BAB II
LANDASAN TEORI

2.1 Tinjauan Studi

Dalam penelitan ini terdapat beberapa penelitian terdahulu yang mengangkat topik menggunakan metode *Naïve Bayes Classifier* pada analisis sentimen, dimana setiap penelitian memiliki kriteria data, jumlah data, dan tingkat keakuratan yang berbeda. Dalam penelitian ini penulis menggunakan beberapa tinjauan studi dari peneliti terdahulu yang akan diajukan sebagai acuan dalam penelitian.

Adapun tinjauan studi dari penelitian terdahulu terdapat pada tabel berikut :

Tabel 2. 1 Penelitian Sejenis

No.	Nama Peneliti	Tahun	Judul	Metode	Perbandingan Hasil Penelitian
1	Nur Adinda Salsabila	2022	Analisis Sentimen Pada Media Sosial Twitter Terhadap Tokoh Gus Dur Menggunakan Metode Naïve Bayes dan Support Vector Machine	Naïve Bayes, Support Vector Machine	Penelitian sebelumnya : Peneliti lebih memfokuskan perbandingan dari hasil dua algoritma yang digunakan dan jumlah data set yang digunakan lebih sedikit. Lalu pada metode pengujian model menggunakan Cross Validation Penelitian sekarang : Peneliti lebih memfokuskan hasil klasifikasi yang dihasilkan oleh model dan jumlah data set lebih banyak.

					Metode pengujian model menggunakan split data
2	Muhammad Enggar Aziz Hibbannuari, Hartatik	2023	Analisis Sentimen Pengguna Twitter Terhadap Layanan Provider IndiHome Menggunakan Algoritma Naïve Bayes	Naïve Bayes	Peneitian sebelumnya : Menggunakan metode K-Fold Cross Validation yang lebih robust dalam evaluasi model. Distribusi data tidak seimbang dalam pengujian kedua. Penelitian sekarang : Menggunakan metode SEMMA sebagai pendekatan dalam analisis data. Akurasi lebih rendah dikarenakan data postifi lebih sedikit dibandingkan negatif yang menyebabkan ketidakseimbangan data
3	Tri Prasetyo, Hsdi Zakaria, Pandu Wiliantor	2022	Analisis Layanan Pelanggan PT PLN Berdasarkan Media Sosial Twitter Dengan Menggunakan Metode Naïve Bayes Classifier	Naïve Bayes Classifier	Peneitian sebelumnya : Menggunakan metode Consfusion Matrix dan mendapatkan akurasi yang lebih tinggi yaitu 80%. Penelitian sekarang : Menggunakan metode yang sama yaitu Confusion Matrix namun mendapatkan akurasi yang lebih kecil yaitu 70.12% dikarenakan distribusi data yang tidak seimbang

4	Johanes Florensius Sianipar, Yudhi Raymond Ramadhan, Irsan Jaelani	2023	Analisis Sentimen Pembangunan Kereta Cepat Jakarta-Bandung di Media Sosial Twitter Menggunakan Metode Naïve Bayes	Naïve Bayes	Penelitian sebelumnya : Menggunakan tiga kelas klasifikasi yaitu positif, negatif, dan netral. Rasio data training dan testing 80%:20% dan mendapatkan nilai akurasi sebesar 71%. Penenilitan sekarang : Menggunakan dua kelas klasifikasi saja yaitu positif dan negatif. Rasio data training dan testing 30%:70% dan mendapatkan nilai akurasi sebesar 70.12%
5	Muhammad Zidan	2022	Analisis Sentimen Kenaikan Harga Bahan Bakar Minyak (BBM) Berdasarkan Respon Pengguna Media Sosial Twitter di Indonesia Menggunakan Metode Naïve Bayes	Naïve Bayes	Penelitian sebelumnya : Menggunakan tiga kelas klasifikasi positif, negatif, dan netral. Metode pengujian menggunakan Split Validation dengan rasio 80% training, dan 20% testing dengan jumlah dataset sebanyak 1.500 data. Mendapatkan nilai akurasi sebesar 81%. Penelitian sekarang : Menggunakan dua kelas klasifikasi saja yaitu positif dan negatif. Metode pengujian menggunakan Split Validation dengan rasio 30% training dan 70% testing dengan jumlah

					datset sebanya 2.000 data. Mendapatkan nilai akurasi yang lebih rendah yaitu 70.12%
6	Muhammad Ammarullah Ridho	2021	Klasifikasi Pengaduan Layanan Pengguna Indihome Pada Media Sosial Twitter Menggunakan Metode Support Vector Machine Dengan Seleksi Fitur Information Gain	Support Vector Machine	Penelitian sebelumnya : Metode yang digunakan SVM. Kata kunci yang digunakan lebih luas yang terdiri dari dua kata “Indihome” dan “IndihomeCare”. Dataset yang digunakan lebih sedikit yaitu 1.000 data karena lebih fokus pada klasifikasi pengaduan. Eksraksi fitur menggunakan dua metode yaitu TF-IDF dan Information Gain sehingga akurasi yang didapatkan cukup tinggi Penelitian sekarang : Menggunakan metode Naïve Bayes dengan hanya menggunakan satu kata kunci saja “Indihome”. Dataset lebih banyak dengan 2.000 data namun tedapat ketidakseimbangan distribusi data positif sehingga model sulit mengenali kalimat yang memiliki nilai positif.

