BAB III
METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Objek penelitian ini mencakup opini-opini yang disampaikan oleh masyarakat terkait pandangan, dan, pengalaman mereka terhadap layanan *Paylater*. Layanan *Paylater* semakin populer di masyarakat, namun demikian, penting untuk mengetahui bagaimana opini masyarakat terhadap layanan tersebut, baik dari sisi kemudahan, manfaat, maupun potensi risiko yang mungkin timbul. Opini publik memberikan umpan balik yang sangat berguna untuk mengevaluasi kinerja penyedia layanan *Paylater* serta memperbaiki aspek-aspek layanan yang masih dirasa kurang optimal. Analisis sentimen dilakukan untuk mengeksplorasi sentimen yang muncul dari opini publik, seperti sentimen positif, atau, negatif terkait dengan layanan *Paylater* yang diungkapkan di *platform* media sosial X (Twitter).

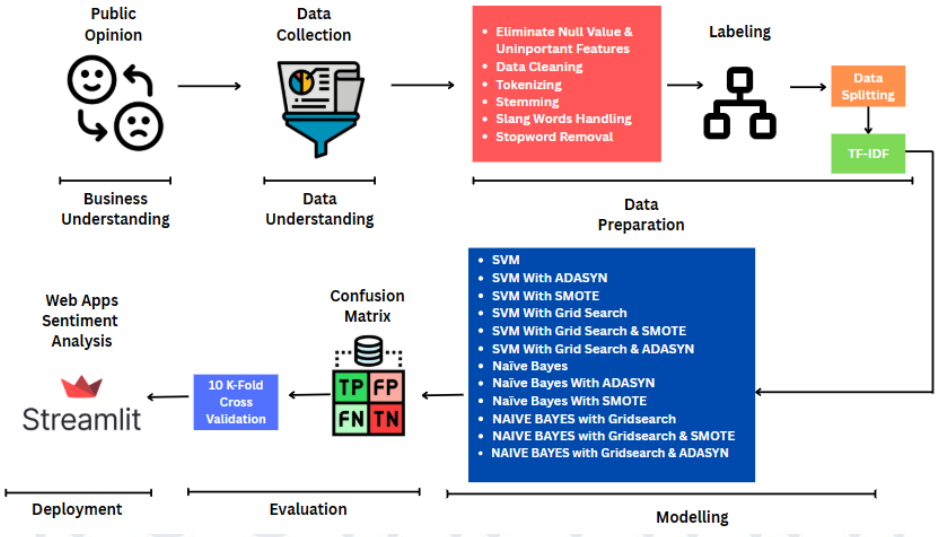
3.2 Metode Penelitian

Metode Penelitian yang digunakan pada penelitian adalah *CRISP-DM.* Merupakan metodologi dalam bidang penambangan data yang terdiri dari 6 tingkatan. Tahapan *CRISP-DM* terdiri dari *Business Understanding, Data Understanding, Data Preparation, modeling, Evaluation, Deployment* [40].

Tabel 3. 1 Perbandingan Framework Data Mining

CRISP DM	KDD	SEMMA
Memiliki 6 tahapan pengembangan *Data Mining* yaitu *Business Understanding, Data Understanding, Data Preparation, modeling, Evaluation, Deployment*	Memiliki 5 tahapan pengembangan *Data Mining* yaitu *selection, pre processing, transformation, Data Mining, interpretation/Evaluation*.	5 tahapan *Sample, Explore, Modify, model, Assess*
Terdapat tahapan penerapan *model* yang dibuat pada tahapan *Deployment*	Tidak terdapat tahap penerapan *model* atau *Deployment* .	Tidak terdapat tahap penerapan *model* atau *Deployment* .
Tahapan awal dilakukan pembelajaran dari tujuan dari projek yang merupakan tahapan pada *Data Understanding*.	Tahapan awal dilakukan *selection* yaitu melakukan pembuatan *dataset* untuk pembuatan *model*.	Fokus awal pada pengambilan sampel data

