BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Setelah melakukan kajian mendalam terdahulu analisis sentimen, ditemukan sejumlah jurnal dan penelitian sebelumnya yang memiliki keterkaitan dengan studi ini. Berbagai penelitian tersebut mencakup beragam aspek dan metode yang menjadi dasar bagi penelitian ini. Beberapa penelitian yang dianggap relevan disajikan dalam Tabel 2.1.

Table 2.1 Penelitian Terdahulu

Nama Jurnal	Judul Artikel	Nama Peneliti	Hasil Penelitian
Scientific Journal of Informatics Vol. 11, No. 3, Aug 2024 [7]	Comparative Performance of SVM and Multinomial Naïve Bayes in Sentiment Analysis of the Film 'Dirty Vote'	Aisha Shakila Iedwan, Nia Mauliza, Yoga Pristyanto, Anggit Dwi Hartanto, Arif Nur Rohman	Penelitian ini menganalisis sentimen komentar YouTube terhadap film dokumenter “Dirty Vote” dengan menggunakan algoritma SVM dan Multinomial Naïve Bayes. Dari 1000 komentar yang dianalisis, SVM menghasilkan akurasi sebesar 88%, sedangkan Multinomial Naïve Bayes menghasilkan akurasi 86%. Hasil ini menunjukkan bahwa SVM sedikit lebih unggul, terutama dalam mendeteksi komentar positif, dan dinilai lebih andal untuk analisis opini sosial-politik.
JISA (Jurnal Informatika dan	Naive Bayes and Support Vector Machine Algorithm for	Laurenzius Julio Anreaja, Norma Nobuala Harefa, Julius	Penelitian ini menganalisis sentimen ulasan pengguna aplikasi Opensea di Google Play

Sains) Vol. 05, No. 01, June 2022 [8]	Sentiment Analysis Opensea Mobile Application Users in Indonesia	Galih Prima Negara, Venantius Nathan Hermanu Pribyantara, Agung Budi Prasetyo	Store menggunakan algoritma Naïve Bayes dan Support Vector Machine (SVM). Dari 1028 ulasan, dilakukan labeling manual dan praproses teks. Hasilnya, SVM menunjukkan kinerja lebih baik dengan akurasi 90,78%, precision 94,23%, dan recall 71,96%, dibandingkan Naïve Bayes yang memperoleh akurasi 89,81%, precision 87,31%, dan recall 71,02%. Disimpulkan bahwa SVM lebih unggul untuk analisis sentimen pada studi ini.
JOURNAL OF MULTIDISCIPLINARY ISSUES 2(2) 1 - 21(2022) [9]	SENTIMENT ANALYSIS OF COMMENTS ON SEXUAL HARASSMENT IN COLLEGES ON FOUR POPULAR SOCIAL MEDIA	Vinson Phoan, Johan Setiawan	Penelitian oleh Vinson Phoan dan Johan Setiawan menganalisis komentar tentang pelecehan seksual di UMN dari empat media sosial menggunakan algoritma SVM dan FastText. Dari 287 komentar, hasil menunjukkan 54,7% netral, 36,6% negatif, dan 8,7% positif, dengan akurasi model sebesar 55,14%. Twitter menjadi platform dengan respons paling beragam.
TEKNIKA : JURNAL SAINS DAN TEKNOLOGI VOL 19 NO 02 (2023) [10]	Sentiment analysis on public opinion of electric vehicles usage in Indonesia using support vector machine algorithms	Naufal Avilandi Poedjimartojo, Dita Pramesti, Riska Yanu Fa'rifah	Model SVM mencapai akurasi 94.8%, precision 95.5%, recall 99.1%, dan F1-score 97.2%. Penelitian ini menganalisis sentimen masyarakat Indonesia terhadap penggunaan kendaraan listrik.

