BAB II

LANDASAN TEORI

2.1. Tinjauan Pustaka

2.1.1. *Data Mining*

Data mining adalah sebuah proses pencarian informasi secara otomatis yang terdapat dalam tempat penyimpanan data berukuran besar, Teknik *data mining* digunakan untuk memeriksa basis data berukuran besar sebagai cara untuk mengidentifikasi pola yang tidak terlihat, hubungan, atau informasi yang berguna lainnya yang dapat membantu dalam pengambilan keputusan yang lebih baik *Metode data mining* melibatkan berbagai teknik statistik, matematika, dan kecerdasan buatan untuk menganalisis data secara mendalam dan mengekstraksi wawasan yang berguna (Gede Aditra Pradinyana dkk, 2020).

Tahapan *data mining* meliputi penerapan algoritma untuk klasifikasi. Sebelum menerapkan algoritma, Dataset harus di bagi menjadi dua bagian, Data *training* dan data *testing*. Data *training* digunakan untuk melatih model, sementara data *testing* digunakan untuk menguji keakuratan model tersebut(Hondro, 2023).

2.1.2. *Crisp-DM*

Crisp-DM (Cross Industry Standard Process for Data Mining) adalah kerangka kerja yang populer dalam analisis dan penambangan data, Selain itu *Crisp-DM* merupakan sebagai *standar metodologi data mining* untuk industri.

Dengan *crisp-DM*, proses *data mining* bisa dilakukan secara cepat, ekonomis, dan terstruktur (Nandang Iriadi dkk, 2020).

Setiap fase memiliki tujuan dan kegiatan yang spesifik, memastikan bahwa proyek *data mining* dapat diselesaikan dengan hasil yang optimal. Dengan demikian, *Crisp-DM* membantu perusahaan dalam memaksimalkan nilai dari data mereka melalui pendekatan yang terorganisir dan efisien. Berikut adalah tahapan-tahapan dalam *crisp-DM :*

2.1.2.1. **Tahapan Pemahaman Bisnis**

Tahapan pemahaman bisnis merupakan tahapan pertama, yang juga dikenal sebagai tahapan pemahaman bisnis atau pemahaman penelitian, melibatkan pemahaman yang jelas terhadap tujuan dan persyaratan proyek secara keseluruhan, baik dari segi unit bisnis maupun penelitian.

Langkah-langkah dalam tahapan ini mencakup menerjemahkan tujuan dan pembatasan proyek ke dalam perumusan permasalahan pada data mining, serta menyiapkan strategi awal untuk mencapai tujuan penelitian atau bisnis.

2.1.2.2. **Tahapan Pemahaman Data**

Tahapan pemahaman data merupakan langkah-langkah seperti pengumpulan data, penggunaan analisis data eksploratif untuk mendapatkan pemahaman awal tentang data, evaluasi terhadap kualitas data, dan pemilihan subset data yang menarik untuk ditindaklanjuti.

Langkah-langkah ini penting dalam memastikan bahwa data yang digunakan untuk analisis atau pengembangan aplikasi memenuhi standar yang diperlukan. Dengan melakukan pemahaman data secara menyeluruh,

pengembang atau analis dapat mengidentifikasi potensi masalah, pola menarik, atau kesempatan yang mungkin tidak terlihat pada pandangan pertama. Dengan demikian, tahapan pemahaman data merupakan fondasi yang kuat untuk membangun analisis yang relevan dan aplikasi yang efektif berdasarkan data yang tersedia.

2.1.2.3. **Tahapan Persiapan Data**

Tahap persiapan data merupakan penyediaan data awal dan pengolahan hingga menjadi dataset akhir yang akan digunakan di seluruh tahap berikutnya. Proses ini melibatkan pemilihan kasus dan variabel yang relevan untuk analisis, serta transformasi variabel yang diperlukan.

Selain itu, data mentah juga perlu dibersihkan agar siap digunakan dalam pemodelan. Pembersihan data mencakup penanganan nilai-nilai yang hilang, deteksi dan penghapusan outlier, serta pengubahan format data yang tidak konsisten. Proses persiapan data yang baik sangat penting untuk memastikan kualitas dan keakuratan hasil analisis.

2.1.2.4. **Tahapan Pemodelan**

Tahapan pemodelan merupakan pemilihan dan penerapan teknik pemodelan yang tepat, serta kalibrasi model untuk memastikan hasil yang optimal. Seringkali beberapa teknik berbeda dapat diterapkan untuk menyelesaikan masalah data mining yang sama.

Jika hasil pemodelan kurang bagus, kembali ke tahap persiapan data bisa menjadi langkah yang diperlukan untuk memastikan bahwa data sudah siap digunakan sesuai dengan persyaratan teknik data mining yang dipilih. Pengulangan proses ini memastikan bahwa model yang dikembangkan tidak

hanya sesuai dengan data yang ada tetapi juga mampu memberikan prediksi atau klasifikasi yang akurat dalam berbagai skenario.