Tabel 3.1, menyajikan perbandingan beberapa kerangka kerja dalam Data Mining. Penelitian ini memilih pendekatan CRISP-DM karena memiliki orientasi yang kuat terhadap pemahaman mendalam atas tujuan serta kebutuhan bisnis, sekaligus mendukung penerapan *model* secara optimal. CRISP-DM memiliki keunggulan dalam menangani data yang telah terstruktur, menjadikannya lebih sesuai untuk proses analisis dan pemodelan yang sistematis. Berdasarkan tiga aspek teratas yang dibandingkan, peneliti memilih framework CRISP-DM sebagai pendekatan dalam pengembangan proyek Data Mining karena dinilai lebih unggul dari segi kelengkapan dan keterstrukturan tahapan. CRISP-DM memiliki enam tahapan utama, mulai dari pemahaman bisnis hingga *Deployment*, yang tidak hanya berfokus pada proses teknis seperti pemodelan, tetapi juga memastikan bahwa hasil akhir dari proses Data Mining dapat selaras dengan tujuan bisnis dan dapat diimplementasikan secara nyata. Jika dibandingkan, KDD tidak memiliki tahapan *Deployment*, sedangkan SEMMA lebih menekankan pada proses teknis tanpa mempertimbangkan aspek bisnis secara mendalam. CRISP-DM dinilai lebih komprehensif dan aplikatif, serta lebih sesuai digunakan dalam penelitian ini yang menggabungkan analisis data dengan penyusunan solusi berbasis kebutuhan nyata. Gambar 3.1, Menggambarkan Tahapan yang dilakukan dalam penelitian bedasarkan enam fase utama dari tahapan CRISP-DM.



Gambar 3. 1 Tahapan Penelitian Menggunakan Framework Crisp-Dm

3.2.1 Business Understanding

Proses *Business Understanding* merupakan langkah pertama dalam metodologi *CRISP-DM* (*Cross-Industry Standard Process for Data Mining*). Penelitian ini berfokus pada upaya memahami tujuan dan konteks bisnis di balik penggunaan layanan *Paylater* di Indonesia, serta bagaimana analisis sentimen dapat memberikan wawasan terhadap opini masyarakat. Dalam beberapa tahun terakhir, *Paylater* menjadi salah satu metode pembayaran yang populer karena kemudahannya dalam memberikan akses kredit tanpa jaminan.

Di balik kemudahannya, terdapat kekhawatiran mengenai pengaruhnya terhadap kebiasaan finansial masyarakat, terutama terkait utang konsumtif dan literasi keuangan. Penelitian ini bertujuan untuk mengidentifikasi dan memahami opini publik yang diungkapkan melalui komentar di media sosial X (Twitter), sehingga dapat memberikan gambaran yang lebih akurat mengenai penerimaan masyarakat terhadap layanan *Paylater*. Hasil analisis ini diharapkan dapat membantu pihak-pihak yang berkepentingan, seperti perusahaan *fintech*, *regulator* keuangan, dan peneliti, dalam mengambil keputusan yang lebih tepat berdasarkan opini pengguna. Melalui perbandingan performa algoritma *Naïve Bayes Classifier* dan *Support Vector Machine* (*SVM*), penelitian ini bertujuan untuk mengevaluasi metode analisis sentimen yang paling efektif digunakan untuk melakukan klasifikasi sentimen mengenai layanan *Paylater*.

3.2.2 Data Understanding

Pada langkah ini, penelitian ini melakukan pengambilan data dengan menggunakan kata kunci terkait *Paylater*. *Dataset* yang diperoleh mencakup sekitar 1052 data, yang menjadi sumber informasi signifikan dalam analisis sentimen terhadap opini publik mengenai layanan *Paylater*. Data tersebut dikumpulkan dari *platform* media sosial X dengan melakukan metode *web scraping* menggunakan *library TweetHarvest* dengan bahasa pemrograman *Python*. Proses pengambilan data dilakukan dengan *Google Colab*. Tahap *data*

understanding juga memahami karakteristik data yang telah diperoleh seperti struktur teks.