Journal of Engineering and Scientific Research (JESR) Volume 6, Issue 1, June 2024 [11]	Comparison of SVM & Naïve Bayes Methods in Sentiment Analysis of Electric Vehicle Subsidy Policy Based on X Data	IWD Wiguna, DV Waas, IKAG Wiguna, ML Radhitya	Metode SVM dengan kernel RBF menunjukkan akurasi tertinggi sebesar 83.02%. Penelitian ini membandingkan kinerja SVM dan Naïve Bayes dalam analisis sentimen terkait kebijakan subsidi kendaraan listrik.
Journal of Computer Networks, Architecture and High Performance Computing Vol. 6 No. 4 (2024) [12]	Analysis Of Opinion Sentiment Towards Electric Vehicle Tax On Social Media X Using The Support Vector Machine Method	Dara Taqa Assajidah Jusli, Rakhmat Kurniawan	Metode SVM mencapai akurasi 79%, precision 85%, recall 89%, dan F1-score 87%. Penelitian ini menganalisis sentimen masyarakat terhadap pajak kendaraan listrik di media sosial.
International Research Journal of Humanities and Interdisciplinary Studies (IRJHIS) Volume 4 Issue 3 March 2023 [13]	A STUDY ON SENTIMENT ANALYSIS OF THE TWO-WHEELER ELECTRIC VEHICLE USERS IN INDIA	Sakshi Chaturvedi, Mrs. Kalyani Gorti	Metode SVM menunjukkan akurasi 84%. Penelitian ini menemukan bahwa pengguna kendaraan listrik dua roda di India memiliki sikap positif terhadap kendaraan ini.
Jurnal Teknik Informatika (JUTIF) Vol. 5, No. 1, February 2024 [5]	COMPARISON OF RANDOM FOREST, SUPPORT VECTOR MACHINE AND NAIVE BAYES ALGORITHMS TO ANALYZE SENTIMENT TOWARDS MENTAL	Putri Elisa, Auliya Rahman Isnain	Hasil penelitian ini menunjukkan bahwa dari 3.095 tweet tentang stigma kesehatan mental, algoritma SVM memberikan akurasi tertinggi sebesar 86,11%, diikuti Random Forest 82,55%, dan Naive Bayes 78,19%. Ini menandakan SVM paling akurat dalam mengklasifikasikan

	HEALTH STIGMA		sentimen negatif, positif, dan netral terhadap isu tersebut.
JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi) olume 5 Number 2 of 2021 [14]	Sentiment Analysis of Public Acceptance of Covid - 19 Vaccines Types in Indonesia using Naïve Bayes, Support Vector Machine, and Long Short-TermMemory (LSTM)	Dinar Ajeng Kristiyanti, Sri Hardani	penelitian ini menganalisis sentimen masyarakat terhadap berbagai jenis vaksin Covid-19 di Indonesia menggunakan algoritma Naïve Bayes, Support Vector Machine (SVM), dan Long Short-Term Memory (LSTM). Dari 2000 data tweet, hasil evaluasi menunjukkan bahwa SVM memiliki akurasi tertinggi sebesar 84,89%, disusul Naïve Bayes (84,65%) dan LSTM (82,97%). SVM dinilai paling optimal dalam klasifikasi sentimen terhadap vaksin.
Kalbi Scientia, Jurnal Sains dan Teknologi, Volume 11, No. 02, September 2024 [6]	Perbandingan Performa Algoritma Naïve Bayes, SVM dan Random Forest: Studi Kasus A nalisis Sentimen Pengguna Sosial Media X	Putri Cahyani, Lufty Abdillah	Hasil penelitian ini menemukan bahwa dari 10.000 tweet tentang Ibu Kota Nusantara (IKN), algoritma SVM juga memberikan performa terbaik dengan akurasi 87%, mengungguli Random Forest (85%) dan Naive Bayes (73,9%) setelah penerapan teknik SMOTE. SVM terbukti paling efektif dalam mengklasifikasikan sentimen publik terhadap topik IKN.

Penelitian terdahulu analisis sentimen dengan pendekatan machine learning telah banyak dilakukan pada berbagai domain. Beberapa di antaranya fokus pada pembandingan performa algoritma klasifikasi teks seperti Support Vector Machine (SVM), Naïve Bayes (NB), dan Random Forest (RF). Kajian ini bertujuan tidak

hanya menyebutkan hasil tiap studi, tetapi juga menunjukkan keterkaitan antar penelitian tersebut dan bagaimana penelitian ini melengkapi celah (gap) yang belum diteliti sebelumnya.

Penelitian oleh Elisa dan Isnain [5] mengkaji klasifikasi sentimen terhadap stigma kesehatan mental di media sosial. Mereka membandingkan tiga algoritma: SVM, Naïve Bayes, dan Random Forest. Hasilnya menunjukkan bahwa SVM memiliki akurasi tertinggi sebesar 86,11%, diikuti oleh Random Forest (82,55%) dan Naïve Bayes (78,19%). Penelitian ini mengindikasikan bahwa SVM memiliki keunggulan dalam menangani data teks berdimensi tinggi. Namun, penelitian ini tidak menangani isu ketidakseimbangan kelas (class imbalance) dalam data, sehingga hasil klasifikasinya belum merepresentasikan distribusi sentimen yang proporsional.

Menanggapi kelemahan tersebut, Cahyani dan Abdillah [6] dalam penelitiannya mengenai opini publik terhadap pemindahan Ibu Kota Negara (IKN) menggunakan data komentar YouTube, menerapkan metode SMOTE (Synthetic Minority Over-sampling Technique) untuk menyeimbangkan distribusi kelas. Hasil klasifikasi menunjukkan bahwa setelah penerapan SMOTE, algoritma SVM mencatat akurasi tertinggi sebesar 87%, diikuti oleh Naïve Bayes dan Random Forest. Penelitian ini menegaskan bahwa teknik balancing data sangat berpengaruh terhadap peningkatan performa klasifikasi, terutama dalam konteks data sosial yang cenderung tidak seimbang.