2.1.2.5. **Tahapan *Evaluasi***

Tahapan evaluasi merupakan model-model yang dibangun selama tahap pemodelan akan dievaluasi untuk kualitas dan efektivitasnya sebelum diterapkan di lapangan. Evaluasi mencakup penentuan apakah model berhasil mencapai tujuan yang ditetapkan sebelumnya dan apakah aspek penting dari masalah bisnis atau penelitian telah diperhitungkan.

Hasil evaluasi ini menjadi dasar untuk mengambil keputusan tentang penggunaan hasil analisis data. Proses evaluasi yang menyeluruh memastikan bahwa model tidak hanya memberikan hasil yang akurat tetapi juga relevan dan bermanfaat bagi pemangku kepentingan. Dengan demikian, evaluasi yang baik membantu menghindari potensi risiko dan memastikan bahwa solusi yang dihasilkan benar-benar dapat diandalkan untuk pengambilan keputusan di dunia nyata.

2.1.2.6. **Tahapan Penyebaran**

Tahapan penyebaran merupakan langkah penting untuk melakukan penyebaran hasil analisis data. Contoh penyebaran sederhana meliputi penghasilan laporan hasil analisis. Namun, ada juga contoh penggunaan yang lebih rumit, seperti melaksanakan proses dalam penambangan data secara paralel di departemen lain.

Terutama dalam konteks bisnis, seringkali penyebaran dilakukan oleh pelanggan berdasarkan model yang telah dibuat. Penyebaran yang efektif memastikan bahwa hasil analisis dapat diterapkan secara optimal dan

memberikan nilai tambah yang nyata bagi organisasi. Selain itu, penyebaran yang baik juga melibatkan pelatihan dan dukungan bagi pengguna akhir untuk memastikan bahwa mereka dapat memanfaatkan hasil analisis dengan maksimal.

2.1.3. **Analisis Sentimen**

Analisis sentimen merupakan salah satu algoritma *Natural Language Processing* (NLP), Yang memanfaatkan ulasan atau komentar dari pelanggan untuk mengidentifikasi, mengekstrak, dan memahami respon mereka yang terkandung dalam teks terhadap suatu layanan, jasa, atau produk tertentu, Analisis sentimen umum nya di gunakan untuk klasifikasi ulasan atau komentar pelanggan menjadi sentimen negatif,positif atau netral(Gede Aditra Pradinyana dkk, 2020).

Dengan menggunakan analisis sentimen, perusahaan dapat memperoleh wawasan mendalam tentang persepsi pelanggan mereka dan membuat keputusan yang lebih baik berdasarkan data yang ada. Hal ini sangat berguna dalam mengembangkan strategi bisnis, meningkatkan layanan pelanggan, dan meningkatkan kualitas produk atau layanan yang ditawarkan.

2.1.4. ***Scraping Data***

Teknik *web scraping* merupakan cara untuk mengambil data dalam jumlah besar dari situs *website*. Data yang diekstraksi kemudian disimpan dalam file lokal di komputer atau database dalam format tabel, mirip dengan spreadsheet.

Setelah itu, data dapat diproses menggunakan alat seperti *RapidMiner* untuk melakukan pra-pemrosesan atau pra-pengolahan data sebelum memasuki tahap berikutnya(Herlinawati et al., 2020).

Pada penelitian ini *scraping* data yang di gunakan memakai tools yang sudah di sediakan oleh situs *website apify* untuk mengambil data ulasan atau komentar pelangan *Tik Tok scentplus official* dan *moris Indonesia*.

2.1.5. ***Python***

Python adalah salah satu bahasa pemrograman tingkat tinggi sama layak nya seperti bahasa pemrograman lain, misalnya, *Java*, *PHP*, ***C***++, dan lain - lain, Sebagai bahasa pemrograman, *Python* tentu memiliki struktur dan kosakata unik, atau kata kunci (*keyword*), dan aturan tersendiri yang jelas berbeda dengan bahasa pemrograman lainnya, Dalam bahasa pemrograman *Python*, terdapat berbagai *library* dan kerangka kerja yang memudahkan analisis data. Mulai dari pemrosesan data dasar hingga eksplorasi data yang kompleks (Budi Raharjo, 2019),

Dalam penelitian ini menggunakan *library python* seperti *pandas, matplotib* dan *numpy* untuk pembuatan atau melatih model regresi logistik, Berikut pengertian dan penerapan dari library yang digunakan:

2.1.5.1. ***Pandas***

Pandas merupakan *library open-source* yang sangat bagus dan populer di kalangan praktisi data, *Pandas* digunakan untuk melakukan *preprocessing* dan analisis data dalam bahasa pemrograman *Python*,

Library pandas dapat digunakan untuk menampilkan, membaca dan mengolah dataset yang berformat file csv, txt, excel, dan sebagainya (Teguh Wahyono, 2021).

