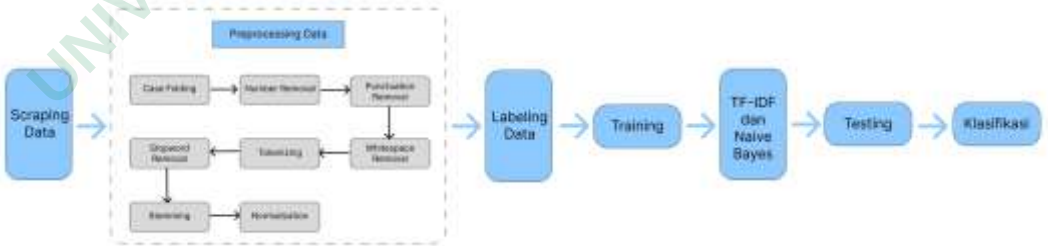
BAB 3
METODE PENELITIAN

3.1 METODE PENELITIAN

Dalam penelitian ini menggunakan Metode *Naïve Bayes Classification* untuk mendapatkan nilai prediksi dari data tweet yang sudah dilabeli kemudian hasil dari klasifikasi *Naïve Bayes* dianalisis nilai presisi dari hasil model klasifikasi yang dibuat. Penelitian ini merupakan sebuah penelitian analisis sentiment positif dan negatif. Penelitian ini menggunakan data *tweet* dan *re-tweet* yang didapatkan dari *Twitter* dengan menggunakan keyword Eril Ridwan Kamil pada tanggal 1 Juni – 14 Juni 2022 dengan data tweet sebanyak 10.000, kemudian Pemilu Ridwan Kamil, dan Elektabilitas Ridwan Kamil pada tanggal 01 Januari – 14 Juni 2022 dengan jumlah data 1016 data *tweet* menggunakan teknik *scraping*. Selanjutnya melakukan *preprocessing* data untuk mendapatkan hasil yang diinginkan. Nantinya data tersebut digunakan untuk memetakan informasi atau sentiment dari *netizen* di *Twitter* mengenai Meninggalnya Almarhum Emmeril Kahn dan dari sentiment tersebut apakah mempunyai pengaruh terhadap elektabilitas Pemilu Ridwan Kamil sehingga didapatkan informasi yang lebih jelas.

Penelitian ini dilakukan karena permasalahan masalah yang ada, sehingga didapatkan hasil analisis yang sesuai apa yang dibutuhkan, berikut merupakan alir diagram Metoda Penelitian dapat dilihat pada Gambar 3.1.



Gambar 3.1 Alur Diagram Metode Penelitian

Berdasarkan Gambar 3.1, Penelitian dimulai dengan melakukan pengambilan data dari *Twitter* menggunakan *Scraping Data* menggunakan tools Snscrape, kemudian data tersebut dilabelkan secara manual dan dimuat kedalam file csv agar

mempermudah dalam melakukan *Preprocessing Data*. Setelah melakukan *Preprocessing Data* kemudian melakukan *Labelng data* terhadap 1000 data tweet, kemudian data tersebut diuji kedalam Training dan Testing data menggunakan metode *Naïve Bayes Classification* dan *TF-IDF* supaya mendapatkan sebuah klasifikasi sentiment untuk dijadikan prediksi. setelah itu melihat hasil dan tingkat keakuratan pemodelan yang telah dibuat.

3.2 BAHAN PENELITIAN

Bahan penelitian yang akan dibutuhkan adalah data *tweet*, *re-tweet* maupun komentar oleh netizen di *Twitter* yang berkaitan dengan Meninggalnya Almarhum Emmeril Kahn. Data yang diambil merupakan data *tweet* dengan keyword "Eril Ridwan Kamil" dari tanggal 01 Juni – 15 Juni 2022 dengan jumlah 10.000 tweet, "Pemilu Ridwan Kamil" dan "Elektabilitas Ridwan Kamil" pada tanggal 01 Januari – 15 Juni 2022 dengan jumlah 1016 tweet, data yang diambil pada tanggal tersebut akan digunakan agar dapat melihat opini-opini masyarakat mengenai Almarhum Eril dan Elektabilitas Pemilu Ridwan Kamil.

