BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Penelitian yang telah dilakukan sebelumnya penting sebagai landasan dalam melaksanakan penelitian yang akan dilakukan. Tabel 2.1 merupakan beberapa penelitian terdahulu.

Tabel 2.1 Penelitian terdahulu

1	Judul Artikel	Analisis sentimen mobil listrik menggunakan metode Naïve Bayes Classifier [17]
	Jurnal dan Tahun	INFOTECH: Jurnal Informatika & Teknologi, 2024
	Metode	CRISP-DM dan Naïve Bayes Classifier
	Permasalahan	Minimnya pemahaman tentang bagaimana masyarakat merespons keberadaan mobil listrik di Indonesia
	Hasil	Mayoritas masyarakat memiliki sentimen positif terhadap mobil listrik, diikuti netral, dan hanya sedikit negatif. Akurasi model mencapai 94.4% dengan AUC 0.997.
2	Judul Artikel	Analisis Sentimen Terhadap Mobil Listrik di Indonesia pada Twitter: Penerapan Naïve Bayes Classifier [18]
	Jurnal dan Tahun	Just IT: Jurnal Sistem Informasi, Teknologi Informasi dan Komputer, 2024
	Metode	Naïve Bayes Classifier
	Permasalahan	Tingginya polusi dan pertumbuhan mobil listrik belum diimbangi pemahaman sentimen masyarakat
	Hasil	Sentimen masyarakat: 52% positif, 48% negatif. Akurasi model (RapidMiner): 100% dari 385 data akurat dari total 500 tweet

3	Judul Artikel	Komparasi Algoritma Support Vector Machine dengan Naive Bayes Untuk Analisis Sentimen Pada Aplikasi BRImo [10]
	Jurnal dan Tahun	Bangkit Indonesia, Vol. XI, No.02, 2022
	Metode	SVM, Naïve Bayes + TF-IDF
	Permasalahan	Menentukan algoritma terbaik dalam klasifikasi sentimen ulasan aplikasi BRImo
	Hasil	SVM unggul dengan rata-rata akurasi 97.56%, dibandingkan Naïve Bayes 96.52%.
4	Judul Artikel	Penerapan *Naïve Bayes Classifier*, *K-Nearest Neighbor* (KNN) dan *Decision Tree* untuk Menganalisis Sentimen pada Interaksi Netizen dan Pemerintah
	Jurnal dan Tahun	Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer, Vol. 21, No. 1, 2021
	Metode	*Naïve Bayes Classifier*, *K-Nearest Neighbor* (KNN), *Decision Tree*, menggunakan *RapidMiner* dan data dari Twitter
	Permasalahan	Pemerintah Kota Pekanbaru mendapat berbagai respons atas penerapan teknologi, namun data publik dari Twitter belum dikaji secara mendalam untuk memahami persepsi masyarakat terkait program seperti smart city, command center, dan CCTV.
	Hasil	Penelitian mengumpulkan 115 tweet tentang pemerintah Pekanbaru dan dianalisis dengan tiga metode: Naïve Bayes, KNN, dan Decision Tree. Naïve Bayes mencapai akurasi tertinggi sebesar 100%, KNN 98,25%, dan Decision Tree hanya 62,28%. Sentimen netral mendominasi, menandakan masyarakat belum sepenuhnya peduli terhadap implementasi teknologi pemerintah. Naïve Bayes direkomendasikan untuk data di bawah 200 karena kinerjanya sangat baik. Studi ini

		menyarankan untuk menambah volume data dan menggunakan teknik seleksi fitur agar akurasi metode lain juga meningkat. [9]
5	Judul Artikel	E-Wallet Sentiment Analysis Using Naïve Bayes and Support Vector Machine Algorithm [19]
	Jurnal dan Tahun	Journal of Physics: Conf. Series, 2020
	Metode	Naïve Bayes & SVM + N-Gram + Confusion Matrix + ROC
	Permasalahan	Masih terdapat beberapa permasalahan seperti infrastruktur yang belum memadai, harga yang relatif mahal, dan waktu pengisian baterai yang lama. Untuk meningkatkan penggunaan kendaraan listrik, diperlukan pemahaman dan kesadaran dari masyarakat. Oleh karena itu, penelitian ini bertujuan untuk Mengetahui algoritma terbaik untuk analisis sentimen ulasan pengguna aplikasi OVO dan DANA
	Hasil	Naïve Bayes memiliki akurasi lebih tinggi (OVO: 94.90%), tapi AUC tertinggi dimiliki oleh SVM (OVO: 0.986) menandakan performa klasifikasi yang sangat baik
6	Judul Artikel	Sentiment Analysis on myIndiHome User Reviews Using Support Vector Machine and Naïve Bayes Classifier Method [20]
	Jurnal dan Tahun	International Journal of Industrial Optimization, Vol. 2 No. 2, 2021
	Metode	SVM dan Naïve Bayes
	Permasalahan	Evaluasi kualitas layanan IndiHome berdasarkan ulasan pengguna di Play Store
	Hasil	SVM lebih unggul (86.54%) dibanding Naïve Bayes (84.69%). Ditemukan 12 masalah layanan, dikategorikan dalam 5P: *Price, People, Process, Place, Product.*