Sementara itu, Anreaja et al. [8] melakukan klasifikasi sentimen terhadap opini masyarakat tentang NFT di platform OpenSea. Mereka membandingkan algoritma Naïve Bayes dan Support Vector Machine, dan hasilnya menunjukkan bahwa SVM memiliki akurasi sebesar 83%, lebih tinggi dibandingkan Naïve Bayes yang mencapai 76%. Penelitian ini memperkuat posisi SVM sebagai algoritma yang lebih unggul dalam pengolahan opini berbasis teks panjang dan informal seperti yang terdapat di media sosial.

Dalam konteks yang berbeda, Iedwan et al. [7] mengkaji sentimen masyarakat terhadap film dokumenter *Dirty Vote*. Penelitian ini menggunakan

pendekatan klasifikasi teks pada data komentar YouTube. Meskipun algoritma yang digunakan bukan ketiganya secara bersamaan, penelitian ini menunjukkan bahwa data komentar YouTube dapat merepresentasikan opini publik yang kaya, tetapi juga rentan terhadap noise dan ketidakseimbangan.

2.2 Teori Penelitian

2.2.1 Mobil Listrik

Mobil listrik merupakan jenis kendaraan yang memanfaatkan sumber energi alternatif dalam operasionalnya. Secara definisi, mobil listrik adalah kendaraan yang digerakkan oleh satu atau lebih motor listrik dengan menggunakan energi listrik yang tersimpan dalam baterai yang dapat diisi ulang atau media penyimpanan energi lainnya. Sejarah mencatat bahwa mobil listrik pertama kali dikembangkan pada tahun 1880-an sebagai bentuk awal kendaraan berbasis listrik. Dalam proses kerjanya, energi listrik pada mobil listrik dikonversikan menjadi energi mekanik melalui motor listrik atau dinamo sebagai komponen utama penggeraknya. Tidak seperti kendaraan berbasis mesin pembakaran internal, mobil listrik hanya menggunakan baterai sebagai sumber tenaga, sehingga tidak menghasilkan emisi gas buang dan memiliki daya output sebesar 3000 watt/72 volt. Dibandingkan dengan kendaraan bermesin konvensional, mobil listrik menawarkan berbagai keunggulan potensial. Salah satu keunggulan utamanya adalah tidak adanya emisi gas buang, sehingga lebih ramah lingkungan. Selain itu, mobil listrik tidak bergantung pada bahan bakar fosil sebagai sumber energi utama, sehingga turut berkontribusi dalam mengurangi emisi gas rumah kaca serta ketergantungan terhadap sumber energi tak terbarukan[15].

2.2.2 Analisis Sentimen

Analisis sentimen, yang juga dikenal sebagai *opinion mining*, merupakan bidang studi yang berfokus pada pengidentifikasian dan pengukuran sentimen, opini, sikap, emosi, evaluasi, serta persepsi masyarakat terhadap suatu layanan, produk, individu, organisasi, peristiwa, topik, atribut, maupun isu tertentu. Analisis sentimen umumnya diperoleh dari berbagai sumber, seperti komentar, ulasan, dan umpan balik, yang memberikan wawasan penting

untuk berbagai tujuan, termasuk pengambilan keputusan bisnis dan peningkatan layanan. Sentimen yang dianalisis dapat diklasifikasikan ke dalam tiga kategori utama, yaitu positif, negatif, dan netral, atau dapat pula menggunakan sistem pemeringkatan seperti skala bintang. Secara teknis, analisis sentimen merupakan bagian dari *text mining* yang bertujuan untuk mengidentifikasi opini, emosi, serta sikap yang terkandung dalam suatu teks atau kumpulan teks [16].

2.2.3 TF-IDF

TF-IDF (*Term Frequency-Inverse Document Frequency*) merupakan salah satu metode yang umum digunakan dalam bidang *Natural Language Processing* (NLP) dan analisis teks untuk menentukan tingkat kepentingan suatu kata dalam sebuah dokumen maupun dalam keseluruhan kumpulan dokumen (*corpus*). Teknik ini bekerja dengan menghitung bobot suatu kata berdasarkan dua komponen utama, yaitu frekuensi kemunculan kata dalam dokumen (TF) dan frekuensi kemunculan kata tersebut dalam seluruh korpus (IDF). Melalui penggabungan kedua nilai ini, *TF-IDF* menghasilkan bobot yang merepresentasikan sejauh mana suatu kata dianggap penting dalam konteks dokumen yang sedang dianalisis[17].

