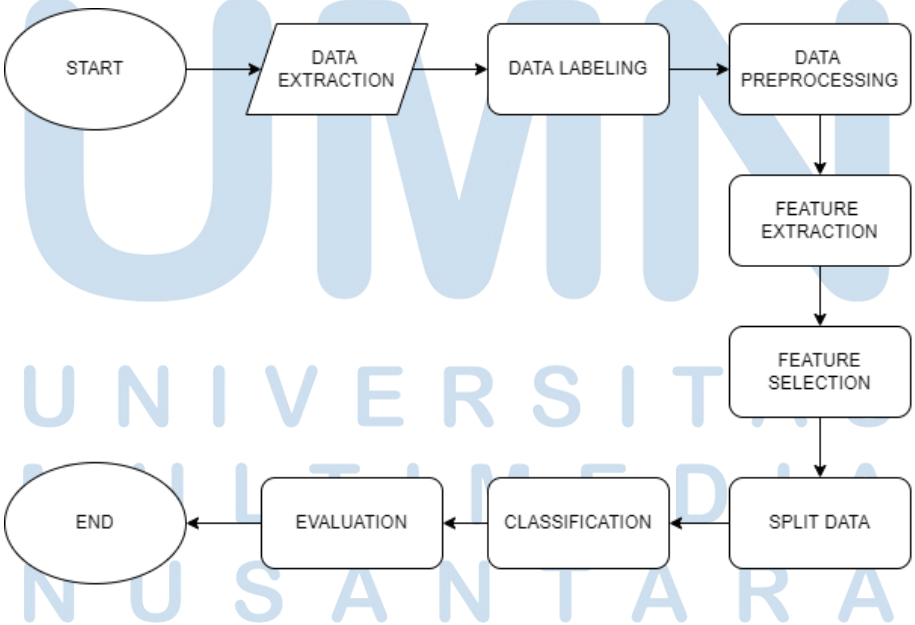
BAB III
METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Penelitian ini dilakukan untuk mengamati sentimen dari objek yang telah ditentukan, yaitu *cryptocurrency*. Pengamatan dilakukan dengan menggunakan sarana media sosial sebagai perantara untuk mengambil komentar – komentar masyarakat mengenai *cryptocurrency*. Komentar – komentar tersebut akan dikelola dan dikelompokkan menjadi sebuah sentimen analisis agar dapat menjadi bahan pertimbangan bagi *trader cryptocurrency* dalam melakukan transaksi *cryptocurrency*. Selain *cryptocurrency*, penelitian ini juga mengamati tiga koin *cryptocurrency* yaitu *Bitcoin*, *Ethereum*, dan *Ripple*.

3.2 Metode Penelitian

Pada penelitian ini metode yang akan digunakan adalah *Support Vector Machine* sebagai model klasifikasi analisis sentimen terhadap data yang telah dikumpulkan dan menggunakan *Chi-square* sebagai *feature selection* untuk mengolah data teks menjadi *numerical features*. Penelitian dilakukan dengan tahapan yang telah digambarkan pada gambar 3.1 di bawah.



Gambar 3.1 Kerangka Kerja

3.2.1 *Data Extraction*

Dataset yang akan digunakan merupakan data primer karena data yang didapatkan berasal dari narasumber langsung [32]. *Dataset* dikumpulkan dari *tweet* masyarakat yang membicarakan mengenai *cryptocurrency* dengan menggunakan teknik *Twitter crawler*. *Twitter crawler* adalah sebuah teknik yang digunakan untuk mengambil atau mengunduh data dari server *Twitter* dengan bantuan *Application Programming Integration* (*API*) *Twitter* baik berupa data *user* maupun data *tweet* [33]. *Tweet* diambil selama tiga bulan setiap minggunya pada tahun 2021, lebih tepatnya dari tanggal 1 Oktober 2021 hingga 31 Desember 2021. *Tweet* diambil selama tiga bulan agar mendapatkan data yang cukup untuk menganalisis sentimen dan pada kuartal 4 tahun 2021 terdapat beberapa ahli yang memprediksi bahwa *cryptocurrency* akan memiliki nilai paling besar sepanjang masa [26], [28], [34] - [35]. Hasil dari *Twiitter crawler* adalah sebanyak 22.327 data dengan kata kunci *cryptocurrency*, 3.269 data dengan kata kunci *Bitcoin*, 880 data dengan kata kunci *Ethereum*, 871 data dengan kata kunci *Ripple*.