7	Judul Artikel	*Fine-Grained Sentiment Analysis on PeduliLindungi Application Using Multinomial Naïve Bayes-SMOTE* [21]
	Jurnal dan Tahun	IEEE, 2022
	Metode	Multinomial Naïve Bayes + SMOTE
	Permasalahan	Ketidakseimbangan jumlah data ulasan positif dan negatif pada aplikasi PeduliLindungi serta kebutuhan untuk memahami reaksi pengguna secara mendalam
	Hasil	Dari total 9021 ulasan, sebanyak 6244 adalah negatif dan 2777 positif. Penelitian menggunakan metode SMOTE untuk menyeimbangkan data dan meningkatkan akurasi klasifikasi. Model Multinomial Naïve Bayes-SMOTE terbukti paling efektif karena menghasilkan AUC tertinggi, menunjukkan bahwa metode ini paling andal untuk klasifikasi sentimen pada data yang tidak seimbang.
8	Judul Artikel	Perbandingan Metode Klasifikasi Support Vector Machine Dan Naïve Bayes Pada Analisis Sentimen Kendaraan Listrik [22]
	Jurnal dan Tahun	Jurnal SPEKTRUM, 2023
	Metode	Support Vector Machine, Naïve Bayes
	Permasalahan	Untuk menghadapi permasalahan emisi gas rumah kaca. Peralihan ke kendaraan listrik bisa menjadi solusi yang efektif karena kendaraan listrik memiliki banyak keunggulan. Namun, penerimaan kendaraan listrik di Indonesia tergantung pada opini atau sentimen yang diberikan oleh masyarakat.
	Hasil	Sebanyak 717 data digunakan sebagai data uji SVM didapat 150 data negatif, 152 data netral, dan 277 data positif. Sedangkan Naïve Bayes mengklasifikasikan dengan benar sebanyak 166

		data negatif, 143 data netral, dan 282 data positif. Hasil akurasi SVM sebesar 81% dan Naïve Bayes sebesar 82%.
9	Judul Artikel	Analisis Sentimen dengan SVM, Naïve Bayes dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial Twitter [13]
	Jurnal dan Tahun	PRISMA, Prosiding Seminar Nasional Matematika, Vol. 4, 2021
	Metode	Support Vector Machine (SVM), Naïve Bayes, dan K-Nearest Neighbor (KNN), dengan evaluasi menggunakan 10-Fold Cross Validation
	Permasalahan	Banyaknya opini masyarakat Indonesia terhadap pandemi Covid-19 di media sosial Twitter belum diklasifikasikan secara sistematis menjadi sentimen positif atau negatif. Diperlukan analisis yang dapat mengidentifikasi dan membandingkan metode terbaik dalam klasifikasi data opini publik dalam jumlah besar.
	Hasil	Penelitian ini menggunakan 10.000 tweet dari Maret hingga Juli 2020 dengan teknik text crawling dan preprocessing teks seperti case folding, tokenizing, stemming, dan filtering. Hasil pelabelan menunjukkan 6.128 tanggapan positif dan 3.872 tanggapan negatif. Setelah dilakukan klasifikasi, diperoleh hasil: SVM menghasilkan akurasi tertinggi (90,1%), karena unggul dalam menangani data berdimensi tinggi seperti teks. Naïve Bayes mendapat akurasi 79,2% karena metode ini efisien dan sederhana, namun terkendala asumsi independensi fitur. Sementara itu, KNN mendapat akurasi paling rendah (62,1%) karena sensitif terhadap fitur tidak relevan dan kompleksitas tinggi. Penelitian menyimpulkan

		bahwa SVM paling cocok untuk analisis sentimen Twitter terkait Covid-19 di Indonesia. Visualisasi Wordcloud dan asosiasi kata memperkuat hasil klasifikasi dan menggambarkan isi sentimen publik.
10	Judul Artikel	*The effect of review content and electronic word of mouth on the purchase intention of electric cars (a quantitative study on the youtube channels of* fitra eri, ridwan hanif, *and* oto driver) [23]
	Jurnal dan Tahun	Jurnal Ilmiah Manajemen Informasi dan Komunikasi, 2023
	Metode	kuantitatif dengan 100 responden, menggunakan software SPSS 17 melalui beberapa uji seperti uji instrumen (uji validitas dan uji reliabilitas), uji asumsi klasik (uji normalitas, uji linearitas, uji heteroskedastisitas, uji autokorelasi), uji hipotesis (uji regresi linier berganda, uji R2, uji t, uji f)
	Permasalahan	Menguji pengaruh dari konten *review* dan *electronic-word of mouth* pada Youtube terhadap minat beli mobil listrik
	Hasil	Variabel konten review mobil listrik berpengaruh terhadap minat beli, Variabel eWOM berpengaruh sebesar 70% terhadap minat beli

Pada analisis sentimen yang dilakukan peneliti, terdapat sepuluh (10) jurnal yang digunakan sebagai acuan utama dalam pembuatan penelitian ini, Berikut perbedaan yang akan dilakukan dari penelitian terdahulu.

