BAB II

LANDASAN TEORI

2.1. Data Science

Text Mining adalah cabang penting dari Data Science yang fokus pada ekstraksi informasi dan pengetahuan dari dokumen atau teks [6]. Dalam konteks Data Science, Text Mining membuka peluang besar untuk menganalisis dan memahami teks dalam skala besar, seperti artikel, laporan, pesan, dan lainnya. Tujuannya adalah untuk mengidentifikasi pola, tren, dan wawasan yang tersembunyi dalam data teks, memberikan nilai tambah yang signifikan dalam pengambilan Keputusan [7].

Data Science memanfaatkan berbagai teknik dalam Text Mining untuk mengolah dan menganalisis teks secara efektif. Salah satu teknik utama adalah pemrosesan bahasa alami (Natural Language Processing/NLP), yang memungkinkan komputer untuk memahami, menginterpretasi, dan menghasilkan teks dalam bahasa manusia [8]. Dengan menggunakan algoritma-algoritma NLP, Data Scientists dapat melakukan tugas seperti klasifikasi teks, ekstraksi entitas, dan sentiment analysis, yang semuanya memiliki implikasi penting dalam pemahaman konten teks.

Selain itu, Text Mining juga dapat digunakan untuk membuat ringkasan otomatis, memahami hubungan antar-entitas, dan membangun model prediktif berdasarkan data teks historis. Dengan menggabungkan Text Mining dalam analisis Data Science, organisasi dapat memanfaatkan data teks mereka untuk mendapatkan

pemahaman yang lebih dalam tentang kebutuhan pelanggan, tren industri, dan masalah yang mungkin timbul.

Dengan terus berkembangnya jumlah data teks yang dihasilkan setiap hari, Text Mining menjadi semakin vital dalam memanfaatkan potensi informasi yang terkandung dalam teks. Dengan cara ini, Data Science dan Text Mining saling melengkapi untuk menyediakan alat yang kuat untuk menggali pengetahuan dari sumber data teks yang melimpah.

2.1.1. Text Mining

Text Mining, juga dikenal sebagai Text Analytics atau Natural Language Processing (NLP), adalah cabang dari ilmu data yang fokus pada ekstraksi informasi yang bermanfaat dari teks [9]. Tujuannya adalah untuk mengubah data teks yang tidak terstruktur menjadi struktur yang dapat dianalisis secara lebih efektif [10]. Teknik-teknik dalam Text Mining mencakup pemrosesan bahasa alami, analisis sentimen, klasifikasi teks, dan ekstraksi informasi [11]. Salah satu aspek penting dari Text Mining adalah pemrosesan bahasa alami, yang melibatkan kemampuan komputer untuk memahami, menginterpretasi, dan menghasilkan bahasa manusia. Ini melibatkan tugas seperti tokenisasi (memecah teks menjadi unit-unit seperti kata atau frasa), penghilangan stop words (kata-kata umum yang sering tidak memberikan makna khusus), dan stemming (mengembalikan kata-kata ke bentuk dasarnya).

Analisis sentimen adalah aplikasi Text Mining yang populer, terutama dalam konteks media sosial dan ulasan produk. Melalui analisis sentimen, sistem dapat menentukan apakah teks menyiratkan sentimen positif, negatif, atau netral. Hal ini

dapat membantu perusahaan atau organisasi untuk memahami persepsi dan pendapat pelanggan mereka. Klasifikasi teks melibatkan pengelompokan teks ke dalam kategori atau kelas tertentu berdasarkan karakteristik atau topik tertentu. Contohnya termasuk pengelompokan email ke dalam kategori spam atau bukan spam, atau mengkategorikan artikel berita ke dalam topik-topik tertentu.

Ekstraksi informasi adalah aspek Text Mining yang berfokus pada identifikasi dan pengekstrakan informasi tertentu dari teks. Misalnya, ekstraksi informasi dapat digunakan untuk mengidentifikasi entitas seperti nama orang, tanggal, atau lokasi dari dokumen teks. Text Mining memiliki berbagai aplikasi, termasuk dalam penambangan data web, analisis sentimen untuk merespons umpan balik pelanggan, dan pemahaman konten teks untuk meningkatkan proses pengambilan keputusan. Keberhasilan Text Mining seringkali bergantung pada kualitas pemrosesan bahasa alami dan model machine learning yang digunakan untuk tugas-tugas khususnya.