2.2.4 *SMOTE*

SMOTE (Synthetic Minority Over-sampling Technique) merupakan salah satu metode oversampling yang pertama kali diperkenalkan oleh Nitesh V. Chawla. Teknik ini digunakan untuk menangani permasalahan *class imbalance problem* (CIP), yaitu kondisi ketika jumlah data pada kelas mayoritas jauh lebih besar dibandingkan kelas minoritas. *SMOTE* bekerja dengan menghasilkan data sintetis baru pada kelas minoritas, bukan sekadar menduplikasi data yang ada, sehingga dapat membantu meningkatkan kinerja algoritma klasifikasi. Namun, penerapan *SMOTE* juga memiliki risiko, salah satunya adalah potensi terjadinya overfitting, karena model dapat terlalu menyesuaikan diri terhadap data sintetis yang dihasilkan[18].

2.2.5 Confusion Matrix

Confusion Matrix merupakan konsep yang digunakan untuk mengevaluasi kinerja suatu model klasifikasi dengan mengukur tingkat akurasi berdasarkan hasil prediksi. Metode ini sering diterapkan dalam Data Mining dan sistem pendukung keputusan untuk menilai efektivitas model dalam mengklasifikasikan data[19].

Confusion Matrix terdiri dari empat komponen utama yang merepresentasikan hasil klasifikasi, yaitu:

***True Positive* (TP)**: Data sebenarnya bernilai positif dan berhasil diklasifikasikan sebagai positif oleh model.

***False Positive* (FP)**: Data sebenarnya bernilai negatif, tetapi salah diklasifikasikan sebagai positif (dikenal sebagai kesalahan tipe I).

***False Negative* (FN)**: Data sebenarnya bernilai positif, tetapi salah diklasifikasikan sebagai negatif (dikenal sebagai kesalahan tipe II).

***True Negative* (TN)**: Data sebenarnya bernilai negatif dan berhasil diklasifikasikan sebagai negatif oleh model.

Confusion Matrix akan menghasilkan nilai accuracy, precision, Recall, dan F1-Score seperti pada table 2.2 dibawah ini:

Table 2.2 *Confusion Matrix*

	True	False
True (*Positive*)	TP (*True Positive*)	FP (*False Positive*)
True (*Negative*)	FN (*False Negative*)	TN (*True Negative*)

2.2.6.1 *Accuracy*

Accuracy adalah metrik evaluasi yang mengukur sejauh mana model klasifikasi dapat membuat prediksi yang benar terhadap seluruh data yang tersedia. Akurasi dihitung sebagai rasio antara jumlah prediksi yang terhadap total keseluruhan data yang diuji.

$$Accuracy = \frac{TN+TP}{TN+FP+FN+TP}$$

Rumus 2.1 Perhitungan Nilai *Accuracy*

2.2.6.2 *Recall*

Recall merupakan metrik evaluasi yang digunakan untuk mengukur sejauh mana model mampu mendeteksi data positif dengan benar. Metrik ini dihitung sebagai rasio antara jumlah True Positive (TP) dengan total keseluruhan data yang seharusnya diklasifikasikan sebagai positif.

$$Recall = \frac{TP}{TP+FN}$$

Rumus 2.2 Perhitungan Nilai *Recall*

2.2.6.3 *Precision*

Precision merupakan metrik evaluasi yang mengukur sejauh mana model dapat memberikan prediksi positif yang benar dibandingkan dengan seluruh prediksi positif yang dihasilkan. Precision dihitung sebagai rasio antara True Positive (TP) dengan total jumlah prediksi positif.

$$Precision = \frac{TP}{TP+FP}$$

Rumus 2.3 Perhitungan Nilai *Precision*

2.2.6.4 *F1-Score*

F1-Score adalah metrik evaluasi yang menggabungkan Precision dan Recall dalam satu nilai harmonik untuk memberikan gambaran keseimbangan antara keduanya.

$$\text{F1} - score = \text{F1} = score = 2 \text{ x } \frac{precision \text{ x } recall}{precision+recall}$$

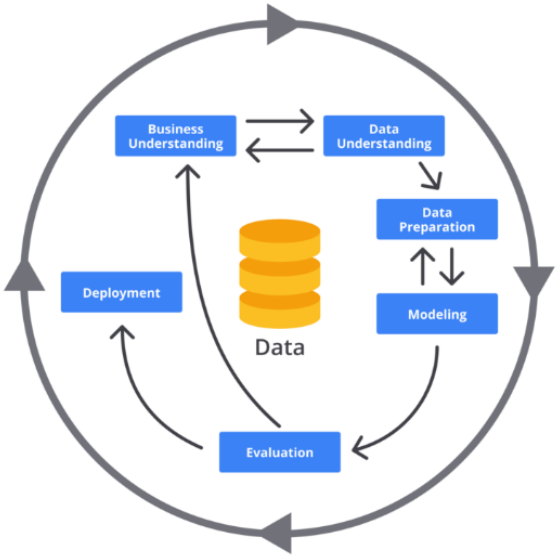
Rumus 2.4 Perhitungan Nilai *F1-score*

2.3 Framework/Algoritma/SDLC Penelitian

2.3.1 CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM) adalah kerangka kerja sistematis yang digunakan dalam proses *Data Mining* dan analisis data untuk membantu perusahaan dalam memahami serta menyelesaikan tantangan bisnis secara terstruktur. Metodologi ini memberikan