3.2.2 *Data Labelling*

Data labelling merupakan tahap dimana dilakukan pelabelan terhadap *tweets* yang telah didapatkan. Pelabelan dilakukan dengan menggunakan metode manual. Metode tersebut dilakukan dengan cara membersihkan dan merapikan *tweet* lalu ditentukan nilai dari *tweet* tersebut, apakah bernilai negatif dalam artian mengandung kata yang berunsur hujatan, hinaan atau berkata kasar dan yang bernilai positif mengandung kata yang sebaliknya.

3.2.3 *Data Preprocessing*

Pada *data preprocessing* dilakukan beberapa tahap untuk membersihkan *tweet* dari berbagai *noise*. Tahapan – tahapan yang

akan dilakukan yaitu c*ase folding* untuk mentransformasi *uppercase* menjadi *lowercase*, d*ata cleaning* untuk menghapus karakter – karakter lain selain alfabet, *tokenization* untuk memecah teks menjadi satuan kata penyusunnya, *stopword removal* untuk menghapus token yang sering muncul dalam data, dan *stemming* untuk mengubah kata berimbuhan menjadi kata dasarnya. Bagian 2.4 telah menjelaskan lebih lanjut mengenai *data preprocessing*.

3.2.4 *Feature Extraction*

Tabel 3.0.1 Perbandingan *Feature Extraction*

Jurnal	**TF-IDF**	***Count Vectorizer***
"*Vectorizer Comparison for Sentiment Analysis on Social Media Youtube: A Case Study*" [36]	97,5%	97,3%
"*Sentiment Analysis on COVID Tweets: An Experimental Analysis on the Impact of Count Vectorizer and TF-IDF on Sentiment Predictions using Deep Learning Models*" [37]	93,15%	93,07%
"*Comparison of Various Machine Learning Models for Accurate Detection of Fake News*" [38]	92,8%	89,1%

Feature extraction adalah proses untuk mentransformasi data teks menjadi *numerical features* yang dapat digunakan dalam model pembelajaran (*learning model*) [39]. Untuk menentukan metode *feature extraction* mana yang akan digunakan dilakukan perbandingan pada tabel 3.1. Pada perbandingan tersebut dapat dilihat bahwa metode TF-IDF memiliki akurasi yang lebih tinggi daripada *Count Vectorizer*. Oleh karena itu pada penelitian ini akan digunakan metode TF-IDF untuk melakukan *feature extraction*.

3.2.5 Feature Selection

Tabel 3.0.2 Perbandingan Metode *Feature Selection*

Feature Selection Method	***Filter***	***Wrapper***	***Embedded***
Technique	*Statistical measures*	*Optimization algorithm*	*Combination of filter and wrapper method*
Computational efficiency	*Efficient*	*Inefficient*	*Inefficient*
Computation time	*Time efficient*	*Slow*	*Slow*
Computational cost	*Cheaper*	*Expensive*	*Expensive*
Computational space	*Less computational space*	*More computational space*	*More computational space*
Complexity	*Low*	*High*	*High*
Generality	*High*	*Less*	*Less*
For high dimensional data	*Suitable*	*Not suitable*	*Not suitable*

Sumber : [40]

Feature selection merupakan sebuah metode untuk menyeleksi data yang tidak relevan, memiliki *noise*, dan memilih *subset* yang dapat menjadi representatif dari semua data untuk meminimalisir kompleksitas klasifikasi [29]. *Feature selection* memiliki beberapa metode yaitu *filter*, *wrapper*, dan *embedded*.