3.3 ALAT PENELITIAN

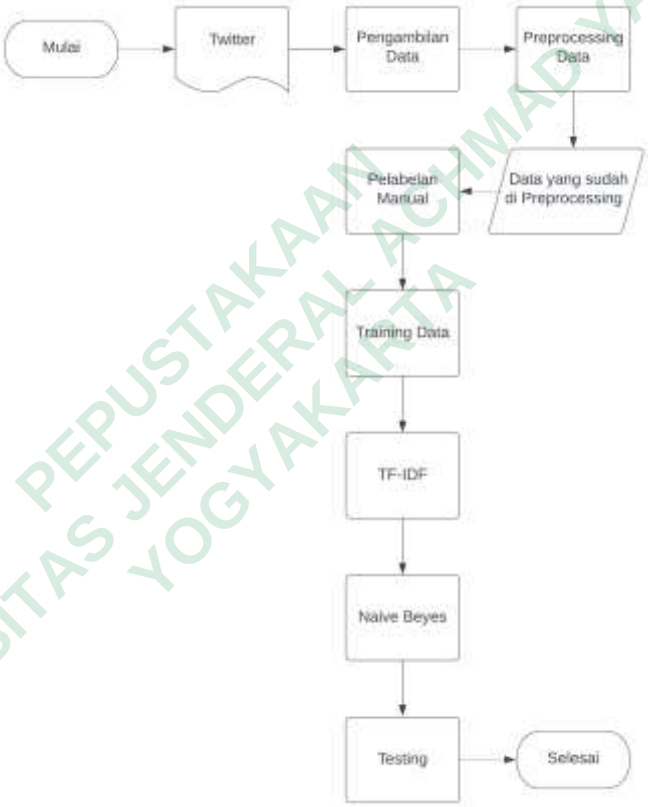
Alat yang digunakan dalam penelitian ini merupakan komputer dengan spesifikasi cukup untuk menjalankan proses pengolahan data serta koneksitas *Internet*. Sistem Operasi dan program-program aplikasi yang dipergunakan dalam dalam pengembangan aplikasi ini adalah:

1. Sistem Operasi: *Windows 10 64-bit*
2. *Microsoft Office Excel 2013*
3. Bahasa Pemrograman *Python 3.9.13*
4. *Jupyter Lab*

3.4 JALAN PENELITIAN

Penelitian ini menggunakan Bahasa pemrograman *Python* dan aplikasi *Jupyter Lab* untuk melakukan pengambilan data dari *Twitter* yang akan ditampilkan di *Microsoft Office Excel* dan dimodelkan dengan bantuan *library* yang sudah tersedia pada Bahasa pemrograman *Python*.

Berikut merupakan tahapan-tahapan yang akan digunakan dalam penelitian ini, yaitu :



Gambar 3.2 Flowchart Jalan Penelitian

Berdasarkan pada gambar 3.2 berikut merupakan tahapan-tahapan yang akan digunakan dalam penelitian ini, yaitu :

3.4.1 Pengambilan data

Dalam tahap ini dilakukan untuk mengambil data dari *Twitter* supaya mendapatkan data *tweet* maupun *re-tweet* mengenai Meninggalnya Almarhum Emmeril Kahn dengan keyword pengambilan data “Eril Ridwan Kamil” diambil

pada tanggal 3 Juni – 14 Juni 2022 dengan jumlah data total 10.000, kemudian melakukan pengambilan data mengenai pemilu dan elektabilitas Ridwan Kamil dengan keyword “Pemilu Rdwan Kamil” dan “Elektabilitas Ridwan Kamil” dari tanggal 1 Januari – 14 Juni 2022 sehingga didapatkan data sebanyak 1016 data *tweet*, jika digabungkan semua data menjadi 11016 data tweet. Menggunakan *Jupyter Notebook* dan menggunakan *scraping tools* yang bernama *Snscrape* dengan modulnya yaitu *snscrape.module.twitter* untuk mengambil data dari Twitter dengan menggunakan beberapa atribut untuk memanggil data, selanjutnya data akan ditampilkan pada *Microsoft Office Excel*.

3.4.2 Preprocessing data

Preprocessing data adalah proses pengolahan data teks yang sudah tersedia dengan melakukan tahapan-tahapan untuk memperbaiki data teks yang belum rapih dan masih belum diperbaiki. Pada tahapan preprocessing ini dibutuhkan beberapa library untuk membantu proses preprocessing, seperti library pandas untuk memanipulasi data, library nltk untuk membantu pengolahan natural language, dan library sastrawi untuk mengurangi infleksi kata dalam bahasa indonesia. Berikut merupakan tahapan dalam *Preprocessing data* :

1. Casefolding

Casefolding merupakan tahapan untuk melakukan konversi teks dari bentuk awal menjadi bentuk standar huruf kecil atau *lowercase*.

2. Cleaning data

Cleaning data, berfungsi untuk melakukan pembersihan kata yang tidak penting dalam melakukan klasifikasi nanti dengan membersihkan atau menghilangkan simbol, angka, dan kata-kata yang berada pada *tweet* dan *re-tweet* yang tidak diperlukan.

3. Tokenizing

Pada tahap *Tokenizing* bertujuan untuk melakukan pemecahan atau pemisahan karakter dalam suatu teks yang didefinisikan sebagai pemisah kata atau bukan dan nantinya akan mempermudah melakukan Stopword Removal.

4. Stopword removal

Stopword removal yaitu melakukan penghapusan kata-kata yang tidak memiliki informasi contohnya seperti (“yang”,”dan”,”di”,“dengan”,”sementara” dll).

5. Stemming

Stemming yaitu melakukan penghilangan infleksi kata menjadi bentuk dasarnya, misalkan kata “menjadi” dan dilakukan *stemming* akan menghasilkan kata “jadi”.