Penelitian ini memanfaatkan algoritma *machine learning*, yaitu *Naïve Bayes, Support Vector Machine* (SVM), *dan K-Nearest Neighbor* (KNN) yang telah diseimbangkan kelasnya menggunakan teknik SMOTE, dan mengoptimalkan kinerja model dengan pencarian parameter terbaik melalui metode *Grid Search*. Secara umum telah terbukti mampu memberikan tingkat akurasi yang cukup baik dalam melakukan klasifikasi data. Namun demikian,

hingga saat ini masih terbatas jumlah penelitian yang secara eksplisit membandingkan kinerja ketiga algoritma tersebut dalam konteks analisis sentimen terhadap ulasan mengenai mobil listrik. Berbeda dengan sejumlah penelitian sebelumnya yang cenderung membahas mobil listrik secara umum, penelitian ini secara khusus difokuskan pada dua jenis mobil listrik yaitu BYD M6 dan Wuling Binguo EV, guna memperoleh hasil analisis yang lebih mendalam dan terfokus.

Dataset yang digunakan dalam penelitian ini diperoleh dari komentar pada video YouTube yang diunggah oleh kreator konten terverifikasi dan terpercaya, dengan batasan waktu publikasi berada dalam kurun 1 hingga 2 tahun terakhir, terhitung sejak peluncuran masing-masing mobil. Selain itu, penelitian ini mengklasifikasikan data ke dalam tiga kategori sentimen, yaitu positif, netral, dan negatif. Hal ini menjadi pembeda dari sebagian besar penelitian terdahulu yang umumnya hanya membedakan sentimen ke dalam dua kategori, yakni positif dan negatif.

2.2 Teori Penelitian

2.2.1 Analisis Sentimen

Analisis sentimen merupakan salah satu pembelajaran *machine learning* (komputasi mesin) yang mempelajari mengenai emosi, pendapat, penilaian atau suatu pandangan dari kumpulan teks untuk mengidentifikasi suatu karakteristik dari kumpulan kata. Proses dari analisis sentimen ini adalah mengklasifikasi dokumen berbentuk teks ke dalam suatu kelas positif atau negatif [24]. Dalam melabeli analisis sentimen, diperlukan ulasan sebagai sumber data. Ulasan ini sering mengandung pendapat positif dan negatif tentang perubahan yang terjadi. Pelabelan dilakukan untuk mengklasifikasikan apakah ulasan pengguna bersifat positif, negatif, atau netral dalam penelitian analisis sentimen [23]. Ulasan tersebut memberikan wawasan tentang dampak yang dihasilkan. Hal ini membantu dalam konteks analisis sentimen untuk memahami apakah pendapat terkait dengan masalah atau objek tertentu cenderung positif atau negatif [13]. Selain itu, analisis sentimen juga

merupakan teknik untuk memeriksa bagaimana masyarakat merespons perubahan [14]

2.2.2 Mobil Listrik

Mobil listrik adalah jenis kendaraan bermotor yang menggunakan energi listrik sebagai sumber tenaga [25]. Mobil listrik memiliki beberapa keunggulan dibandingkan dengan mobil konvensional yang menggunakan bahan bakar bensin. Mobil listrik (Electric Vehicle/EV) adalah kendaraan yang digerakkan oleh motor listrik dengan sumber energi dari baterai yang dapat diisi ulang. EV menjadi solusi potensial untuk mengurangi emisi karbon dan ketergantungan terhadap bahan bakar fosil, serta mendukung sistem transportasi berkelanjutan. Kemajuan teknologi penyimpanan energi, terutama teknologi baterai, telah membuat penggunaan kendaraan listrik menjadi populer akhir-akhir ini.

2.2.3 Machine Learning

Machine Learning merupakan salah satu disiplin dalam kecerdasan buatan (*Artificial Intelligence*) yang memberikan kemampuan kepada sistem komputer untuk belajar dari data dan melakukan adaptasi tanpa perlu instruksi pemrograman secara eksplisit. Dalam penerapannya, Machine Learning memanfaatkan algoritma serta model statistik untuk menganalisis data, mengenali pola, dan menemukan hubungan penting yang dapat mendukung proses prediksi maupun pengambilan keputusan. Pendekatan ini secara signifikan berkontribusi terhadap peningkatan akurasi hasil dan performa sistem prediktif [26].

