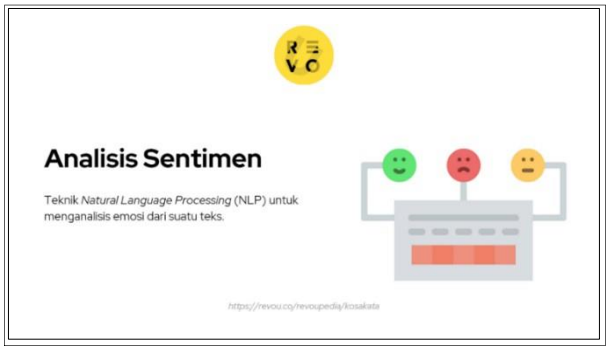
BAB 2
TINJAUAN PUSTAKA

2.1 Landasan Teori

2.1.1 *Analisis Sentimen*

Analisis sentimen merupakan metode yang digunakan untuk mengidentifikasi kecenderungan pendapat atau sikap yang terkandung dalam sebuah teks. Pada penelitian ini, metode tersebut diterapkan untuk menilai tanggapan pengguna terhadap layanan keuangan dengan mengelompokkan feedback ke dalam tiga kelas, yaitu positif, negatif, dan netral. Penerapannya telah digunakan dalam berbagai sektor, seperti pemasaran, pelayanan pelanggan, dan kajian sosial, karena dapat membantu memahami respons pengguna terhadap suatu produk maupun layanan.



Gambar 2.1 Analisis Sentimen

Menurut Purnamasari dkk. (2023) ada 3 metode yang digunakan untuk melakukan analisis sentimen, diantaranya:

1) *Machine Learning Based Approach*, merupakan algoritma *machine learning* yang melatih klasifikasi dari data yang diberi label secara manual.
2) *Lexicon-Based Approach*, merupakan metode analisis sentimen yang digunakan untuk mendeskripsikan polaritas (positif, netral, dan negatif) dari sebuah teks.
3) *Hybrid Approach*, merupakan penggabungan antara *machine learning* dan *metode lexicon based*.

2.1.2 *Text Mining*

Text mining merupakan pendekatan untuk mengekstraksi informasi dari kumpulan teks yang belum terstruktur dan berasal dari topik tertentu. Melalui teknik ini, data teks dapat diolah sehingga menghasilkan informasi yang relevan dan bernilai (Afdal dkk., 2022). Selain itu, Hermawan dkk. (2023) menjelaskan bahwa *text mining* dapat dimanfaatkan untuk menganalisis sentimen dalam kalimat secara cepat, sehingga proses memperoleh informasi dari data tekstual menjadi lebih efektif.

2.1.3 *Text Preprocessing*

Preprocessing merupakan tahap awal untuk mengolah data mentah agar menjadi lebih teratur, bersih, dan siap digunakan pada analisis berikutnya (Salam dkk., 2023). Daniswara dan Nuryana (2023) menyatakan bahwa tahap ini penting dalam data mining karena mencakup kegiatan pembersihan data, penyesuaian format, serta penyiapan data untuk meningkatkan kemudahan dan ketepatan analisis. Dalam pengolahan teks, beberapa tahapan yang umum diterapkan adalah sebagai berikut:

1) *Case folding*, yaitu menyeragamkan seluruh karakter teks ke dalam bentuk huruf kecil.
2) *Tokenization*, yaitu membagi teks menjadi unit-unit kata secara terpisah.
3) *Stopword removal*, yaitu menghilangkan kata umum yang dianggap kurang memberikan informasi bagi analisis sentimen, misalnya “yang”, “dan”, dan “di”.
4) *Stemming*, yaitu mengembalikan kata yang memiliki imbuhan ke bentuk akarnya. Proses ini dapat dilakukan menggunakan Sastrawi, sebuah modul Python yang dirancang untuk mengubah kata berimbuhan bahasa Indonesia menjadi kata dasar (Himawan & Eliyani, 2021).

2.1.4 *Python*

Python dikembangkan oleh Guido van Rossum dan pertama kali diperkenalkan pada tahun 1991 (Alfarizi dkk., 2023). Bahasa pemrograman ini banyak dimanfaatkan dalam pembuatan beragam aplikasi, baik untuk platform desktop, web, maupun perangkat mobile (Nazar R, 2024). Python bekerja secara interpretatif dan menyediakan berbagai fitur yang mendukung pengembangan

aplikasi. Selain itu, rancangan sintaksnya menekankan kemudahan membaca serta memahami kode program (Semendawai dkk., 2021).