Dalam penelitian ini, *library Pandas* berfungsi untuk membaca dataset berupa file CSV dengan mudah dan efisien. *Pandas* menyediakan fungsi yang memungkinkan pengguna untuk mengimpor data dari berbagai sumber, termasuk file CSV, dan melakukan manipulasi data seperti pengindeksan, penggabungan, dan pembersihan data.

2.1.5.2. ***Matplotlib***

Matplotlib merupakan salah satu *library* yang paling populer untuk *visualisasi* data yang dibangun di atas array *numpy*, *Matplotlib* memiliki kemampuan untuk bekerja dengan baik dengan banyak system operasi dan *backend grafis*, *Matplotlib* mendukung berbagai jenis *backend* dan jenis *output* untuk menyesuaikan format luaran yang diinginkan, Dengan *Matplotlib* dapat membuat dan menampilkan berbagai jenis *plot*, seperti *grafik garis*, *scatter plot*, *histogram*, *bar chart*, dan lain - lain (Teguh Wahyono, 2021).

Dalam penelitian ini, *library Matplotlib* digunakan untuk menampilkan grafik sentimen negatif, netral, dan positif dalam bentuk diagram batang. *Matplotlib* menyediakan beragam jenis plot dan visualisasi yang dapat disesuaikan, sehingga memungkinkan peneliti untuk menampilkan data dengan cara yang informatif dan mudah dipahami.

2.1.5.3. ***NumPy***

NumPy merupakan *library* untuk bahasa pemrograman *Python* yang menyediakan dukungan untuk operasi-operasi *array* dan *matriks* besar, Selain itu *NumPy* merupakan *library* yang menyediakan objek *array* multidimensi, Serta berbagai objek turunannya seperti masked *array* dan *matriks*, Dengan menggunakan *NumPy*, Dapat melakukan berbagai operasi pada *array* secara lebih cepat, Termasuk operasi matematika, Operasi logika, *Sorting* atau pengurutan, *Selecting*, *Transformasi fourier*, aljabar linier dasar, stastika dasar dan sebagainya (Teguh Wahyono, 2021).

Dalam penelitian ini, *library NumPy* digunakan untuk memproses pelatihan model regresi logistik. *NumPy* menyediakan array dan operasi matematika yang efisien, yang penting dalam pengolahan data dan implementasi algoritma machine learning seperti regresi logistik. Hal ini memungkinkan peneliti untuk melakukan perhitungan numerik secara cepat dan efisien dalam konteks pengembangan model prediktif.

2.1.6. ***Jupyter Notebook***

Jupyter notebook merupakan *tools open-source* untuk menyusun code – code *Python* dengan kelebihan *user friendly* dan mudah digunakan, Dengan *Jupyter notebook* dapat membuat dan berbagi dokumen yang menggabungkan *live code*, *visualisasi*, narasi teks, dan elemen-elemen lainnya. Dokumen-dokumen ini dapat berisi kode dalam berbagai bahasa pemrograman, seperti *Python*, *R*, *Julia*, dan lainnya, Selain itu *jupyter notebook* dapat mendokumentasikan sebuah pekerjaan, dimana coding dan dokumentasi bisa

dilakukan dalam satu page dan disimpan dalam bentuk presentasi yang menarik (Teguh Wahyono, 2021).

Dalam penelitian ini, *Jupyter notebook* digunakan untuk melatih dan membangun model menggunakan metode regresi logistik. Dengan memanfaatkan *jupyter notebook*, proses pengembangan model menjadi lebih interaktif dan efisien, Selain itu penggunaan jupyter notebook memungkinkan untuk mempermudah melakukan eksplorasi data dan visualisasi hasil.

2.1.7. ***Text Mining***

Text mining merupakan sebuah metode yang dapat mengkumpulkan informasi dan menggali sebuah pola pengetahuan dari dokumen text yang tidak terstruktur, *Text mining* pada dasar nya serupa dengan *data mining*, Tetapi lebih terfokus pada analisis teks atau dokumen yang tidak terstruktur, seperti artikel berita,ulasan masyarakat terhadap suatu layanan,jasa,produk dan lain-lain, Teknik-teknik dalam *text mining* termasuk pengelompokan, klasifikasi, ekstraksi informasi, analisis sentimen dan sebagai nya (Eka Dyar Wahyuni dkk, 2020).

Proses *text mining* melibatkan pencarian informasi di mana pengguna berinteraksi dengan sekumpulan dataset menggunakan tools analisis. Tools ini merupakan bagian dalam data mining, seperti kategorisasi informasi dan pengelompokan teks. Sumber dataset yang digunakan pada text mining diambil dari sekumpulan pola bahasa alami yang mampu menganalisis data teks semi-terstruktur dan tidak terstruktur (Noviana & Rasal, 2023).