Meskipun banyak potensi dan manfaat, Text Mining juga dihadapkan pada beberapa tantangan, seperti kompleksitas struktur bahasa, variasi gaya penulisan, dan interpretasi konteks. Seiring dengan perkembangan teknologi dan peningkatan dalam pemahaman bahasa alami, Text Mining terus menjadi bidang yang dinamis dan menarik dalam analisis data.

2.1.2. Database dan Data Processing

Database adalah sistem terstruktur yang dirancang untuk menyimpan, mengelola, dan mengakses data dengan efisien. Database berfungsi sebagai penyimpanan data yang terorganisir dalam tabel atau struktur data lainnya, dan dapat diakses oleh pengguna atau aplikasi perangkat lunak untuk mendukung

berbagai operasi, seperti pencarian, penyortiran, dan pembaruan data. Model data yang umum digunakan dalam database termasuk model relasional, di mana data diorganisir dalam tabel dan dihubungkan melalui kunci relasional. Data processing, di sisi lain, merujuk pada serangkaian langkah dan operasi yang dilakukan pada data untuk mengubahnya menjadi informasi yang lebih berguna. Proses ini melibatkan pengumpulan, pemrosesan, penyimpanan, dan pengambilan data. Data processing dapat dilakukan secara manual atau dengan menggunakan sistem komputer dan perangkat lunak khusus. Dalam konteks modern, data processing sering kali melibatkan penggunaan teknologi seperti komputer, server, dan algoritma pemrosesan data.

Sistem database dan data processing saling terkait erat. Database menyediakan tempat untuk menyimpan dan mengelola data, sementara data processing melibatkan operasi yang dilakukan pada data tersebut. Sistem database membantu dalam memudahkan akses, manajemen, dan keamanan data, sementara data processing mengizinkan transformasi data dari bentuk awalnya menjadi bentuk yang lebih bermakna atau berguna. Dalam dunia teknologi informasi, peran database dan data processing sangat penting. Perusahaan dan organisasi menggunakan database untuk menyimpan dan mengelola informasi penting, sementara data processing membantu dalam mengolah data ini untuk mendukung pengambilan keputusan, analisis, dan berbagai fungsi bisnis lainnya. Pengembangan teknologi ini terus berkembang, termasuk pendekatan baru seperti pemrosesan big data dan teknologi database yang dirancang untuk menangani volume data yang sangat besar dan beragam.

2.1.3. Visualisation

Visualisasi data adalah proses representasi grafis informasi dan data untuk mempermudah pemahaman, analisis, dan komunikasi. Tujuannya adalah membuat pola, tren, dan hubungan dalam data lebih jelas dan dapat dipahami dengan menggunakan elemen visual seperti grafik, diagram, dan peta. Visualisasi membantu manusia menjelajahi dan memahami informasi yang kompleks secara lebih efektif daripada hanya berdasarkan pada representasi numerik atau teks. Pentingnya visualisasi data terletak pada kemampuannya untuk menyampaikan cerita yang tersembunyi di balik angka dan fakta. Grafik dan diagram dapat mengungkap pola atau anomali dalam data, membuatnya lebih mudah untuk mengidentifikasi tren atau perubahan. Selain itu, visualisasi memainkan peran penting dalam mendukung pengambilan keputusan dengan menyajikan informasi secara intuitif dan menyederhanakan kompleksitas. Beberapa jenis visualisasi data yang umum digunakan meliputi diagram batang, diagram lingkaran, diagram garis, peta panas, dan diagram pencar. Setiap jenis visualisasi memiliki kegunaan tertentu tergantung pada jenis data dan pesan yang ingin disampaikan. Misalnya, diagram batang sering digunakan untuk membandingkan jumlah kategori, sedangkan diagram lingkaran efektif untuk menunjukkan proporsi. Dengan munculnya teknologi dan perangkat lunak visualisasi data yang lebih canggih, seperti Tableau, Power BI, dan D3.js, para profesional dapat membuat visualisasi data yang interaktif dan dinamis. Ini memungkinkan pengguna untuk menjelajahi data dengan lebih mendalam dan menyesuaikan tampilan sesuai kebutuhan mereka.