2.1.5 *Flask Framework*

Penelitian Haezer dkk. (2022) menjelaskan bahwa *Flask Framework* adalah kerangka kerja pengembangan web yang dibangun menggunakan bahasa pemrograman Python. *Framework* ini membantu pengembang menyusun struktur aplikasi sekaligus mengatur tampilan dan perilaku halaman web. *Flask* termasuk kategori *microframework* karena hanya menyediakan komponen inti yang diperlukan, sedangkan fitur tambahan dapat dipasang sesuai kebutuhan. Pendekatan tersebut membuat aplikasi lebih fleksibel dan menghindari penggunaan komponen yang tidak diperlukan sehingga kinerja *framework* tetap ringan.

2.1.6 Xampp

XAMPP digunakan sebagai lingkungan server lokal untuk mengembangkan, menjalankan, dan menguji aplikasi web. Paket ini telah dilengkapi berbagai komponen, seperti Apache untuk menangani layanan web, PHP sebagai bahasa pemrograman di sisi server, serta MySQL untuk mengelola basis data. Karena seluruh kebutuhan tersebut tersedia dalam satu instalasi, konfigurasi server dapat dilakukan dengan lebih praktis. Melalui akses localhost, aplikasi beserta databasenya dapat dijalankan tanpa bergantung pada server daring. Pada penelitian ini, XAMPP dimanfaatkan selama tahap pengembangan dan pengujian aplikasi, termasuk untuk menjalankan sistem serta mengatur database (Lesmana & Silalahi, 2022).

2.1.7 *Pembobotan Kata*

Menurut penelitian (Wati dkk., 2023) menjelaskan bahwa pembobotan kata diperlukan untuk merepresentasikan teks dalam bentuk angka agar dapat diproses oleh algoritma. Beberapa pendekatan yang dapat digunakan dalam tahap ini antara lain Bag of Words, N-Gram, Word2Vec, dan TF-IDF. Dalam penelitian ini, metode TF-IDF dipilih untuk menentukan tingkat kepentingan suatu kata berdasarkan kemunculannya dalam dokumen. Komponen Term Frequency (TF) menunjukkan seberapa sering sebuah kata muncul pada dokumen tertentu. Perhitungan nilai TF dilakukan menggunakan rumus berikut:

$$\text{TF} = \frac{(Jumlah\ kemunculan\ kata\ dalam\ dokumen)}{(Jumlah\ kata\ dalam\ dokumen)}$$

Meskipun demikian, frekuensi kemunculan saja belum cukup untuk menunjukkan seberapa penting suatu kata dalam kumpulan dokumen. Oleh sebab itu, digunakan komponen Inverse Document Frequency (IDF). Komponen ini memberikan nilai bobot yang lebih besar kepada kata-kata yang kemunculannya relatif jarang pada keseluruhan dokumen. Berikut rumus untuk menghitung bobot IDF:

$$\text{IDF} = \log \frac{N}{n}$$

Dalam perhitungan tersebut, N menyatakan banyaknya dokumen secara keseluruhan, sedangkan n menunjukkan jumlah dokumen yang memuat istilah tertentu. Selanjutnya, nilai TF-IDF diperoleh dengan mengalikan skor Term Frequency (TF) dan Inverse Document Frequency (IDF). Rumus perhitungannya dapat dilihat sebagai berikut:

W = TF x IDF

W = Bobot suatu kata dalam dokumen.

TF = Frekuensi kemunculan kata dalam dokumen.

IDF = Bobot kepentingan kata berdasarkan jumlah dokumen yang mengandung kata tersebut.

Kata dengan skor TF-IDF tinggi menunjukkan bahwa kata tersebut relatif penting dalam dokumen dan memiliki pengaruh lebih besar terhadap representasi atau topik yang dibahas.

2.1.8 *N-Gram*

N-gram merupakan salah satu fitur dalam text preprocessing yang dimanfaatkan untuk menangkap hubungan antarkata dalam klasifikasi sentimen. Fitur ini membentuk kelompok kata yang berdekatan sehingga frasa yang mengandung makna sentimen dapat dikenali dengan lebih baik (Indarbensyah & Rochmawati, 2021). Berdasarkan jumlah kata yang digabungkan, N-gram dibedakan menjadi tiga bentuk, yaitu unigram yang memuat satu kata, bigram yang terdiri atas dua kata, dan trigram yang menggabungkan tiga kata.

