BAB II

LANDASAN TEORI

2.1. Landasan Teori

2.1.1. Analisis Sentimen

Analisis sentimen adalah suatu proses yang digunakan untuk mengidentifikasi opini, emosi, dan sikap yang tercermin dalam teks, yang dimana diklasifikasikan kembali menjadi opini negatif dan positif (Cindo M, 2019). Analisis sentimen pada suatu kalimat mencerminkan aspek penilaian terhadap entitas atau peristiwa tertentu. Analisis sentimen menggunakan pemrosesan bahasa alami untuk mendapatkan pandangan pelanggan mengenai produk atau topik tertentu.

Analisis sentimen merupakan sistem yang digunakan untuk mengumpulkan dan mengevaluasi pendapat mengenai produk atau layanan yang terdapat pada postingan web, blog, atau komentar di platform media sosial. Penggunaan analisis sentimen dapat digunakan pada beragam jenis opini, termasuk opini negatif, opini positif, opini umum, ulasan film, isu politif, merek produk, dan lainnya yang ada di media sosial.

Analisis sentimen dimulai dengan membuat data latih, yang selanjutnya diproses menggunakan suatu algoritma klasifikasi khusus untuk mengembangkan model klasifikasi. Pembuatan data latih dimulai dari informasi awal berupa sejumlah komentar yang telah melalui proses pelabelan pada tahap preprocessing.

2.1.2. Data Mining

Data mining adalah proses pengumpulan informasi penting dari suatu data. Informasi penting ini diperoleh melalui proses yang kompleks seperti penggunaan kecerdasan buatan, teknik statistik, matematika, *machine learning*, dan lain sebagainya (Sudarsono B G, 2021).

Data mining adalah proses mengekstraksi data tersembunyi dari dalam database (Indriyani F, 2019). Data mining digunakan untuk mengolah data berukuran besar dalam database, dengan tujuan menghasilkan informasi baru yang dapat digunakan dalam strategi bisnis.

Data mining secara umum dibagi menjadi dua kategori utama, yaitu:

1. Deskriptif Mining
   Deskriptif Mining digunakan untuk mengidentifikasi karakteristik data, sementara
2. Prediktif Mining
   Prediktif Mining digunakan untuk menemukan pola data.

2.1.3. Instagram

Instagram adalah sebuah platform komunikasi yang relatif baru yang memungkinkan pengguna untuk berbagi foto dan video (Finandra S, 2021). Sejak diluncurkan pada bulan Oktober 2010, terjadi pertumbuhan yang pesat dalam jumlah pengguna di seluruh dunia. Saat ini, Instagram tidak hanya berperan sebagai sarana hiburan dan interaksi, tetapi juga dalam bisnis. Menurut Al Ghamdi dan Reilly , 83% pemasar sangat menghargai platform sosial ini karena peran penting mereka dalam menjangkau dan mempertahankan pelanggan serta menciptakan peluang bisnis..

2.1.4. *Crawling*

Crawling adalah metode pengumpulan data dari sebuah situs web dengan memasukkan alamat Uniform Resource Locator (URL) (Arsi P, 2021). URL ini digunakan sebagai acuan untuk menemukan semua tautan yang terdapat di situs web. Selanjutnya, dilakukan indeksasi untuk mencari kata-kata dalam dokumen yang terkait dengan setiap tautan yang ada.

2.1.5. *Preprocessing* Data

Preprocessing Data adalah tahap dimana teks yang akan diklasifikasi akan dibersihkan dan dipersiapkan sebelum dilakukan analisis dokumen (Habibi M, 2019). Hal ini dilakukan untuk mencegah kekurangan data, gangguan pada data, dan ketidak konsistenan dalam data. Data yang telah di ambil dari Instagram adalah data mentah yang perlu diproses sebelum dapat digunakan lebih lanjut. Oleh karena itu, tahapan *preprocessing* diperlukan untuk mempersiapkan data tersebut sebelum memasukkannya ke dalam proses berikutnya. Tahapan *preprocessing* yang dilakukan yaitu:

1. *Cleaning*

   Cleaning adalah proses menghilangkan tanda baca, simbol, dan huruf besar menjadi huruf kecil dan angka. Hal ini bertujuan untuk meningkatkan data menjadi lebih efisien dan dapat diolah dengan lebih baik.

