BAB 3
METODE PENELITIAN

Penelitian ini merupakan studi analisis sentimen yang memfokuskan pada aspek positif dan negatif dari data media sosial *X*. Metode yang digunakan yaitu *Naïve Bayes Classification.* Penelitian memerlukan data tweet yang diambil dari *X* dengan kata kunci yang berkaitan dengan *Jersey* Tim Nasional Indonesia. Selanjutnya, melakukan proses *preprocessing* untuk memperoleh data yang sesuai dengan kebutuhan. Data-data tersebut kemudian digunakan dalam menganalisis sentimen atau pendapat dari pengguna Twitter mengenai *Jersey* Tim Nasional Indonesia.

Penelitian dimulai dengan mengidentifikasi latar belakang permasalahan, melakukan pengolahan data yang ada, membentuk model topik, dan mencari nilai-nilai sentimen yang optimal untuk memperoleh informasi sesuai dengan tujuan yang diinginkan (Putra Aziztiya et al., 2022). Berikut ini adalah komponen-komponen penelitian analisis sentimen *Jersey* Tim Nasional Indonesia beserta dengan langkah-langkah yang diambil dalam proses analisis sentimen menggunakan data tweet.

3.1 BAHAN DAN ALAT PENELITIAN

Bahan penelitian yang diperlukan mencakup koleksi *tweet, re-tweet,* dan komentar yang berkaitan dengan *jersey* Tim Nasional Sepak Bola Indonesia di platform media sosial *X.*

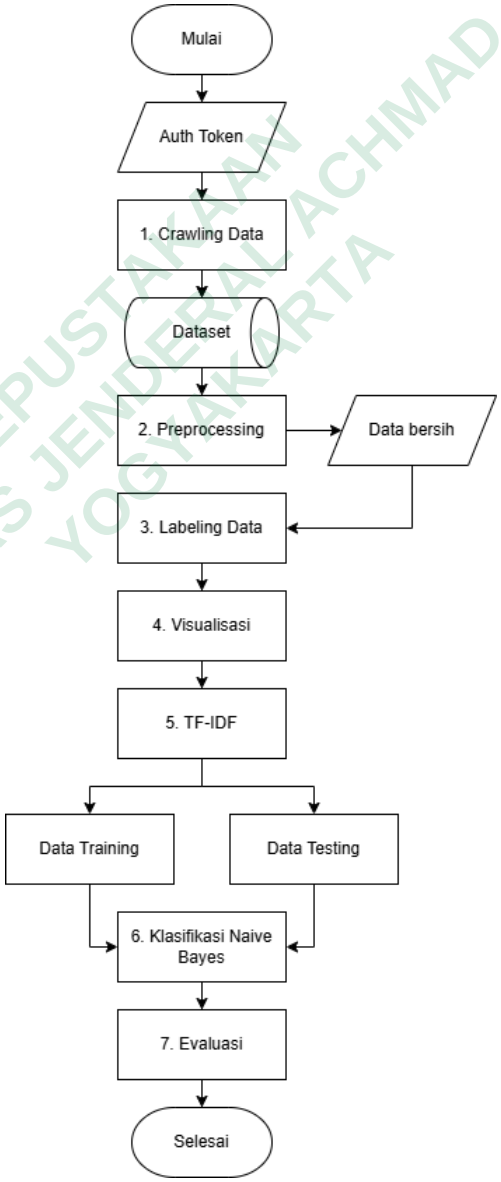
Perangkat yang digunakan dalam penelitian adalah laptop model *Lenovo IPS 340* yang memiliki kemampuan terhubung ke jaringan internet. Dan menggunakan beberapa sistem operasi serta program-program aplikasi dalam pengolahan data analisis sentimen, diantaranya:

1. Sistem operasi *: Windows* 11
2. *Processor* : *AMD Ryzen 3 3200U*
3. *Memory* : 8 GB RAM
4. Bahasa Pemrograman *: Phyton 3.9.1*

5. *Google Colab*
6. *Microsoft Excel 2019*
7. *Media Sosial X*

3.2 JALAN PENELITIAN

Penelitian ini menggunakan bahasa pemrograman Python, Google Colab untuk pengambilan data, yang disajikan dalam format Microsoft Excel. Pengolahan data dan pembuatan model dilakukan dengan berbagai library Python. Langkah-langkah penelitian tercantum dalam **Gambar 3.1.**