2.1.9 SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) merupakan metode yang digunakan untuk menyeimbangkan jumlah data pada kelas sentimen yang memiliki jumlah data berbeda. Metode ini dilakukan dengan membuat data baru berdasarkan karakteristik data dari kelas yang jumlahnya lebih sedikit. Dalam

penelitian ini, SMOTE diterapkan pada data training sebelum proses klasifikasi menggunakan algoritma Naïve Bayes. Penerapan SMOTE digunakan untuk membantu mengurangi ketimpangan jumlah data antar kelas sehingga model dapat melakukan klasifikasi terhadap setiap kategori sentimen dengan lebih baik.

2.1.10 *Lexicon Based*

Lexicon Based merupakan metode yang digunakan untuk melakukan analisis sentimen dengan menggunakan pendekatan kamus dan korpus. *Lexicon Based* dapat menghitung polaritas suatu kata dengan rumus sederhana. Metode ini dapat mengkategorikan sebuah kalimat berdasarkan sifat positif, netral, atau negatif (Subagio dkk., 2024). Dimana untuk dapat mengkategorikan kalimat dilihat dari hasil polarity score nya, berikut rumus perhitungan polarity:

$$S = \{Negatif,\ jika\ P < 0\ Netral, jika\ P = 0\ Positif,\ jika\ P > 0$$

Keterangan:

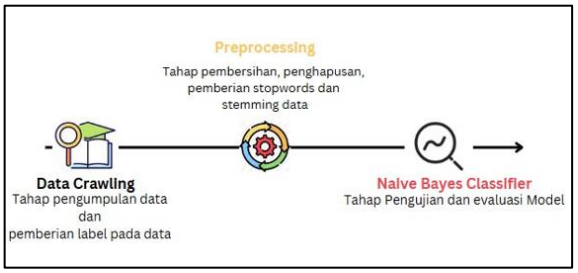
S = Sentimen

P = Polarity Score

1) Jika Polarity kecil dari 0, maka sentimen adalah Negatif.
2) Jika Polarity sama dengan 0, maka sentimen adalah Netral.
3) Jika Polarity besar dari 0, maka sentimen adalah Positif.

2.1.11 *Algoritma Naïve Bayes*

Algoritma Naïve Bayes merupakan teknik klasifikasi yang menerapkan Teorema Bayes dengan menganggap setiap fitur tidak saling bergantung. Pada analisis sentimen, metode ini menghitung peluang suatu teks untuk kemudian menentukan kelas sentimen yang paling sesuai. Walaupun menggunakan asumsi independensi yang sederhana, Naïve Bayes tetap mampu menghasilkan performa yang baik, khususnya ketika digunakan pada kumpulan data berukuran besar. Selain itu, proses komputasinya yang ringan dan cepat menjadikan algoritma ini banyak digunakan dalam pengelompokan teks.



Gambar 2.2 *Naïve Bayes*

2.2 Penelitian Terdahulu

Berdasarkan jurnal Andriawan & Ernawati (2024) yang berjudul "Penggunaan Algoritma Naïve Bayes dan Support Vector Machine untuk Analisis Sentimen Konflik Palestina dan Israel pada Platform X", penelitian ini membandingkan dua algoritma dalam klasifikasi sentimen terkait konflik Palestina dan Israel, yaitu Naïve Bayes dan Support Vector Machine (SVM). Hasil penelitian menunjukkan bahwa Naïve Bayes lebih unggul dalam menangkap pola sentimen yang terkandung dalam data. Algoritma ini memberikan hasil klasifikasi yang lebih baik dibandingkan SVM, terutama dalam menangani isu-isu sensitif seperti konflik Palestina dan Israel. Dengan demikian, Naïve Bayes dapat digunakan sebagai dasar yang lebih kuat untuk mendukung pengambilan kebijakan yang lebih tepat dalam menangani konflik ini.

Selanjutnya, dalam jurnal Kumala & Wibowo (2023) berjudul "Product Overview Sentiment Analysis Using Lexicon Hybrid-Based Approach and Machine Learning", empat algoritma machine learning dibandingkan, yaitu SVM, Naïve Bayes, Logistic Regression, dan Random Forest, untuk analisis sentimen pada ulasan game. Hasil eksperimen menunjukkan bahwa klasifikator terbaik untuk ulasan game adalah kombinasi hybrid lexicon dan Random Forest, dengan tingkat akurasi mencapai 95%. Selain itu, penelitian ini membuktikan bahwa kombinasi antara metode lexicon-based dan machine learning menghasilkan hasil yang lebih baik.