3.2.3 Data Preparation

Tahap *preprocessing*, bertujuan mempersiapkan data agar sesuai dengan kebutuhan dalam pembuatan *model*, sehingga informasi yang diperoleh menjadi lebih terstruktur, relevan, dan siap digunakan untuk proses analisis sentimen secara optimal.

1. *Eliminate Null value & Uninportant Features*, merupakan proses yang melakukan baris dengan nilai null dihapus dan fitur-fitur yang tidak mendukung proses pemodelan atau tidak memiliki kontribusi signifikan terhadap hasil analisis dihilangkan, sehingga dapat meningkatkan kualitas data yang digunakan dan memastikan *model* yang dihasilkan lebih akurat dan efisien.

2. *Data Cleaning* merupakan proses penghapusan *username*, *URL*, *Hashtag,* Tanda baca, Angka*,* simbol, dan mengubah seluruh huruf menjadi huruf kecil. Langkah pembersihan dilaksanakan untuk mengurangi keberadaan karakter atau simbol yang tidak diinginkan atau tidak memiliki makna.

3. *Tokenizing*, adalah sebuah proses pemecahan sekumpulan karakter pada suatu kalimat menjadi satuan kata.

4. *Stemming*, adalah proses yang dilakukan untuk menemukan kata dasar dari setiap kata dalam sebuah kalimat dengan cara menghapus imbuhan awal atau imbuhan akhir.

5. *Slang Words Handling,* adalah proses yang digunakan untuk mengubah kata-kata gaul atau tidak baku menjadi bentuk kata yang baku atau formal. Proses ini penting dalam analisis teks, terutama pada media sosial, agar makna dari setiap kata dapat dipahami dengan lebih akurat.

6. *Stopword Removal,* adalah proses yang digunakan untuk melakukan penghilangan kata umum yang tidak memiliki arti atau makna.

7. *Labeling*, adalah proses pemberian label atau kategori pada data untuk mempersiapkannya dalam analisis atau pelatihan *model* pembelajaran mesin. Pada penelitian ini, proses *Labeling* dilakukan dengan menetapkan dua kelas sentimen, yaitu positif dan negatif. Teknik *semi-supervised learning* digunakan untuk melabeli data, memanfaatkan kombinasi data berlabel dan tidak berlabel, sehingga memungkinkan pelatihan *model* dengan efisiensi tinggi meskipun data berlabel terbatas.

8. *Train test splitting*, adalah proses pembagian data menjadi dua bagian yaitu *Train* dan *test* dengan proporsi rasio 80:20.

9. *TF-IDF*, tahapan ini melakukan pembobotan dengan *TF-IDF* yang bertujuan untuk menghitung seberapa sering sebuah kata muncul dalam suatu dokumen, dan menghasilkan vektor skor dari setiap kata pada setiap komentar.

3.2.4 Data Modelling

Pada tahap ini, digunakan dua algoritma yaitu *SVM* dan *Naïve Bayes*. Pemilihan algoritma dilakukan untuk mendapatkan *model* dengan performa terbaik dalam klasifikasi. Untuk mengoptimalkan *model*, dilakukan *Tuning hyperparameter* menggunakan metode *Grid search*, dengan penyesuaian parameter seperti C, kernel, dan Gamma pada *SVM*, serta *alpha* fit_prior, class_prior pada *Naïve Bayes*. Untuk mengatasi ketidakseimbangan kelas, digunakan teknik *balancing* berupa *oversampling* dengan metode *SMOTE* dan *ADASYN*, yang bertujuan memperbanyak data kelas minoritas secara sintetis agar distribusi kelas menjadi lebih seimbang. Setelah *model* dilatih, dilakukan evaluasi awal menggunakan metrik seperti akurasi, *precision*, *recall*, dan *f1-score* untuk mengukur kemampuan awal *model* dalam mengenali data.

