BAB III
METODE PENELITIAN

3.1. Metode Pengumpulan Data

3.1.1. Studi Pustaka

Peneliti melakukan studi pustaka untuk mendapatkan referensi yang nantinya akan menjadi acuan dalam penelitian ini, seperti metode atau variabel yang digunakan dan saran-saran yang dapat diterapkan agar penelitian ini dapat lebih baik dari penelitian sebelumnya. Peneliti mencari referensi dari jurnal-jurnal, sumber internet atau buku terkait dengan penelitian ini, yaitu mengenai analisis sentimen, *machine learning*, algoritma SVM dan *Naïve Bayes*.

3.1.2. *Crawling* Data

Pada tahapan ini penulis mengambil data dari *twitter* yang berkaitan dengan kuliah luring. Penulis melakukan pencarian menggunakan kata kunci "kuliah luring" dan "kuliah *offline*". Data yang penulis ambil berupa *tweet* dari tanggal 1 Oktober 2022 – 31 Januari 2023. Proses pengambilan *tweet* ini memanfaatkan library yang terdapat pada situs https://colab.research.google.com/drive/1jQhAGKanGZ290rIaf06705xmhAB3vpD9. Data *twitter* yang telah berhasil didapat kemudian disimpan ke dalam format csv. Total data yang penulis dapatkan sebanyak 3544 *tweets*, dengan rincian sebanyak 340 data merupakan hasil dari kata kunci "kuliah luring" dan 3204 data merupakan hasil dari kata kunci "kuliah *offline*". Berikut ini merupakan *sample* dari data yang telah penulis dapatkan.

Tabel 3.1 *Sample Crawling* Data *Tweet*

created_at	id_str	full_text	quote_count	Reply_count	Retweet_count	Favorit_count	lang	user_id_str	conversation_id_str	User name	tweet_url
Mon Jan 30 17:00:04 +0000 2023	1,62E+18	suamiku kuliah offline, serem kalo dia ngegaet katingnya ntar di kampus hufttt	0	1	0	0	in	1,1E+18	1,62E+18	caramelone_	https://twitter.com/caramelone_/status/1620104572118396928

created_at	id_str	full_text	quote_count	Reply_count	Retweet_count	Favorit_count	lang	user_id_str	conversation_id_str	User name	tweet_url
Mon Jan 30 11:37:01 +0000 2023	1,62E+18	ga bakal tau temen kuliah yang alhamdulilah seruu, ga bakal tau ospek offline yang menyeramkan	0	1	0	0	in	1,3E+18	1,62E+18	btyksm12	https://twitter.com/btyksm12/status/1620232746621583336
Sun Jan 29 14:26:04 +0000 2023	1,62E+18	kuliah offline bingung pake baju apa alias GAPUNYA BAJU YAALLAAAAH	0	1	0	0	in	9,7E+17	1,62E+18	pesvgihan	https://twitter.com/pesvgihan/status/1619703428124651520
Sun Jan 29 14:18:20 +0000 2023	1,62E+18	inget struggle kuliah offline di smt 5 kemarin bikin makin ngga semangat masuk kuliah smt 6	0	1	0	1	in	8,1E+17	1,62E+18	oursdhiverr	https://twitter.com/oursdhiverr/status/1619701481204240384
Sun Jan 29 14:16:59 +0000 2023	1,62E+18	@tanyarlfes Aku kuliah sehari di bawain duit 27 ribu. 15 ribu buat makan siang, 12 ribu buat bensin pertalite. jujur itu kurang, karena kuliah offline, tugas kayak makalah, essay, artikel itu harus di print, belum lagi kalau di ajak nugas di cafe.	0	1	0	0	in	1,5E+18	1,62E+18	Wildani__	https://twitter.com/Wildani__/status/1619701414824099889

3.2. Pengolahan Data

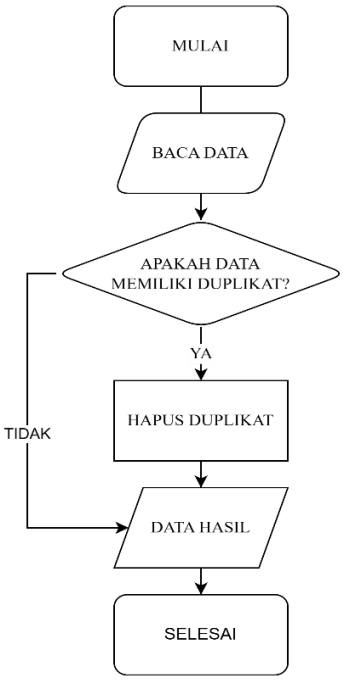
Pengolahan data dilakukan melalui proses pemecahan teks menjadi *term* (*preprocessing*), melakukan pelabelan data, menghitung bobot dari setiap *term*, dan melakukan klasifikasi menggunakan Algoritma *Support Vector Machine* dan *Naïve Bayes*. Dalam proses pengolahan data, dibantu dengan *Google Colaboratory* dan bahasa pemrograman *Python*.