7	Kirana Aldina Larasati	2023	Analisis Sentimen Masyarakat pada Media Sosial Twitter Terhadap Sistem Kerja Work From Anywhere (WFA) Menggunakan Naïve Bayes dan Particle Swarm Optimiztion	Naïve Bayes, Particle Swarm Optimization	Penelitian sebelumnya : Penambahan metode Particle Swarm Optimization yang bersifat sebagai teknik optimasi lanjutan untuk meningkatkan performa keseluruhan model dengan meyesuaikan parameter atau fitur masukan Penelitian sekarang : Penambahan metode Laplace Smoothing digunakan sebagai teknik untuk mengatasi probabilitas nol

2.2 Tinjauan Pustaka

2.2.1 Analisis Sentimen

Analisis sentimen adalah cabang pengolahan bahasa alami NLP (*Natural Language Processing*) yang bertujuan untuk mengidentifikai, mengekstrak, dan mengklasifikasikan opini atau emosi dalam bentuk suatu teks. Analisis sentimen dilakukan untuk mengetahui apakah pembicara atau penulis opini berkenan dengan suatu topik, produk, layanan, organsasi, tokoh atau individu, ataupun kegiatan tertentu (Ardiani, et al., 2020).

Pada umumnya pendekatan untuk melakukan anlisis sentimen terbagi menjadi dua, yaitu menggunakan pendekatan *Supervised Learning* dan *Unsupervised Learning. Supervised Learning* merupakan pendekatan menggunakan algoritma yang bergantung pada data pelatihan. Model yang digunakan dalam proses klasifikasi berdasarkan data latih yang telah

ditentukan labelnya dalam satu domain. Beberapa algoritma yang menggunakan pendekatan ini adalah *Naïve Bayes, Support Vector Machine,* dan *Maximum Entropy*. Sedangkan *Unsupervised Learning* merupakan pendekatan yang melakukan klasifikasi sentimen tanpa memiliki label atau dengan data set input yang tidak berlabel. Pendekatan *Unsupervised Learning* memungkinkan model beroperasi sendiri untuk menemukan pola dan informasi yang sebelumnya tidak teramati (Larasati, 2023).

2.2.2 Metode *SEMMA*

Metode *SEMMA* (*Sample, Explore, Modify, Model, Assess*) berfokus pada tugas modifikasi terhadap proyek data mining serta pemodelan dan dirancang untuk membantu pengguna perangkat lunak *SAS enterprise miner* (Saputra, et al., 2022).

Proses data mining diterapkan di industri dan menyediakan metodologi untuk berbagai masalah bisnis seperti deteksi penipuan, pengelolaan asset, analisis risiko, kepuasan pelanggan, prediksi kebangkrutan, dan analisis portofolio (SAS Institute, 2025). Metode ini digunakan dalam berbagai bidang untuk menmukan pola dan tren yang bermanfaat dari data mentah yang besar dan kompleks.

Berikut merupakan tahapan – tahapan dalam metode *SEMMA* (Alizah, et al., 2020):

a. *Sample*

   Langkah pertama adalah tahap pengumpulan atau pengambilan sampel data yang representatif.

b. *Explore*

   Tahap eksplorasi dilakukan untuk memahami karakteristik data, mendeteksi anomali, pola, atau korelasi antar variabel yang mungkin ada.

c. *Modify*

Pada tahap modifikasi, data diolah untuk memenuhi kebutuhan analisis. Proses ini bertujuan untuk memaksimalkan akurasi hasil dengan meminimalkan noise dan variabel yang tidak relevan dalam data.

d. *Model*

Pada tahap ini, berbagai teknik pemodelan diterapkan untuk mengidentifikasi pola dalam data. Algoritma machine learning digunakan untuk membangun model prediksi atau deskripsi data yang telah dimodifikasi.

e. *Assess*

Tahap terakhir adalah mengevaluasi model yang dibangun. Pada tahap ini, model diuji untuk mengukur performanya dalam memprediksi atau menggambarkan data yang melibatkan penggunaan data uji dan metrik evaluasi.