Secara garis besar, *Machine Learning* diklasifikasikan ke dalam tiga kategori utama, yaitu *supervised learning*, *unsupervised learning*, dan *semi-supervised learning*. *Supervised learning* menggunakan data berlabel sebagai dasar pelatihan model agar mampu melakukan prediksi atau klasifikasi secara akurat. Di sisi lain, *unsupervised learning*

mengolah data tanpa label untuk mengungkap struktur tersembunyi yang terdapat dalam data kompleks. Teknik ini banyak digunakan dalam analisis data seperti teks, gambar, suara, maupun video. Adapun *semi-supervised learning* menggabungkan penggunaan data berlabel dan tidak berlabel dalam proses pelatihan, sehingga memaksimalkan manfaat dari kedua jenis data tersebut [26].

2.2.4 Youtube

Media sosial merupakan salah satu platform online yang memudahkan pengguna untuk berinteraksi dan berbagi informasi dari jarak jauh. Media sosial tidak mengenal batasan jarak dan waktu, sehingga memungkinkan pengguna terhubung dengan orang lain di seluruh dunia [27]. Media sosial juga memberikan kesempatan bagi pengguna untuk berbagi informasi dan menyampaikan pendapat secara cepat, seperti melalui komentar atau forum diskusi di dalam platform tersebut [23].

Ada beberapa keuntungan bagi pengguna dalam menggunakan media sosial, diantaranya:

a. Menjalin suatu hubungan pertemenan dengan berbagai pengguna dari seluruh dunia melalui media sosial [27].

b. Mendapatkan motivasi bagi pengguna untuk dapat mengembangkan keterampilan diri dari media sosial [27].

c. Mempererat hubungan antar sesama pengguna dalam bersahabat, berempati, dan memiliki perhatian dengan satu sama lain dalam media sosial [27].

Dari keuntungan-keuntungan diatas, media sosial dapat menjadi wadah dalam menjembatani para pengguna dalam berinteraksi dan mengekspresikan dirinya, serta menyampaikan pendapat dengan mudah.

2.2.5 *Cross-Industry Standard Process for Data Mining* (CRISP-DM)

CRISP-DM merupakan sebuah pendekatan sistematis yang berfungsi sebagai kerangka kerja dalam pelaksanaan proses data mining. Kerangka ini dikembangkan sebagai solusi atas tantangan yang sering dihadapi dalam konteks bisnis maupun kegiatan penelitian [28]. Terdapat enam tahap utama dalam struktur CRISP-DM, yaitu:

1. ***Business Understanding***

   Tahap ini berfokus pada pencapaian pemahaman yang mendalam mengenai tujuan dan kebutuhan dari sudut pandang bisnis. Setelah itu, informasi yang relevan akan dikumpulkan untuk merumuskan masalah yang akan diselesaikan, kemudian dirancang strategi guna mencapai tujuan yang telah ditentukan [28].

2. ***Data Understanding***

   Proses ini diawali dengan pengumpulan data dan dilanjutkan dengan analisis awal untuk memahami karakteristik data. Selain mengevaluasi kualitas data, tahap ini juga bertujuan mengidentifikasi potensi permasalahan, menemukan pola, dan mengungkap informasi tersembunyi dalam data yang telah dikumpulkan [28].

3. ***Data Preparation***

   Data Preparation adalah tahap di mana data mentah diolah, dibersihkan, dan disiapkan sehingga membentuk dataset akhir yang siap digunakan dalam proses pemodelan [28].

4. ***Data Modelling***

   Pada tahap ini, berbagai teknik pemodelan diterapkan dan diuji untuk memilih metode terbaik. Proses ini mempertimbangkan parameter optimal guna mendapatkan performa analisis yang maksimal. Evaluasi dilakukan secara cermat terhadap setiap pendekatan model yang digunakan untuk memperoleh solusi paling efektif dan efisien dalam konteks permasalahan data [28].

5. ***Evaluation***

Tahap Evaluation melibatkan analisis kritis terhadap model yang telah dibangun serta peninjauan menyeluruh atas kemampuannya dalam mencapai tujuan penelitian. Penilaian menyeluruh dilakukan untuk memastikan model sesuai dan layak diterapkan dalam konteks yang telah dirancang [28].

6. ***Deployment***

Deployment merupakan fase di mana model yang telah dikembangkan mulai diterapkan secara nyata bagi pengguna. Biasanya, ini dilakukan melalui pengembangan aplikasi atau platform yang dapat diakses serta dimanfaatkan oleh pengguna akhir [28].

