BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Objek dalam penelitian adalah komentar atau opini yang terdapat pada platform X yang membahas mengenai cyptocurrency. Komentar-komentar tersebut akan dikumpulkan dengan menggunakan teknik *web scrapping*. *Cryptocurrency* merupakan aset digital dalam bentuk mata uang virtual yang dilindungi dengan sistem kriptografi. *Cryptocurrency* menjadi perbincangan hangat ditengah masyarakat dikarenakan cara kerja aset digital tersebut terbilang cukup unik dan dapat menguntungkan. *Cryptocurrency* tidak terikat dalam regulasi otoritas pusat, seperti perbankan dan dapat diperjualbelikan secara bebas di internet dengan sistem yang terdesentralisasi dengan *blockchain*. Namun disatu sisi, aset digital ini tegolong investasi dengan resiko yang tinggi. Harga pasar *cryptocurrency* sangat bergantung pada sentimen yang beredar ditengah masyarakat terutama di media sosial. Data yang digunakan pada penelitian ini akan menggunakan kata kunci yang mengandung unsur tagar #cryptocurrency, #bitcoin, #crypto, #ethereum, dan #binance pada media sosial X. Data akan mulai diambil dari periode Desember 2023 sampai dengan Januari 2024.

3.2 Metode Penelitian

Metode yang digunakan dalam penelitian ini adalah kualitatif. Metode penelitian kualitatif merupakan sebuah metode yang digunakan untuk memahami isi dan konteks dari sebuah fenomena yang dialami oleh subjek penelitian. Penelitian dengan metode kualitatif menekankan kualitas informasi, bersifat fleksibel, dan adaptif. Selain itu, metode ini juga sering menggunakan landasan teori ataupun penelitian terdahulu untuk membentuk kerangka kerja analisis data. Namun, hasil analisis yang dihasilkan dapat membuat sebuah wawasan baru dan berkontribusi dalam pengembangan teori yang sudah ada sebelumnya [72]. Penelitian ini menggunakan metode kualitatif dikarenakan peneliti ingin menghasilkan dan memberikan wawasan penting mengenai respon masyarakat terkait topik yang dibahas.

3.2.1 Metode Data Mining

Data mining merupakan proses menghimpun dan mengelola data dengan tujuan untuk mengambil informasi utama dari data. Proses ini melibatkan penggunaan perangkat lunak yang menggunakan perhitungan statistik, matematika, dan teknologi kecerdasan buatan [73]. Terdapat beberapa metode yang digunakan dalam *data mining*, seperti CRISP-DM (*Cross-Industry Standard Process for Data Mining*), SEMMA (*Sample, Emplore, Modify, Model, and Assess*), dan KDD (*Knowledge Discovery in Database Process*). Berikut adalah tabel perbandingan dari metode data mining yang telah disebutkan diatas.