Metode *SEMMA* membantu peneliti dan praktisi memastikan bahwa model yang dihasilkan tidak hanya akurat tetapi juga relevan untuk pengambilan keputusan berbasis data.

2.2.3 Text Mining

Text mining atau analisis teks adalah proses ekstraksi informasi berkualitas tinggi dari data teks yang tidak terstruktur melaui penggunaan pembelajaran mesin, statistik, dan linguistik (Agung, 2024). Teknik ini digunakan untuk mengubah teks yang tidak terstruktur menjadi format terstruktur guna mengidentifikasi pola, topik, kata kunci, dan atribut lainya dalam data.

Text mining pada dasarnya memerlukan pendekatan kuantitatif untuk data tekstual yang banyak. Hal ini dapat mempercepat penemuan pengetahuan dengan meningkatkan jumlah data yang dapat dianalisis. Teknik analisis yang digunakan pada *text mining* diantaranya tentang *dimensionality reduction*, *distance and similarity computing*, *clustering*, pemodelan topik, dan klasifkasi (Kobayashi, et al., 2018). *Text mining* juga merupakan bagian dari data mining, bedanya bentuknya lebih tidak

terstruktur sedangkan data mining datanya lebih terstruktur (Sholekha, et al., 2022).

2.2.4 Text Processing

Text processing adalah tahap awal dalam pengolahan bahasa alami (*Natural Language Processing/NLP*) yang bertujuan untuk mengolah data teks mentah menjadi bentuk yang terstruktur dan dapat dianalisis lebih lanjut. Dalam proses ini, data teks dibersihkan, dipilah dan disederhanakan untuk mengurangi kompleksitas dan menghilangkan unsur-unsur yang tidak relevan (Kharade, 2021).

Hal ini dilakukan untuk mempermudah algoritma analisis dalam memahami pola dan informasi penting yang terkandung dalam teks, seperti saat menerapkan analisis sentimen pada ulasan produk. *Text processing* dilakukan dalam data melalui beberapa tahapan secara berurutan, sebagai berikut (Krisdayanto & Nurhayanto, 2021) :

a. *Cleansing*

   Cleansing merupakan suatu proses untuk menghapus semua karakter dalam data set tidak termasuk alfabet, sehingga akan mengurangi karakter yang tidak diinginkan atau tidak memiliki arti. Karakter tersebut seperti angka, #, @, emoji, maupun tautan dari situs web yang ada didalam suatu data set

b. *Case Folding*

   Case folding merupakan suatu proses untuk mengubah setiap huruf dalam data set menjadi huruf kecil secara keseluruhan.

c. *Tokenization*

   Tokenization merupakan proses memotong data teks menjadi beberapa token atau potongan kata. Proses ini bertujuan untuk membedakan karakter-karakter tertentu yang dapat dilakukan sebagai pemisah kata atau bukan.

d. *Transform Case*

Transform case adalah tahapan merubah kalimat data teks menjadi terks yang seragam. Dengan adanya tahapan ini dapat berperan dalam penyamarataan dalam penggunaan huruf kecil dan huruf kapital. Sebagai contoh pada kata “Kecepatan” dan “kecepatan” akan terbaca sebagai dua kata yang berbeda, sehingga melalui proses ini sistem akan dapat membaa secara efektif.

e. Stopword Removal

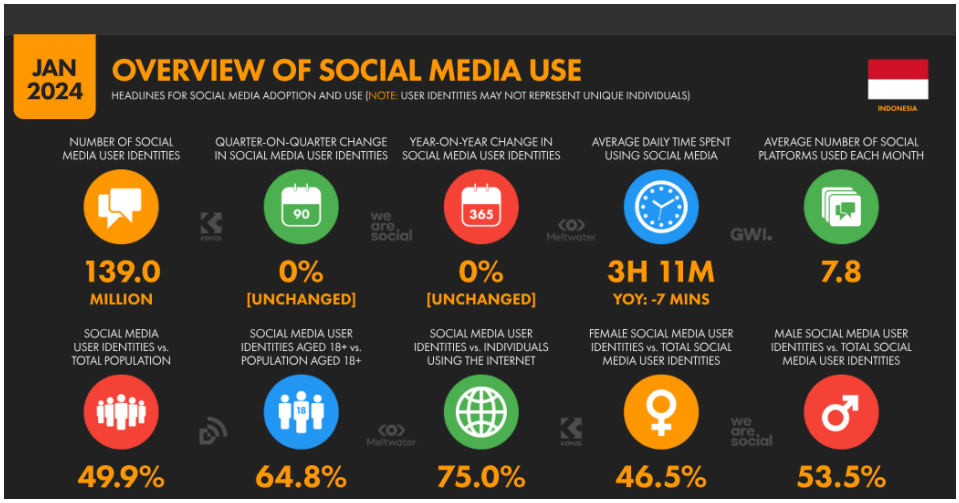
Stopword Removal adalah tahapan untuk menghilangkan kata yang sering muncul dan tidak memiliki makna. Pada tahapan ini, kata akan disaring menggunakan database dari kumpulan kata-kata *stopword*. Jenis kata yang akan di *stopword* seperti kata keterangan, kata seru, kata ganti, kata depan dan kata hubung.

f. Filtering

Filtering adalah tahapan untuk menghapus kata-kata yang terlalu pendek dan terlalu panjang dengan minimal 3 huruf dan maksimal 25 huruf.