pendekatan terorganisir dalam mengelola proyek *Data Mining* dari tahap awal hingga implementasi solusi, sehingga memungkinkan pengambilan keputusan yang lebih efektif berdasarkan wawasan yang diperoleh dari data. CRISP-DM terdiri dari enam tahapan utama, yaitu Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment[20].



Gambar 2.1 Tahapan pada CRISP-DM [20]

a. Business Understanding

Tahap ini bertujuan untuk memahami kebutuhan bisnis dan pelanggan secara menyeluruh. Aktivitas utama pada tahap ini mencakup penentuan tujuan bisnis, penilaian situasi dan ketersediaan sumber daya, penetapan tujuan pengumpulan data, serta penyusunan rencana proyek yang akan dijalankan.

b. Data Understanding

Tahap ini berfokus pada eksplorasi dan pemahaman data yang akan digunakan dalam proyek. Aktivitas yang dilakukan meliputi pengumpulan data awal, analisis karakteristik data, eksplorasi lebih lanjut untuk menemukan pola awal, serta verifikasi kualitas data guna memastikan bahwa data yang digunakan memenuhi standar yang diperlukan.

c. Data Preparation

Pada tahap ini, dilakukan proses persiapan data agar siap digunakan dalam pemodelan. Data yang telah dikumpulkan sebelumnya akan dibersihkan, diolah, dan disesuaikan dengan kebutuhan analisis. Proses ini mencakup penanganan data yang hilang, penghapusan duplikasi, normalisasi, serta transformasi data ke dalam format yang sesuai untuk tahap pemodelan berikutnya.

d. Modeling

Tahap pemodelan bertujuan untuk membangun dan mengevaluasi berbagai model berdasarkan teknik *machine learning* atau *data mining* yang berbeda. Aktivitas utama dalam tahap ini meliputi pemilihan teknik pemodelan yang sesuai, perancangan skenario pengujian, pembangunan model, serta penilaian performa model untuk menentukan model terbaik yang dapat digunakan dalam analisis.

e. Evaluation

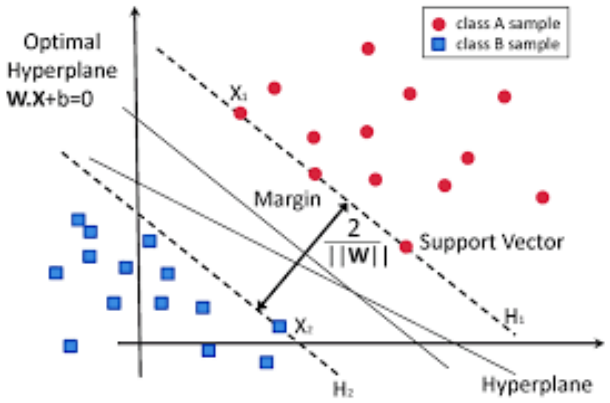
Tahap evaluasi dilakukan untuk memastikan bahwa model yang telah dibangun sesuai dengan tujuan bisnis yang telah ditetapkan. Evaluasi dilakukan dengan meninjau hasil model, membandingkan performa model berdasarkan berbagai metrik evaluasi, serta menentukan langkah selanjutnya untuk meningkatkan akurasi dan relevansi model terhadap permasalahan yang dianalisis.

f. Deployment

Tahap akhir dari proses CRISP-DM adalah implementasi model ke dalam sistem bisnis. Pada tahap ini, perencanaan deployment yang telah dimulai sejak tahap *Business Understanding* diwujudkan dalam bentuk penerapan model untuk mendukung pengambilan keputusan. Selain itu, tahap ini juga mencakup strategi konversi hasil model menjadi wawasan yang dapat ditindaklanjuti serta integrasi model ke dalam proses operasional bisnis.

