BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Objek penelitian yang dilakukan dalam penelitian ini adalah analisis sentimen terhadap opini publik tentang Kementerian Keuangan, dengan fokus pada pemahaman dan klasifikasi sentimen masyarakat terhadap lembaga tersebut, yang tercermin dalam platform media sosial X. Kementerian Keuangan, sebagai salah satu lembaga di bawah pengawasan dan bertanggung jawab kepada Presiden Republik Indonesia, memiliki fungsi melaksanakan tugas pemerintahan terkait keuangan negara dan kekayaan negara. Tugas utama Kementerian Keuangan mencakup pelaksanaan urusan pemerintahan di sektor keuangan negara dan pengelolaan kekayaan negara[44]. Dalam konteks ruang lingkup pemerintahan, media sosial seperti X dapat dijadikan sebagai alat penting untuk menyampaikan berbagai kebijakan dan informasi yang dikeluarkan oleh pemerintah. X dipilih karena popularitasnya yang global memungkinkan akses ke beragam opini dari berbagai latar belakang dan wilayah. Platform ini juga memungkinkan informasi disampaikan secara langsung dan real-time, memastikan data yang digunakan selalu terkini [45].

Objek penelitian ini mencakup opini-opini yang dinyatakan oleh masyarakat terkait pandangan, evaluasi terhadap kinerja, dan kebijakan yang diterapkan oleh Kementerian Keuangan. Kementerian Keuangan membutuhkan opini publik karena kebijakan dan tindakan yang diambil oleh Kementerian Keuangan dapat memiliki dampak yang signifikan terhadap masyarakat secara keseluruhan. Opini publik memberikan umpan balik yang penting bagi Kementerian Keuangan untuk memahami pandangan, kebutuhan, dan harapan masyarakat terkait dengan kebijakan fiskal, pengelolaan keuangan publik, dan kegiatan ekonomi lainnya yang berkaitan dengan keuangan negara [46]. Analisis sentimen dilakukan untuk mengeksplorasi sentimen yang muncul, seperti opini yang bersifat positif, negatif, atau netral, terhadap Kementerian Keuangan di *platform* media sosial X. Analisis

sentimen dilakukan untuk mengeksplorasi sentimen yang muncul, seperti opini yang bersifat positif, negatif, atau netral, terhadap Kementerian Keuangan di platform media sosial X. Data opini publik Kemenkeu diperoleh menggunakan keyword kemenkeu dengan rentang waktu Agustus 2023 hingga Agustus 2024. Penggunaan algoritma *Naive Bayes* dan *Support Vector Machines (SVM)* dipilih untuk mengklasifikasikan opini secara otomatis dalam analisis sentimen, memanfaatkan data yang tersedia di X terkait dengan Kementerian Keuangan.

3.2 Metode Penelitian

Metode penelitian mengacu pada pendekatan yang terstruktur dan sistematis yang digunakan oleh peneliti untuk merancang, melaksanakan, dan mengevaluasi suatu studi atau penelitian. Dalam penelitian analisis sentimen opini publik terkait Kementerian Keuangan, pendekatan kualitatif digunakan dengan memanfaatkan data dari media sosial X. Proses pengumpulan data dilakukan melalui *scraping*, untuk mengakses dan mengumpulkan informasi opini masyarakat secara langsung dari *platform* tersebut. Pemilihan metode kualitatif ini bertujuan untuk memberikan pemahaman yang lebih terukur dan statistik terkait sentimen pengguna X terhadap Kementerian Keuangan.

Metode penelitian terdapat beberapa jenis, metode yang paling sering digunakan adalah metode kualitatif dan kualitatif. Metode kualitatif adalah metode yang menggunakan pengamatan, wawancara. Metode kuantitatif adalah metode yang menggunakan analisis statistik terhadap data angka. Penelitian ini menggunakan metode kuantitatif yang dikarenakan menggunakan dataset di ambil dari *social media* X. Dataset diambil menggunakan teknik *scraping* dengan bantuan bahasa pemograman *Python*. Pada penelitian dengan metode kuantitatif, akan digunakan perbandingan antara algoritma *Naive Bayes*, *Support Vector Machines (SVM*) untuk analisis sentiment.