Filter merupakan metode dimana metode tersebut bekerja dengan memberikan rangking pada *subset selection*. Pemberian rangking didasarkan pada evaluasi fungsi seperti jarak, informasi, ketergantungan, dan konsistensi terhadap klasifikasi [12]. Metode *Wrapper* merupakan metode *feedback*, dimana metode ini menggabungkan algoritma *machine learning* dalam proses pemilihan *feature*. Metode ini mengandalkan kinerja pengklasifikasi untuk mengevaluasi kualitas serangkaian *feature* [12]. *Embedded* merupakan metode yang menggabungkan kualitas metode *filter* dan *wrapper*. Metode *embedded* pada umumnya diimplementasikan oleh algoritma yang memiliki metode *feature selection* sendiri. Dengan membandingkan berbagai bidang seperti teknik, efisiensi, waktu, dan lain – lain seperti pada tabel 3.2, maka dipilih *Chi-square* sebagai feature selection yang akan digunakan pada penelitian ini. *Chi-square* merupakan salah satu metode *filter* yang memiliki tingkat kesesuaian yang paling baik dibandingkan dengan metode lainnya [12].

3.2.6 *Split Data*

Setelah dilakukan *data preprocessing* maka data akan di *split* menjadi dua (2) yaitu data latih (*data training*) dan data tes (*data testing*). *Data training* adalah data yang digunakan untuk melatih model. Sedangkan *data testing* adalah data yang digunakan untuk mengetes performa model. Pada penelitian ini proporsi *split data* yang akan digunakan adalah 80% untuk *training* dan 20% untuk *testing*. Proporsi ini digunakan karena merupakan metode yang popular digunakan untuk melakukan *split data* [25], [18], [36].

3.2.7 *Classification*

Tabel 3.0.3 Perbandingan Algoritma Analisis Sentimen

Algorithm	***Naïve Bayes***	***SVM***	***Maximum Entropy***	***Random Forest***
Understanding complexity	*Very less*	*High*	*Moderate*	*Moderate*
Theoretical accuracy	*Low*	*High*	*Moderate*	*High*
Theoretical training speed	*High*	*High*	*Moderate*	*Low*
Performance with small number of observatins	*High*	*Low*	*Low*	*Low*
Classifier	*Probabilistic*	*Linear*	*Probabilistic*	*Tree based*

Sumber : [10]

Pada tahap ini dilakukan klasifikasi dengan menggunakan algoritma atau metode tertentu yang telah ditentukan. Pada penelitian ini, telah dipilih algoritma *Support Vector Machine (SVM)* untuk melakukan klasifikasi terhadap *cryptocurrency*.

Dalam pemilihan algoritma yang akan digunakan dalam menganalisis sentimen maka dibuat tabel komparasi terhadap algoritma – algoritma yang telah dipilih. Dengan perbandingan dari masing – masing algoritma yang menunjukkan kompabilitas dari segi volume data, kemampuan klasifikasi, dan parameter penggunaan, diputuskan untuk menggunakan *SVM*. Selain itu *SVM* juga dipilih berdasarkan tingkat pemahaman kompleksitas yang tinggi dan tingkat akurasi yang tinggi berdasarkan jumlah *data training* yang dimiliki dalam penelitian ini. Pada Tabel 3.3 dapat

dilihat perbandingan antara beberapa algoritma yang telah diteliti untuk melakukan analisis sentimen:

3.2.8 *Evaluation*

Tahap *evaluation* merupakan tahapan dimana hasil pemodelan dinilai berdasarkan kriteria keberhasilan tertentu. *Evaluation* memiliki peran sebagai parameter untuk mengukur tingkat keberhasilan atau keoptimalan dari pembangunan model. Dalam melakukan *evaluation*, digunakan *confusion matrix* untuk melihat nilai dari *accuracy*, *f-measure*, *recall*, dan *precision*. *Confusion matrix* dipilih sebagai metode evaluasi karena metode tersebut merupakan metode yang popular digunakan untuk melakukan validasi *learning model* [11], [28], [27].