1. *Modelling* Algoritma *SVM*(Support Vector Machine), tanpa dilakukan Hyper Parameter tuning, dan data balancing untuk melakukan prediksi sentimen pada komentar pengguna.
2. *Modelling Algoritma SVM Dengan ADASYN,* menerapkan metode *data balancing* menggunakan *ADASYN* untuk menyamaratakan proporsi fitur label *positive*, dan *negative* menjadi seimbang.

3. *Modelling Algoritma SVM Dengan SMOTE,* menerapkan metode *data balancing* menggunakan *SMOTE* untuk menyamaratakan proporsi fitur label *positive*, dan *negative* menjadi seimbang.
4. *Modelling* Algoritma *SVM Dengan Grid search*, menggunakan *Hyper Parameter tuning Grid search*. untuk melakukan prediksi sentimen pada komentar pengguna.
5. *Modelling* Algoritma *SVM Dengan ADASYN & Grid search*, dengan menggunakan *Hyper Parameter tuning Grid search*, dan menerapkan metode *data balancing* menggunakan *ADASYN* untuk menyamaratakan proporsi fitur label *positive*, dan *negative* menjadi seimbang, pada saat pemodelan melakukan prediksi sentimen pada komentar pengguna.
6. *Modelling* Algoritma *SVM Dengan SMOTE & Grid search*, dengan menggunakan *Hyper Parameter tuning Grid search*, dan menerapkan metode *data balancing* menggunakan *ADASYN* untuk menyamaratakan proporsi fitur label *positive*, dan *negative* menjadi seimbang, pada saat pemodelan melakukan prediksi sentimen pada komentar pengguna.
7. *Modelling* Algoritma *Naïve Bayes Classifier* , tanpa dilakukan *Hyper Parameter tuning*, dan *data balancing*. untuk melakukan prediksi sentimen pada komentar pengguna.
8. *Modelling* Algoritma *Naïve Bayes* Dengan *ADASYN*, menerapkan metode *data balancing* menggunakan *ADASYN* untuk menyamaratakan proporsi fitur label *positive*, dan *negative* menjadi seimbang.
9. *Modelling* Algoritma *Naïve Bayes* Dengan *SMOTE*, menerapkan metode *data balancing* menggunakan *SMOTE* untuk menyamaratakan proporsi fitur label *positive*, dan *negative* menjadi seimbang.
10. *Modelling* Algoritma *Naïve Bayes Classifier Dengan Grid search*, dengan menggunakan *Hyper Parameter tuning Grid search*. untuk melakukan prediksi sentimen pada komentar pengguna.
11. *Modelling* Algoritma *Naïve Bayes Classifier dengan ADASYN & Grid search*, dengan menggunakan *Hyper Parameter tuning Grid search*, dan menerapkan metode *data balancing* menggunakan *ADASYN* untuk

menyamaratakan proporsi fitur label *positive*, dan *negative* menjadi seimbang, pada saat pemodelan melakukan prediksi sentimen pada komentar pengguna.

12. *Modelling* Algoritma *Naïve Bayes Classifier Dengan SMOTE & Grid search*, dengan menggunakan *Hyper Parameter tuning Grid search*, dan menerapkan metode *data balancing* menggunakan *SMOTE* untuk menyamaratakan proporsi fitur label *positive*, dan *negative* menjadi seimbang, pada saat pemodelan melakukan prediksi sentimen pada komentar pengguna
13. *Load model*(Menyimpan *model*) dengan format .pkl untuk *model* yang telah dibuat dapat digunakan untuk *Deployment*

3.2.5 Evaluation

Tujuan dilakukan tahap *Evaluation*, adalah untuk memastikan sistem telah berjalan sesuai dengan kebutuhan yang ditetapkan. Selanjutnya, evaluasi dilakukan melalui penerapan *Confusion matrix* untuk mengukur performa sistem yang telah dibuat, termasuk nilai akurasi, presisi, *recall, dan f1-score*. Pada tahapan ini juga dilakukan validasi *model* menggunakan *Cross-Validation* dengan 10 lipatan(*K-fold)*.