Pada proses *data mining*, terdapat beberapa metode yang sering digunakan, antara lain *Cross Industry Standard Process (CRISP-DM), Knowledge Discovery in Database (KDD) [41], dan SEMMA (Sample, Explore, Modify, Model, and*

Assess)[47]. Setiap metode ini memiliki tahapan-tahapan yang berbeda. Berikut adalah rangkuman tentang tahapan dari masing-masing metode.

Tabel 3. 1 Perbandingan *CRISP-DM, SEMMA, KDD*

Metode	Kelebihan	Kekurangan	Tahapan
CRISP-DM	*CRISP-DM* memiliki struktur yang terorganisir dengan tahapan-tahapan yang jelas namun juga fleksibel, serta panduan yang detail untuk setiap tahapan proses *data mining* memudahkan alur kerja	Memerlukan waktu yang lebih lama untuk menyelesaikan seluruh tahapan proses	*Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment*
SEMMA	Metodologi ini fokus pada langkah-langkah kunci dalam proses *data mining* dan mudah untuk diimplementasikan serta cocok digunakan untuk analisis data yang langsung, tidak terlalu komplek	SEMMA tidak cocok untuk proyek-proyek *data mining* memerlukan analisis yang lebih mendalam atau lebih banyak tahapan yang terstruktur dan tidak terlalu rumit	*Sample, Explore, Modify, Model, Assesment*
KDD	Metodologi ini memungkinkan untuk pemahaman yang lebih mendalam terhadap data dan masalah yang dihadapi karena pemahaman data dan seleksi data yang cermat	KDD seringkali memakan waktu dan memerlukan sumber daya yang cukup besar terutama pada tahap pemrosesan data serta model yang dihasilkan kompleks untuk diinterpretasikan	*Pre KDD, Selection, Pro processing, Transformation, Data mining, Interpretation/Evaluation, Post KDD*

Pada tabel 3.1 dari perbandingan *CRISP-DM, SEMMA, KDD,* menggunakan *CRISP-DM*. Metodologi *CRISP-DM* memberikan panduan langkah-demi-langkah yang mudah dipahami dan terstruktur dalam melakukan proses *data mining*. Struktur yang terorganisir dapat mengikuti alur kerja yang sistematis dari pemahaman masalah hingga evaluasi hasil. Hal ini membantu mengelola proyek-proyek *data mining* dengan lebih efisien dan memastikan tidak ada tahapan yang terlewatkan. Teknik *CRISP-DM* (*Cross Industry Standard Process for Data mining*) sebagai metodologi. *CRISP-DM* ini terdiri dari enam tahap utama yang disusun secara sistematis untuk membimbing penelitian atau proyek *data mining*. Tahap-tahap tersebut meliputi Pemahaman Bisnis (*Business Understanding*), Pemahaman Data (*Data Understanding*), Persiapan Data (*Data Preparation*), Pembuatan Model (*Modeling*), Evaluasi (*Evaluation*), dan Penyebaran (*Deployment*). *CRISP-DM* memberikan kerangka kerja yang terstruktur untuk memahami masalah bisnis, mengumpulkan dan mempersiapkan data, mengembangkan model, mengevaluasi hasil, dan mengimplementasikan solusi.

3.3 Variabel Penelitian

Penelitian ini memanfaatkan dataset yang diperoleh dari *platform* media sosial X. Dataset ini dihasilkan melalui proses *scraping* data. Dalam kerangka variabel penelitian, terdapat dua jenis variabel, yaitu variabel independen dan variabel dependen.

3.3.1 Variabel Independen

Variabel independen adalah variabel yang tidak terikat atau bersifat bebas, mampu memberikan pengaruh pada variabel lainnya. Dalam konteks penelitian ini, variabel independen terutama terfokus pada variabel "full_text". Variabel ini menjadi fokus dalam menganalisis dampaknya terhadap variabel lainnya.