Gambar 3.1 Alur Penelitian

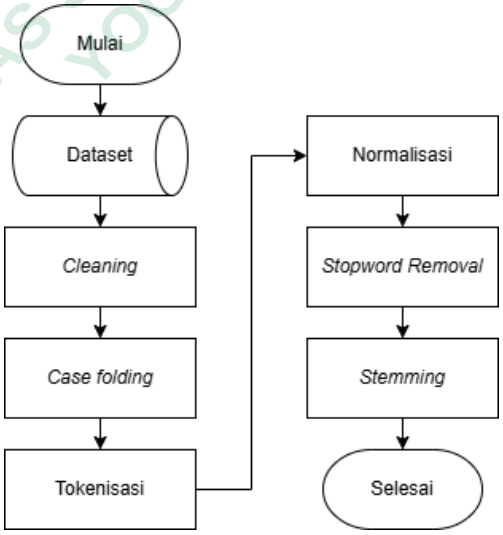
1. *Crawl Data* (Pengumpulan Data)

Proses pengambilan dan pengumpulan data merupakan langkah untuk mendapatkan informasi dari media sosial *X* dengan mengambil *tweet, retweet,* ataupun *reply* yang terdapat kata kunci seperti "*jerseybarutimnas*", "*jerseyerspo*", dan lain-lain. Pengambilan data dilakukan menggunakan *Tweet Harvest* pada *Google Colab*, dengan hasil akhir akan disimpan ke *Microsoft Excel* dengan format csv. Untuk alur crawl data bisa dilihat pada **Gambar 3.2.**

Gambar 3.2 Alur *crawl data*

2. *Preprocessing Data*

Tahap ini akan melibatkan rangkaian tindakan untuk membersihkan, menata, dan menyiapkan data mentah agar siap digunakan dalam proses analisis lebih lanjut. Tujuannya adalah mempersiapkan data sehingga dapat dilakukan analisis secara akurat dan efisien.. Berikut merupakan alur tahapan dalam *Preprocessing* pada **Gambar 3.3.**



Gambar 3.3 Alur *preprocessing*

a. *Cleaning*

Cleaning adalah proses membersihkan teks dari karakter atau elemen yang tidak diinginkan yang dapat mengganggu analisis data. Ini termasuk menghapus tanda baca, angka, karakter khusus, spasi ganda, dan elemen non-teks lainnya seperti HTML tags atau emojis. Tujuannya adalah untuk memperoleh teks yang bersih dan konsisten.

b. *Case Folding*

Case folding adalah proses mengubah semua huruf dalam teks menjadi huruf kecil. Contohnya, kata "Erspo", "ERSPO", dan "erspo" akan diubah menjadi "erspo". Ini dilakukan untuk memastikan bahwa perbedaan huruf besar dan kecil tidak mempengaruhi analisis, sehingga kata-kata yang seharusnya sama tidak dianggap berbeda.

c. Tokenisasi

Tokenisasi adalah proses memecah teks menjadi unit-unit kecil yang disebut token. Token biasanya berupa kata, tetapi bisa juga berupa frasa atau simbol tergantung pada kebutuhan analisis. Misalnya, kalimat "I love *Erspo*" akan di-tokenisasi menjadi ["I", "love", "Erspo"]. Tokenisasi memudahkan analisis dengan memungkinkan penanganan teks pada level yang lebih mendetail.

d. Normalisasi

Normalisasi adalah proses standarisasi teks untuk memastikan konsistensi. Ini mencakup penggantian bentuk kata yang berbeda tetapi memiliki makna yang sama dengan bentuk standar atau dasar. Contohnya, mengubah "aren't" menjadi "are not" atau "running" menjadi "run". Normalisasi juga dapat mencakup konversi teks slang atau singkatan ke bentuk yang lebih formal atau lengkap.

e. *Stopwords removal*

Stopword removal adalah proses menghapus kata-kata umum yang dianggap tidak membawa makna signifikan dalam analisis teks, seperti "and", "the", "is", dan "in". Stopword removal membantu mengurangi kebisingan dalam data dan meningkatkan efisiensi analisis dengan fokus pada kata-kata yang lebih penting secara kontekstual.

f. Stemming

Stemming adalah proses mengurangi kata ke bentuk dasarnya (stem) dengan menghapus akhiran (suffix) dan kadang-kadang awalan (prefix). Misalnya, kata "running", "runner", dan "ran" akan diubah menjadi "run". Stemming membantu mengelompokkan varian dari kata yang sama sehingga analisis dapat lebih konsisten dan efektif.