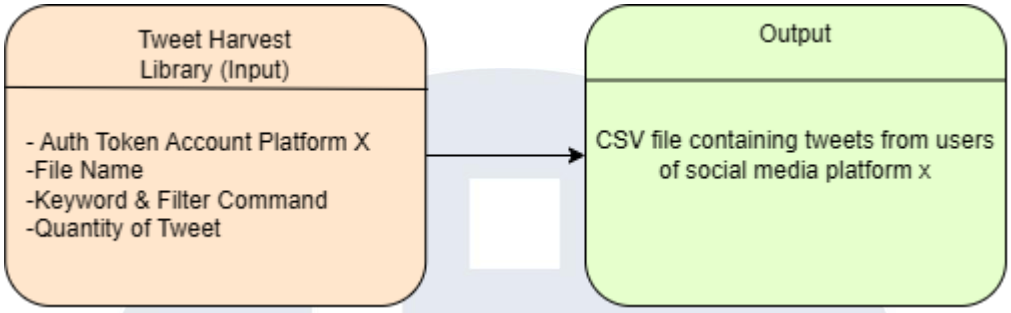
3.2.6 Deployment

Pada tahap *Deployment* , *model* yang telah dibangun diimplementasikan dalam bentuk aplikasi web menggunakan *library Streamlit* dan bahasa pemrograman *Python*. Aplikasi ini dirancang agar dapat memprediksi sentimen pengguna terhadap layanan *Paylater* secara otomatis. Proses prediksi dilakukan dengan memanfaatkan algoritma terbaik yang telah dipilih berdasarkan evaluasi kinerja *model* sebelumnya, sehingga hasil analisis dapat digunakan sebagai bahan pertimbangan dalam pengambilan keputusan. Dilakukan pula *User Acceptance Testing* (UAT) untuk memastikan bahwa aplikasi dapat berjalan dengan baik sesuai kebutuhan pengguna, baik dari segi fungsionalitas, kemudahan penggunaan, maupun akurasi hasil prediksi.

3.3 Teknik Pengumpulan Data

Penelitian ini menggunakan data primer yang diperoleh melalui proses *web scraping* dari media sosial Twitter dengan memanfaatkan pustaka (*library*) *TweetHarvest* menggunakan bahasa pemrograman *Python* [36]. *TweetHarvest* merupakan salah satu alat yang dirancang untuk memudahkan pengambilan data dari Twitter secara otomatis, yang memungkinkan peneliti untuk mengakses, menyaring, dan mengunduh *tweet* berdasarkan kata kunci tertentu, waktu publikasi, serta parameter relevan lainnya. endekatan ini dipilih karena kemampuannya dalam mengumpulkan data dalam jumlah besar secara cepat dan efisien, tanpa perlu melakukan interaksi manual dengan antarmuka Twitter. Proses pengumpulan data pada penelitian ini dapat dilakukan secara sistematis dan konsisten, yang penting dalam menjaga kualitas dan validitas data penelitian. Penggunaan metode *web scraping* memungkinkan peneliti untuk menyesuaikan parameter pencarian sesuai kebutuhan, seperti rentang waktu, kata kunci, dan bahasa. Hal ini memberikan fleksibilitas tinggi dalam memperoleh data yang benar-benar relevan dengan topik yang diteliti.