3.2.1. Data *Preprocessing*

Proses data *preprocessing* yang dilakukan pada penelitian ini melalui 7 tahapan yaitu cek duplikat, *case folding*, *cleaning*, normalisasi, *tokenizing*, *remove stopwords*, dan *stemming*. Tahapan *preprocessing*akan menghasilkan kumpulan atribut atau *keyword* dalam bentuk *term* (kata per kata). Tahapan *preprocessing* sebagai berikut:

a. Cek Duplikat

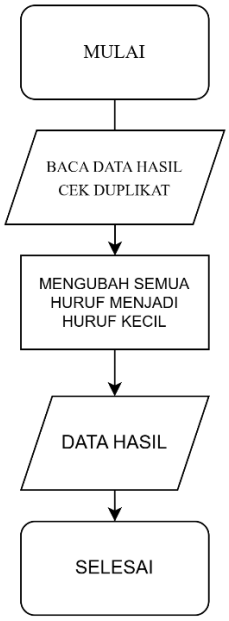
Cek Duplikat merupakan proses mendeteksi apakah ada *tweet* atau data yang berjumlah lebih 1 dari database. Jika ada, maka data yang mempunyai duplikatnya akan dihapus sehingga hanya tersisa 1 data saja. Cek Duplikat dilakukan agar nantinya model tidak mempelajari teks yang sama. Gambar 3.1 merupakan gambar dari proses cek duplikat.



Gambar 3.1 Flowchart *Proses Cek Duplikat*

b. Case Folding

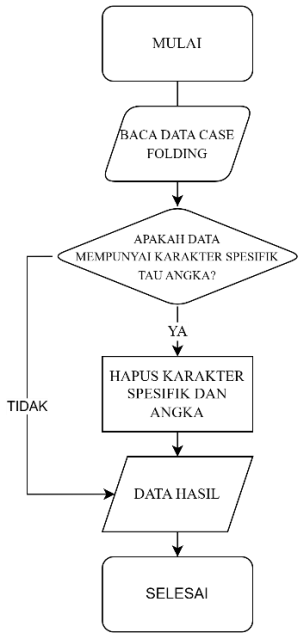
Dalam proses ini, untuk menyamakan karakter dari setiap dokumen maka semua huruf dalam kalimat atau kata dari setiap dokumen dikonversikan menjadi huruf kecil (*lowercase*). Tahap-tahap dalam proses *case folding* di tunjukan pada gambar berikut:



Gambar 3.2 *Flowchart* Proses *Case Folding*

c. *Cleaning* Data

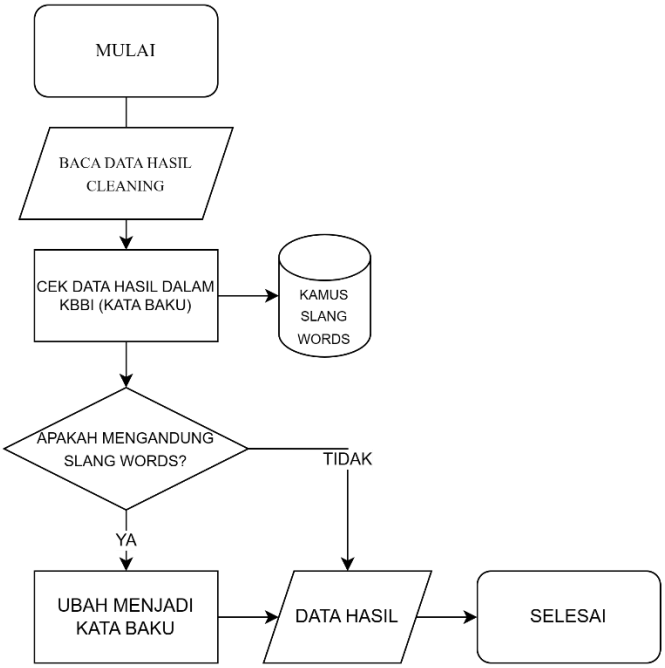
Cleaning merupakan proses menghapus karakter-karakter tertentu seperti *hashtag*, *mention, emoticon,* karakter angka, tanda baca, URL, *website*, *tag*, HTML, dan lain-lain dalam sebuah *tweet.* Tujuan dari proses ini adalah untuk mengurangi *noise* atau kesalahan acak pada data. Gambar 3.3 merupakan gambar dari proses *cleaning* data.