3.3.2 Variabel Dependen

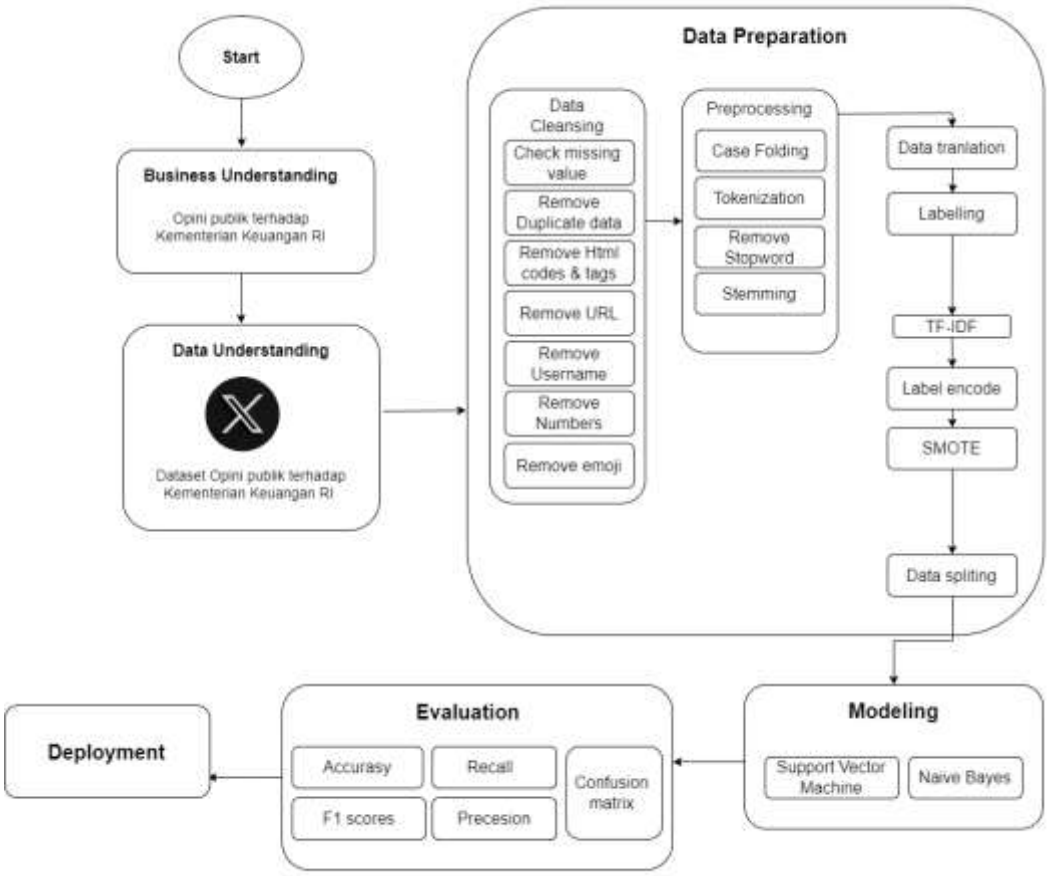
Variabel dependen merupakan variabel yang bersifat terikat atau dipengaruhi oleh variabel independen. Dalam penelitian ini, variabel dependen terfokus pada variabel "label" yang memiliki kategori positif, negatif, dan netral. Variabel ini menjadi objek penelitian untuk mengidentifikasi bagaimana variabel independen, yaitu " *full_text* ", mempengaruhi atau berhubungan dengan label-label tersebut.

3.3 Teknik Pengumpulan Data

Teknik pengumpulan data mengacu pada cara atau metode yang digunakan untuk mengumpulkan informasi atau data dalam sebuah penelitian atau studi. Data primer merujuk pada informasi yang dikumpulkan secara langsung dari sumber aslinya untuk keperluan penelitian tertentu. Data sekunder adalah informasi yang telah dikumpulkan oleh pihak lain untuk tujuan yang berbeda dari penelitian yang sedang dilakukan. Data tersier adalah data yang telah diolah atau diinterpretasikan oleh pihak ketiga, yang kemudian disajikan dalam bentuk yang siap digunakan oleh peneliti. Pada penelitian ini, teknik pengumpulan data dilakukan secara primer yang diperoleh melalui *scraping* data dari media sosial X dengan data teks yang dikumpulkan dari bulan Oktober 2023 hingga Februari 2024. Dataset ini berisi opini publik mengenai pandangan masyarakat terhadap Kementerian Keuangan.