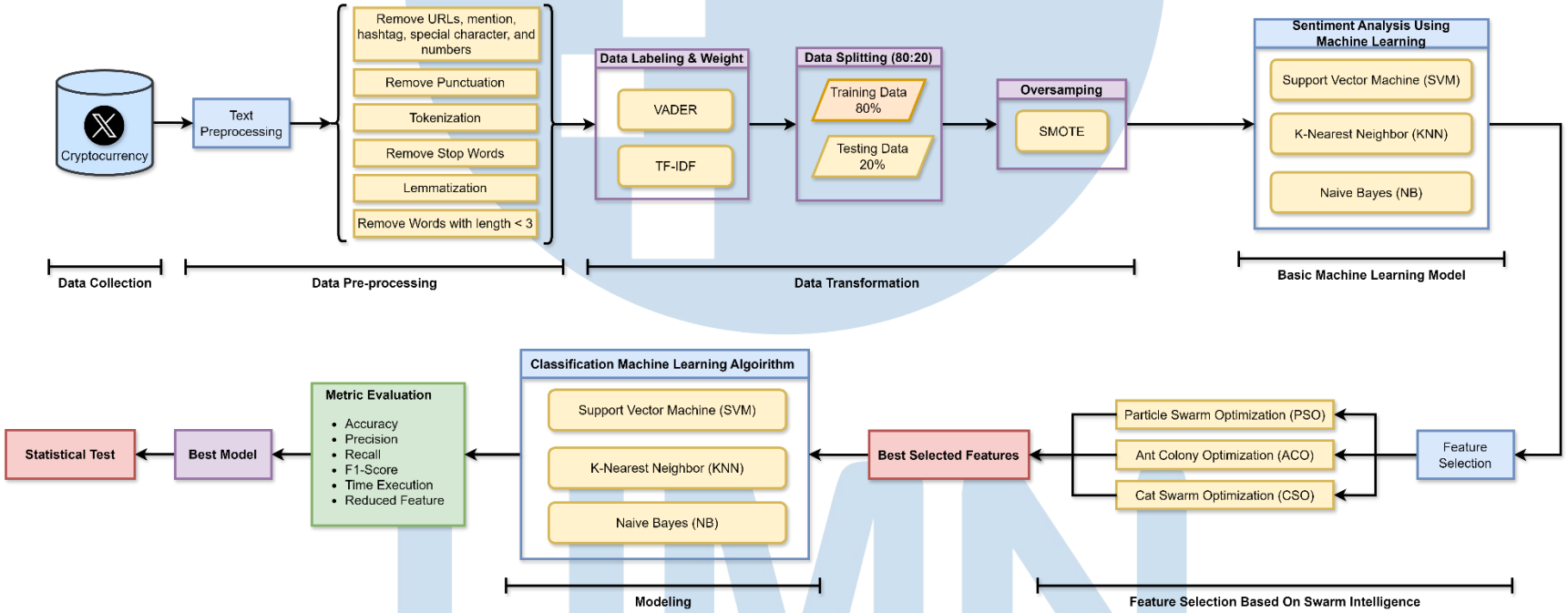
Tabel 3. 1 Perbandingan Metode *Framework Data Mining*

Indikator	KDD	CRISP-DM	SEMMA
Alur Proses	1. Data Selection 2. Preprocessing 3. Transformation 4. Data Mining 5. Evaluation	1. Business Understanding 2. Data Understanding 3. Data Preparation 4. Modeling 5. Evaluation 6. Deployment	1. Sample 2. Explore 3. Modify 4. Model 5. Assess
Kelebihan	Berfokus pada proses dan evaluasi	Struktur penelitian yang sistematis	Berfokus pada proses dan fleksibel dalam penilaian
Kekurangan	Kurang berfokus pada tahapan implementasi / *depployment* [74]	Membutuhkan waktu dan sumber data yang besar [75]	Kurang dapat menangani volume data yang besar [76]

Berdasarkan Tabel 3.1 Perbandingan Metode *Framework Data Mining*, penelitian ini menggunakan *framework* KDD. Pemilihan tersebut dikarenakan penelitian tidak melakukan tahapan *deployement* sehingga *framework* CRISP-DM dinilai kurang cocok untuk digunakan dalam penelitian ini. Penelitian ini menggunakan teknik *swarm intelligence* untuk optimasi parameter dan seleksi fitur, seringkali membutuhkan eksperimen yang iteratif sehingga *framework* SEMMA dinilai kurang cocok diterapkan. Selain itu, *sentiment analysis* membutuhkan prosedur yang komprehensif untuk mengolah data. Rangkaian

prosedur tersebut ialah *cleaning data*, *text preprocessing*, *oversampling*, *machine learning*, dan *feature selection*. Oleh karena itu, digunakan framework KDD karena dinilai lebih mendukung dibandingkan dengan *framework* CRISP-DM dan SEMMA.