Visualisasi data tidak hanya terbatas pada bisnis dan analisis statistik; itu juga memiliki peran penting dalam ilmu pengetahuan, riset, dan pemberitaan. Melalui penggunaan visualisasi, informasi kompleks dapat disajikan dengan cara yang mudah dipahami oleh berbagai pemirsa, dari ahli hingga orang awam, membantu memperkuat komunikasi dan memperluas pemahaman kita tentang data dan fenomena di sekitar kita.

2.1.4. Statistik

Statistik adalah cabang ilmu matematika yang berkaitan dengan pengumpulan, analisis, interpretasi, presentasi, dan pengorganisasian data. Tujuan utama statistik adalah menyajikan data dalam bentuk yang bermakna dan memberikan wawasan yang dapat digunakan untuk membuat keputusan atau menyimpulkan informasi tentang suatu populasi. Statistik digunakan di berbagai bidang, termasuk ilmu sosial, ekonomi, kedokteran, sains alam, dan bisnis, untuk mengidentifikasi pola, tren, dan hubungan dalam data. Ada dua jenis statistik utama: statistik deskriptif dan inferensial. Statistik deskriptif digunakan untuk merangkum dan menggambarkan karakteristik dasar dari suatu dataset, seperti mean (rata-rata), median (nilai tengah), dan deviasi standar. Di sisi lain, statistik inferensial melibatkan penggunaan sampel data untuk membuat inferensi atau prediksi tentang populasi yang lebih besar. Ini melibatkan konsep probabilitas dan pembuatan keputusan berdasarkan data yang terbatas.

Metode statistik sering melibatkan pengujian hipotesis, analisis regresi, dan distribusi probabilitas. Pengujian hipotesis digunakan untuk menentukan apakah perbedaan atau hubungan yang diamati dalam sampel bersifat signifikan secara

statistik. Analisis regresi digunakan untuk memahami hubungan antara variabel-variabel yang berbeda. Distribusi probabilitas memainkan peran penting dalam membuat prediksi berdasarkan data dan mengukur ketidakpastian. Statistik juga merupakan alat penting dalam pengambilan keputusan. Bisnis menggunakan statistik untuk analisis pasar dan perencanaan bisnis. Dalam kedokteran, statistik digunakan untuk penelitian klinis dan pengambilan keputusan berdasarkan data kesehatan. Dalam ilmu sosial, statistik membantu merinci dan memahami perilaku manusia dan masyarakat. Dengan kemajuan teknologi, analisis statistik semakin ditingkatkan melalui penggunaan perangkat lunak statistik dan algoritma kecerdasan buatan. Dengan peran yang mendalam dan luas di berbagai bidang, statistik tidak hanya menjadi alat analisis, tetapi juga alat untuk memahami dunia di sekitar kita dan mengambil keputusan yang informasional dan terinformasi.

2.1.5. Pattern Recognition

Pattern Recognition, atau pengenalan pola, adalah cabang ilmu komputer dan kecerdasan buatan yang berkaitan dengan identifikasi pola dalam data. Tujuan utama dari pengenalan pola adalah mengembangkan model atau algoritma yang dapat mengenali dan mengklasifikasikan pola atau informasi yang tersembunyi dalam data. Ini melibatkan pemahaman dan ekstraksi fitur yang dapat membedakan antara kelas atau jenis pola yang berbeda.

Proses pengenalan pola melibatkan beberapa tahap, termasuk pengumpulan data, ekstraksi fitur, pemilihan model, pelatihan model, dan pengujian model. Data yang digunakan dalam pengenalan pola dapat berupa gambar, suara, teks, atau data multidimensi lainnya. Ekstraksi fitur melibatkan identifikasi fitur atau karakteristik

khusus dalam data yang dapat digunakan untuk membedakan pola. Pemilihan model melibatkan pemilihan algoritma atau struktur model yang paling sesuai untuk tugas pengenalan pola tertentu. Pengenalan pola memiliki berbagai aplikasi di berbagai bidang. Di bidang pengolahan citra, pengenalan pola digunakan untuk mengenali objek, wajah, dan karakter dalam gambar. Di bidang pengenalan suara, teknologi ini digunakan untuk mentranskripsi ucapan menjadi teks. Dalam konteks kecerdasan buatan, pengenalan pola seringkali digunakan dalam pembangunan sistem yang dapat memahami dan merespons terhadap data kompleks.