3. Labeling Data

Pada tahap pelabelan, data yang telah diperoleh sebelumnya akan diberi label. Pelabelan data dilakukan menggunakan perangkat lunak Excel dengan menambahkan kolom baru yang diberi nama "Label". Dalam kolom "Label" ini, setiap ulasan negatif akan diberi angka 0, sedangkan ulasan positif akan diberi angka 1 (Alpin Rizaldi et al., 2023). Proses tersebut penting agar model Naive Bayes dapat mempelajari pola dalam data latih dan memberikan prediksi akurat pada data uji. Setelah dilatih, model dapat mengklasifikasikan data baru berdasarkan probabilitas kelas yang paling mungkin sesuai dengan fitur yang diamati.

4. Visualisasi

Visualisasi dalam analisis sentimen adalah representasi grafis dari hasil analisis sentimen, yang bertujuan untuk memudahkan pemahaman pola, tren, dan distribusi sentimen dari data teks. Terdapat beberapa jenis visualisasi yang umum digunakan pada analisis sentimen seperti diantaranya Word Cloud, Bar Chart, Pie Chart, Time Series Plot, Heatmap, dll.

5. TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) adalah teknik yang digunakan dalam analisis teks untuk menilai pentingnya sebuah kata dalam satu dokumen relatif terhadap kumpulan dokumen lainnya. Dalam analisis sentimen, TF-IDF membantu mengidentifikasi dan memberi bobot pada kata-kata yang lebih informatif dan relevan untuk analisis, bukan hanya kata-kata yang sering muncul tetapi mungkin tidak memiliki banyak arti (seperti "the" atau "is").

6. Klasifikasi Naïve Bayes

Proses klasifikasi menggunakan *Naive Bayes*, data awalnya dibagi menjadi dua set: data latih dan data uji. Langkah pertama adalah membersihkan dan memproses data untuk memastikan konsistensi dan kualitas. Fitur teks diubah menjadi representasi numerik menggunakan teknik seperti *TF-IDF*. Model *Naive Bayes* dilatih menggunakan data latih, di mana ia mempelajari pola dan hubungan dalam data untuk mengklasifikasikan sentimen. Setelah pelatihan, model diuji dengan data uji yang belum pernah dilihat sebelumnya. Hasil prediksi dievaluasi menggunakan metrik seperti akurasi, precision, recall, dan F1-score untuk mengevaluasi kinerja model dalam mengklasifikasikan sentimen secara efektif. Alur ini memastikan bahwa model mampu memprediksi sentimen pada data baru dengan tingkat keakuratan yang optimal.

7. Evaluasi

Evaluasi akan disajikan hasil klasifikasi menggunakan algoritma *Naive Bayes* yang diukur melalui *confusion matrix.* Confusion matrix akan memberikan gambaran mendetail tentang kinerja model dalam mengklasifikasikan data ke dalam kategori yang benar dan salah, serta menunjukkan jumlah prediksi yang benar dan salah untuk setiap kelas. Selanjutnya, proses *cross-validation* akan dilakukan untuk memastikan keandalan model dengan membagi data menjadi beberapa *fold*, di mana model diuji secara berulang untuk mengukur performanya secara lebih robust. Hasil

akhir evaluasi akan mencakup akurasi keseluruhan model, yang mengukur seberapa baik model dalam mengklasifikasikan data secara akurat.

Analisis juga akan menyertakan evaluasi sentimen keseluruhan data, memberikan wawasan tentang distribusi sentimen dalam dataset. Evaluasi ini mencakup penampilan data yang menunjukkan sentimen positif dan negatif, memungkinkan pemahaman yang lebih baik mengenai pola sentimen dalam data. Beberapa contoh data dengan sentimen positif maupun negatif akan ditampilkan untuk memberikan gambaran yang jelas tentang bagaimana model menangani dan mengkategorikan berbagai jenis sentimen dalam konteks yang relevan. Dengan pendekatan ini, kita dapat memperoleh pemahaman mendalam tentang performa model serta distribusi dan karakteristik sentimen dalam dataset yang dianalisis.