BAB III
METODELOGI PENELITIAN

3.1 Metode Penelitian

Metode penelitian merupakan tahapan-tahapan yang dilakukan oleh peneliti untuk memperoleh gambaran yang jelas mengenai penelitian. Tahapan yang dilakukan dalam metode penelitian ini adalah sebagai berikut.

3.1.1 Pengumpulan Data

Data yang dikumpulkan merupakan data yang diambil dari media sosial *twitter, tweet* yang memiliki kata kunci sesuai dengan topik yang diangkat yaitu tentang vaksin *booster*. Data tersebut diperoleh dengan cara melakukan *crawling* data di twitter melalui rapid miner.

3.1.2 Studi Literatur

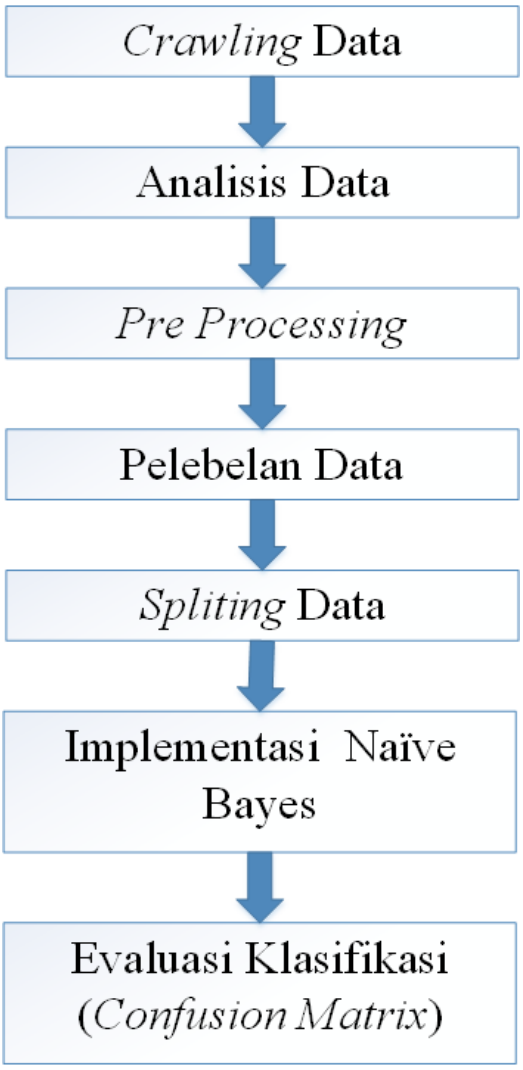
Studi literatur dilakukan dengan cara mengumpulkan dan mempelajari jurnal, literatur, paper, makalah sebagai sumber pustaka yang berkaitan dengan materi penelitian khususnya analisis sentimen menggunakan algoritma naïve bayes.

3.1.3 Perancangan

Pada proses ini dilakukan perancangan terhadap perangkat lunak yang akan dibuat berdasarkan hasil studi literatur seperti desain struktur data, desain aliran informasi dan desain algoritma.

3.2 Alur Penelitian

Alur penelitian ini secara garis besar dapat dilihat pada Gambar 3.1.



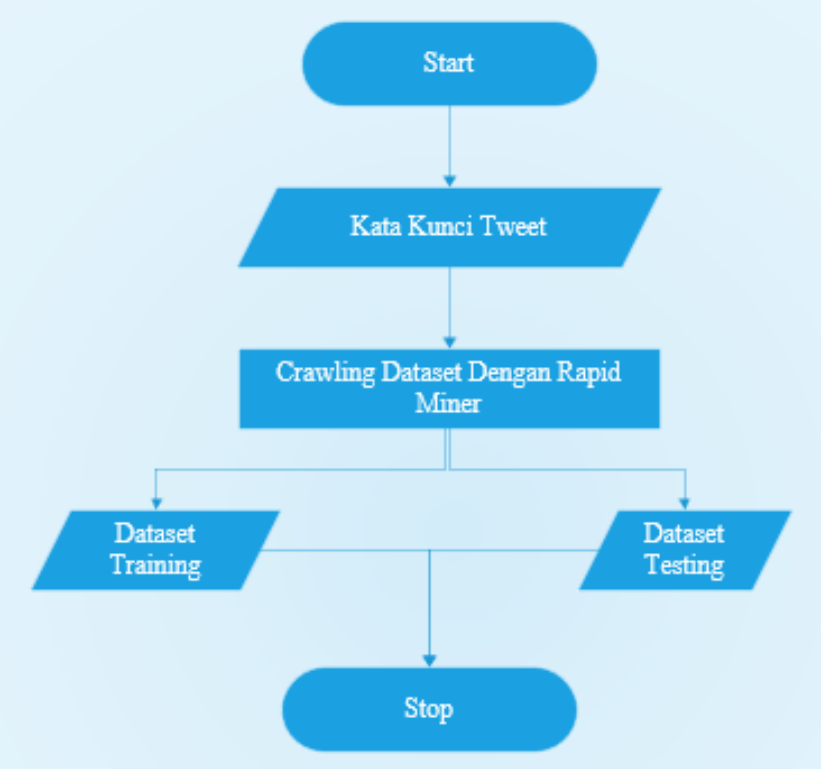
Gambar 3.1 Alur penelitian secara umum (Laurensz & Eko Sediyono, 2021)

Penjelasan dari masing-masing tahapan pada alur penelitian tersebut diuraikan pada beberapa sub bab di bawah ini.