2.1.8. ***Text Preprocessing***

Text preprocessing merupakan sebuah proses yang dilakukan sebelum proses *data mining* diterapkan pada sebuah dataset, Prapemrosesan data ini bertujuan untuk memastikan bahwa data yang di gali atau di olah oleh *metode data mining* adalah data yang tepat dan berkualitas, Dengan melakukan prapemrosesan ini dapat membersihkan data dari *noise*, mengurangi dimensi, dan mengoptimalkan representasi teks untuk hasil yang lebih akurat dan efisien (Gede Aditra Pradinyana dkk, 2020),

Dalam penelitian ini, proses *text preprocessing* digunakan untuk membersihkan dan membobotkan kata-kata dalam dataset. Langkah-langkah seperti penghapusan tanda baca, konversi ke huruf kecil, dan penghapusan kata-kata yang tidak relevan merupakan bagian dari tahap *preprocessing* yang penting untuk meningkatkan kualitas analisis teks dan pengolahan informasi selanjutnya, Terdapat 5 proses yang terdiri dari :

2.1.8.1. ***Cleansing***

Tahap *cleansing* atau pembersihan merupakan langkah-langkah untuk menghilangkan elemen-elemen yang tidak diinginkan atau tidak relevan dari teks, seperti tag, tanda baca, URL, atau kode yang tidak terbaca.

Tahapan ini bertujuan untuk mengurangi noise dan mempersiapkan teks agar siap untuk proses analisis lebih lanjut, seperti ekstraksi fitur atau pemodelan. Dengan membersihkan teks dari elemen-elemen yang tidak relevan, peneliti dapat meningkatkan kualitas data yang akan diolah, sehingga meningkatkan akurasi dan relevansi hasil analisis yang dilakukan.

2.1.8.2. ***Case Folding***

Tahap *case folding* atau juga dikenal sebagai *case conversion*, adalah langkah penting dalam standarisasi huruf dalam dokumen. Tidak semua dokumen memiliki teks yang konsisten dalam penggunaan huruf kapital. Oleh karena itu, tahapan *case folding* diperlukan untuk mengubah seluruh huruf menjadi huruf kecil menggunakan fungsi *lowercase*.

Proses ini memastikan konsistensi dalam representasi teks sehingga memudahkan analisis lebih lanjut seperti pencarian teks dan pemrosesan informasi berbasis teks. Dengan menerapkan *case folding*, peneliti dapat menghindari ambigu dalam data dan memastikan bahwa data siap untuk tahap selanjutnya dalam proses analisis data atau pengolahan bahasa alami.

2.1.8.3. ***Stopword***

Tahap *stopword* adalah langkah penting dalam pemrosesan teks yang bertujuan menghilangkan kata-kata yang sering muncul tetapi tidak memberikan pengaruh besar dalam analisis teks. *Stopwords* seperti "dan", "atau", "yang", dan kata umum lainnya tidak memberikan nilai tambah yang signifikan dalam memahami makna atau pola dalam teks. Dengan menghapus *stopwords*, data teks dapat dibersihkan dari elemen yang tidak relevan, sehingga meningkatkan kualitas analisis teks dan akurasi klasifikasi.

Proses penghapusan *stopwords* ini tidak hanya menyederhanakan representasi teks, tetapi juga dapat mempercepat waktu komputasi dan meminimalkan noise dalam hasil analisis. Selain itu, dalam konteks analisis sentimen atau klasifikasi dokumen, penghapusan *stopwords* membantu model untuk lebih fokus pada kata-kata kunci yang lebih informatif dan

berkontribusi signifikan terhadap penentuan sentimen atau klasifikasi yang akurat. Dengan demikian, tahap stopword merupakan salah satu dari beberapa teknik *preprocessing* teks yang penting untuk memastikan bahwa data yang digunakan dalam analisis tidak terbebani oleh kata-kata yang tidak relevan secara semantik atau kontekstual.

2.1.8.4. ***Stemming***

Tahap *stemming* merupakan proses untuk mengubah kata-kata yang berhimbuan menjadi kata dasar. Proses *stemming* berfungsi dalam memperbaiki konsistensi kata-kata dalam dokumen teks sehingga analisis teks lebih mudah dilakukan.

Stemming membantu mengurangi variasi kata yang memiliki akar kata yang sama tetapi dengan bentuk yang berbeda, seperti "berlari", "berlari-lari", dan "lari". Dengan mengubah kata-kata ini menjadi bentuk dasarnya, seperti "lari", proses analisis teks dapat lebih fokus pada makna dasar kata-kata tanpa terganggu oleh variasi morfologis yang tidak relevan. Hal ini penting dalam memastikan bahwa representasi teks yang digunakan dalam analisis adalah konsisten dan sesuai dengan tujuan analisis yang dilakukan.

2.1.8.5. ***Tokenizing***

Tahap *tokenizing* adalah langkah penting dalam pemrosesan teks di mana kalimat dipecah menjadi unit-unit kata yang lebih kecil yang disebut token. Proses ini memungkinkan komputer untuk memahami dan menganalisis teks dengan lebih baik.