Dalam penelitian ini, data yang dikumpulkan berupa opini publik yang mengandung kata kunci terkait layanan *Paylater* [37], sehingga hanya *tweet* yang relevan dengan topik penelitian yang disimpan untuk dianalisis lebih lanjut. Data tersebut kemudian diekspor dan disimpan dalam format *.csv* (Comma-Separated Values), yang merupakan format populer dan kompatibel dengan berbagai alat analisis data seperti *Microsoft Excel*, *Python Pandas*, maupun perangkat lunak statistik lainnya. Penyimpanan dalam format ini bertujuan untuk dapat memudahkan proses transformasi data, pembersihan data, serta analisis data yang dilakukan pada tahap *Data Preparation*. Proses atau tahapan dalam pengumpulan data tersebut dapat dilihat secara rinci pada Gambar 3.2, yang menggambarkan alur kerja dari proses *web scraping* hingga data siap digunakan untuk analisis.



Gambar 3. 2 Teknik Pengumpulan Data

Metode pengambilan sampel yang digunakan dalam penelitian ini adalah *purposive sampling*, yaitu teknik pengambilan sampel di mana peneliti secara selektif memilih data yang memenuhi kriteria tertentu dan relevan dengan tujuan penelitian [26]. Dalam konteks penelitian ini, data yang diambil berupa opini publik di platform *X* (Twitter) yang membahas layanan *Paylater*.

Referensi kata kunci merujuk pada penelitian yang dilakukan oleh Safira et al. [11], dengan menggunakan kombinasi seperti *"Paylater (pakai OR pake) (to:worksfess)"*. Kombinasi ini digunakan untuk menyaring *tweet* yang berisi opini terkait layanan *Paylater* dan ditujukan kepada atau menyebut akun *@worksfess*, sebuah akun *autobase* yang banyak membagikan kiriman anonim dari pengguna, khususnya mengenai topik pekerjaan dan keuangan [11]. Pada penelitian ini *query* pencarian dikembangkan lebih lanjut menjadi *"Paylater" (to:worksfess OR to:tanyakanrl) -gestun -cair -gesek -tunai since:2024-02-01 until:2025-02-28*. Pengembangan ini bertujuan untuk memperoleh data yang lebih relevan dan bersih dengan memperluas sumber opini melalui penambahan akun *@tanyakanrl*, yang juga merupakan akun *autobase* aktif yang kerap membahas topik kehidupan.

Pengecualian kata-kata seperti *gestun*, *cair*, *gesek*, dan *tunai* dilakukan untuk menyaring *tweet* yang cenderung mengarah pada penyalahgunaan layanan *Paylater* sebagai alat pencairan dana instan yang tidak sesuai konteks penelitian. Rentang waktu pengambilan data dibatasi dari tanggal 1 Februari 2024 hingga 28 Februari 2025 agar data yang diperoleh tetap aktual dan relevan dengan periode analisis. Berdasarkan query tersebut, terkumpul sebanyak 1.052 data

yang berasal dari tweet pengguna yang mengirimkan opini ke akun *@worksfess* dan @tanyakanrl. *Detail* penjelasan *search query* dapat dilihat pada tabel 3.2.

Tabel 3. 2 Penjelasan Search Query

Bagian Query	Fungsi	Keterangan
"*Paylater*"	Mencari tweet yang mengandung kata *"Paylater"*.	Kata kunci utama
(to:worksfess OR to:tanyakanrl)	digunakan untuk mengambil tweet yang ditujukan ke akun *@worksfess* atau *@tanyakanrl*.	*Target* akun
-gestun	Menghapus tweet yang mengandung kata *"gestun"*.	Filter negatif
-cair	Menghapus tweet yang mengandung kata *"cair"*.	Filter negatif
-gesek	Menghapus tweet yang mengandung kata *"gesek"*.	Filter negatif
-tunai	Menghapus tweet yang mengandung kata *"tunai"*.	Filter negatif
since:2024-02-01	Mencari tweet mulai 1 Februari 2022.	Rentang waktu awal
until:2025-02-28	Mencari tweet hingga 28 Februari 2025.	Rentang waktu akhir

3.4 Teknik Analisis Data

Penelitian ini bertujuan untuk menganalisis sentimen publik terhadap layanan *Paylater* berdasarkan data yang dikumpulkan dari media sosial X. Dalam konteks implementasi *Machine Learning*, metode yang paling sering digunakan untuk analisis sentimen adalah *Support Vector Machine* (*SVM*) dan *Naïve Bayes*. Tabel 3.3, menyajikan perbandingan antara algoritma *Support Vector Machine* dan *Naïve Bayes*[24].