2. *Case Folding*

   Case Folding adalah proses pengubahan seluruh huruf kapital *(uppercase)* menjadi huruf kecil *(lowercase)* agar sama.

3. *Tokenization*

   Tokenization adalah proses yang digunakan dengan cara memecah kalimat menjadi beberapa bagian yang dikenal

sebagai token. Jika kata tersebut dibuang maka tidak akan mengubah ataupun mengurangi informasi yang terkandung dalam kalimat tersebut. Misalnya kata hubung “yang”, “akan”, “pada”, “di”, dan lain-lain.

4. Normalisasi

   Normalisasi adalah proses untuk menormalkan bentuk kata. Pada tahap normalisasi yaitu mengubah bentuk kata yang kurang normal seperti singkatan, bahasa asing, dan kata yang kurang baku dalam *dataset* menjadi kata baku.

5. *Stopword Removal*

   Stopword Removal adalah proses di mana kata-kata yang dianggap tidak penting atau kurang relevan dihapus dari teks. Misalnya kata-kata seperti “dan”, “atau”, “kamu”, dan “saya”.

6. *Stemming*

   Stemming adalah proses yang mengubah kata ke dalam struktur dasarnya dengan menghapus gabungan sebelumnya dan menempatkan kata setelahnya. Pada proses ini akan mengubah kembali semua kata menjadi kata dasar. Misalnya contoh kata “pembohong” menjadi “bohong”.

2.1.6. *Naive Bayes Classifier*

Naive Bayes adalah algortima yang digunakan untuk mengklasifikasikan suatu variable tertentu dengan memanfaatkan metode probabilitas dan statistic (Khotimah, 2022). Dengan bergantung pada model probalitasnya, klasifikasi *Naive Bayes* dapat dilatih secara efisien dalam *supervised learning.* Salah satu keunggulan *Naive Bayes* adalah bahwa ia tidak membutuhkan jumlah data *training* (data pelatihan) yang besar.

Adapun beberapa manfaat dari penggunaan metode *Naïve Bayes Classifier* meliputi:

a. Pengklasifikasian teks dokumen seperti berita atau teks akademis.
b. Penggunaan probabilitas dalam metode *machine learning.*
c. Otomatisasi diagnosis medis.
d. Deteksi dan penyaringan spam.
e. Memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya.

Algortima ini berdasarkan Teorema Bayes dan mengasumsikan bahwa setiap pasangan fitur dalam data pelatihan adalah independen satu sama lain. Proses algoritma Naive Bayes melibatkan pemrosesan data pelatihan untuk menciptakan model klasifikasi, yang nantinya dapat digunakan untuk memprediksi label kelas yang sesuai untuk data yang belum pernah dilihat sebelumnya.

Secara umum teorema Bayes dinotasikan pada persamaan berikut ini:

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

Keterangan:

B = Class data yang dimasukkan.

A = Data hipotesis.

P(A|B) = Probabilitas hipotesis A yang mengacu pada kondisi B.

P(A) = Probabilitas hipotesis A.

P(B|A) = Probabilitas B berdasarkan kondisi pada hipotesis A.

P(B) = Probabilitas B.

2.1.7. TF-IDF

TF-IDF (*Term Frequency-Inverse Document Frequency*) adalah algoritma yang memberikan nilai bobot untuk kata-kata dalam teks. Frekuensi suatu kata dalam suatu dokumen (TF) dan nilai invers dari jumlah dokumen yang mengandung kata tersebut (IDF) dihitung terlebih dahulu. Hasil perkalian antara TF dan IDF menghasilkan nilai bobot yang mencerminkan signifikansi kata dalam teks. Hal itu dapat dilihat pada persamaan berikut ini:

$$TF - IDF(d,t) = TD(d,t) \times IDF(t)$$

Dimana:

1. *Term Frequency* (TF)