Meskipun pengenalan pola telah mencapai banyak kemajuan, masih ada beberapa tantangan yang harus diatasi, termasuk variabilitas data, skalabilitas, dan keterbatasan data pelatihan. Peningkatan teknologi komputasi, seperti penggunaan deep learning, telah meningkatkan kemampuan pengenalan pola dalam menangani data yang kompleks dan besar. Dengan perkembangan teknologi dan aplikasi yang semakin meluas, pengenalan pola terus menjadi area penelitian yang aktif dan menarik, memainkan peran penting dalam memungkinkan sistem untuk mengenali dan memahami pola dalam data, mendukung berbagai aplikasi mulai dari pengolahan citra hingga kecerdasan buatan.

2.2. Model Klasifikasi

Model klasifikasi adalah suatu bentuk dari model prediktif dalam dunia machine learning yang dirancang untuk mengklasifikasikan atau mengelompokkan data ke dalam kategori atau kelas tertentu [12]. Tujuan utama dari model klasifikasi adalah membuat prediksi berdasarkan pola atau fitur yang ada dalam data pelatihan, dan kemudian mengaplikasikan pemahaman tersebut untuk mengelompokkan data

yang belum pernah dilihat sebelumnya [13]. Model klasifikasi sangat berguna dalam mengambil keputusan otomatis dan memberikan wawasan terhadap data yang kompleks. Beberapa jenis model klasifikasi yang umum digunakan termasuk Decision Trees, Naive Bayes, Support Vector Machines (SVM), dan Neural Networks. Decision Trees menggunakan serangkaian keputusan berbasis aturan untuk mengklasifikasikan data, sementara Naive Bayes mengaplikasikan teorema probabilitas untuk memprediksi kelas. SVM menciptakan batas keputusan untuk memisahkan kelas-kelas yang berbeda dalam ruang fitur, dan Neural Networks menggunakan struktur jaringan saraf tiruan untuk pembelajaran mesin yang kompleks.

Proses pelatihan model klasifikasi melibatkan penggunaan data pelatihan yang berisi contoh-contoh dengan label yang sudah diketahui. Model belajar dari pola dalam data ini dan kemudian dapat digunakan untuk membuat prediksi pada data yang belum pernah dilihat sebelumnya. Evaluasi kinerja model klasifikasi dilakukan dengan menggunakan metrik seperti akurasi, presisi, recall, dan F1-score, yang memberikan gambaran tentang sejauh mana model dapat mengenali dan mengklasifikasikan data dengan benar.

Model klasifikasi diterapkan dalam berbagai bidang, termasuk pengenalan wajah, deteksi spam email, analisis sentimen, dan diagnostik medis. Dalam konteks kesehatan, misalnya, model klasifikasi dapat digunakan untuk mengklasifikasikan gambar medis sebagai gambar normal atau patologis, membantu dokter dalam proses diagnosis. Meskipun model klasifikasi memiliki kegunaan yang besar, pemilihan model yang tepat dan penanganan data yang baik sangat penting untuk

mencapai hasil yang optimal. Seiring dengan kemajuan teknologi, model klasifikasi terus mengalami perkembangan, dengan penerapan teknik seperti deep learning yang telah memungkinkan model-model yang lebih kompleks dan mampu menangani data yang semakin besar dan beragam.