3.2.1 *Crawling Data*

Pada penelitian ini pengumpulan data dilakukan dengan cara *crawling* data pada media sosial *twitter* dengan memanfaatkan *Application Programming Interface* (API) yang telah disediakan aplikasi tersebut untuk memperoleh kumpulan data text

yang telah diunggah oleh pengguna *twitter*. Prosesnya dilakukan menggunakan *Rapid miner* dengan memasukkan *access token* yang didapat melalui pendaftaran di *twitter developer*. Kata kunci yang digunakan untuk keperluan ini adalah "*vaccine booster*".



Gambar 3.2 Alur Crawling data

3.2.2 Analisis Data

Pada tahap ini pengolahan data menggunakan *tool rapid miner*. Ada tujuh tahapan dalam analisis data yaitu: *crawling* data, pelabelan data, *pre processing*, klasifikasi algoritma *naïve bayes*, validasi *K-fold cross validation*, evaluasi *confusion matrix* dan visualisasi data.

3.2.3 *Pre Processing*

Pre processing merupakan tahapan yang sangat penting sebelum dilakukan proses *data mining*. *Pre processing* dilakukan untuk mengeliminasi sejumlah

permasalahan pada data, diantaranya: *cleaning, case folding, tokenizing* ataupun format data yang tidak sesuai dengan sistem dan berpotensi dapat mengurangi performa proses *data mining*. *Pre processing* dilakukan agar data dapat direpresentasikan dalam kondisi ideal sebelum diproses lebih lanjut.

1. *Cleaning*

   Data yang diperoleh dari crawling twitter ada yang terdapat unsur yang tidak terpakai untuk digunakan pada tahap penelitian selanjutnya seperti hashtag, url, Rt, username dan tanda baca (*punctuation*).

2. *Case folding*

   Pada tahap ini dilakukan penyetaraan dimana data masih ada yang pada kalimatnya terdapat huruf besar dan kecil. Sehingga harus disamakan standar dari huruf dengan mengubah menjadi huruf kecil.

3. *Tokenizing*

   Pada tahap ini dilakukan memotong kata yang ada pada kalimat menjadi kata yang terpisah antara satu dengan yang lainya. Sehingga setiap kata pada kalimat dapat di cek satu persatu pada langkah berikutnya.

4. *Term Frequency-Inverse Document Frequency* (TF-IDF)

   Langkah selanjutnya adalah pembobotan kata menggunakan tf-idf. Tf-idf merupakan cara pemberian bobot hubungan suatu kata (*term*) terdadap kata.

3.2.4 Pelabelan Data

Pelabelan data *tweet* dalam proses ini setiap dokumen teks akan dilabeli dengan sentimen positif, netral dan negatif. Sentimen positif adalah komentar atau ulasan dari pengguna *twitter* yang setuju dengan penggunaan vaksin *booster*, sedangkan sentimen negatif merupakan komentar atau ulasan dari pengguna *twitter* yang tidak setuju dengan penggunaan vaksin *booster*.

3.2.5 *Spliting Data*

Pada tahap ini dilakukan pembagian data ke dalam *data training* dan *data testing*. Berikut adalah penjelasan lebih lanjut mengenai keduanya.

1. Data *Training*

   Merupakan bagian dataset yang dilatih untuk membuat prediksi, klasifikasi atau menjalankan fungsi dari sebuah algoritma *machine learning* lainnya sesuai tujuan yang diharapkan. Pada tahap ini petunjuk diberikan melalui algoritma agar mesin yang dilatih dapat mencari korelasinya sendiri.

2. Data *Testing*

   Merupakan bagian *dataset* yang akan diuji untuk melihat keakuratannya, atau dengan kata lain melihat performanya.