3.4 Teknik Analisis Data

Dalam penelitian ini, akan menerapkan pendekatan model *CRISP-DM* (*Cross Industry Standard Process for Data mining*). Berdasarkan gambar 3.1 menampilkan enam tahap dalam *CRISP-DM* akan diterapkan dalam penelitian ini, dengan penjelasan sebagai berikut [46].



Gambar 3. 1 Teknik Analisis Data

3.4.2 *Business Understanding*

Tahapan *Business Understanding* adalah langkah awal dalam metodologi *CRISP-DM* (*Cross-Industry Standard Process for Data mining*). Pada penelitian ini tujuan utama adalah untuk analisis sentimen terhadap opini publik tentang pandangan masyarakat terhadap Kementerian Keuangan dengan pemilihan algoritma *Naive Bayes* dan *Support Vector Machine*. Pemahaman terhadap perspektif masyarakat mengenai pemerintahan, kebijakan keuangan, dan peran Kementerian Keuangan. Analisis sentimen secara spesifik, seperti mengidentifikasi pola sentimen positif, negatif, atau netral, terkait pandangan masyarakat terhadap Kemenkeu.

3.4.3 *Data Understanding*

Pada langkah ini, penelitian ini melakukan pengambilan data dengan menggunakan kata kunci terkait Kementerian Keuangan dan Kemenkeu. Dataset yang dihasilkan mencakup sekitar 2354 data, menjadi sumber informasi yang signifikan untuk dilibatkan dalam analisis sentimen terhadap opini publik terkait Kementerian Keuangan. Data yang digunakan berasal dari *platform* media sosial X dan diperoleh melalui proses *scraping* dengan Google collab.

3.4.4 *Data Preparation*

Pada tahapan ini, dilaksanakan proses pembersihan data yang dimulai dengan cleansing data dan tahap *text preprocessing*. Pada tahap *text preprocessing*, beberapa proses dilakukan, termasuk *case folding, tokenizing, stemming,* dan *stopword*. Proses ini bertujuan untuk memastikan data yang digunakan dalam analisis lebih terstruktur dan bersih, memungkinkan hasil yang lebih akurat dalam langkah-langkah analisis selanjutnya. Tahapan *data preparation* terdiri dari *data cleansing, preprocessing, data translation, labeling, TF-IDF, label encode, SMOTE, data splitting.*

3.4.4.1 *Data Cleansing*

Data cleansing merupakan proses membersihkan data dari kesalahan, duplikasi, atau informasi yang tidak lengkap, untuk memastikan data yang digunakan akurat, konsisten, dan siap untuk dianalisis atau diproses lebih lanjut. Tujuannya adalah meningkatkan kualitas data sehingga analisis atau keputusan yang dibuat berdasarkan data tersebut menjadi lebih valid. Berikut merupakan tahapan data cleansing yang digunakan pada penelitian ini *check missing value, remove duplicate data, remove html codes & tags, remove URL, remove username, remove numbers, remove emoji.*

3.4.4.2 *Preprocessing*

Tahapan preprocessing merupakan proses persiapan data mentah, khususnya data teks, sebelum dilakukan analisis atau digunakan dalam model machine learning. Langkah ini bertujuan untuk membersihkan dan menyederhanakan data agar lebih mudah dipahami oleh algoritma. Preprocessing penting untuk meningkatkan kualitas data dan hasil analisis.

1. Case Folding: Mengubah semua huruf dalam teks menjadi huruf kecil (lowercase) agar konsisten, sehingga perbedaan huruf besar dan kecil tidak mempengaruhi analisis. Misalnya, "Kemenkeu" dan "kemenkeu" dianggap sama.
2. Tokenization: Memecah teks menjadi unit-unit kecil seperti kata atau frasa. Misalnya, kalimat "Kemenkeu mengelola anggaran" akan dipecah menjadi ["Kemenkeu", "mengelola", "anggaran"].
3. Remove Stopwords: Menghilangkan kata-kata umum yang tidak memiliki makna penting dalam analisis, seperti "dan", "yang", "atau". Tujuannya untuk fokus pada kata-kata kunci yang lebih bermakna.
4. Stemming: Mengubah kata-kata menjadi bentuk dasarnya. Misalnya, kata "mengelola" akan diubah menjadi "kelola", sehingga berbagai bentuk kata dianggap sebagai satu entitas.