2.3.2 *Support Vector Machine*

Support Vector Machine (SVM) merupakan algoritma klasifikasi yang bekerja berdasarkan konsep *margin* maksimal, sehingga sering disebut sebagai *maximum margin classifier*. Algoritma ini termasuk dalam kategori *Supervised Learning*, di mana model dilatih menggunakan data yang telah memiliki label sebelumnya. Dalam proses klasifikasi, SVM membangun sebuah *hyperplane* atau batas keputusan yang memisahkan data dari dua kelas berbeda dengan margin maksimal. Ketika menerima data *testing*, model akan mengkategorikannya ke dalam kelas yang sesuai berdasarkan karakteristik yang telah dipelajari dari data *training* . Dibawah ini merupakan gambar 2.2 contoh *Hyperline* SVM:



Gambar 2.2 *Hyperline Support Vector Machine* [21]

Algoritma *Support Vector Machine* (SVM) bekerja dengan menentukan batas pemisah atau *hyperplane* yang optimal antara dua kelas data, dengan tujuan memaksimalkan margin dari data terdekat yang disebut support vectors. Semakin besar margin antara kelas, semakin baik model dalam melakukan klasifikasi terhadap data baru. Untuk menemukan *hyperplane* terbaik, SVM mengukur margin dan mencari titik maksimal yang memisahkan kelas-kelas data. Dalam kasus di mana data tidak dapat dipisahkan secara linear, SVM menggunakan konsep kernel trick untuk mentransformasikan data ke dalam ruang berdimensi lebih tinggi, sehingga memungkinkan pencarian *hyperplane* optimal dalam kasus non-linear. *Hyperplane* dalam SVM berfungsi sebagai garis atau bidang pemisah yang membedakan dua kelas, yang umumnya

dinyatakan sebagai +1 dan -1. Setiap kelas memiliki pola atau karakteristik tersendiri yang digunakan oleh model untuk melakukan prediksi terhadap data baru secara akurat.

2.3.3 *Naïve Bayes*

Algoritma *Naïve Bayes* adalah metode klasifikasi yang didasarkan pada prinsip probabilitas dan statistik untuk mengklasifikasikan data. Pendekatan ini bekerja dengan menghitung probabilitas bersyarat P(x|y) berdasarkan probabilitas kelas X yang telah diketahui. Proses klasifikasi dilakukan dengan menentukan kelas yang memiliki nilai probabilitas maksimum dari P(x|y), sehingga data baru dapat dikategorikan secara optimal. Salah satu keunggulan utama dari metode ini adalah efisiensinya dalam menangani data, di mana hanya diperlukan sejumlah kecil data pelatihan untuk memperkirakan parameter yang dibutuhkan dalam proses klasifikasi. Hal ini menjadikan *Naïve Bayes* sebagai algoritma yang cepat dan efektif, terutama dalam analisis teks seperti analisis sentimen terhadap komentar YouTube mengenai mobil listrik[22].

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

Rumus 2.5 *Naïve Bayes*

Berikut penjelasan rumus *Naïve Bayes* pada rumus 2.5 :

Keterangan :

X = Data dengan class yang belum diketahui

H = Hipotesis data X merupakan suatu class spesifik

P(H|X) = Probabilitas hipotesis H berdasarkan kondisi x

P(H) = Probabilitas hipotesis H

P(X|H) = Probabilitas X berdasarkan kondisi tersebut

P(X) = Probabilitas dari X

2.3.4 Random Forest

Algoritma *Random Forest* merupakan salah satu metode *ensemble learning* yang memanfaatkan sejumlah *decision tree* sebagai *base classifier* yang dibentuk melalui proses pembelajaran terarah. Salah satu karakteristik utama dari algoritma ini adalah penggunaan teknik *guided sampling* dalam membangun setiap pohon prediksi, di mana masing-masing pohon dilatih dengan variabel acak untuk meningkatkan keberagaman model. Proses prediksi dilakukan dengan menggabungkan hasil dari seluruh pohon menggunakan pendekatan *majority voting* pada kasus klasifikasi dan *average voting* pada regresi. Dibandingkan dengan metode *ensemble* lainnya seperti *bagging* dan *boosting*, *Random Forest* dikenal memiliki tingkat akurasi yang tinggi, tahan terhadap *outlier* dan *noise*, serta lebih efisien dalam hal waktu komputasi[16].

2.3.5 Text Mining

Tahap pengumpulan teks dalam Text Mining bertujuan untuk memperoleh data dari berbagai sumber, seperti dokumen digital, situs web, atau media sosial, guna mendukung analisis yang sesuai dengan tujuan penelitian. Text Mining sendiri merupakan bagian dari Data Mining yang berfokus pada analisis teks dalam jumlah besar dengan menerapkan metode kategorisasi dan teknik pemrosesan bahasa alami (*Natural Language Processing*). Dalam konteks analisis sentimen, *Text Mining* digunakan untuk mengidentifikasi dan mengklasifikasikan sentimen yang terkandung dalam teks, baik positif, negatif, maupun netral. Analisis ini dapat diterapkan pada berbagai jenis data tekstual, seperti opini, ulasan produk, kritik, dan komentar di media sosial. Apabila dikelola dengan baik, hasil dari *Text Mining* dapat memberikan wawasan yang berharga bagi individu maupun organisasi. Wawasan ini dapat dimanfaatkan untuk mendukung pengambilan keputusan bisnis, meningkatkan layanan pelanggan, serta memahami persepsi publik terhadap suatu produk atau layanan[23].