Misalnya, kalimat "Saya suka makan nasi goreng" akan di *tokenize* menjadi "Saya", "suka", "makan", "nasi", "goreng". Dengan demikian,

tokenizing membantu dalam mengurai teks menjadi bagian-bagian yang lebih mudah diolah dan dipahami oleh mesin. Proses ini penting dalam berbagai aplikasi seperti analisis sentimen, pemodelan bahasa alami, dan pengolahan teks lainnya di mana representasi kata per kata sangat diperlukan untuk pemrosesan lebih lanjut.

2.1.9. **TF-IDF**

TF-IDF (*Term Frequency-Inverse Document Frequency*) adalah sebuah ukuran statistik yang mengevaluasi seberapa penting suatu kata dalam suatu dokumen dalam konteks kumpulan dokumen yang lebih besar. Metode ini sangat berguna dalam menentukan tingkat *relevansi* kata terhadap suatu dokumen. Selain itu, TF-IDF merupakan salah satu metode pembobotan yang sering digunakan, *Metode* ini memiliki keunggulan utamanya terletak pada efisiensinya, kemudahan *implementasi*, dan memiliki hasil akurasi yang akurat (Eka Dyar Wahyuni dkk, 2020),

Dalam penelitian ini, TF-IDF digunakan untuk membobotkan kalimat yang akan diproses oleh algoritma regresi logistik, TF-IDF membantu menyoroti kata-kata yang sering muncul yang paling relevan dan signifikan dalam konteks analisis teks atau klasifikasi kalimat menggunakan regresi logistik, Berikut adalah rumus dari TF-IDF :

2.1.9.1. **TF**

TF (*Term frequency*) merupakan ukuran berapa kali kata muncul dalam sebuah dokumen. Cara menghitungnya adalah dengan membagi jumlah kemunculan kata tersebut dalam dokumen dengan jumlah total kata dalam

dokumen itu, Berikut adalah rumus dari TF yang digunakan di penelitian ini:

TF = Jumlah kata pada kalimat / Jumlah total kata pada kalimat

2.1.9.2. **IDF**

IDF (*Inverse document frequency*) merupakan ukuran seberapa umum sebuah istilah muncul di seluruh kumpulan dokumen. Cara menghitungnya adalah dengan menghitung total jumlah kalimat dalam kumpulan data tersebut, dibagi dengan jumlah kalimat yang mengandung kata tersebut, Berikut adalah rumus dari IDF yang digunakan di penelitian ini:

IDF = Jumlah total kalimat / Jumlah total kata pada seluruh kalimat

2.1.9.3. **TF-IDF**

TF-IDF singkatan dari (*Term Frequency-Inverse Document Frequency*) merupakan metode yang menggabungkan nilai frekuensi sebuah kata (TF) dengan kemampuan uniknya untuk membedakan pentingnya kata tersebut dalam suatu dokumen (IDF). Secara sederhana, perhitungan TF-IDF dilakukan dengan mengalikan nilai TF dengan nilai IDF dari setiap kata dalam dokumen, Berikut adalah rumus TF-IDF yang digunakan di penelitian ini :

TF-IDF = TF * IDF

2.1.10. **Algoritma Regresi Logistik**

Regresi logistik adalah salah satu jenis *regresi* yang sangat berguna dalam mengatasi masalah klasifikasi, Regresi logistik memungkinkan untuk

memahami dan memodelkan hubungan antara *variabel input* dan *output* dengan cara yang mudah dipahami (Anis Zubair, 2022),

Dalam penelitian ini rumus regresi logistik *multinomial* akan diterapkan karena rumus regresi logistik *multinomial* merupakan generalisasi dari regresi logistik *biner* untuk masalah multikelas. Dalam regresi logistik *biner*, hanya ada dua kemungkinan hasil, seperti "Positif" atau "Negatif". Dalam regresi logistik *multinomial*, terdapat lebih dari dua kemungkinan hasil, seperti "Positif", "Negati", atau "Netral", Cara menghitung nya adalah dengan menghitung probabilitas masing-masing kategori (positif, negatif, dan netral) berdasarkan frekuensi kemunculan kata-kata, Untuk setiap kategori, nilai probabilitas ditentukan dengan mengalikan koefisien *regresi* (misalnya, B1positif, B1negatif, B1netral) dengan frekuensi kemunculan kata-kata positif, negatif, atau netral, Berikut adalah rumus regresi logistik *multinomial* yang digunakan di penelitian ini :

$$g_j(x) = \beta_j 0 + \beta_j 1 x 1 + \beta_j 2 x 2 + \cdots + \beta_j p x p$$

2.1.11. ***Website***

Website merupakan salah satu layanan yang memanjakan pengguna internet dengan menggunakan *hypertext* untuk menampilkan beragam konten, mulai dari teks, gambar, suara, animasi dan multimedia lainnya, Pada dasarnya *website* merupakan semacam ruang informasi di dalam dunia internet. Melalui teknologi *hypertext*, pengguna dapat dengan mudah menjelajahi informasi dengan mengikuti tautan yang disediakan di dalam dokumen *website*(Yeni Kustiyahningsih dan Devie Rosa Anamisa, 2011).