   Untuk mengukur seberapa sering sebuah term yang muncul dalam suatu dokumen, dapat dilihat pada persamaan berikut:

   $$TF(d,t) = \frac{\text{Jumlah kata t pada dokumen d}}{\text{Total kata pada dokumen d}}$$

2. *Inverses Document Frequency* (IDF)

   Untuk mengukur keunikan suatu kata dengan memperhitungkan jumlah dokumen dalam kumpulan data yang memuat istilah tersebut dapat dilihat pada persamaan berikut:

   $$ID(t) = \log \frac{\text{Total dokumen}}{\text{Jumlah dokumen yang mengandung kata t}}$$

   Keterangan:

   t = kata

   d = dokumen

2.2.Penelitian Terdahulu

Tabel 2.1. Penelitian Terdahulu

NO	Peneliti (Tahun Penelitian)	Judul Penelitian	Dataset	Hasil
1	Finandra S, et al. (2021).	Penerapan Analisis Sentimen Melalui Data Instagram Untuk Mengetahui Reputasi Wisata Kuliner Di Kota Bandung Menggunakan Metode Klasifikasi *Naive Bayes*.	Menggunakan 864 dataset komentar dan Kiriman postingan	Algoritma Naive Bayes menunjukkan hasil akurasi yang lebih tinggi daripada algoritma *Decision Tree* dan K-NN, dengan tingkat akurasi mencapai 86.87% yang telah diuji menggunakan perangkat RapidMiner. Evaluasi performansi algoritma ini juga menghasilkan hasil yang memuaskan, termasuk Precision sekitar 93%, Recall sekitar 57.50%, yang akhirnya menghasilkan F1-Measure sekitar 75.39%. Hasil tersebut diperoleh

				melalui pembagian data dengan rasio perbandingan 70:30.
2	Khotimah, A.C dan Utami, E (2022).	*Comparison Naïve Bayes Classifier, K-Nearest Neighbor And Support Vector Machine In The Classification Of Individual On Twitter Account*	Menggunakan dataset 130 akun *twitter*.	Hasil evaluasi perhitungan dengan menggunakan confusion matrix menunjukkan bahwa persentase akurasi dari perbandingan antara hasil klasifikasi psikolog secara manual dan hasil klasifikasi sistem adalah 31.5% untuk algoritma Naïve Bayes Classifier, 23.8% untuk algoritma K-Nearest Neighbor, dan 28.5% untuk algoritma Support Vector Machine.
3	Syaifudin Y. W, I. R. (2018)	JMO Clustering Dan Sentimen Data Twitter Pada Opini Wisata Pantai Menggunakan Metode *K*-	Menggunakan dataset dari 10 pantai yang ada di Indonesia sebanyak 500 tweet. Data	Algoritma Support Vector Machine mencapai tingkat akurasi sebesar 74,39% dalam proses klasifikasi. Selanjutnya, data

		Means	tersebut diperoleh melalui Twitter dalam Bahasa Indonesia.	opini yang diperoleh dari kuesioner digunakan untuk mengklasifikasikan pantai berdasarkan berbagai faktor seperti ketersediaan sumber daya, fasilitas, aksesibilitas, kesiapan masyarakat, potensi pasar, dan posisi pariwisata.Proses pengelompokan data ini dilakukan menggunakan metode K-Means.
4	Putra, A. D & Juanita, S (2021)	Analisis Sentimen Pada Ulasan Pengguna Aplikasi Bibit Dan Bareksa Dengan Algoritma KNN	Dataset yang diambil yaitu berupa data kumpulan opini seperti review agen travel, review restoran, review produk, dan lain-lain	Hasil penelitian menunjukkan bahwa dalam proses klasifikasi model menggunakan algoritma k-nearest neighbors, dengan pembagian data pada rasio 60:40 terhadap dataset ulasan pengguna di aplikasi Bibit dan Bareksa,