3.2.2 Alur Penelitian



Gambar 3. 1 Alur Penelitian

Gambar 3.1 Alur Penelitian, merupakan rangkaian prosedur atau proses yang dijadikan sebagai landasan dasar dalam melakukan sebuah penelitian ilmiah. Acuan alur penelitian didasarkan pada *framework* KDD yang dibagi menjadi 5 tahapan, yakni:

a) Data Selection

Tahapan ini memfokuskan kepada memilih sumber data yang relevan dan memahami isi data. Sumebr data yang digunakan harus relevan dengan topik yang sedang diteliti. Topik pada penelitian ini adalah *sentiment analysis cryptocurrency* sehingga perlu mencari opini/komenttar yang berkaitan dengan hal tersebut. Kemudian, dalam memhami data, diperlukan pemahaman terhadap karakteristik dan meseleksi variabel data. Seleksi dan pemilihan data sangat krusial karena dapat mempengaruhi nilai dari akurasi model yang dibuat. Langkah pertama yang perlu dilakukan dalam *sentiment analysis* adalah *scraping data*. *Scraping data* menggunakan tool Google Colab dan dengan kata kunci "cryptocurrency", "crypto", "bitcoin", "$btc"," binance", dan "eth". Hasil scraping data menghasilkan 9884 data yang terdiri dari 4 kolom numerik, 1 kolom tanggal, dan 7 kolom string.

a) Preprocessing

Tahapan ini merupakan pra-pemrosesan untuk meningkatkan kualitas data. Proses yang dilakukan pada tahapan ini adalah *cleansing data* untuk menghilangkan *noise* dan *missing value* agar hasil analisis menjadi lebih akurat. *Preprocessing* yang dilakukan mencakup *remove urls*, *remove hashtags*, *remove special characters and numbers*, *excluding punctuation*, *convert to lower case*, *remove stop words*, *lemmatization of words*, dan *remove words with length < 3*. Setelah itu, semua kata yang sudah melalui proses *cleansing data* akan disatukan kembali dengan fungsi "*join*(*words*)". Semua langkah tersebut dilakukan agar nantinya menghasilkan data bersih dan dapat digunakan dalam pemodelan algoritma. Kemudian, dilakukan *random screening* secara manual untuk memastikan bahwa semua *tweets* sudah berfokus pada topik penelitian.

b) *Transformation*

Pada tahapan transformasi, data diubah atau dikonversi menjadi format yang sesuai dalam proses pengolahan data selanjutnya. Data yang sudah melalui tahap *preprocessing* akan dilakukan *labeling data* untuk memberikan sentimen terhadap *tweet* yang sudah dibersihkan menggunakan *library* VADER. Pengelompokan sentimen dibagi menjadi 2, yaitu positif dan negatif. Hal tersebut ditujukan untuk memudahkan dalam klasifikasi sentimen dalam masyarakat terkait opini yang tersebar di media sosial yang bersasngkutan. Selanjutnya, akan dilakukan perhitungan untuk bobot per kata menggunakan metode TF-IDF agar dapat menghasilkan nilai per kata yang optimal. Penelitian ini juga menerapkan teknik SMOTE untuk mengatasi masalah *data imbalance*. Penerapa teknik tesebut didasarkan pada teknik *oversampling* untuk mencegah *overfitting* pada data yang akan dibuat pemodelannya.

c) *Data Mining*

Tahapan ini merupakan tahap utama dalam proses KDD untuk mengekstrak pola pada data penelitian. Tahap ini berfokus pada melakukan implementasi model algoritma yang telah melalui tahapan *preprocessing* dan *transformation*. Penelitian ini akan melakukan 2 modeling yang bekerja secara terpisah, yaitu

- Algoritma *machine learning* tunggal, seperti SVM, Naïve Bayes, dan KNN
- Algoritma *machine learning* dengan optimasi parameter dan *feature selection*, seperti PSO-SVM, PSO-Naïve Bayes, PSO-KNN, ACO-SVM, ACO-Naïve Bayes, ACO-KNN, CSO-SVM, CSO-Naïve Bayes, dan CSO-KNN

Pada penelitian ini akan menghasilkan total 12 output yang dibagi menjadi 2 pemodelan utama. Hasil dari pemodelan pertama akan berfokus pada kinerja dari algoritma machine learning tunggal. Kemudian, akan dilanjutkan pembuatan pemodelan kedua untuk melihat tingkat signifikansi dari implementasi optimasi *feature selection* dan parameter

pada model *machine learning*. Tingkat kinerja model akan dilihat dari nilai *performance* yang terdiri dari *accuracy*, AUC, *time execution*, *reduced feature*, *f1-score*, dan *recall*. Dilihat dari nilai performance model yagn dibuat, akan dicari hasil terbaik dari masing-masing teknik yang telah diterapkan.

