BAB III

METODOLOGI PENELITIAN

3.1. Gambaran Umum Objek Penelitian

Objek dari penelitian ini adalah mengumpulkan opini masyarakat yang didapatkan dari Tweet mengenai pelaksanaan vaksinasi COVID-19 yang berlangsung di Indonesia. Opini yang dicari tersebut adalah komentar masyarakat baik yang bersifat positif, negatif maupun netral terkait pelaksanaan vaksinasi COVID-19 baik yang ditujukan terhadap kinerja pemerintah dan kepercayaan masyarakat terhadap vaksin yang digunakan. Sentimen masyarakat dikatakan berperan penting dalam melakukan suatu prediksi dimana salah satu sumber data untuk membaca sentimen dari masyarakat adalah media sosial [27].

Tweet yang diambil menggunakan bantuan kata kunci dan hashtag yang disebutkan pada sub bab 3.3.1 dibawah

3.2. Metode Penelitian

3.2.1. Pengumpulan Data

Penelitian ini menggunakan data *tweet* dan harga saham, dimana untuk *tweet* akan dikumpulkan menggunakan *library Tweepy* pada bahasa pemrograman *Python*. Data *tweet* dikumpulkan dalam periode 1 minggu dengan menggunakan kata kunci yang berhubungan dengan vaksin. Dalam penelitian ini, vaksin yang akan dijadikan sebagai objek adalah beberapa vaksin yang beredar di Indonesia, maka dari itu tag

yang akan digunakan sebagai parameter pengumpulan *tweet* adalah kata kunci yang berhubungan dengan vaksin.

3.2.2. Variabel Independen

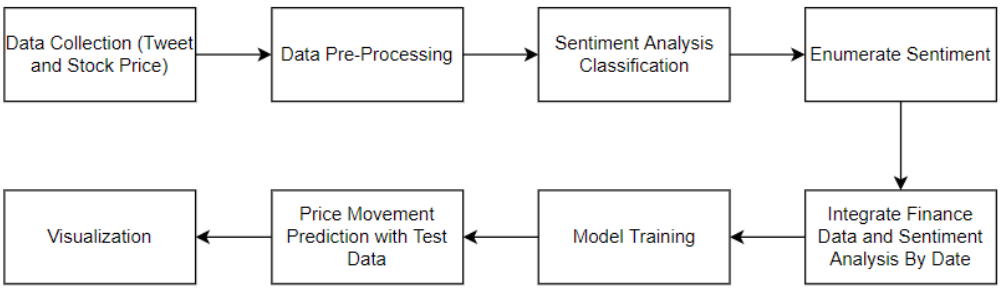
Dalam penelitian ini, objek yang dijadikan sebagai variabel independen adalah *tweet* dari twitter yang berisikan opini dengan kata kunci yang berhubungan dengan vaksin yang beredar di Indonesia dari masyarakat tentang perusahaan vaksin tersebut.

3.2.3. Variabel Dependen

Dalam penelitian ini, variabel penelitian yang dijadikan sebagai variabel dependen adalah jenis klasifikasi opini dari *tweet* yang terbagi menjadi 3 kategori yaitu positif, negatif dan netral.

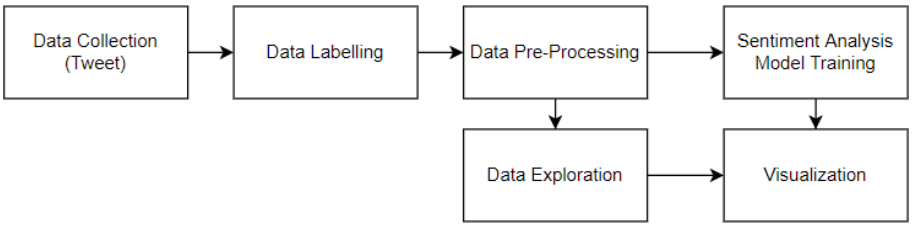
3.3. Alur Penelitian

Dalam penelitian ini, kerangka pikir yang digunakan hampir serupa dengan kerangka pikir penelitian yang juga berhubungan dengan *sentiment analysis*, yaitu "*Sentiment analysis of Twitter data for predicting stock market movements*" [28]. Berikut *flowchart* dari penelitian terdahulu:



Gambar 3.1 Alur Penelitian Sebelumnya [29]

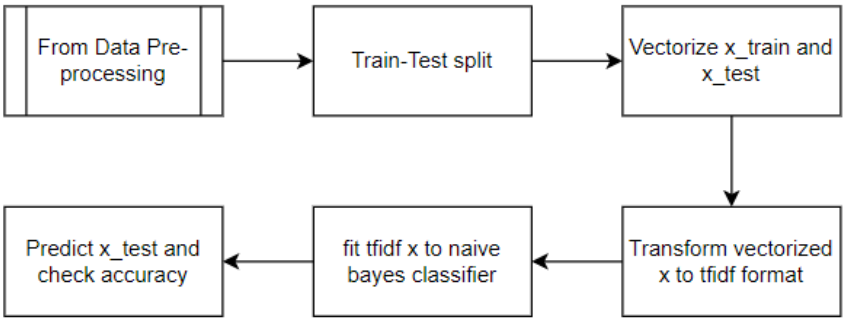
Berikut sedikit modifikasi dari kerangka pikir untuk disesuaikan dengan kebutuhan penelitian karena pada penelitian sebelumnya diagram kerangka pikir yang digunakan untuk melakukan sentimen analisis yang kemudian dilakukan untuk memprediksi pergerakan saham, namun dalam penelitian ini akan melihat apakah model *machine learning* dapat digunakan untuk melakukan sentimen analisis dalam bahasa Indonesia. Berikut kerangka pikir untuk untuk penelitian secara keseluruhan:



Gambar 3.2 Alur Penelitian Modifikasi

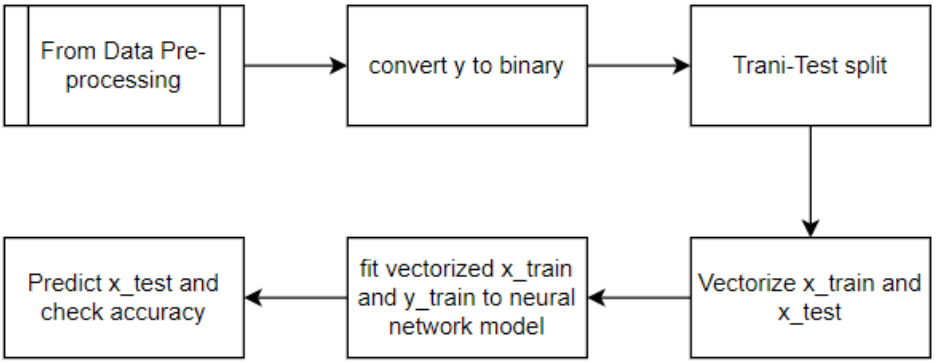
Dikarenakan dalam penelitian ini akan membuat 2 model yang berbeda, maka tahap saat akan melakukan training model klasifikasi dan *testing* model klasifikasi akan sedikit berbeda juga karena input untuk *training* kedua model

berbeda. Berikut alur untuk melakukan *training* dan *testing* algoritma Naïve Bayes:



Gambar 3.3 Alur Training Model Naïve Bayes

Berikut alur untuk melakukan *training* dan *testing* algoritma *Neural Network*:



Gambar 3.4 Alur Training Model *Neural Network*

3.3.1. Data Collection

Pada langkah *data collection*, terdapat tahap-tahap sebagai berikut:

1. Pembuatan akun *Twitter* dan *Twitter developer* pada halaman *website* developer.twitter.com. Akun ini dibuat dengan tujuan mendapatkan akses berupa *authentication key* ke *Application Programming Interface* (API) dari Twitter untuk menarik data *tweet* yang ada pada Twitter. Data yang ditarik akan berjumlah kurang lebih sebanyak 10.000 baris data dimana pada penelitian lain yang ditemukan baris data yang digunakan 4.400 baris data [30].
2. Data *tweet* yang akan diambil dari Twitter menggunakan bahasa pemrograman *Python* dengan *library Tweepy* untuk mengakses API dari Twitter.
3. Pada pembuatan *script* untuk mengambil data dari *Twitter*, akan menggunakan kata kunci dan dan *hashtag* yang memiliki kemungkinan hubungan dengan vaksinasi di Indonesia. Berikut kata kunci dan *hashtag* yang digunakan untuk penarikan *tweet*:
    - #VaksinUntukKita
    - #Vaksin
    - Vaksin.
    - Sinovac
    - CoronaVac

- NovaVax
- AstraZeneca
- Pfizer
- Covax
- Vaksin Merah Putih
- Vaksin Nusantara
- #VaksinIndonesia

Kata kunci yang telah disebutkan diatas menjadi alasan dijadikan search word untuk penarikan tweet karena merupakan nama-nama vaksin yang beredar dan digunakan di Indonesia, selain itu beberapa kata seperti Vaksin yang merupakan kata dasar dalam bahasa Indonesia juga hashtag yang pada saat tweet akan diambil merupakan 3 teratas hashtag di twitter.

3.3.2. Data Labelling

Data-data *tweet* memiliki *sentiment* positif, negatif dan netral. Untuk keperluan pengujian akurasi, maka data *tweet* tersebut perlu diberikan label agar pada saat pengujian model yang telah dibuat, dapat diketahui performa dari masing-masing model.

3.3.3. Data Preprocessing

Data *Preprocessing* adalah tahap yang perlu untuk dilakukan untuk menghilangkan *noise* atau *outlier* yang ada pada data [28]. Tujuan dari menghilangkan *noise* tersebut untuk menstandarisasi antara satu data dengan data yang lain. Pada tahap ini akan dilakukan beberapa hal kepada data *tweet*, yaitu:

3.3.3.1. Remove Duplicate Data

Pada saat pengambilan data, untuk menghindari terdapat data yang duplikasi maka jika terdapat data yang duplikat maka data tersebut akan disisakan, sementara yang lain akan dihapus.