2.2.5 Media Sosial

Perkembangan dunia amat sangat pesat dan selalu menunjukkan pertumbuhan yang sangat besar pula terhadap segala aspek, terkhusus pada aspek teknologi, informasi, dan komunikasi. Perkembangan ini membuat manusia tidak lagi cemas akan adanya batas, jarak, ruang, dan waktu. Dengan menggunakan internet masyarakat dapat mengakses segala informasi mengenai sesuatu dengan lebih mudah dan cepat, baik melalui *smartphone* mauapun komputer. Perkembangan ini dapat dilihat pada pemakaian media sosial yang terus mengalami kemajuan (Armayani, et al., 2021). Berikut merupakan gambaran umum penggunaan media sosial di Indonesia (Kemp, 2024).



Gambar 2. 1 Ringkasan Pengguna Media Sosial

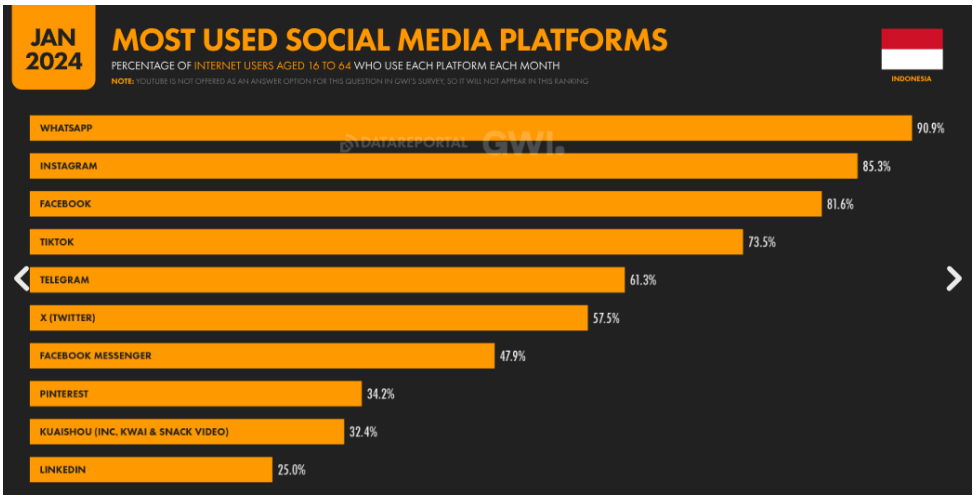
Media sosial telah menjadi platform yang dominan dalam menyampaikan informasi, opini, dan pengalaman konusmen terhadap suatu produk atau layanan. Platfrom ini memungkinkan pengguna untuk dengan mudah mengekspresikan pendapatnya secara terbuka, sehingga dapat dijadikan sumber data yang berharga. Hal ini dapat membantu dalam memprediksi sentimen berbagai macam orang pada peristiwa tertentu yang menarik seperti ulasan produk ataupun pendapat mereka tentang berbagai topik yang berkeliaran di seluruh dunia yang mana sangat penting untuk pengambilan keputusan (Parveen & Pandey, 2017).

2.2.6 Twitter

Saat ini, remaja dan orang dewasa pun tak dapat menghindar dari platform salah satu situs media sosial teratas yaitu Twitter. Twitter menyediakan berbagai fasilitas menarik yang dapat dimanfaatkan untuk berbagai keperluan, termauk mencari informasi. Twitter adalah platform media sosial jenis *microblogging* yang memberikan kemudahan bagi penggunanya menulis dan menyatakan suatu tindakan atau pendapat secara realtime (Rosdiana, et al., 2019).

Dengan fitur-fitur canggih yang tersedia di platform ini, mencari informasi tidak lagi menjadi tantangan bagi siapa saja yang menggunakanya

(Al Assyam & Hasan, 2023). Di Indonesia, penggunaan Twitter untuk penelitian ataupun sekedar untuk mencari informasi telah meningkat seiring dengan popularitas platform tersebut sebagai alat untuk menyuarakan opini publik. Twitter termasuk dalam salah satu platform media sosial yang populer digunakan masyarakat dengan menduduki peringkat ke-6 di tahun 2024, seperti pada statistik berikut (Kemp, 2024).