d) *Evalution*

Tahapan terakhir pada proses KDD yang bertujuan untuk mengevaluasi pola dan model agar dapat menentukan kegunaan dan validitasnya. Tahap ini juga ingin menguji model terkait memiliki nilai yang signifikat yang sesuai dengan tujuan awal penelitian. Tahap evaluasi dari *sentiment analysis* pada penelitian adalah dengan menggunakan *confusion matrix* yang meliputi *accuracy*, AUC, *time execution*, *reduced feature*, *f1-score*, dan *recall*. Kemudian, dilakukan *T-test* untuk pengujian statistik terhadap efektivitas fitur-fitur yang berbeda dalam model dan menguji perbedaan perbedaan peforma antara dua set parameter model.

3.3 Teknik Pengumpulan Data

Metode pengumpulan data yang diterapkan dalam penelitian ini adalah *web scrapping*. Data diambil dan dikumpulkan dalam bentuk *tweets* pada media sosial X. *Tools* yang digunakan dalam *web scrapping* adalah Google Colab dengan bahasa pemograman Python. *Scrapping* data dilakukan secara mandiri sehingga data yang diperoleh dapat dikategorikan sebagai data primer dengan jumlah data sebanyak 9884 baris. Scrapping data menghasilkan format data CSV dengan total 12 kolom. Berikut adalah daftar atribut data yang dapat dilihat pada Tabel 3.2 Daftar Atribut Data.

Tabel 3. 2 Daftar Atribut Data

Nama Atribut	Tipe Data	Deskripsi
created_at	Date	Tanggal pembuatan *tweet*
id_str	String	Kode unik *user* X
full_text	String	*Tweet user*
quote_count	Numeric	Total *tweet* dikutip *user* lain

reply_ count	Numeric	Total *tweet* dibalas *user* lain
retweet_count	Numeric	Total *tweet* diunggah *user* lain
favorite_count	Numeric	Total *tweet* difavoritkan *user* lain
lang	String	Bahasa *tweet*
user_id_str	String	*User* ID pemilik *tweet*
conversation_id_str	String	Kode unik percakapan
username	String	Nama *user* X
tweet_url	String	Tautan *tweet*

3.3.1 Populasi dan Sampel

Populasi adalah objek yang dipilih karena memiliki karakteristik dan jumlah tertentu yang telah ditentukan oleh peneliti untuk digali lebih dalam sehingga dapat diambil kesimpulannya berdasarkan topik penelitian yang bersangkutan. Kemudian, sampel adalah sebagian kecil dari sebuah populasi yang digunakan sebagai sumber data dalam penelitian [77]. Populasi dalam penelitian ini adalah seluruh data *tweets* yang diambil dari media sosial X dengan kata kuncinya adalah segala hal yang berhubungan dengan *cryptocurrency*. Sampel pada penelitian ini adalah data *tweets* dengan topik *cryptocurrency* yang diambil dari periode 31 Desember 2023 sampai dengan 31 Januari 2024.

3.3.2 Periode Pengambilan Data

Proses pengambilan data media sosial X melalui *tools* Jupyter Lab dengan bahasa pemograman *Python* diambil selama kurun waktu 1 bulan. Pengambilan data dimulai dari periode 31 Desember 2023 sampai dengan 31 Januari 2024. Periode pengambilan data dilakukan pada rentang waktu tersebut dikarenakan topik *cryptocurrency* menjadi perbincangan hangat dikalangan masyarakat [2]. Hasil dari pengambilan data menghasilkan *ouptut* data berbentuk CSV yang berisikan 9884 data.