2.3. Algoritma Naïve Bayes

Algoritma Naive Bayes adalah sebuah metode klasifikasi yang didasarkan pada teorema Bayes, yang mengambil namanya dari asumsi "naive" atau sederhana yang dibuat mengenai hubungan antar-variabel [14]. Meskipun sederhana, algoritma ini telah terbukti sangat efektif dalam berbagai aplikasi, termasuk analisis sentimen, klasifikasi teks, dan pengenalan pola. Secara mendasar, Naive Bayes bekerja dengan memanfaatkan prinsip probabilitas [15]. Algoritma ini memperkirakan probabilitas suatu instance data termasuk ke dalam suatu kelas berdasarkan fitur-fitur yang ada. Asumsi dasar Naive Bayes adalah bahwa setiap fitur adalah independen, artinya nilai fitur satu tidak bergantung pada nilai fitur lainnya [16]. Meskipun asumsi ini seringkali tidak sepenuhnya sesuai dengan dunia nyata, kelebihan Naive Bayes terletak pada kinerja yang baik dan efisiensinya dalam klasifikasi meski dengan dataset yang besar.

Rumus utama yang digunakan dalam Naive Bayes adalah Teorema Bayes. Pada dasarnya, teorema ini menggambarkan cara menghitung probabilitas kelas suatu instance data berdasarkan probabilitas prior (sebelum pengamatan data) dan likelihood (probabilitas mengamati data jika instance tersebut berasal dari kelas tertentu). Naive Bayes memiliki tiga variasi utama: Bernoulli Naive Bayes, Multinomial Naive Bayes, dan Gaussian Naive Bayes. Masing-masing cocok untuk

tipe data dan kondisi tertentu. Bernoulli Naive Bayes umumnya digunakan untuk data biner, Multinomial Naive Bayes untuk data kategori diskrit, sementara Gaussian Naive Bayes cocok untuk data kontinu yang terdistribusi normal.

Meskipun kesederhanaannya, Naive Bayes tetap menjadi salah satu pilihan populer dalam klasifikasi data, terutama ketika data terdiri dari banyak fitur atau ketika ada kebutuhan untuk pemrosesan yang cepat dan efisien. Keunggulan Naive Bayes terletak pada kemampuannya untuk memberikan hasil yang memadai dengan mengabaikan ketergantungan antar-fitur, menjadikannya algoritma yang kuat untuk berbagai aplikasi di dunia kecerdasan buatan.

2.4.1. Uji Performa

Confusion matrix adalah sebuah tabel matriks yang digunakan dalam evaluasi kinerja model klasifikasi dalam machine learning dan statistika. Matriks ini memberikan gambaran rinci tentang sejauh mana model dapat mengklasifikasikan instance-data ke dalam kategori yang benar dan sejauh mana kesalahan klasifikasi tersebut terjadi. Confusion matrix terdiri dari empat sel: True Positive (TP) yang mewakili jumlah instance yang benar-benar diklasifikasikan dengan benar, True Negative (TN) yang mewakili jumlah instance yang benar-benar diklasifikasikan dengan benar sebagai negatif, False Positive (FP) yang mewakili jumlah instance yang salah diklasifikasikan sebagai positif, dan False Negative (FN) yang mewakili jumlah instance yang salah diklasifikasikan sebagai negatif. Dengan menyajikan informasi ini, confusion matrix menjadi alat yang kuat untuk mengukur presisi, recall, akurasi, dan nilai lainnya yang memberikan wawasan mendalam tentang kualitas klasifikasi model.

		Kelas Prediksi	
	Kelas	**Benar**	**Salah**
Kelas Atribut	Benar	True Positive (TP)	False Positive (FP)
	Salah	False Negative (FN)	True Negative (TN)

Dimana tabel 1 berisi:

- TP (True Positive), yaitu jumlah data positif yang memiliki nilai benar.
- TN (True Negative), yaitu jumlah data negatif yang memiliki nilai benar.
- FN (False Negative), yaitu jumlah data negatif tetapi yang memiliki nilai salah.
- FP (False Positive), yaitu jumlah data yang positif tetapi yang memiliki nilai salah.

$$\boldsymbol{Acuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FN+FP}} \times \mathbf{100\%} \quad [17]$$

$$\boldsymbol{Presisi} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \times \mathbf{100\%} \quad [18]$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \times \mathbf{100\%} \quad [19]$$