3.3 Variabel Penelitian

Variabel penelitian dapat membedakan atau membawa variasi pada suatu nilai tertentu. Pada penelitian variabel dibagi menjadi dua (2) yaitu:

3.3.1 Variabel Independen

Variabel independen adalah variabel yang menjadi sebab terjadinya atau terpengaruhnya variabel terikat [41]. Dalam penelitian ini yang menjadi variabel independen adalah *tweets* yang berasal dari *Twitter* yang berisikan opini dengan kata kunci yang berhubungan dengan *cryptocurrency*, *Bitcoin*, *Ethereum*, dan *Ripple*.

3.3.2 Variabel Dependen

Variabel dependen adalah variabel terikat yang dipengaruhi karena adanya variabel bebas [41]. Dalam penelitian ini yang menjadi variabel dependen adalah *labels* berisi hasil klasifikasi opini dari *tweet* yang terbagi menjadi dua kategori yaitu positif dan negatif.

3.4 Teknik Pengumpulan Data

Data yang dikumpulkan pada penelitian ini merupakan hasil pengumpulan *tweet* yang berasal dari media sosial *Twitter* dengan menggunakan teknik *Twitter crawler*. Pengumpulan data dilakukan selama 90 hari, yaitu sejak tanggal 1 Oktober 2021 hingga 31 Desember 2021. Data diambil dalam jangka waktu tiga bulan agar dapat membantu dalam pembelajaran model [42]. Dalam melakukan *Twitter crawler* digunakan beberapa parameter yang dapat membantu dalam menyempitkan pencarian data yang dibutuhkan. Parameter yang digunakan merupakan kumpulan *hashtag* dan *keywords* umum yang berhubungan dengan *cryptocurrency* seperti yang dapat dilihat pada Tabel 3.4. Proses pengumpulan data menggunakan bahasa pemrograman *Python* dan *library tweepy* dengan *output* berupa *file .csv*. Dengan menggunakan *library tweepy* maka data dapat dikumpulkan dengan menggunakan OAtuh 1 sebagai akses [43].

Tabel 3.0.4 *Hastag* dan *Keywords*

Hastag (#) dan k*eywords* (kata kunci) yang digunakan
cryptocurrency
cryptocurrencies
bitcoin
BTC
ethereum
ETH
ripple
XRP

3.5 Teknik Pengambilan Sampel

Pengambilan sampel dilakukan ketika penyeleksian hasil *Twitter crawler* berhasil dilakukan dengan menghasilkan nilai sentimen berupa positif atau negatif. Dari 22.327 *tweet* yang berasal dari kata kunci *cryptocurrency*, data yang digunakan untuk proses analisis sentimen adalah sebanyak 2.482 *tweet*. Dari kata kunci *Bitcoin* didapatkan 3.269 *tweet* dan 1.559 *tweet* akan digunakan untuk proses analisis sentimen. Sedangkan *Ethereum* menghasilkan 880 *tweet* dan 343 *tweet* digunakan dalam proses analisis sentimen. Dan untuk kata kunci *Ripple* didapatkan *tweet* sebanyak 871 dan digunakan sebanyak 379 *tweet* untuk analisis sentimen.

3.6 Teknik Analisis Data

Pada penelitian ini teknik analisis data yang digunakan dalam penelitian ini adalah kualitatif karena berfokus untuk menganalisis opini yang dimiliki oleh masyarakat mengenai *cryptocurrency*. Analisis data dilakukan dengan menggunakan bahasa pemrograman *Python* dan model analisis sentimen dengan algoritma *Support Vector Machine* dengan menambahkan *Chi-square* sebagai *feature selection*. Pengukuran tingkat kesuksesan penelitian dapat diukur dengan menggunakan c*onfusion matrix.* Pada bagian 2.8 telah dijelaskan lebih lanjut mengenai *confusion matrix*.