3.4.4.3 *Data Translation*

Data translation adalah proses mengubah data yang telah dikumpulkan dan dibersihkan menjadi bahasa inggris dengan bantuan *library deep translator*. Data yang di *translate* adalah data yang telah dibersihkan dan diproses sebelumnya. Tahap ini penting untuk mempersiapkan data sebelum pelabelan menggunakan *library VADER*.

3.4.4.4 *Labeling*

Labeling merupakan proses memberi label atau kategori pada data berdasarkan kriteria tertentu. Analisis sentimen, labeling biasanya dilakukan untuk mengklasifikasikan teks, seperti tweet atau ulasan, ke

dalam kategori seperti positif, negatif, atau netral. Proses ini penting untuk memudahkan analisis lebih lanjut dan pengembangan model pembelajaran mesin. Pada proses pelabelan data, menggunakan *library VADER (Valence Aware Dictionary Sentiment Reasoner)*. *VADER* menilai setiap teks dan memberikan skor positif, negatif, atau netral. Skor-skor ini kemudian dijumlahkan untuk menghasilkan nilai skor komposit. *Compound score* ukuran yang mempertimbangkan semua skor yang dinormalisasi dalam rentang -1 hingga +1. Nilai komposit di atas 0,05 dianggap sebagai sentimen positif, sedangkan nilai di bawah -0,05 dianggap negatif. Jika nilai berada di antara -0,05 dan 0,05, maka dikategorikan sebagai netral. Pada tabel 3.2 perbandingan yang bisa menjelaskan perbedaan dan kegunaan utama VADER dan SMOTE dalam konteks analisis sentimen:

Tabel 3. 2 Perbandingan Vader danSMOTE

Aspek	**VADER**	**SMOTE**
Jenis Analisis	Sentimen berbasis **lexicon** (daftar kata)	Teknik **resampling** untuk penyeimbangan data yang tidak seimbang dalam klasifikasi
Penelitian	Digunakan untuk **menganalisis sentimen tweet** terkait Kemenkeu (positif, negatif, netral)	Membantu dalam **penyeimbangan data** untuk melatih model SVM dan Naive Bayes pada data tweet
Kata Kunci Utama	"Positive," "Negative," "Neutral," "Compound Score"	"Oversampling," "Synthetic Minority," "Data Imbalance"
Pendekatan	**Lexicon-based**: Menilai sentimen teks pendek seperti tweet	**Data Resampling**: Menyeimbangkan dataset agar model tidak bias ke kelas mayoritas

3.4.4.5 *TF-IDF*

TF-IDF (Term Frequency-Inverse Document Frequency) adalah metode yang digunakan dalam pengolahan teks untuk menilai seberapa penting suatu kata dalam sebuah dokumen dalam konteks koleksi dokumen lainnya. Metode ini memberikan bobot pada setiap kata dan menghitung nilai invers berdasarkan kemunculannya dalam kalimat. *TF-IDF* bertujuan untuk mengubah teks menjadi vektor numerik dengan mempertimbangkan frekuensi kata dalam dokumen serta frekuensi kemunculannya di seluruh kumpulan data.

3.4.4.6 *Label Encode*

Label Encoding adalah metode pengkodean yang mengubah label kategori menjadi format numerik. Dalam hal ini, label sentimen seperti positif, negatif, dan netral diubah menjadi angka, yaitu ‘negatif: 0’, ‘netral: 1’, dan ‘positif: 2’. Proses ini dilakukan setelah tahap TF-IDF untuk mempermudah model dalam memproses data numerik.

3.4.4.7 *SMOTE*

SMOTE (Synthetic Minority Over-sampling Technique) merupakan teknik yang digunakan untuk mengatasi masalah ketidakseimbangan kelas dalam dataset. *SMOTE* membantu meningkatkan performa model dalam mengklasifikasikan data, sehingga model menjadi lebih baik dalam mengenali pola dari kelas yang kurang. Dataset yang telah dilabeli menggunakan *VADER*, langkah selanjutnya adalah menyesuaikan parameter untuk memastikan jumlah sampel sintetis yang dihasilkan sebanding dengan kelas mayoritas. Berikut merupakan kelebihan dan kekurangan menggunakan Teknik *SMOTE, ROS* dan *Data augmentation* [48], [49]:

Tabel 3. 3 Kelebihan dan kekurangan SMOTE, ROS, Data augmentation

Teknik	Kelebihan	Kekurangan
SMOTE	- Menghasilkan sampel sintetis yang menambah variasi dalam kelas minoritas. - Cocok untuk data numerik dan bermanfaat dalam klasifikasi.	- Berisiko overfitting jika sampel sintetis terlalu mirip dengan sampel asli.
ROS	- Dapat digunakan untuk semua jenis data tanpa modifikasi tambahan. - Mudah diterapkan dan tidak membutuhkan komputasi berat.	- Risiko overfitting tinggi karena hanya menyalin data minoritas secara acak. - Tidak menambah variasi dalam kelas minoritas.
Data augmentation	- Meningkatkan ukuran data secara signifikan untuk memperkaya variasi - Cocok untuk data visual, teks, dan suara.	- Berisiko menambahkan noise atau distorsi yang tidak diinginkan pada data. - Memerlukan metode khusus untuk setiap tipe data (misalnya, rotasi gambar, penggeseran teks).

SMOTE digunakan karana menghasilkan sampel sintetis melalui interpolasi antara titik data yang ada, menambah variasi dan kedalaman pada kelas minoritas. Sementara ROS hanya menyalin data yang sudah ada, berisiko menyebabkan model terjebak pada pola yang sama dan meningkatkan kemungkinan *overfitting*. Selain itu, meskipun data augmentation efektif untuk data visual dan teks, teknik ini memerlukan pemahaman mendalam tentang transformasi yang relevan, yang mungkin tidak selalu mudah diterapkan pada semua jenis data. Dengan demikian, *SMOTE* terbukti lebih efektif dalam meningkatkan kinerja model pada dataset yang tidak seimbang, membantu model mengenali pola dalam kelas minoritas, dan meningkatkan metrik evaluasi seperti akurasi dan *recall*.

3.4.4.8 *Data Splitting*

Data splitting merupakan proses membagi dataset menjadi dua atau lebih subset untuk tujuan pelatihan dan pengujian model. Tujuan utama dari pemisahan ini adalah untuk mengevaluasi seberapa baik model yang dibangun dapat memprediksi data yang tidak terlihat sebelumnya. Rasio pembagian data, seperti 90:10, 80:20, atau 70:30, angka pertama menunjukkan persentase data yang digunakan untuk pelatihan (training set), sementara angka kedua menunjukkan persentase data yang digunakan untuk pengujian (test set). Hasil penelitian tersebut menunjukkan bahwa rasio 80:20 menghasilkan tingkat akurasi yang tinggi

3.4.5 Modeling

Pada langkah ini, akan dilakukan analisis terhadap pemodelan menggunakan algoritma klasifikasi yang telah dipilih yaitu *Naive Bayes* dan *Support Vector Machines (SVM)*. Pemilihan algoritma-algoritma ini berdasarkan hasil penelitian terdahulu yang telah menggunakan *Naive Bayes* dan *SVM* dalam analisis sentimen. Proses pemodelan akan menggunakan bahasa pemrograman *Python* untuk mengimplementasikan algoritma-algoritma tersebut secara efektif.

3.4.6 *Evaluation*

Pada langkah ini, akan dilakukan evaluasi kinerja dari model yang telah dibuat. Evaluasi ini mencakup penyajian *akurasi, precision, recall, dan F1-score* dan *confusion matrix* yang telah dihasilkan oleh algoritma *Naive Bayes,* dan *Support Vector Machines (SVM)* serta hasil dari pengujian model.

3.4.7 *Deployment*

Langkah terakhir dalam proses ini adalah menerapkan hasil yang telah didapatkan. Dalam penelitian ini, tahap implementasi akan mencakup pembuatan sebuah *dashboard* yang akan menampilkan visualisasi data terkait analisis sentimen terhadap pendapat publik tentang Kementerian Keuangan. *Dashboard* ini akan memberikan gambaran yang jelas dan mudah dipahami mengenai pola opini yang

beredar di masyarakat terkait lembaga tersebut. Demikian, informasi yang disajikan dapat digunakan sebagai dasar untuk mengambil keputusan yang lebih baik di masa depan.