				menghasilkan nilai akurasi, presisi, dan recall sebesar 85,14%, 91,91%, dan 76,44% untuk Bibit. Sedangkan untuk Bareksa, nilai tersebut adalah 81,70%, 87,15%, dan 75,73%.
5	Sari S, et al (2021)	*Sentiment Analysis Against Beauty Shaming Comments on Twitter Social Media Using SentiStrength Algorithm*	Dataset yang digunakan yaitu data tweet yang berjumlah 300 tweets yang diambil dari proses crawling data	Hasil dari penelitian menggunakan algoritma *SentiStrength* menunjukkan tingkat akurasi sebesar 60%.
6	Aditia Yudhistira, R. A. (2023)	Pengelompokan Data Nilai Siswa Menggunakan Metode *K-Means Clustering*	Dataset yang diambil yaitu data siswa dalam satu semester.	Hasil penelitian tentang pengelompokan data nilai siswa menggunakan metode *K-Means clustering* menunjukkan bahwa berdasarkan hasil *cluster* data siswa yang digunakan

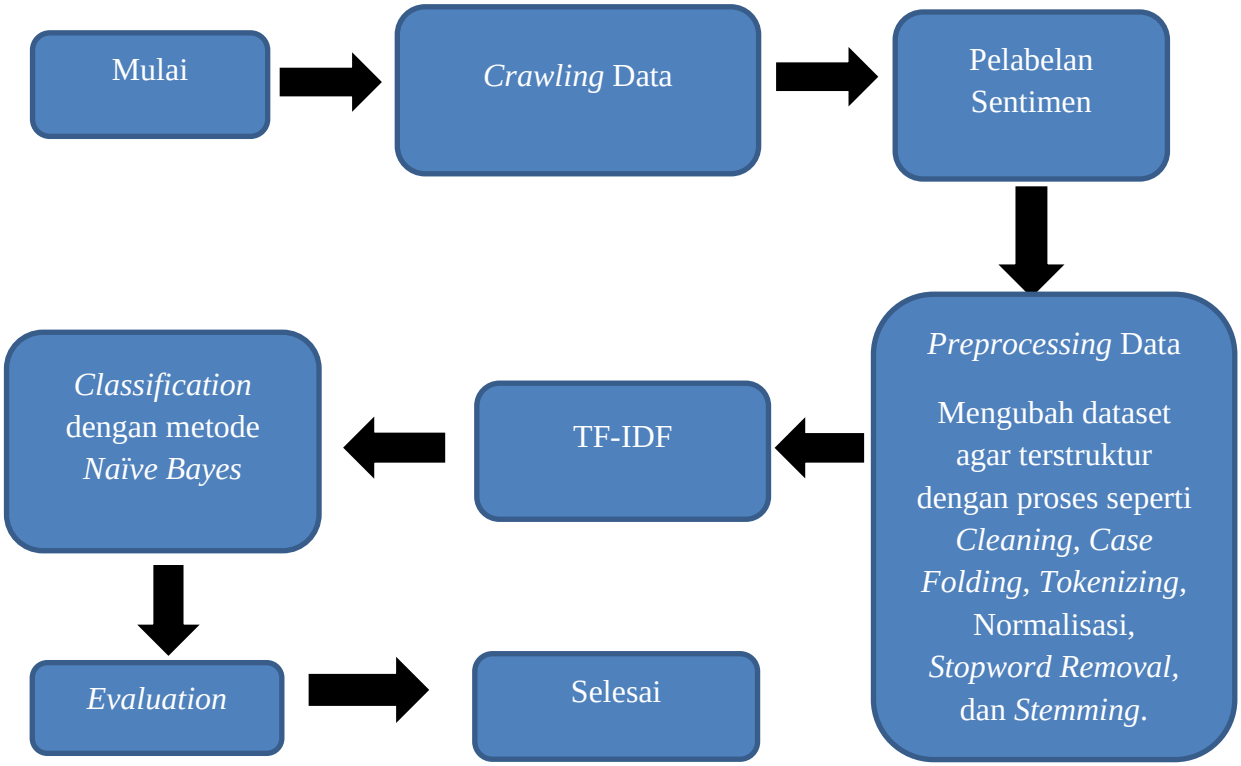
				selama satu semester, *cluster* 0 terdiri dari 59 siswa, *cluster* 1 terdiri dari 94 siswa, dan *cluster* 2 terdiri dari 1 siswa. Berdasarkan hasil pengujian dengan metode elbow, jumlah *cluster* terbaik yang digunakan adalah tiga *cluster*, sehingga dalam penelitian ini digunakan *cluster* 0, *cluster* 1, dan *cluster* 2.
7	Alita D, (2020)	Pendeteksian Sarkasme pada Proses Analisis Sentimen Menggunakan *Random Forest Classifier*	Dataset yang diambil yaitu data mengenai jaringan telekomunikasi seluler, layanan BPJS, dan layanan PLN dengan jumlah 2027 data tweet.	Hasil pengujian deteksi sarkasme menggunakan metode *Random Forest Classifier* menunjukkan nilai akurasi 60,61%.
8	Fikri M, et	Perbandingan	Dataset yang	Hasil penelitian ini