Gambar 2. 2 Grafik Media Sosial Yang Sering Digunakan di Indonesia

2.2.7 Indihome

Indihome (Indonesia Digital Home) merupakan layanan penyedia internet atau *Internet Service Provider* (ISP) yang dikelola oleh PT Telkom Indonesia, perusahaan telekomunikasi milik negara. Indihome menyediakan layanan digital terdepan menggunakan teknologi serat optik yang menawarkan layanan *Triple Play* yang mengintegrasi tiga jenis layanan berbeda meliputi *data*, *voice*, dan media dimana konsumen dapat menggunakan layanan internet (*Internet On Fiber* atau *High Speed Internet*), *IPTV* (*usee* TV) dan telepon secara bersamaan dan terintegrasi satu sama lain dalam satu produk (Indihome, 2025).

Sejak diluncurkannya Indihome, pelanggan yang berlangganan paket internet Speedy satu per satu diminta transmigrasi ke Indihome, karena layanan dengan Speedy akan segera diberhentikan pada tahun 2015 (Yulianti , et al., 2024). Indihome resmi diluncurkan pada awal tahun 2015. Dalam penyelenggaraanya, Telkom menggandeng sejumlah pengembangan teknologi komunikasi untuk membangun rumah konsep digital. Pelayanan Indihome hanya bisa diterapkan pada rumah yang wilayahnya tersedia jaringan serat optik dari Telkom FTTH (*Fiber to the Home*) dan area yang masih menggunakan kabel tembaga.

2.2.8 *Data Crawling*

Data crawling, atau dikenal sebaga *web scraping* atau *spidering*, adalah metode otomatis yang memanfaatkan *bot* (*crawler*) untuk mengekstrak data dar internet (Claussen & Peukert, 2019). Dalam konteks analisis data, *data crawling* sering digunakan untuk mengumpulkan data yang terstruktur dan tidak terstruktur dari internet yang nantinya bisa diproses lebih lanjut untuk tujuan tertentu, seperti analisis sentimen atau pemetaan tren.

Berikut adalah beberapa metode utama yang digunakan dalam *data crawling*, terutama dalam konteks pencarian dan pengumpulan data dari internet :

a. *Breadth-First Search (BFS) Crawling*

   Metode ini bekerja dengan memulai dari satu *URL* utama dan menelusuri semua tautan pada halaman tersebut sebelum melanjutkan ke halaman lain.

b. *Depth-First Search (DFS) Crawling*

   Berbeda dengan *BFS*, metode ini menyelami satu jalur tautan hingga kedalaman tertentu sebelum beralih ke jalur lain

c. *Incremental Crawling*

   Dalam metode ini, *crawler* hanya mengambil data yang baru atau diperbarui sejak kunjungan terakhir ke halaman tersebut.

d. *Politeness and Throttling*

Ini adalah pendekatan tambahan yang diterapkan pada *crawler* agar tidak membebani server. Dengan mengatur kecepatan akses dan jeda waktu antara pengambilan halaman, *crawler* dapat menghindari masalah akses yang mungkin muncul karena terlalu sering mengunjungi situs yang sama.

e. *Distributed Crawling*

Metode ini memecah tugas *crawling* menjad beberapa bagian yang dijalankan secara paralel oleh banyak mesin.

f. *API Crawling*

Ketika situs menyediakan *Application Programming Interface* (API), *crawler* dapat langsung menarik data dari API yang disediakan tanpa harus mengakses halamn web. Ini adalah metode yang sangat efisien dan lebih sesuai untuk situs yang memiliki data terstruktur, seperti Twitter atau Youtube.

2.2.9 *Naïve Bayes*

Naïve Bayes merupakan salah satu metode pada teknik klasifikasi untuk mengatasi ketidakpastian data dan termasuk dalam *classifier* statistik yang dapat memprediksi probabilitas keanggotaan class (Syarah, et al., 2022). Ciri utama dari metode ini adalah asumsi yang kuat akan independensi dari masing masing kondisi (Fairuz, et al., 2021).

Dalam konteks analisis sentimen, *Naïve Bayes* digunakan untuk mengklasifikasikan teks ke dalam kategori tertentu, sepert positif, negatif, atau netral berdasarkan pola yang ditemukan dalam data latih. Keunggulan metode ini terletak pada kemampuanya untuk menangani teks dengan cepat, meskipun terdapat asumsi independensi antar fitur.