2.3.4.1 Data Pre-processing

Data Preprocessing merupakan tahap awal dalam proses Text Mining yang bertujuan untuk mengubah teks tidak terstruktur menjadi

format yang lebih terorganisir dan siap untuk dianalisis. Text Preprocessing melibatkan serangkaian teknik pemrosesan teks guna meningkatkan kualitas data sebelum digunakan dalam analisis lebih lanjut, seperti klasifikasi atau analisis sentiment ada beberapa langkah utama dalam *text preprocessing* meliputi *case folding, tokenizing, stemming dan stopward removal*[24].

2.3.4.2 Case Folding

Case Folding adalah salah satu langkah dalam text preprocessing yang berfungsi untuk menyeragamkan teks dengan mengubah semua huruf menjadi huruf kecil serta menghapus karakter non-alfanumerik dan tanda baca yang tidak relevan.

Dalam konteks analisis sentimen terhadap komentar YouTube tentang mobil listrik, penerapan *case folding* membantu memastikan bahwa perbedaan penggunaan huruf kapital tidak mempengaruhi hasil analisis. Sebagai contoh, komentar "MOBIL LISTRIK ITU MAHAL BANGET!!!" setelah melalui proses case folding akan menjadi "mobil listrik itu mahal banget". Dengan cara ini, teks lebih terstruktur dan lebih mudah dianalisis oleh algoritma *machine learning* seperti SVM, Naïve Bayes, dan Random Forest dalam klasifikasi sentimen.

2.3.4.3 Tokenizing

Tokenisasi adalah salah satu tahap dalam text preprocessing yang bertujuan untuk memecah teks menjadi unit-unit kecil yang disebut token. Token dapat berupa kata, frasa, atau bahkan karakter tertentu yang memiliki makna dalam analisis teks.

Dalam proses tokenisasi, tanda baca, angka, serta karakter lain yang tidak relevan dihilangkan agar teks lebih terstruktur dan mudah diproses oleh algoritma Natural Language Processing (NLP) atau model machine learning. Sebagai contoh, dalam konteks analisis sentimen komentar YouTube tentang mobil listrik, kalimat: "Mobil listrik ini keren banget, tapi harganya masih mahal!" Setelah melalui proses tokenisasi, akan diubah

menjadi: ["mobil", "listrik", "ini", "keren", "banget", "tapi", "harganya", "masih", "mahal"]

Dengan pemisahan ini, teks dapat dianalisis lebih lanjut, seperti pengkodean kata (*vectorization*), analisis sentimen menggunakan SVM, Naïve Bayes, dan Random Forest, atau pengelompokan kata berdasarkan pola tertentu.

2.3.4.4 Stemming

Stemming adalah salah satu tahap dalam text preprocessing yang bertujuan untuk mengubah kata-kata menjadi bentuk dasarnya sesuai dengan kaidah Bahasa Indonesia. Proses ini dilakukan dengan menghilangkan afiks (awalan, sisipan, akhiran) sehingga kata-kata yang memiliki makna serupa dapat dianalisis lebih efektif.

Dalam analisis sentimen terhadap komentar YouTube tentang mobil listrik, *stemming* sangat penting untuk memastikan bahwa variasi kata tetap dikenali sebagai entitas yang sama. Sastrawi, salah satu library *Natural Language Processing (NLP)* di Python, sering digunakan untuk melakukan stemming dalam bahasa Indonesia.

Sebagai contoh:

- "mengisi" → "isi"
- "terkenalnya" → "terkenal"
- "penjualan" → "jual"

Dengan menerapkan stemming, analisis teks dapat menjadi lebih akurat karena mengurangi redundansi kata dan meningkatkan efisiensi pemrosesan data dalam model SVM, Naïve Bayes, dan Random Forest.

2.3.4.5 Stopword Removal

Stopword Removal adalah tahap dalam text preprocessing yang bertujuan untuk menyaring kata-kata yang kurang penting dari hasil tokenisasi, sehingga hanya kata-kata yang memiliki makna signifikan yang

dipertahankan. Dalam analisis sentimen terhadap komentar YouTube tentang mobil listrik, teknik ini membantu meningkatkan akurasi dengan menghilangkan kata-kata yang tidak berkontribusi pada analisis, seperti kata hubung dan kata umum yang sering muncul.