3.3.3.2. Case Folding

Didalam data *tweet* yang telah diambil, terdapat huruf kapital dan huruf kecil, maka untuk menyamakan bentuk dari setiapdkata maka semua kata akan diubah menjadi huruf kecil.

3.3.3.3. Filtering

Tahap *filtering* ini bertujuan untuk mengurangi *noise* dari *tweet* dengan cara menghapus kata-kata atau atribut yang mungkin dapat membuat pembangunan model kurang maksimal. Berikit kata-kata yang akan dihilangkan:

- *Username*
- *URL*
- *Query* (kata kunci saat mencari tweet)

3.3.3.4. Removing Stopwords

Pada tahap ini, akan menghapus beberapa kata stopwords seperti “hahaha”, “lol”, “wkwkwk” untuk mengurangi *noise* dan kata-kata yang tidak begitu signifikan untuk dijadikan sebagai data *training* atau pun sebagai data *test*.

3.3.3.5. Tokenize

Pada tahap ini, data *tweet* dari twitter tadi akan dilakukan *tokenize* dimana *tweet* yang sudah dipotong menjadi per kata agar lebih mudah untuk diolah.

3.3.3.6. Stemmer

Dalam *tweet*, terdapat kata-kata yang bentuknya bukanlah kata dasar melainkan kata yang memiliki imbuhan. *Stemmer* berguna untuk memotong kata berimbuhan tersebut menjadi kata dasar. Untuk melakukan ini akan menggunakan *library* Sastrawi yaitu *stemmer text* untuk bahasa Indonesia.

3.3.4. Sentiment Analysis Classification

Pada tahap ini, dilakukan *sentiment analysis* terhadap *tweet* untuk mengetahui *tweet* termasuk kedalam kategori *sentiment* yang positif, negatif atau netral.

3.3.5. Model Training

Data yang sudah diolah pada tahap *pre-processing* akan dilakukan *fitting* kedalam model yang akan digunakan pada penelitian ini, yaitu Naïve Bayes dan *Neural Network*. Berikut perbedaan dari masing-masing model:

Tabel 3.1 Perbandingan Model

Kategori	**Naïve Bayes**	***Neural Network***
Tipe Data Algoritma	*Supervised Learning*	*Supervised/Unsupervised Learning*
Cara memperlakukan data	Semua variabel dalam dataset diperlakukan dan dianggap sebagai variabel independen	Data dianalisa pattern dan dicocokkan dengan label
Tipe Algoritma	Probabilitas	*Pattern learning*
Kelas Algoritma	*Generative Model* untuk klasifikasi	Model dengan neuron

Tabel 3.1 menjelaskan perbedaan dan karakteristik dari 2 model machine learning yang akan digunakan dalam penelitian ini. Dari tipe data untuk melatih model dimana naïve bayes membutuhkan data yang sudah diberikan label sementara *Neural Network* tidak harus memiliki

label untuk melakukan training. Selain dari tipe data, cara masing-masing model memperlakukan data juga berbeda dimana naïve bayes akan menganggap semua variabel berdiri sendiri atau independent sementara *Neural Network* akan mencoba untuk mencari pola dari data. Tipe algoritma dari naïve bayes dan *Neural Network* juga berbeda, naïve bayes menggunakan algoritma probabilitas sementara *Neural Network* merupakan pattern learning yaitu mempelajari karakteristik dan pola dari masing-masing baris data yang digunakan untuk melatih model. Kelas algoritma keduanya juga berbeda, dimana naïve bayes merupakan generative model, sementara karena *Neural Network* merupakan model dengan pattern learning maka jumlah neuron dapat mempengaruhi kemampuan algoritma *Neural Network* dalam mempelajari data yang dimasukkan.

3.3.6. Pengujian Akurasi Model

Setiap model akan diuji akurasinya dengan memprediksi apakah label dari *row* data pada data test. Akurasi dapat diketahui dengan menggunakan *confusion matrix* yang merepresentasikan hasil akurasi dari model. Dengan *confusion matrix*, akurasi didapatkan dari menjumlahkan *true positives* dan *true negatives* yang kemudian dibagi dengan total data pada data *test*.

	Actual Positive	Actual Negative	Actual Neutral
Predicted Positive	True Positive	False Positive	False Positive
Predicted Negative	False Negative	True Negative	False Negative
Predicted Neutral	False Neutral	False Neutral	True Neutral

Gambar 3.5 Ilustrasi Confusion Matrix

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{True\ Positive + True\ Negative + True\ Neutral}}{\boldsymbol{n}}$$

Rumus 3.1 Rumus Confusion Matrix

3.3.7. Visualisasi

Pada tahap ini, dilakukan visualisasi untuk menampilkan performa dari masing-masing model dalam melatih data dengan *random state* saat pemisahan *training* dan *testing data* dalam melakukan klasifikasi sentimen. Selain menampilkan performa masing-masing model juga memperlihatkan ekplorasi data untuk memperlihatkan kata apa yang paling sering muncul dari masing-masing label *positive*, *negative* dan netral. Visualisasi tersebut ditampilkan dalam bentuk grafik menggunakan *library* matplotlib dari Python.