Klasifikasi *Naïve Bayes* dibangun oleh data pelatihan untuk memperkirakan probabilitas dari setiap kategori yang terdapat pada ciri dokumen yang diuji. Sistem akan dilatih dengan menggunakan data baru (data latih dan data uji) dan selanjutnya diberi tugas untuk menebak nilai

fungsi target dari data tersebut (Prasetyo, et al., 2022). Secara umum, proses klasifikasi dengan menggunakan *Naïve Bayes* dapat dilihat dari persamaan sebagai berikut (Prasetyo, et al., 2022):

$$P(cj|wi) = \frac{P(wi|cj)\ x\ P(cj)}{P(wi)} \quad (1)$$

- $P(cj|wi)$: Peluang suatu teks atau dokumen diklasifikasikan ke dalam kategori j ketika terdapat kemunculan kata i
- $P(wi|cj)$: Peluang kemunculan kata I dalam kategori j, atau seberapa umum kata I muncul dalam dokumen yang termasuk dalam kategori tersebut
- $P(cj)$: Peluang dasar suatu dokumen masuk ke dalam kategori j
- $P(wi)$: Peluang kemunculan sebuah kata i secara keseluruhan

Peluang kemunculan sebuah kata bisa dihilangkan pada perhitungan klasifikasi karena peluang kemunculan kata tidak akan berpengaruh pada perbandingan hasil klasifiasi setiap kategori. Proses klasifikasi dapat disederhanakan sebagai berikut :

$$P(cj|wi)\ x\ P(wi|cj)\ x\ P(cj) \quad (2)$$

Untuk menghitung prior atau peluang kemunculan suatu kategori pada semua dokumen dapat dilakukan dengan menggunakan persamaan :

$$P(c) = \frac{Nc}{N} \quad (3)$$

- *Nc* : Banyak kategori c pada dokumen latih
- *N* : Banyak keseluruhan dokumen yang digunakan

2.2.10 *Laplace Smoothing*

Laplace Smoothing merupakan salah satu teknik perataan yang digunakan dalam metode *Naïve Bayes*. Dalam proses klasifikasi menggunakan algoritma *Naïve Bayes Classifier* pada parameter $P(cj|wi)$, kemungkinan munculnya nilai probabilitas nol dapat terjadi, maka akan terjadi kesalahan dalam proses klasifikasi. Untuk mengatasi masalah tersebut, *Laplace Smoothing* digunakan dengan cara menambahkan nilai positif terkecil pada proses perhitungan probabilitas. Penambahan ini befungsi untuk menghindari terjadinya nilai nol yang dapat menyebabkan kesalahan klasifikasi (Noto & Saputro, 2022).

Metode *Laplacian Smoothing* dilakukan dengan cara menambahkan nilai 1 pada setiap perhitungan data dalam himpunan data latih. Penambahan ini tidak memberikan perubahan signifikan terhadap estimasi probabilitas, namun efektif untuk mencegah munculnya nilai probabilitas nol dengan menyesuaikanya agar cenderung ke arah nilai tertentu, baik itu ulasan positif maupun negatif (Ramadhani & Fajarianto, 2020). Berikut merupakan persamaan *Naïve Bayes* dengan menggunakan *Laplace Smoothing* :

$$P(wi \mid C) = \frac{jumlah\ kata\ wi\ dalam\ kelas\ C + 1}{total\ kata\ dalam\ kelas\ C + |V|} \qquad (4)$$

Di mana :

- wi = kata ke-i
- C = kelas (positif atau negatif)
- $|V|$ = jumlah kata unik di seluruh dataset

2.2.11 *Term Frequency-Inverse Document Frequency* (TF-IDF)

TF-IDF adalah teknik yang banyak digunakan dalam *text mining* dan *information retrieval* untuk menilai seberapa penting suatu kata dalam sebuah dokumen relatif terhadap kumpulan dokumen lain. Metode ini

merupakan algoritma yang melakukan penggabungan dua metode yaitu konsep frekuensi kemunculan *Term* dalam sebuah dokumen dan *Inverse* frekuensi dokumen yang mengandung kata tersebut, akan mampu meningkatkan proporsi jumlah dokumen yang dapat ditemukan kembali dan yang dianggap relevan secara sekaligus. Sehingga kriteria *Term* yang paling tepat adalah *Term* yang sering muncul dalam dokumen secara individu, namun jarang dijumpai pada dokumen lainnya (Nurjannah, et al., 2013).

TF-IDF memberikan bobot yang lebih tinggi untuk istilah yang penting dan bobot yang lebih rendah untuk istilah yang tidak penting. Nilai vektornya terletak di antara 0 hingga 1. 0 berarti istilah tesebut tidak penting dalam konteks dokumen yang kita cari dan 1 berarti istilah tersebut relevan (Larasati, 2023). Pemberian bobot yang lebih tinggi juga dapat diberikan pada kata-kata yang sering muncul dalam dokumen tertentu tetapi jarang muncul dalam keseluruhan koleksi dokumen.