Pada penelitian ini, *website* dipilih sebagai tempat deployment karena *website* dapat diakses dengan mudah oleh semua orang melalui internet. Memilih *website* sebagai platform deployment memungkinkan hasil penelitian dapat diakses dari berbagai perangkat dengan koneksi internet, memperluas jangkauan penggunaan dan aksesibilitas informasi yang disajikan. Dengan demikian, *website* menjadi pilihan yang ideal untuk menyajikan dan mempublikasikan hasil penelitian atau aplikasi kepada masyarakat luas secara efisien dan efektif.

2.1.12. ***MySQL***

MySQL merupakan salah satu database paling populer, karena hampir semua aplikasi berbasis *website* seperti *WordPress* memanfaatkan teknologi nya, Selain itu *MySQL* juga di tawarkan dalam berbagai versi termasuk versi gratis(Jubilee Enterprise, 2018).

Pada penelitian ini, *MySQL* digunakan sebagai database untuk menyimpan hasil klasifikasi kalimat serta data akun admin pemasaran dan manajer. *MySQL* dipilih karena kehandalannya dalam penyimpanan dan pengelolaan data, serta kemampuannya untuk menangani volume data yang besar dengan performa yang baik.

2.1.13. **UML**

UML (*Unified Modeling Language*) merupakan bahasa visual universal bagi para pengembang perangkat lunak. Ini menggunakan meta-model tunggal untuk memberikan gambaran yang jelas dan terstruktur tentang bagaimana

sistem perangkat lunak dibangun, terutama yang dikembangkan dengan pendekatan berorientasi objek. Dengan UML, kompleksitas desain direduksi menjadi representasi grafis yang mudah dipahami, memfasilitasi komunikasi antara tim pengembangan dan memastikan keselarasan visi (Martin Fowler, 2014).

Pada penelitian ini, UML yang digunakan adalah use case diagram dan activity diagram. Use case diagram digunakan untuk menggambarkan interaksi antara aktor-aktor dengan sistem atau aplikasi yang sedang dikembangkan, sementara activity diagram digunakan untuk memodelkan alur kerja atau proses dalam aplikasi tersebut secara grafis. Kedua jenis diagram ini membantu dalam memvisualisasikan dan merancang sistem secara lebih terstruktur dan sistematis, memudahkan pengembang dalam memahami serta mengkomunikasikan kebutuhan dan desain aplikasi kepada tim pengembangan dan pemangku kepentingan lainnya.

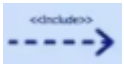
2.1.14. *Use Case Diagram*

Penggunaan *use case* merupakan hal yang sangat penting untuk memahami kebutuhan fungsional suatu sistem. Oleh karena itu, Penting bagi pengembang sistem untuk merancang *use case* sejak awal. Selain itu, versi *use case* yang lebih terperinci harus disusun sebelum memulai pengembangan sistem tersebut (Martin Fowler, 2014),

Dalam penelitian ini, *use case diagram* digunakan untuk membantu dalam pembuatan aplikasi. *Use case diagram* merupakan salah satu jenis diagram dalam UML (*Unified Modeling Language*) yang digunakan untuk

menggambarkan interaksi antara aktor-aktor tertentu (pengguna atau sistem lain) dengan sistem atau aplikasi yang sedang dikembangkan. *Diagram* ini membantu dalam mengidentifikasi dan menggambarkan fungsionalitas utama atau skenario penggunaan yang dimiliki oleh aplikasi tersebut.Berikut merupakan simbol – simbol dan keterangan *use case diagram* (Ariffud Muhammad, 2023).

Tabel 2. 1 Simbol *Use Case Diagram*

No	Simbol	Nama	Keterangan
1		Actor	Representasi pengguna atau sistem eksternal yang berinteraksi dengan sistem.
2		User Case	Aktivitas atau tindakan yang bisa dilakukan oleh aktor dalam sistem.
3		Association	Hubungan aktor dalam kasus penggunaan.
4		System	Sistem yang sedang di kembangkan
5		Include	Hubungan antara dua kasus penggunaan di mana satu kasus penggunaan memasukkan atau mengandung fungsionalitas dari yang lain

6		Extend	Merupakan hubungan antara dua kasus penggunaan, Di mana satu kasus penggunaan memasukkan atau mengandung fungsionalitas dari yang lain
7		Generalization	Hubungan di mana satu kasus penggunaan mewarisi perilaku dari kasus penggunaan lain
8		Collaboration	Dua atau lebih actor atau use case yang terhubung
9		Note	Penjelasan tambahan tentang aktivitas yang terjadi
10		Anchor	Hubungan teks note dengan simbol diagram lain