2.2.6 KDD

Knowledge Discovery in Database (KDD) adalah suatu kerangka kerja dalam data mining yang digunakan untuk mengekstraksi informasi yang bermakna dan bermanfaat dari kumpulan data berukuran besar [26]. KDD mencakup keseluruhan proses yang lebih luas dalam mengubah masalah atau pertanyaan menjadi solusi berbasis data. berikut penjelasan dari metode KDD:

1. *Selection*

   Selection merupakan langkah awal dalam metode penelitian dengan KDD Process yang digunakan untuk mengumpulkan data sesuai dengan kebutuhan penelitian ini. Proses pemilihan data mencakup pengumpulan informasi dari berbagai sumber dengan mempertimbangkan seleksi data yang relevan sesuai dengan tujuan penelitian.

2. *Pre-Processing*

   Proses selanjutnya menekankan pada pembersihan data sebelum dilakukan proses data mining. Proses cleansing data melibatkan langkah-langkah seperti membersihkan unstructured data dan menghilangkan data yang terdapat missing value untuk menjaga

integritas data dengan relevan agar dapat digunakan dalam penelitian. Dataset yang telah disiapkan sebelumnya harus mengikuti tahap preprocessing terlebih dahulu.

3. *Transformation*

   Tahap ini dilakukan melalui proses seleksi diubah menjadi format yang sesuai agar dapat digunakan dalam proses data mining secara efektif. Terdapat beberapa proses yang akan dijalani seperti melakukan labelling data serta melakukan pembobotan kata dengan tujuan mendapatkan nilai yang lebih optimal.

4. *Data Mining*

   Proses ini akan melibatkan penggunaan teknik data mining untuk membentuk model yang sudah direpresentasikan pada tahap Transformation [27]. Tahapan ini mencakup proses seperti *classification, clustering, association rule learning, dan regression,* serta eksplorasi data untuk mengoptimalkan kinerja algoritma.

5. *Interpretation/Evaluation*

   Langkah terakhir adalah mengevaluasi model yang telah berhasil dibentuk pada tahap Data Mining. Evaluasi dapat dilakukan dengan menggunakan berbagai metode seperti *Cross-Validation* atau *Confusion Matrix* guna menghitung metrik seperti *accuracy*, *precision, recall,* dan *F1-score.*

2.2.7 Web Scrapping

Web scraping merupakan teknik otomatisasi yang digunakan untuk mengambil data dari situs web secara terstruktur menggunakan perangkat lunak bernama *web scraper*. Perangkat ini mampu menjelajahi halaman web dan mengekstrak informasi penting dari berbagai sumber secara sistematis. Metode ini sering diterapkan untuk mengakses data dari platform e-commerce, media berita, maupun sumber informasi publik lain, yang kemudian dianalisis untuk kepentingan seperti riset pasar, pengembangan produk, dan analisis tren [29].

Salah satu contoh penerapan scraping adalah pada platform YouTube, di mana informasi seperti judul video, jumlah penonton, komentar, dan metadata lainnya dapat dikumpulkan untuk keperluan analisis konten, perilaku pengguna, atau pengembangan sistem rekomendasi. Meskipun demikian, proses ini harus dilakukan dengan memperhatikan batasan yang ditetapkan oleh kebijakan platform, termasuk penggunaan API resmi dan aspek legalitas akses data [29].

2.2.8 TF-IDF

TF-IDF adalah singkatan dari "*Term Frequency-Inverse Document Frequency*" yang merupakan metode untuk menghitung bobot kata dalam suatu dokumen atau koleksi dokumen. Teknik ini dapat digunakan buat menimbang hubungan antara kata atau istilah terhadap data [30]. Metode ini mengukur sejauh mana kata-kata dalam dokumen memiliki bobot penting dalam konteks keseluruhan koleksi dokumen, seperti mengidentifikasi kata-kata yang paling penting dalam dokumen dan menghitung bobot kata-kata dalam dokumen. Berikut ini adalah rumus untuk menghitung Term Frequency:

$$tf(t,d) = \frac{n_i j}{\sum_k n_i, j} \quad (1)$$

Rumus 2. 1 Formula *TF*

Deskripsi:

$tf(t,d)$ = Frekuensi *term.*
$n_i j$ = Total kemunculan seluruh kata pada dokumen.
$\sum_k n_i,\ j$ = total seluruh kata pada dokumen.

Rumus IDF

$$idf = \log \frac{N}{df_j} \quad (2)$$

Rumus 2. 2 Formula *IDF*

Deskripsi:

N = Total kelas.
df_j = Total kelas j yang memiliki isi kata i.

Rumus untuk menjalankan TF-IDF

$$W_{ij} = tf_{ij} \times idf \quad (3)$$

Rumus 2. 3 Formula *TF-IDF*

Deskripsi:

W_{ij} = Bobot kata I pada kelas j

tf_{ij} = Total kemunculan kata I pada kelas j.

df_j = Total Kelas j yang berisi kata i.