3.4 Variabel Penelitian

Variabel penelitian merupakan sebuah karakteristik, atribut, atau nilai yang bervariasi pada objek yang ditetapkan oleh peneliti untuk diinvestigasi dan dianalisis untuk mendapatkan kesimpulan akhir. Variabel penelitian dibagi menjadi 2 jenis, yaitu variabel independen dan variabel dependen. Variabel independen pada penelitian ini adalah komentar pada media sosial X yang berkaitan dengan *cryptocurrency*. Dikatakan independen karena isi dari komentar tersebut berupa opini yang mengandung sentimen tentang topik yang dibahas. Variabel dependen pada penelitian ini adalah hasil labeling sentimen berdasarkan atribut data *tweets* yang digunakan pada penelitian.

3.5 Teknik Analisis Data

Teknik analisis data pada penelitian ini dibagi menjadi beberapa tahapan. Tahapan pertama adalah *web scrapping*, tahapan ini merupakan proses pengambilan data dari laman situs yang berada di internet. Pengambilan data diambil dari media sosial X dengan topik pembahasan *cryptocurrency*. Proses pengambilan data menggunakan *tools* Google Colab dan melakukan instalasi terhadap Nodejs untuk melakukan *scrapping data tweets*.

Data yang diperoleh dari hasil *scrapping* pada media sosial X merupakan data teks yang belum terstruktur sehingga perlu dilakukan strukturarisasi data teks agar akurasi yang dihasilkan dapat optimal. Tahapan kedua yang perlu dilakukan adalah *text preprocessing*. *Text preprocessing* terdiri dari beberapa tahapan, yakni *remove urls, remove hashtags, remove special characters and numbers, excluding punctuation, convert to lower case, remove stop words, lemmatization of words*, dan *remove words with length < 3*. Tujuan dari diterapkannya text preprocessing tersebut adalah untuk membersihkan data dari kata-kata yang tidak memiliki arti dan membagi kalimat menjadi kata-kata independen untuk kemudian dilakukan *labeling* dan pembobotan kata.

Tahap ketiga adalah *labeling* dan pembobotan kata, diberikan label pada setiap kata, dimana labelnya adalah positif dan negatif. Labeling menggunakan *library* VADER. *Library* VADER merupakan tool analisis sentimen berbasis *lexicon*

yang digunakan sebagai kamus dalam penilaian sentimen kata sehingga peneliti dapat mengetahui tingkat positif ataupun negatif dari suatu sentimen. TF-IDF merupakan metode yang digunakan untuk menghitung nilai dari sebuah kata atau yang biasa disebut sebagai pembobotan kata. Tujuan dari penggunaan *library* VADER dan TF-IDF adalah memberikan label positif ataupun negatif agar dapat diterapkan klasifikasi berdasarkan seleksi fiturnya.

Namun, terdapat beberapa kendala yang perlu diperhatikan sebelum melakukan implementasi *machine learning* dan *feature selection*, yakni *imbalance data*. *Imbalance data* merupakan suatu kondisi ketika jumlah data antar atribut tidak seimbang yang menyebabkan penilaian akurasi menjadi tidak maksimal. Oleh karena itu, perlu diterapkan metode *oversampling* SMOTE untuk mengatasi kendala tersebut. Selanjutnya, masuk kedalam tahapan ketiga yakni, implementasi *machine learning* tunggal untuk algoritma SVM, *Naïve Bayes*, dan KNN. Implementasi algoritma *machine learning* tunggal dilakukan sebagai bahan pembanding untuk algoritma *machine learning* dengan optimasi *feature selection* dan parameter. Proses *feature selection* digunakan untuk menentukan fitur-fitur yang relevan, terutama mengingat data dari media sosial memiliki dimensi yang tinggi. Optimasi parameter dilakukan untuk menemukan nilai paramter yang sesuai dalam implementasi model algoritma agar dapat menghasilkan akurasi yang maksimal. Penerapan optimasi *feature selection* dan parameter dilakukan menggunakan metode berbasis *swarm intelligence*, dengan menerapkan algoritma PSO, ACO, dan CSO. Hasil dari *feature selection* adalah jumlah fitur yang berhasil dikurangi. Ketiga algoritma ini kemudian akan dibandingkan berdasarkan fitur-fitur yang berhasil dikurangi dan nilai performance tertinggi untuk kemudian menghasilkan model terbaik dari kategori tersebut.