Sementara itu, penelitian Parhusip dkk. (2023) yang berjudul "Sentiment Analysis of the Public Towards the Kanjuruhan Tragedy with the Support Vector Machine Method" menganalisis sentimen publik terhadap tragedi Kanjuruhan menggunakan Support Vector Machine (SVM) dengan data dari Twitter (15.224 tweet) dan YouTube (3.999 komentar). Sentimen diklasifikasikan ke dalam kategori positif, negatif, dan netral, dengan hasil terbaik diperoleh menggunakan kernel RBF, mencapai akurasi 76,40% untuk tiga label dan 81,54% untuk dua label. Dibandingkan dengan Naïve Bayes, Random Forest, Decision Tree, dan K-NN, SVM terbukti lebih unggul. Penelitian ini juga mempertimbangkan pendekatan hybrid yang menggabungkan metode lexicon-based dan *machine learning* untuk meningkatkan akurasi.

Penelitian Samsir dkk. (2021)yang berjudul "Analisis Sentimen Pembelajaran Daring pada Masa Pandemi COVID-19 di Twitter Menggunakan Algoritma Naïve Bayes" menunjukkan hasil yang signifikan dalam mengidentifikasi opini publik mengenai pembelajaran daring selama pandemi COVID-19. Dengan memanfaatkan algoritma Naïve Bayes untuk mengolah data dari Twitter, sentimen publik terhadap pembelajaran daring dapat diklasifikasikan menjadi positif, negatif, dan netral. Hasil analisis ini memberikan wawasan mendalam tentang reaksi masyarakat terhadap kebijakan Belajar Dari Rumah (BDR), sekaligus menunjukkan adanya tantangan dan kontroversi dalam pelaksanaan pembelajaran daring. Temuan ini diharapkan dapat memberikan informasi yang relevan untuk meningkatkan adaptasi pembelajaran daring serta mengidentifikasi dampaknya terhadap motivasi belajar siswa selama pandemi.

Tabel 2.1 Perbandingan penelitian

	Penelitian 1	**Penelitian 2**	**Penelitian 3**	**Penelitian 4**
Nama	Andriawan & Ernawati, 2024	Daniel Kumala dan Antoni Wibowo, 2023	Parhusip dkk., 2023	Samsir dkk., 2021
Metode	*Hybrid Lexicon* dan *Random Forest*	*Hybrid Lexicon* SVM, Naïve Bayes, Logistic Regression, dan Random Forest	*SVM, Naïve Bayes, Random Forest, Decision Tree, K-NN*	*Hybrid Lexicon* SVM, Naïve Bayes, Logistic Regression, dan Random Forest
Lingkungan Penerapan	*Platform* X	*Game*	*Platform* Media Sosial	Lingkungan *Twitter.*
Teknologi	Google Colab, Visualisasi Data	Hybrid Lexicon, Machine Learning	*Text Mining*	*Text Mining*
Hasil	Analisis sentimen konflik Palestina-Israel Menunjukkan hasil yang lebih baik dibandingkan metode lain yang dibandingkan	Analisis sentimen ulasan game menunjukkan tingkat Akurasi tertinggi sebesar 95%.	Analisis sentimen tragedi Kanjuruhan menunjukkan Akurasi tertinggi sebesar 76,40% untuk tiga label dan 81,54% untuk dua label..	Analisis sentimen pembelajaran daring selama pandemi COVID-19 berhasil diklasifikasikan dengan baik.

Penelitian pada Tabel 2.1 memiliki perbedaan dengan penelitian terdahulu dalam hal objek, metode, dan lingkungan penerapan. Penelitian ini berfokus pada analisis sentimen *feedback* pengguna layanan keuangan di Politeknik Caltex Riau (PCR), sementara penelitian terdahulu lebih banyak membahas sentimen di media sosial dan *e-commerce*. Metode yang digunakan dalam penelitian ini adalah kombinasi Naïve Bayes dan *Lexicon Based*, dengan *preprocessing* N-Gram, TF-IDF, dan Sastrawi, serta evaluasi menggunakan *Confusion Matrix* dan *User Acceptance Test* (UAT). Sementara itu, penelitian terdahulu membandingkan berbagai algoritma seperti Naïve Bayes, SVM, dan *Hybrid Lexicon-Machine Learning*, serta menerapkan teknik visualisasi data seperti *word cloud*. Dari sisi penerapan, penelitian ini bertujuan untuk meningkatkan kualitas layanan keuangan di PCR, sementara penelitian terdahulu lebih berfokus pada analisis opini publik di media sosial dan *marketplace*.