Oleh karena itu, metode ini bertujuan untuk menemukan representasi numerik dari setiap bagian data, yang menghasilkan pembuatan vektor antara data dan frekuensi yang ditunjukkan oleh representasi numerik dari istilah frekuensi yang berada dalam data. Jumlah bobot perhitungan yang dihasilkan akan meningkat, dan hasil kemiripan data dengan frekuensi juga akan meningkat.

2.2.9 Text preprocessing

Tahapan *text preprocessing* merupakan lanjutan dari tahap penambangan data. Hasil penambangan data belum rapi dan belum terstruktur. Oleh karena itu dibutuhkan tahap *text preprocessing* dimana tahap ini akan mengolah informasi yang masih mentah tadi, akan melalui serangkaian proses seperti berikut untuk

menghasilkan data yang dapat digunakan untuk penelitian, prosesnya sebagai berikut :

- *Data Cleansing*

  Pada tahap ini, dilakukan penghapusan karakter-karakter selain yang sudah ditentukan seperti huruf atau karakter di luar dari daftar alfabet a sampai dengan z termasuk tanda baca pada data komentar [23]. Contohnya seperti “mEsinnya bagus enggaa,,??” menjadi “mEsinnya bagus engga”.

- *Data labeling*

  Tahap *labeling data* akan dilakukan secara manual. Data komentar yang sudah terkumpul, komentar tersebut akan

diberikan label yang terbagi atas tiga kelas yaitu positif, netral, dan negatif [23].

- *Case Folding*

  Pada tahap ini, data komentar akan dilakukan pemrosesan berupa pengubahan seluruh teks huruf kapital atau huruf besar menjadi huruf kecil [23].

- *Tokenizing*

  Pada proses tokenizing ini akan akan dilakukan pemisahan sebuah teks panjang berupa paragraf atau kalimat menjadi teks terpisah yang berbentuk suatu bagianbagian kecil yang biasanya disebut token untuk dianalisa [23]. Contohnya seperti berikut:

  Kalimat = "ada mobil listrik baru nih"

  menjadi = ['ada', 'mobil', 'listrik', 'baru', 'nih']

Pada tahap tokenizing ini juga akan dilakukan penghapusan tanda baca, karena tanda baca akan mengganggu proses perhitungan dalam algoritma yang akan digunakan.

Tabel 2. 2 Rumus kernel

Jenis *Kernel*	Model
Linear	$K\ (x.x') = x.x$
Polynomial	$K\ (x.x') = (x.x' + \mathrm{c})'$
RBF Gaussian	$K\ (x.x') = \exp(-y\|\|x - x'\|\|^2)$
Siqmoid	$K\ (x.x) = \tanh(\alpha x.x + \beta)$

2.2.10 SMOTE

SMOTE (*Synthetic Minority Oversampling Technique*) merupakan teknik yang dirancang untuk menangani permasalahan ketidakseimbangan data, terutama ketika jumlah data pada salah satu kelas jauh lebih sedikit dibandingkan kelas lainnya [31]. Metode ini adalah pengembangan dari teknik oversampling, dengan tujuan untuk menyetarakan distribusi antara kelas minoritas dan mayoritas, sehingga algoritma pembelajaran mesin dapat belajar secara lebih seimbang dan akurat [31]. SMOTE menghasilkan data sintetis baru dengan mengambil

sampel dari kelas minoritas dan membuat data baru berdasarkan tetangga terdekat dari titik tersebut. Hasil dari proses ini adalah peningkatan jumlah representasi data minoritas. Agar teknik ini dapat diterapkan, data harus dalam format numerik serta memiliki label biner. Melalui pendekatan ini, proses analisis menjadi lebih representatif karena distribusi kelas dalam dataset telah diseimbangkan.

2.3 *Framework* dan Algoritma Penelitian

2.3.1 Naïve Bayes Classifier

Algoritma *Naïve Bayes* ditemukan pada pertengahan abad ke-18 oleh Reverend Thomas Bayes [32]. Metode *Naïve Bayes* populer digunakan untuk pengelompokan dan pengkategorian teks berdasarkan frekuensi kata-kata. *Naïve Bayes* digunakan untuk metode klasifikasi statistik yang memprediksi keanggotaan kelas dari sampel yang ada [32]. Pada penelitian ini, algoritma ini akan digunakan untuk mengklasifikasikan data teks yang diambil dari YouTube menjadi kelas positif atau negatif. *Naïve Bayes* merupakan metode klasifikasi yang didasarkan pada teorema Bayes untuk memprediksi peluang di masa depan berdasarkan data masa lalu. Kelebihan algoritma *Naïve Bayes* adalah memiliki akurasi yang baik dalam mengolah data yang berjumlah besar seperti analisis sentimen.