TF-IDF bekerja dengan menghitung dua komponen utama, yaitu :

1. *Term Frequency* (TF) : mengukur seberapa sering suatu kata muncul dalam dokumen tertentu. Semakin tinggi frekuensi suatu kata, semakin besar pula nilai TF.
2. *Inverse Document Frequency* (IDF) : mengukur seberapa jarang kata tersebut muncul dalam kumpulan dokumen. Semakin jarang kata tersebut muncul dalam seluruh dokumen, semakun tinggi pula nila IDF-nya.

Kedua nilai ini dikalikan untuk menghasilkan bobot akhir bagi setiap kata dalam dokumen. Adapun rumus umum TF-IDF adalah sebagai berikut :

Term Frequency (TF) :

$$TFij = \frac{fij}{total\ jumlah\ kata\ dalam\ dokumen\ i} \qquad (5)$$

- $TFij$: Frekuensi term j dalam dokumen i.
- fij : Jumlah kemunculan term j dalam dokumen i.

Inverse Document Frequency (IDF) :

$$IDFj = \log\left(\frac{N}{dfj}\right) \quad (6)$$

- $IDFj$: Bobot logaritmik untuk term j.
- N : Total jumlah dokumen dalam korpus.
- dfj : Jumlah dokumen yang mengandung term j.

Pada TF-IDF terdapat rumus untuk menghitung bobot (W) setiap dokumen untuk kunci.

$$wij = TFij\ x\ IDFj \quad (7)$$

Setelah bobot (W) masing-masing dokumen diketahui, selanjutnya dilakukan proses pengurutan dimana semakin besar nilainya maka semakin besar pula tingkat kemiripan dokumen dengan kata kunci, begitu juga sebaliknya.

2.2.12 *Confusion Matrix*

Confusion Matrix adalah matriks yang digunakan untuk mengevaluasi kinerja model klasifikasi dalam machine learning. Matriks ini memberikan gambaran tentang prediksi model dibandingkan dengan nilai aktual atau data yang sebenarnya. Data prediksi merupakan nilai yang didapatkan dari hasil pemodelan machine learning, sedangkan data aktual adalh nilai sebenarnya yang dimiliki.

Secara umum tabel *Confusion Matrix* (matrik klasifikasi atau tabel kontigensi) digunakan untuk menentukan performa model *Naïve Bayes* untuk mengklasifikasikan data, yang diterapkan untuk mengevaluasi hasil pengujian pada sistem analisis sentimen (Zidan, 2022).

Evaluasi menggunakan *Confusion Matrix* menghasilkan nilai *Accuracy*, *Precision*, serta *Recall*. *Accuracy* dalam klasifikasi merupakan presentasi ketetapan *record* data diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi. *Precision* merupakan proposikasi yang diprediksi positif yang juga positif benar pada data sebenarnya. *Recall* merupakan proporsi kasus positif yang sebenarnya diprediksi positif secara benar (Putra & Wibowo, 2020).

Proses perhitungan ini diimplementasikan ke dalam tabel yang terdiri dari dua kelas yang bersifat positif dan negatif, seperti pada tabel berikut :

Tabel 2. 2 Confusion Matrix

		True Clas	
		Positive	***Negative***
Predicted Class	***Positive***	TP	FP
	Negative	FN	TN

Dalam *Confusion Matrix* tersebut, terdapat empat nilai yang dijadikan acuan dalam perhitungan, yaitu :

1. TP (*True Postive)*, yaitu data yang diprediksi positif dan faktanya data itu positif (Sesuai).
2. TN (*True Negative*), yaitu data yang diprediksi negatif dan faktanya data itu negatif (Sesuai).

3. FP (*False Positve*), yaitu data yang diprediksi positif dan faktanya data itu negatif (Tidak sesuai).
4. FN (*False Negative*), yaitu data yang diprediksi negatif dan faktanya data itu positif (Tidak sesuai).

Nilai-nilai yang dihasilkan dari pengukuran performansi dari sebuah algoritma adalah sebagai berikut :

1. *Accuracy* adalah jumlah perbandingan data yang benar dengan jumlah keseluruhan data, dengan rumus sebagai berikut :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

2. *Precision* digunakan untuk mengukur seberapa besar proporsi dari kelas positif yang berhasil diprediksi dengan benar dibandingkan dengan keseluruhan hasil prediksi kelas positif, berikut merupakan rumusnya :

$$Precission = \frac{TP}{FP + TP} \quad (9)$$

3. *Recall* digunakan untuk menunjukkan presentase kelas data positif yang berhasil diprediksi benar dibandingkan dengan keseluruhan data kelas yang benar positif, berikut merupakan rumusnya :

$$Recall = \frac{TP}{FN + TP} \quad (10)$$

4. F1 Score adalah perbandingan rata-rata precission dan recall yang dibobotkan, dengan rumus sebagai berikut :

$$F1\ Score = \frac{2\ x\ Recall\ x\ Precission}{Recall + Precission} \quad (11)$$

2.2.13 Rapid Miner

Rapid miner merupakan perangkat lunak komputasi statistik yang dikembangkan dan berhasil diterapkan pada berbagai data untuk dianalisis dan memantau prosesnya (Syahbiddin & Mailangkay, 2021). Perangkat lunak ini dibuat oleh Dr. Markus Hofmann dari Institute of Technology Blanchardstown dan Ralf Klinkenberg dari rapid-i.com dengan tampilan GUI (*Graphical User Interface*) yang bersifat *open source* dan dibuat menggunakan program Java yang dapat dijalankan di sistem operasi manapun.