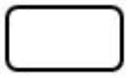
2.1.15. ***Activity Diagram***

Activity diagram merupakan teknik untuk menggambarkan logika prosedural, proses bisnis dan alur kerja, Dalam beberapa hal Mirip dengan *diagram alir*, *activity diagram* membantu menggambarkan langkah-langkah logis dalam suatu proses, tetapi keunggulannya terletak pada kemampuannya untuk menggambarkan perilaku paralel. Ini seperti memperluas kapasitas jalan

dengan menambah jalur baru dalam navigasi, memungkinkan proses bisnis untuk berjalan lebih efisien dan efektif (Martin Fowler, 2014),

Dalam penelitian ini, *activity diagram* digunakan untuk membuat alur aplikasi. Activity diagram merupakan salah satu jenis diagram dalam UML (*Unified Modeling Language*) yang digunakan untuk menggambarkan aktivitas atau langkah-langkah yang terjadi dalam suatu proses atau aplikasi. Diagram ini membantu dalam memvisualisasikan urutan aktivitas, keputusan, dan garis waktu yang terlibat dalam suatu proses secara grafis. Dengan menggunakan *activity diagram*, pengembang dapat mengidentifikasi interaksi antara berbagai komponen atau bagian dari aplikasi, serta memahami alur kerja secara keseluruhan. Berikut merupakan simbol – simbol dan keterangan *activity diagram* (Dicoding Intern, 2021a).

Tabel 2. 2 Simbol *Activity Diagram*

1		Start	Simbol untuk menandai awal aktivitas sitem di mulai
2		End	Simbol untuk menandai akhir aktivitas sitem
3		Activity	Simbol untuk memberikan informasi aktivitas atau tindakan yang sedang di lakukan oleh sistem

4		Decision	Simbol untuk percabangan dimana ada pilihan aktivitas lebih dari satu
5		Join	Simbol untuk penggabungan aktivitas

2.1.16. ***Streamlit***

Streamlit merupakan *library python* yang memungkinkan untuk membangun aplikasi dengan cepat menggunakan *Python* murni. Dengan *Streamlit*, Pengguna dapat membuat prototipe dan menguji aplikasi dengan mudah tanpa harus memiliki pengetahuan mendalam tentang alat-alat *front-end* untuk antarmuka pengguna. Hal ini merupakan solusi yang sederhana namun bermanfaat untuk menciptakan aplikasi dengan kecepatan dan efisiensi (Partha Mishra, 2023).

Pada penelitian ini, *Streamlit* digunakan untuk melakukan deployment model yang sudah jadi agar dapat digunakan oleh semua orang dengan mudah. Streamlit dapat mengubah model machine learning atau analisis data yang kompleks menjadi aplikasi web interaktif dengan cepat dan efisien.

2.1.17. ***GitHub***

GitHub merupakan platform multifungsi yang menggabungkan manajemen proyek, sistem versioning code, dan jaringan sosial bagi para developer di seluruh dunia. Dengan beragam fitur yang disediakan, GitHub mempermudah developer dalam mengelola, berbagi, dan mengembangkan proyek-proyek

inovatif mereka. Melalui platform ini, kolaborasi menjadi lebih efisien dan inspirasi semakin mudah ditemukan(Dicoding Intern, 2021b).

GitHub umum nya di gunakan untuk menjadi repositori Open Source. Tujuan utamanya adalah untuk mempelajari proyek perangkat lunak sebelumnya agar dapat meningkatkan pengembangan proyek baru. Dalam beberapa tahun terakhir, GitHub telah menjadi sumber data yang kaya untuk penelitian, dengan banyak makalah yang diterbitkan berdasarkan analisis data dari platform ini, yang membantu komunitas memahami lebih dalam tentang rekayasa perangkat lunak (Cosentino et al., 2017).

Dalam penelitian ini, GitHub digunakan untuk menyimpan file-file analisis sentimen dan menghubungkan file-file tersebut ke Streamlit sebagai platform untuk melakukan deployment analisis sentimen. Dengan memanfaatkan GitHub, seluruh proses pengelolaan dan pengembangan proyek menjadi lebih terorganisir dan mudah diakses.

2.1.18. *Confusion Matrix*

Confusion Matrix merupakan perhitungan yang menggambarkan performa sebuah classifier model dengan cara membandingkan hasil prediksi terhadap nilai label sebenarnya dari dataset yang dijadikan data uji (Studi et al., 2024).