Naïve Bayes adalah algoritma pengklasifikasian probabilistik yang menghitung sekumpulan probabilitas dengan mengkombinasikan nilai dari dataset yang diberikan lalu hasilnya didapatkan dari frekuensi yang dijumlahkan. Algoritma teorema bayes mengasumsikan bahwa semua atribut yang ada merupakan atribut independen pada nilai yang diberikan pada variabel kelas. Kelebihan algoritma *Naïve Bayes* salah satunya adalah menghasilkan akurasi yang cukup baik dalam mengolah data besar seperti sentimen analisis. Berikut merupakan dari rumus *Naïve Bayes* [32].

$$P(c \mid x) = \frac{P(x|c).P(c)}{P(x)} \quad (4)$$

Rumus 2. 4 Formula *Naïve Bayes*

Deskripsi:

$P(c|x)$ = sebuah probabilitas kata x muncul pada kelas c.

$P(c)$ = probabilitas kata pada kelas c.

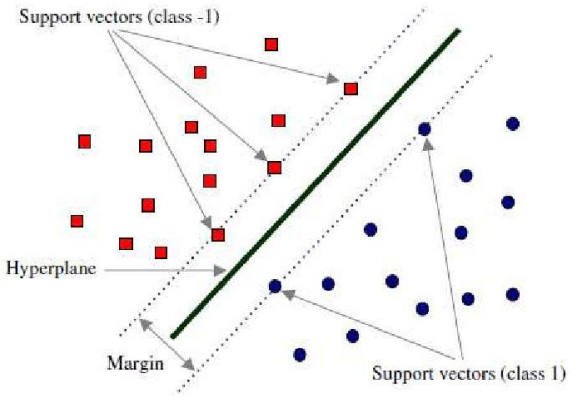
$P(x) =$ probabilitas kemunculan pada kata x.

2.3.2 Support Vector Machine

Support Vector Machine (SVM) merupakan algoritma supervised learning yang digunakan dalam tugas klasifikasi dan pertama kali dikenalkan pada tahun 1992. Dalam implementasinya, SVM dimulai dengan proses pelatihan untuk mempelajari karakteristik dan pola dari data yang akan dianalisis, kemudian dilanjutkan dengan proses pengujian guna mengukur performa model [21].

Prinsip kerja SVM adalah dengan menentukan sebuah garis atau hyperplane yang paling optimal untuk memisahkan dua kelas data yang berbeda, seperti kelas positif (+1) dan kelas negatif (−1). Gambar 2.1 memperlihatkan ilustrasi pemisahan data dengan margin maksimum, di mana hyperplane dengan jarak terluas dari kedua kelas diharapkan mampu memberikan hasil klasifikasi yang lebih baik [31].

Dalam praktiknya, SVM menyelesaikan persoalan klasifikasi melalui sistem persamaan atau pertidaksamaan linear. Namun, dengan bantuan fungsi kernel, SVM mampu mentransformasikan data ke ruang berdimensi lebih tinggi. Hal ini memungkinkan pemisahan data secara optimal menggunakan hyperplane, meskipun data awalnya tidak dapat dipisahkan secara linear dalam ruang asli.

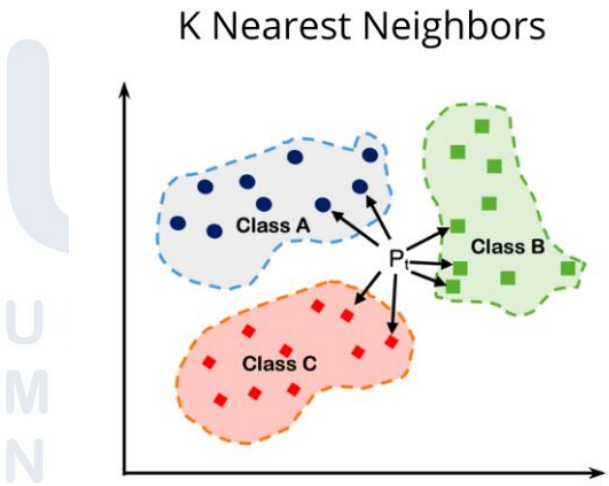


Gambar 2.1 Algoritma SVM [33]

SVM digunakan untuk menyelesaikan masalah klasifikasi dengan menggunakan persamaan atau pertidaksamaan linear. Dengan menerapkan teknik kernel pada SVM, model ini dapat memetakan data ke dalam ruang vektor berdimensi lebih tinggi, sehingga memungkinkan pemisahan data dengan hyperplane yang optimal [33]. Dengan demikian, masalah yang awalnya tidak dapat diselesaikan secara linear dapat diatasi.