2.4. Alat Bantu Program/Tools Pendukung

2.4.1. Orange

Orange adalah sebuah platform open-source yang dirancang khusus untuk analisis data visual dan pemodelan prediktif. Merupakan alat yang sangat populer di kalangan peneliti, ilmuwan data, dan profesional yang bekerja dalam bidang ilmu komputer dan analisis data. Orange menyediakan antarmuka pengguna grafis (GUI) yang intuitif, memungkinkan pengguna untuk melakukan tugas analisis data kompleks tanpa memerlukan keterampilan pemrograman yang mendalam. Salah satu fitur utama Orange adalah kemampuannya dalam membangun model machine

learning. Platform ini menyediakan berbagai algoritma machine learning seperti regresi linear, klasifikasi, clustering, dan lainnya, yang dapat diakses dan dikonfigurasi melalui antarmuka grafis. Ini membuatnya sangat berguna bagi mereka yang mungkin tidak memiliki latar belakang pemrograman yang kuat, tetapi tetap ingin memanfaatkan kekuatan analisis data dan machine learning.

Selain itu, Orange menawarkan berbagai alat visualisasi yang dapat membantu pengguna menjelajahi dan memahami data mereka. Fitur ini mencakup berbagai jenis plot, diagram, dan representasi visual lainnya untuk membantu analisis data yang lebih baik. Dengan menggunakan antarmuka drag-and-drop, pengguna dapat dengan mudah membangun alur kerja analisis data yang sesuai dengan kebutuhan mereka. Orange juga mendukung integrasi dengan berbagai sumber data, termasuk CSV, Excel, dan database. Hal ini memungkinkan pengguna untuk mengimpor dan mengolah data dari berbagai sumber dengan mudah, memperluas fleksibilitas platform untuk bekerja dengan berbagai jenis dataset.

Selain kegunaannya dalam dunia akademis dan riset, Orange juga dapat diterapkan dalam berbagai konteks bisnis, seperti analisis pasar, prediksi bisnis, dan pengambilan keputusan berbasis data. Dengan kombinasi antarmuka yang ramah pengguna dan kemampuan analisis yang kuat, Orange terus menjadi alat yang diminati untuk pemodelan prediktif dan eksplorasi data dalam berbagai industri.

Orange merupakan platform analisis data open-source yang memiliki aplikasi yang luas dalam berbagai bidang. Dikembangkan untuk memberikan solusi analisis data yang mudah digunakan dan dapat diakses oleh berbagai kalangan, Orange

menawarkan lingkungan visual untuk eksplorasi, pemodelan, dan visualisasi data tanpa memerlukan pengetahuan mendalam tentang pemrograman atau statistik.

Salah satu aplikasi utama Orange adalah di bidang ilmu data, di mana platform ini menyediakan alat dan fungsi untuk melakukan tugas seperti pembersihan data, pemodelan prediktif, dan clustering. Dengan antarmuka grafis yang intuitif, pengguna dapat dengan mudah membangun model, memilih algoritma machine learning, dan menginterpretasi hasilnya tanpa perlu menulis kode.

Selain itu, Orange juga sering digunakan dalam bidang ilmu sosial, biologi, dan kesehatan untuk menganalisis data percobaan dan penelitian. Dengan dukungan untuk berbagai jenis tugas analisis data, mulai dari eksplorasi data dasar hingga pemodelan kompleks, Orange memberikan fleksibilitas yang dibutuhkan oleh peneliti dan profesional dalam berbagai disiplin ilmu.

Aplikasi Orange mencakup beragam modul dan widget, termasuk alat untuk visualisasi data, pemilihan fitur, klasifikasi, regresi, dan masih banyak lagi. Oleh karena itu, Orange memiliki peran penting dalam mendemokratisasi akses ke analisis data, menjadikannya alat yang berharga untuk pembelajaran mesin dan pemodelan statistik yang dapat digunakan oleh berbagai kalangan.