Tabel 3. 3 Perbandingan Svm, Dan Naïve Bayes
Sumber : [24]

Perbandingan	***Support Vector Machine(SVM)***	***Naïve Bayes*(NB)**
Kemampuan algoritma dalam mengolah data	Cenderung memberikan kinerja yang baik pada *dataset* kecil hingga menengah	Cenderung memberikan kinerja yang baik pada *dataset* kecil hingga besar
Training data	Untuk mencapai hasil yang optimal, terutama pada *model* kompleks, dibutuhkan beberapa ribu hingga jutaan data pelatihan.	Meski jumlah data pelatihan yang lebih sedikit, *model* ini tetap dapat menghasilkan kinerja yang baik meski hanya dengan ratusan hingga beberapa ribu data.

Perbandingan	*Support Vector Machine(SVM)*	*Naïve Bayes*(NB)
Kecepatan pelatihan dan prediksi	memiliki waktu proses yang lama, khususnya pada *dataset* yang memiliki volume besar dalam melakukan *training* data	Memiliki waktu proses yang cepat karena kesederhanaannya.

Pemilihan *Support Vector Machine* (*SVM*) dan *Naïve Bayes* untuk analisis sentimen opini masyarakat indonesia mengenai layanan *Paylater* di X (Twitter) didasarkan pada keunggulan masing-masing metode yang saling melengkapi. *SVM* terkenal karena kemampuannya dalam menangani data yang kompleks dan besar, serta kemampuannya membedakan kelas dengan presisi tinggi meskipun data memiliki banyak dimensi dan fitur, yang sering muncul dalam *tweet* yang bervariasi. Sementara itu, *Naïve Bayes* adalah algoritma yang lebih sederhana dan lebih cepat, dengan performa yang baik dalam mengolah teks. Kecepatan dan efisiensi *Naïve Bayes* sangat penting untuk analisis *tweet* dalam jumlah besar, sehingga hasil analisis dapat diperoleh dengan cepat dan tepat waktu.

3.5 Teknik Pengujian Atau Validasi Sistem

Teknik pengujian dalam penelitian ini dilakukan terhadap dua aspek utama, yaitu model *machine learning* dan sistem berbasis web. Untuk model *machine learning*, digunakan metode evaluasi *confusion matrix* dan *k-fold cross validation*. *Confusion matrix* digunakan untuk menilai performa klasifikasi model dengan menampilkan jumlah prediksi benar dan salah dalam tiap kelas, serta menghasilkan metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *f1-score* [23]. *k-fold cross validation* digunakan untuk menghindari *overfitting* dan mengevaluasi kestabilan model [42]. Dalam Penelitian ini digunakan *10-fold cross validation*, di mana data dibagi menjadi sepuluh bagian, dan model diuji sebanyak sepuluh kali dengan setiap bagian bergiliran menjadi data uji, sementara sisanya digunakan sebagai data latih. Nilai performa akhir diperoleh dari rata-rata seluruh iterasi [43].

Untuk menguji sistem berbasis web pada fase *deployment*, digunakan metode *User Acceptance Test (UAT)* dengan menggunakan skala *Likert* [45].

Pengguna diminta memberikan penilaian terhadap 5 aspek yaitu menjadi 5 fitur yaitu *Input data, Data Preparation, Sentimen Analisis,* Visualisasi, dan *Summary Insight*. Skala penilaian dilakukan dengan menggunakan skala likert terdapat 5 skala yaitu sangat tidak setuju, tidak setuju, netral, setuju, dan sangat setuju, hasilnya dianalisis untuk mengetahui sejauh mana sistem telah memenuhi kebutuhan dan ekspektasi pengguna[46].