2.3.3 K-Nearest Neighbours

Algoritma *K-Nearest Neighbor* (KNN) merupakan metode klasifikasi yang bekerja dengan mengelompokkan data berdasarkan sejumlah tetangga terdekat (K) untuk menentukan kelas dari data baru [34]. Proses klasifikasi dilakukan dengan memetakan data ke dalam ruang berdimensi tinggi. KNN bersifat non-parametrik, artinya tidak memerlukan parameter tetap dalam menjalankan proses klasifikasinya, sehingga jumlah parameternya tidak bersifat pasti atau tetap [34]. Berikut gambaran untuk memberikan ilustrasi mengenai cara kerja algoritma ini.



Gambar 2.2 Ilustasi cara kerja algoritma KNN [34]

Pada Gambar diatas diperlihatkan bagaimana algoritma *K-Nearest Neighbor* (KNN) mengklasifikasikan sebuah data uji, yang ditunjukkan dengan titik P, berdasarkan jarak kedekatannya terhadap data latih di

sekitarnya. Algoritma menghitung jarak antara titik P dan seluruh data latih, kemudian memilih K data latih yang paling dekat. Prediksi kelas untuk titik P ditentukan oleh mayoritas kelas dari tetangga-tetangga terdekat tersebut. Dalam ilustrasi tersebut, titik P dikelilingi oleh data dari tiga kelas yang berbeda (A, B, dan C), dan akan diklasifikasikan ke dalam kelas yang paling dominan di antara tetangga tersebut. Beberapa metode populer yang digunakan untuk mengukur kedekatan antar data antara lain adalah euclidean distance, hamming distance, manhattan distance, dan minkowski distance [34].

2.4 Tools Penelitian

2.4.1 Python

Python merupakan bahasa pemrograman tingkat tinggi yang dirancang oleh Guido Van Rossum pada tahun 1989 di *National Research Institute*, Belanda [20]. Python memiliki dukungan untuk pemrograman berbasis grafis seperti GUI Programming dan *Object Oriented Programming* (OOP), dan fitur utama yang menjadi ciri khas bahasa ini adalah alokasi memori dinamis yang efisien. Selain itu, Python sering digunakan sebagai bahasa utama untuk prototyping dalam melakukan pengembangan *web back-end* dan juga diterapkan dalam berbagai bidang, seperti data science dan analisis data besar [20].

2.4.2 Google Colaboratory

Google Colaboratory, yang dikenal dengan sebutan Google Colab, merupakan sebuah platform komputasi awan yang dikembangkan oleh Google Research untuk mendukung pengembangan dan eksekusi kode pemrograman dalam bahasa Python [26]. Platform ini secara khusus dirancang guna menunjang aktivitas di bidang pembelajaran mesin (*machine learning*), analisis data, serta kegiatan edukatif lainnya. Google Colab merupakan bentuk terintegrasi dari Jupyter Notebook yang dapat diakses secara daring tanpa memerlukan instalasi atau konfigurasi awal, sehingga memungkinkan pengguna untuk langsung menulis dan menjalankan kode secara efisien. Selain kemudahan akses,

Google Colab juga menyediakan sumber daya komputasi secara cuma-cuma, termasuk dukungan untuk *Central Processing Unit* (CPU) dan *Graphics Processing Unit* (GPU) [26].

2.4.3 Microsoft Excel

Microsoft Excel, yang lebih dikenal sebagai Excel, adalah sebuah produk perangkat lunak dari Microsoft yang digunakan untuk mengelola dan melakukan perhitungan terhadap data. Tidak hanya mengolah data, tetapi pengguna excel juga dapat membuat visualisasi data serta melakukan analisis statistic [29].

2.4.4 Visual Studio Code

Visual Studio Code (VS Code) merupakan editor kode sumber bersifat open-source yang bersifat lintas platform dan dikembangkan oleh Microsoft [35]. Editor ini dirancang untuk memenuhi kebutuhan pengembangan perangkat lunak modern, dengan menawarkan antarmuka yang intuitif, ringan, dan sangat fleksibel dalam hal kustomisasi. VS Code mendukung berbagai bahasa pemrograman serta dilengkapi dengan beragam ekstensi dan integrasi langsung dengan sistem kontrol versi seperti Git, yang secara signifikan meningkatkan efisiensi dan produktivitas dalam proses penulisan kode. Beberapa fitur unggulan dari VS Code mencakup IntelliSense yang menyediakan saran kode secara cerdas, kemampuan untuk melakukan debugging langsung dari editor, serta ekosistem ekstensi yang luas. Fitur-fitur tersebut menjadikan VS Code sebagai salah satu pilihan utama dalam berbagai skenario pengembangan, mulai dari pembangunan aplikasi web, layanan berbasis *cloud*, hingga pengembangan perangkat lunak lintas *platform* [35].