6. Normalization

Normalization yaitu mengubah kata yang sebelumnya merupakan singkatan dan kata tidak baku diubah menjadi kata baku. Dalam melakukan normalisasi terdapat beberapa kata yang harus diubah, contoh pada Gambar 3.11 merupakan kata yang dinormalisasi

gk	tidak
ga	tidak
gua	saya
gue	saya
pake	pakai
bgt	banget
liat	lihat
utk	untuk

Gambar 3.3 Contoh Kata Normalisasi

Setelah melakukan proses preprocessing data disimpan kedalam bentuk file Excel yang nantinya akan dilakukan pelabelan data secar manual

3.4.3 Pelabelan Data

Tahap selanjutnya adalah pelabelan data dengan memberikan nilai sentiment terhadap data tweet yang sudah diambil melalui proses scraping data, sehingga dapat dilakukan analisis lebih lanjut mengenai sifat sentiment yang positif, netral atau negatif. Pelabelan ini dilakukan secara manual dengan menggunakan 450 data dan nantinya akan didapatkan nilai masing-masing sentiment 150 data. Pelabelan data juga dibantu menggukana library dari *Python* yaitu *VaderSentiment* dengan menggunakan *VaderSentiment* bisa mendapatkan score polaritas dari sebuah dokumen, sehingga memungkinkan untuk mengetahui sentiment sebuah dokumen apakah Positif, Netral atau Negatif.

3.4.4 Training Data

Tahap *Training Data* merupakan proses latih (*training)* pada data dengan menggunakan metode *Naïve Bayes Classification.* Tahapan dalam proses *training* diawali dengan ekstrasi pada data teks menggunakan TF-IDF(*Term Frequency-Inverse Document Frequency*) untuk mengitung bobot pada setiap kata dan memudahkan proses Naïve Bayes dalam melakukan prediksi pada setiap kata, kemudian dilakukan proses training data untuk membuat model klasifikasi sentiment.

1. TF-IDF

Melakukan perhitungan TF-IDF dilakukan untuk mendapatkan bobot dari setiap kata yang terdapat dalam sebuah dokumen yang hasilnya digunakan dalam metode *Naïve Bayes Classification* untuk mendapatkan hasil klasifikasi.

Klasifikasi yang dilakukan pada penelitian ini menggunakan fitur ekstrasi TF-IDF yang melakukan perhitungan secara otomatis dan menghasilkan pembobotan pada kata yang berada dalam dokumen pada data training. Perhitungan TF-IDF ini menggunakan library pada Python yaitu Sklearn dan TfidfVectorizer untuk melakukan perhitungan TF-IDF secara otomatis.

Dengan menggunakan fitur ekstrasi Tfidfvectorizer membuat fungsi untuk menjalankan TfidfVectorizer dengan nama feature_extraction untuk mengkonversi dokumen menjadi data metrix fitur TF-IDF kemudian menghitung bobot dokumen

keseluruhan dari sebuah kata, kemudian menyesuaikan data yang diambil dan mengkonversi data tersebut menggunakan fungsi ‘fit.transform’ untuk mendapatkan hasil data metrix.

2. Naïve Bayes Classification

Menghitung nilai probabilitas dengan menggunakan bobot kata yang sudah dihitung menggunakan TF-IDF. Kemudian menggunakan library Sklearn untuk import modul MultinomialNB untuk melakukan perhitungan Naïve Bayes secara otomatis. Menggunakan MultinomialNB untuk menghitung nilai probabilitas setiap kelas pada sebuah dokumen dengan memprediksi jumlah kemunculan kata dalam dokumen yang bersifat positif, netral atau negatif secara otomatis. Setelah itu melakukan prediksi pada tweet yang masuk kedalam sentiment positif dengan menggunakan *predict_proba.*

3. Model Klasifikasi

Setelah melakukan perhitungan TF-IDF kemudian dilanjutkan dengan pembuatan model klasifikasi dengan menggunakan variabel X dan y dengan data training yang sudah disediakan. Model dibuat sebagai fungsi supaya bisa memudahkan dalam pemanggilan dan eksekusi pada tahap berikutnya. Kemudian model akan disimpan kedalam bentuk file pickle agar dapat dibuka kembali dan digunakan lagi.

3.4.5 Testing

Pada tahap *testing* ini merupakan tahapan untuk melihat tingkat akurasi pemodelan yang telah dibuat pada tahap *training* yang digunakan untuk memprediksi label atau kelas dari data uji yang tersedia.

Model yang sudah didapatkan dihitung menggunakan beberapa metode pada confusion matrix untuk mengetahui persentase setiap dilakukan pengujian. Dengan menggunakan library sklearn yang melakukan import confusion_matrix untuk melakukan testing dan prediksi. Metode yang digunakan meliputi:

1. accuracy untuk mengetahui jumlah klasifikasi dibagi dengan total sampel testing yang diuji

2. precision untuk mengetahui klasifikasi kategori positif yang benar dibagi dengan total sampel klasifikasi positif
3. recall untuk mengetahui sampel yang diklasifikasi kategori positif dibagi total sampel dalam testing yang berkategori positif
4. f-measure untuk menghitung rata-rata dari precision dan recall.