3.2.6 *Implementasi Naïve Bayes*

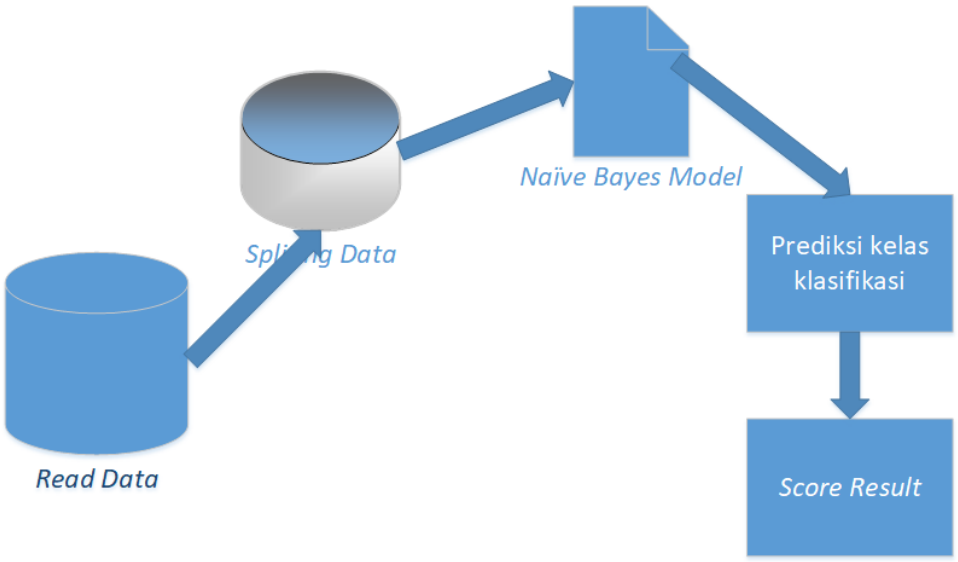
Pada tahapan ini metode klasifikasi *Naive Bayes* diterapkan untuk menentukan kelas atau label dari suatu *tweet* data uji. Klasifikasi *naïve bayes* merupakan klasifikasi secara statistik, klasifikasi ini dapat memprediksi peluang keanggotaan kelas seperti probabilitas suatu *tupel* merupakan milik kelas tertentu. Klasifikasi ini berdasarkan *teorama bayes*. Bentuk umum teorema *bayes* dapat dilihat pada Persamaan 2 berikut ini.

$$P(\mathrm{H}|\mathrm{X}) = \frac{P(X|H)P(H)}{P(X)} \_\_\_\_\_ (2)$$

Berikut adalah penjelasan dari notasi-notasi yang digunakan pada Persamaan 2 tersebut.

a. “H” merupakan hipotesa data X merupakan suatu kelas spesifik.
b. “X” merupakan data dengan kelas yang belum diketahui.
c. P(H|X) merupakan probabilitas hipotesis H berdasarkan kondisi X (*posterior probability*).
d. P(H) merupakan probabilitas hipotesis H (*prior probability*).

Algoritma *naïve bayes* juga merupakan algoritma yang cukup efektif dan efesien yang bisa digunakan pada machine learning data mining seperti *rapid miner*. Proses implementasi algoritma *naïve bayes* dapat dilihat pada gambar 3.2.



Gambar 3.4 Proses implementasi *naïve bayes* (Handayani & Pribadi, 2015).

3.2.7 *Evaluasi Klasifikasi*

Setelah proses klasifikasi selesai dilakukan maka tahap selanjutnya adalah tahap evaluasi klasifikasi. Pada tahap ini dilakukan pengujian menggunakan *confusion matrix* untuk melihat performa dari algoritma *Naïve Bayes*.

Tabel 3.1 Tabel Evaluasi *Confusion Matrix*

Kelas	Klasifikasi Positif	Klasifikasi Negatif
Positif	TP (True Positif)	FN (False Negatif)
Negatif	FP (False Negatif)	TN (True Negatif)

Berikut adalah penjelasan dari tabel visualisasi *confusion matrix*.

a. TP (*true positive*), yaitu jumlah data positif yang terklasifikasi dengan benar oleh sistem.

b. TN (*true negative*), yaitu jumlah data negatif yang terklasifikasi dengan benar oleh sistem.
c. FP (*false positive*), yaitu jumlah data positif namun terklasifikasi salah oleh sistem
d. FN (*false negative*)**,** yaitu jumlah data negatif namun terklasifikasi salah oleh sistem.

Dengan kata lain, nilai akurasi merupakan perbandingan antara data yang terklasifikasi benar dengan keseluruhan data. Nilai akurasi dapat diperoleh dengan persamaan berikut ini.

$$Akurasi = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \; X \; 100\%$$

Nilai presisi menggambarkan jumlah data kategori positif yang diklasifikasi secara benar dibagi dengan total data yang diklasifikasi positif, *presisi* dapat diperoleh dengan persamaan berikut ini.

$$Presisi = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \; X \; 100\%$$

Sementara itu, *recall* menunjukkan beberapa persen data kategori positif yang terklasifikasi dengan benar oleh sistem.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \; X \; 100\%$$

3.3 Kebutuhan Perangkat Keras dan Lunak

3.3.1 *Perangkat Keras*

Perangkat keras (*hardware)* yang digunakan dalam penelitian ini menggunakan satu buah laptop dengan spesifikasi laptop sebagai berikut.

1. Processor : Intel Core i5-8250U

2. Ram : 4 GB
3. Tipe sistem : 64-Bit sistem operasi

3.3.2 *Perangkat Lunak*

Perangkat lunak *(software)* yang digunakan dalam penelitian ini sebagai berikut.

1. Windows 10
2. Google Chroome
3. Rapid miner
4. Ms. Excel