Rapid Miner dikhususkan untuk penggunaan *data mining*. Model yang disediakan juga cukup banyak dan lengkap, seperti Model *Bayesian*, *Modelling*, *Tree Induction*, *Neural Network* dan lain-lain. Banyak metode yang disediakan oleh Rapid Miner mulai dari klasifikasi, *clustering*, asosiasi dan sering kali digunakan untuk menganlisis jumlah data yang besar (Sudarsono , et al., 2021).

Pemilihan Rapid Miner sebagai perangkat lunak untuk penerapan model dalam penelitian ini didasarkan pada kemampuan yang mendukung proses analisis data secara lengkap, mulai dari tahap preprocessing, ekstraksi fitur, pembentukan model klasifikasi, dan evaluasi kinerja model termasuk algoritma *Naïve Bayes* yang digunakan pada penelitian ini. Kemudahan dalam mengatur parameter, memvisualisasikan hasil, serta mendukung format data yang beragam menjadikan Rapid Miner sebagai pilihan yang efisien dan tepat untuk mengimplementasikan analisis sentimen berbasis teks secara terstruktur dan sistematis.

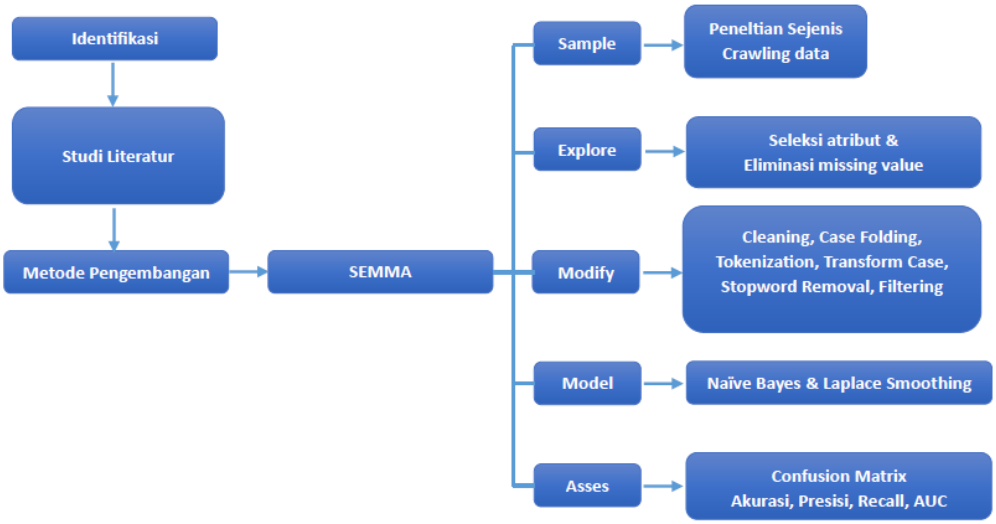
2.3 Kerangka Pemikiran

Kerangka pemikiran adalah gambaran sistematis yang menjelaskan langkah-langkah dalam penelitian berdasarkan teori, fakta, dan studi literatur. Kerangka pemikiran befungsi sebagai panduan dalam menjelaskan

hubungan antara variabel penelitian serta membantu peneliti dalam mengembangkan hipotesis dan analisis data berdasarkan teori yang relevan,

Penelitian ini berawal dari meningkatnya opini pengguna mengenai layanan penyedia internet Indihome yang disampaikan melalui media sosial, khususnya twitter. Dikarenakan besarnya jumlah data yang ada maka tidak memungkinkan untuk mendeteksi opini tersebut apakah bersifat positif atau negatif secara manual, maka dibutuhkan cara yang sistematis untuk menganalisis pola sentimen tersebut yaitu dengan menggunakan metode klasifikasi *Naïve Bayes*.

Dalam penelitian ini, digunakan metode *SEMMA* (*Sample, Explore, Modify, Model, Asses*) sebagai kerangka pemikiran dalam proses analisis data. Berikut merupakan gambaran kerangka pemikiran yang diterapkan pada penelitian ini :



Gambar 2. 3 Kerangka Pemikiran