	al (2020)	Metode *Naïve Bayes* dan *Support Vector Machine* pada Analisis Sentimen Twitter	digunakan yaitu data pada tweet yang mengandung kata "Universitas Muhammadiyah Malang", "muhammadiyah", "UMM", atau "unmuh" dari tahun 2018-2019	mengindikasikan bahwa Naïve Bayes memiliki tingkat akurasi yang lebih tinggi, mencapai 73,65%, dibandingkan dengan SVM yang memiliki tingkat akurasi sebesar 70,20%.
9	Khairunnisa S, et al (2021)	Pengaruh Text *Preprocessing* terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)	Dataset yang digunakan yaitu data *tweet* yang berkaitan dengan COVID-19. Data tersebut diambil dari bulan Maret-April 2020.	Hasil dari penelitian yang menggunakan *Support Vector Machine* menunjukkan tingkat akurasi sebesar 77,77%.
10	Pilar (2023)	A novel flexible feature extraction algorithm for Spanish tweet	Dataset yang digunakan yaitu 1.899 tweet.	Hasil dari penelitian ini yaitu klasifikasi polaritas 1.899 tweet ke dalam kategori positif, negatif, dan

		sentiment analysis based on the context of words		netral. Dengan menggunakan korpus interTASS diperoleh 54.39% dan akurasi 61,35%.

2.3. Kerangka Pemikiran

Kerangka pemikiran merupakan landasan utama dari suatu penelitian yang diperoleh dari data empiris, pengamatan, dan studi literatur (Syahputri, 2023). Dengan demikian, kerangka berpikir mengandung teori, prinsip, atau konsep-konsep yang menjadi dasar bagi pelaksanaan penelitian. Rancangan Penelitian yang akan dilaksanakan, akan disusun dalam bentuk kerangka pemikiran sebagai berikut:



Gambar 2.1 Kerangka Pemikiran

Keterangan:

1. *Crawling* Data

   Pada tahap ini akan dilakukan penarikan dataset komentar dengan cara *crawling* menggunakan URL postingan Instagram @mie.gacoan untuk menarik data.

2. Pelabelan Sentimen

   Pada tahap ini dataset yang telah diambil akan dilakukan proses pelabelan secara manual dan akan diberi label positif dan negatif.

3. *Preprocessing* Data

   Pada tahap ini data akan diolah dengan *Preprocessing* berupa *Cleaning, Case Folding, Tokenizing,* Normalisasi, *Stopword Removal*, dan *Stemming*. Hal tersebut bertujuan untuk memperbaiki kualitas data tersebut, menghilangkan permasalahan yang bisa saja terjadi pada saat pemrosesan data, dan membuat dataset menjadi lebih efektif dan efisien.

4. TF-IDF

   Pada tahap TF-IDF dilakukan pembobotan terhadap kata-kata dalam komentar untuk mengetahui bobot dari setiap kata tersebut.

5. *Classification*

   Pada tahap *Classification* akan dilakukan pengklasifikasian data dengan menggunakan algoritma *Naïve Bayes.*

6. *Evaluation*

   Pada tahap *Evaluation* dilakukan perhitungan *accuration, precision, recall,* dan F1-Score dengan *confusion matrix* pada komentar yang telah diklasifikasikan dengan algoritma *Naïve Bayes.* Hal ini bertujuan untuk mengetahui keakuratan dokumen setelah diklasifikasi menjadi sentiment positif dan sentiment negatif. BAB III