Referensi Penelitian	1
Judul	Penerapan Text Mining Analisis Sentimen Mengenai Vaksin Covid - 19 Menggunakan Metode Naïve Bayes
Nama Penulis	Fira Fathonah[1)], Asti Herliana[2)]
Tahun	2021
Hasil	Sejak wabah COVID-19 melanda dunia, media sosial telah menjadi saluran utama bagi masyarakat, termasuk pemerintah Indonesia, untuk berinteraksi dan menyebarkan informasi. Dalam konteks ini, penelitian mengenai respons masyarakat terhadap Vaksin COVID-19 melibatkan penerapan text mining dengan metode Naïve Bayes. Dalam penelitian ini, berbagai komentar dari pengguna Twitter yang mencakup spektrum positif dan negatif terhadap vaksin tersebut dianalisis untuk mengukur sentimen secara keseluruhan. Metode Naïve Bayes dipilih karena potensinya dalam mengklasifikasi dokumen dengan akurasi dan efisiensi yang baik. Melalui analisis sentimen terhadap 34 data yang diperoleh melalui teknik data crawling, hasil penelitian menunjukkan persentase akurasi sebesar 100%, menggambarkan keberhasilan metode Naïve Bayes dalam memahami dan mengklasifikasikan respons masyarakat terhadap vaksin COVID-19 di platform Twitter [20].

Referensi Penelitian	2
Judul	Penerapan Text Mining Dengan Menggunakan Metode TF-IDF Untuk Menentukan Genre Dari Komik
Nama Penulis	Windi Sri Utami Saragih[1*], Nelly Astuti Hasibuan[1,] Rivalri Kristianto Hondroo1
Tahun	2020
Hasil	Penelitian ini bertujuan untuk meningkatkan efektivitas pembagian genre komik, yang sering kali kurang tepat akibat keterbatasan kata-kata representatif untuk mewakili genre. Dengan menerapkan text mining dan metode Term Frequency-Inverse Document Frequency (TF-IDF), penelitian ini berusaha untuk mengembangkan suatu sistem otomatis yang dapat menentukan genre komik. Data yang digunakan sebagai acuan terdiri dari empat kategori: horor, inspiratif, misteri, dan romantis. Proses dimulai dengan persiapan dan seleksi dokumen, diikuti oleh pembobotan kata menggunakan TF-IDF pada judul, penulis, dan sinopsis komik. Sinopsis menjadi fokus untuk menentukan genre, dan kemiripan antara teks dengan node di data

Referensi Penelitian	3
Judul	Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia
Nama Penulis	Alvina Felicia Watratan[1], Arwini Puspita. B[2], Dikwan Moeis[3]
Tahun	2020
Hasil	Penelitian ini bersifat antisipatif terhadap pandemi COVID-19 dengan fokus pada prediksi tingkat penyebarannya di Indonesia. Dalam menghadapi krisis kesehatan global ini, penelitian menggunakan metode Naive Bayes sebagai alat analisis untuk memprediksi sejauh mana virus corona menyebar. Dengan menerapkan analisis masalah, studi literatur, dan pengumpulan data, penelitian ini mencoba memberikan gambaran yang lebih akurat terkait penyebaran COVID-19. Hasilnya menunjukkan bahwa dari 33 data yang diuji, metode Naive Bayes berhasil mengklasifikasikan dengan benar 16 data mengenai kasus COVID-19 per provinsi, dengan tingkat akurasi sebesar 48,4848%. Meskipun angka ini mencerminkan tantangan kompleksitas prediksi dalam konteks pandemi, penerapan metode ini diharapkan dapat memberikan kontribusi dalam upaya memahami dan mengendalikan penyebaran virus corona di Indonesia [22]

Referensi Penelitian	4
Judul	FILTERING SPAM EMAIL MENGGUNAKAN METODE NAIVE BAYES
Nama Penulis	Aria Wibisono
Tahun	2023
Hasil	Penelitian ini bertujuan untuk mengatasi masalah spam melalui penerapan metode Naive Bayes dalam konteks klasifikasi email. Spam, yang seringkali berisi konten yang tidak diinginkan seperti promosi produk, pornografi, dan virus, dapat diidentifikasi dan difilter secara otomatis menggunakan metode ini. Naive Bayes, sebagai metode klasifikasi sederhana yang mengandalkan teorema probabilitas, digunakan untuk memprediksi probabilitas sebuah email sebagai spam atau ham (non-spam) berdasarkan informasi dari email sebelumnya. Dengan menguji aplikasinya terhadap lima email, termasuk dua email spam dan tiga email ham, penelitian ini bertujuan untuk mengukur tingkat keberhasilan metode Naive Bayes dalam membedakan dan mengklasifikasikan jenis email dengan akurasi yang diharapkan [23].