Gambar 3.3 *Flowchart* Proses *Cleaning Data*

d. Normalisasi

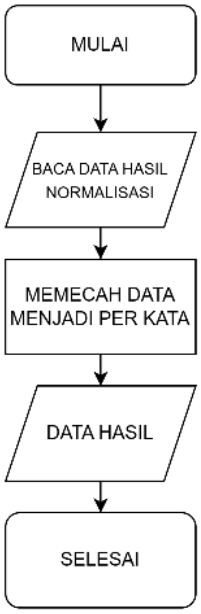
Normalisasi merupakan tahapan mengubah kata yang disingkat dan tidak baku atau *slang word* menjadi kata baku sesuai KBBI (Kamus Besar Bahasa Indonesia). Pada proses normalisasi digunakan kamus yang sudah berisi *slang word*. Gambar 3.4 merupakan gambar tahapan normalisasi



Gambar 3.4 *Flowchart* Proses Normalisasi

e. Tokenizing

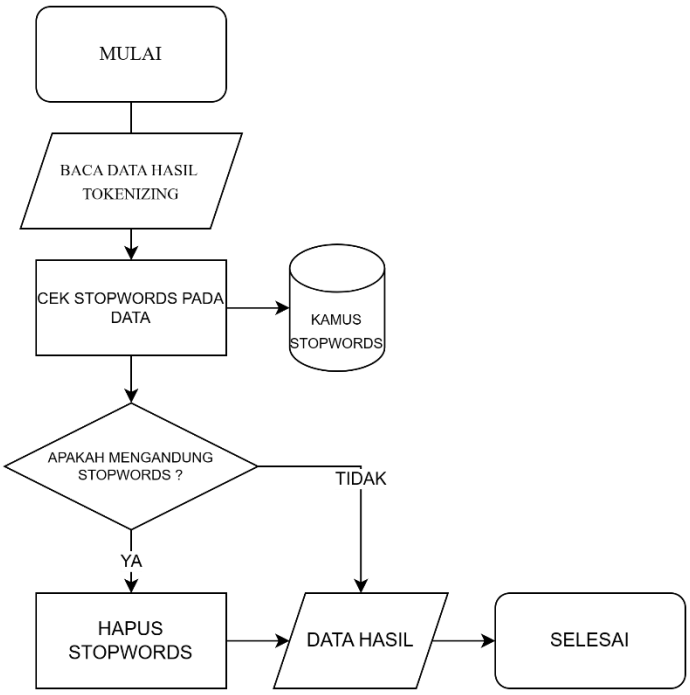
Tokenizing merupakan tahapan dari data *preprocessing* yang bertujuan untuk membuat dokumen menjadi bagian yang lebih kecil dengan cara menghilangkan kalimat yang sama dan memecahkan kalimat atau paragraf. Gambar 3.5 merupakan tahapan *tokenizing*.



Gambar 3.5 *Flowchart* Proses *Tokenizing*

f. *Remove Stopwords*

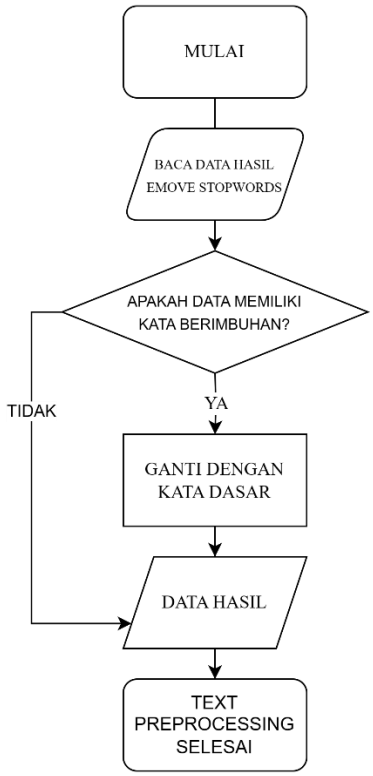
Stopwords merupakan kata-kata yang tidak memiliki makna seperti kata hubung, kata depan, dan lain-lain. Pada tahap ini *remove stopwords* dilakukan untuk menghapus *stopwords* yang ada pada data untuk memudahkan proses pengolahan data. Gambar 3.6 merupakan tahapan *remove stopwords*.