Terdapat dua metode utama dalam Stopword Removal:

1. *Stoplist* adalah daftar kata-kata yang dianggap tidak relevan dalam analisis teks. Kata-kata seperti *"dan"*, *"atau"*, *"yang"*, dan *"ke"* sering muncul dalam teks tetapi tidak memberikan makna signifikan dalam analisis sentimen. Dengan menghapus kata-kata ini, model SVM, Naïve Bayes, dan Random Forest dapat fokus pada kata-kata yang lebih bermakna dalam menentukan sentimen.
   Contoh sebelum Stoplist:
   "Mobil listrik yang mahal dan pengisian dayanya lama"
   Setelah Stoplist:
   "Mobil listrik mahal pengisian daya lama"
2. *Wordlist* adalah daftar kata-kata yang dianggap penting dalam konteks analisis. Dalam penelitian **analisis sentimen mobil listrik**, kata-kata seperti *"baterai"*, *"mahal"*, *"cepat"*, *"kenyamanan"*, dan *"performa"* dapat dimasukkan dalam *wordlist* agar tetap dipertahankan dalam analisis.Dengan menerapkan **Stopword Removal**, proses analisis sentimen menjadi lebih efektif karena hanya kata-kata yang relevan yang dipertimbangkan dalam klasifikasi sentimen positif, negatif, atau netral.

2.4 Tools Penelitian

2.4.1 YouTube

YouTube pertama kali didirikan pada tahun 2005 sebagai platform streaming video. Perusahaan ini awalnya didirikan oleh tiga orang karyawan yang terinspirasi dari nama sebuah kedai pizza dan restoran Jepang di San Mateo, California. Popularitas YouTube meningkat pesat dalam waktu satu tahun setelah peluncurannya. Pada tahun 2006, jumlah unggahan video di

platform ini mencapai 65.000 unggahan baru per hari dan terus bertambah hingga mencapai 100.000 unggahan pada bulan Juli 2006. Sebelumnya, pada bulan Juni 2006, YouTube telah menjalin kerja sama dengan NBC dalam bidang pemasaran dan periklanan. Perkembangan YouTube semakin pesat setelah Google mengakuisisi platform ini pada Oktober 2006 dengan nilai investasi sebesar 1,65 miliar USD. Akuisisi ini berkontribusi pada ekspansi global YouTube sebagai salah satu platform berbagi video terbesar di dunia. Prestasi YouTube juga diakui secara internasional, terbukti dengan penghargaan yang diterimanya dari majalah *PC World*, yang menempatkannya sebagai salah satu dari sepuluh produk terbaik tahun 2006[25].

2.4.2 Python

Python adalah bahasa pemrograman tingkat tinggi yang bersifat interpreted, dikembangkan oleh Guido van Rossum. Bahasa ini sangat populer karena readability-nya yang tinggi dan sintaks yang ringkas. Python menggunakan indentasi whitespace sebagai penanda blok kode, yang membuatnya lebih bersih dan mudah dipahami dibandingkan bahasa pemrograman lain. Python memiliki library standar yang luas, yang dapat digunakan dalam berbagai aplikasi, termasuk pemrosesan bahasa alami (NLP), pembelajaran mesin (Machine Learning), dan analisis data. Dalam konteks analisis sentimen terhadap komentar YouTube tentang mobil listrik, Python menyediakan berbagai alat dan pustaka seperti Sastrawi, scikit-learn, dan NLTK untuk melakukan text preprocessing, pembangunan model klasifikasi, serta evaluasi performa model. Python menjadi pilihan utama untuk proyek-proyek kompleks karena kesederhanaannya, dukungan pustaka yang beragam, dan sifatnya yang dinamis, sehingga memudahkan pengembangan dan eksperimen dalam analisis data[26].

2.4.3 Jupyter

Jupyter Notebook (sebelumnya dikenal sebagai IPython Notebook) adalah aplikasi web interaktif yang digunakan untuk membuat dan berbagi dokumen komputasional. Awalnya dikembangkan dengan nama IPython, proyek ini kemudian berganti nama menjadi Jupyter pada tahun 2014. Jupyter

Notebook sepenuhnya bersifat sumber terbuka, sehingga semua fungsionalitasnya dapat digunakan secara gratis. Aplikasi ini mendukung lebih dari 40 bahasa pemrograman, termasuk Python, R, dan Scala. Setiap notebook dalam Jupyter disimpan dalam format .ipynb, yang bersifat mutable atau dapat diubah. Untuk memudahkan pengelolaan dokumen, Jupyter menyediakan dashboard notebook yang memungkinkan pengguna mengorganisasi berbagai proyek mereka. Selain itu, Jupyter menggunakan kernel, yaitu proses yang menjalankan kode secara interaktif dalam bahasa pemrograman tertentu dan mengembalikan hasil eksekusi kepada pengguna. Kernel juga mendukung fitur seperti tab completion dan introspeksi kode. Jupyter Notebook memiliki fitur konversi dokumen ke berbagai format standar seperti HTML, LaTeX, PDF, Markdown, dan Python melalui opsi "Download As" di antarmuka web. Proses konversi ini juga dapat diotomatisasi menggunakan alat seperti nbconvert, yang memudahkan dokumentasi serta distribusi kode dan analisis data[27].