toko Online JD.ID menggunakan Metode *Naïve Bayes Classifier* Berbasis	Hasil penelitian menunjukkan bahwa metode naïve bayes classifier tanpa

		Konversi Ikon Emosi	penambahan fitur dapat mengklasifikasi sentimen dengan akurasi mencapai 96,44%. Namun, dengan penambahan fitur berupa pembobotan tf-idf dan konversi ikon emosi, terjadi peningkatan signifikan dalam nilai akurasi menjadi 98%.