Dalam perhitungan ini, prediksi kelas positif yang benar disebut True Positive (TP), sementara yang salah disebut False Positive (FP). Sedangkan, untuk kelas negatif, prediksi yang benar disebut True Negative (TN) dan yang salah disebut False Negative (FN). Confusion Matrix memberikan gambaran yang jelas tentang seberapa baik model tersebut mengklasifikasikan data, serta membantu dalam mengevaluasi kekuatan dan kelemahan model tersebut.Berikut merupakan rumus perhitungan akurasi, perisisi dan recal

2.1.18.1.**Akurasi**

Akurasi merupakan Tingkat kesamaan antara hasil prediksi dan nilai sebenarnya merupakan indikator keberhasilan model dalam melakukan prediksi secara tepat. Berikut merupakan rumus untuk menghitung akurasi.

$$Accuracy = \frac{jumlah\ prediksi\ yang\ benar}{jumlah\ data\ test}$$

2.1.18.2.**Persisi**

Presisi merupakan suatu perhitungan untuk mengukur seberapa banyak dari prediksi positif yang benar-benar positif. Ini adalah metrik yang penting ketika biaya kesalahan positif (false positives) tinggi. Berikut merupakan rumus untuk menghitung akurasi.

$$\text{Presisi} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

2.1.18.3.**Recal**

Recall merupakan suatu perhitungan untuk mengukur seberapa baik model mendeteksi semua sampel positif yang ada. Ini adalah metrik yang penting ketika biaya kesalahan negatif (false negatives) tinggi. Berikut merupakan rumus untuk menghitung akurasi.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

2.2. Kajian Pustaka

Tabel 2. 3 Review Paper

No	Judul	Author	Tahun	Klasifikasi
1	Literatur Review: Metode Klasifikasi Pada Sentimen Analisis	Mona Cindo, et all	2019	SAINTEKS 2019, ISBN: 978-602-

				52720-1-1, Januari 2019, Hal : 66 – 70,
Metode: Klasifikasi Hasil: Berdasarkan penelitian pada artikel jurnal tersebut menyimpulkan bahwa metode klasifikasi menggunakan *logistic regression* sebagai *metode* dengan akurasi tertinggi di bandingan metode *lexicon-based*, *Support Vector Machine*, *Naïve Bayes*, dan *Random Forest*.				
2	COMPARISON OF LOGISTIC REGRESSION, MULTINOMIALNB, SVM, AND K-NN METHODS ON SENTIMENT ANALYSIS OF GOJEK APP REVIEWS ON THE GOOGLE PLAY STORE	A. Maulana et all.	2023	Jurnal Teknik Informatika (JUTIF), Vol. 4, No. 6, December 2023, e-ISSN: 2723-3871, Sinta 3
Metode : *LOGISTIC REGRESSION, MULTINOMIALNB, SVM,DAN K-NN* Hasil : Berdasarkan penelitian pada artikel jurnal tersebut menyimpulkan bahwa yang memiliki performa tertinggi adalah *metode Logistic Regression*. Skor *accuracy*, *recall*, dan *precision* dari *metode Logistic Regression* secara berturut-turut adalah 82,45%, 82,49%, 82,45%, dan 82.43%.				

3	ANALISIS PERBANDINGAN SENTIMEN CORONA VIRUS DISEASE2019 (COVID19) PADA TWITTER MENGGUNAKAN METODE LOGISTIC REGRESSION DAN SUPPORT VECTOR MACHINE (SVM)	Kelvin, et all	2022	JUSIKOM PRIMA, Vol. 5 No. 2, Februari 2022, e-ISSN : 2580-2879, Sinta 4
Metode : *LOGISTIC REGRESSION DAN SUPPORT VECTOR MACHINE (SVM)* Hasil : Berdasarkan penelitian pada artikel jurnal tersebut menyimpulkan bahwa *SVM* memiliki nilai *akurasi* 91,15% dalam data test sedangkan *metode Logistic Regression* mendapatkan nilai akurasi sebanyak 87,68% dalam data test.				
4	Analisis Sentimen Pengguna Twitter Terhadap Kasus COVID-19 di Indonesia Menggunakan Metode Regresi Logistik Multinomial	Ridho Prabowo, et all	2023	JUSTIN, Vol. 11, No. 1, Januari 2023, e-ISSN : 2620-8989, Sinta 3

Metode : *Regresi Logistik Multinomial* Hasil : Berdasarkan penelitian pada artikel jurnal tersebut menyimpulkan bahwa regresi logistik *multinomial* menghasilkan *akurasi* sebesar 64%, dengan *precission* untuk sentimen positif 85%, netral 56%, dan negatif, 53% dan *recall* untuk sentimen positif 74%, netral 67%, dan negatif 50%.				
5	Analisis Sentimen Evaluasi Pembelajaran Tatap Muka 100 Persen pada Pengguna Twitter menggunakan Metode Logistic Regression	Saiful Anwar.A, Prabowo, et all	2022	Jurnal Pendidikan Tambusai, Volume 6 Nomor 2 Tahun 2022, ISSN: 2614-3097, Sinta 3
Metode : *Logistic Regression* Hasil : Berdasarkan penelitian pada artikel jurnal tersebut menyimpulkan bahwas analisis Sentimen menggunakan *metode Logistic Regression* pada penelitian ini menghasilkan nilai *accuracy* sebesar 78,57%, *precision* 76,92%, *recall* 83,3 % dan nilai *F1-Score* sebesar 80%.				