Gambar 3.6 *Flowchart* Proses *Remove Stopwords*

g. *Stemming*

Kata-kata yang terdapat dalam proses sebelumnya diubah ke dalam bentuk kata dasar dengan menghilangkan kata imbuhan. Pada penelitian ini, proses *stemming* menggunakan *library Sastrawi*. *Flowchart stemming* ditunjukkan pada gambar 3.7.



Gambar 3.7 *Flowchart* Proses *Stemming*

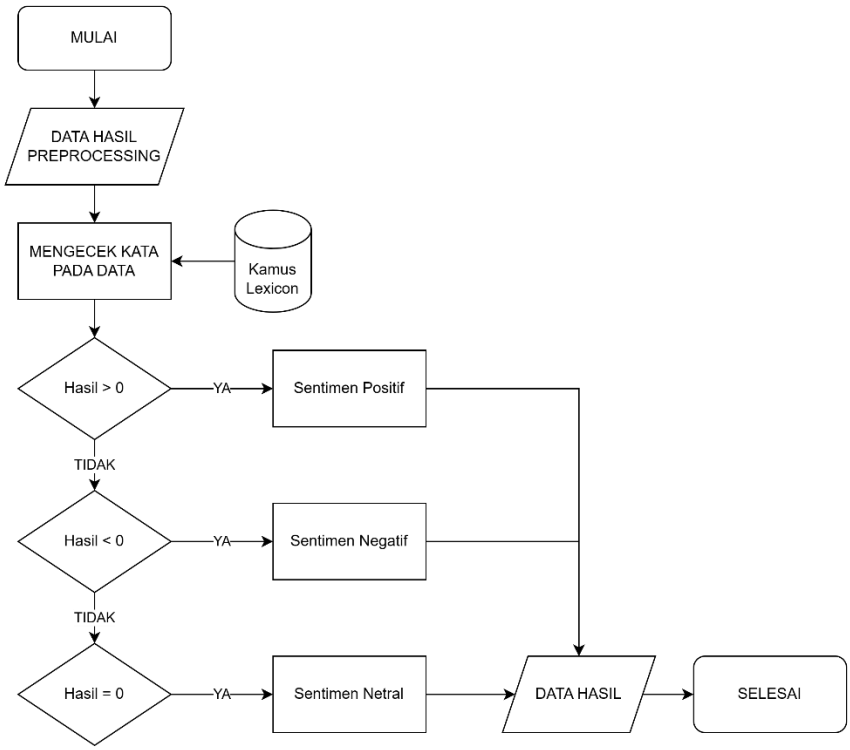
3.2.2. Pelabelan Data

Pelabelan data digunakan menggunakan metode *Lexicon Based*. Data *tweet* yang akan digunakan untuk melatih metode SVM dan *Naïve Bayes* membutuhkan label kelas, apakah *tweet* itu positif, negatif atau netral. Untuk menentukan label kelas dari suatu data, metode yang sering digunakan adalah dengan cara memberikan label kelas manual berdasarkan pendapat sendiri, menilai dari jumlah *rating* yang diberikan jika ada fitur *rating* atau dengan menggunakan metode klasifikasi teks dengan pendekatan kamus.

Pada penelitian ini, peneliti akan menggunakan metode klasifikasi dengan pendekatan kamus yaitu peneliti akan menggunakan kamus *lexicon* untuk menilai apakah *tweet* termasuk ke dalam kategori sentimen positif, negatif atau netral.

Sebelum menentukan label kelas diperlukan perhitungan terhadap masing-masing sentimen yang terdapat di dalam *tweet* agar dapat diberikan label sesuai skornya.

Jika skor lebih besar dari 0 maka *tweet* akan dinilai positif, jika skor lebih kecil dari 0 maka akan dinilai negatif dan jika skor sama dengan 0 makan akan dinilai netral. Gambar 3.8 merupakan tahapan pelabelan data menggunakan metode *Lexicon Based*.



Gambar 3.8 Tahapan Pelabelan Data

3.2.3. Pembobotan TF-IDF

Pembobotan dapat diperoleh berdasarkan jumlah dokumen (d), jumlah kemunculan suatu *term* pada setiap dokumen yang disebut *term frequency* (tf), banyaknya dokumen dimana suatu *term* muncul disebut *document frequency* (df) dan inversi jumlah dokumen yang mengandung sebuah *term* yang disebut *inverse document frequency* (idf).

Pembobotan menggunakan TF-IDF dijelaskan pada persamaan berikut:

$$w_{i,j} = tf_{i,j}\, x \log\left(\frac{N}{df_i}\right)$$

Dimana

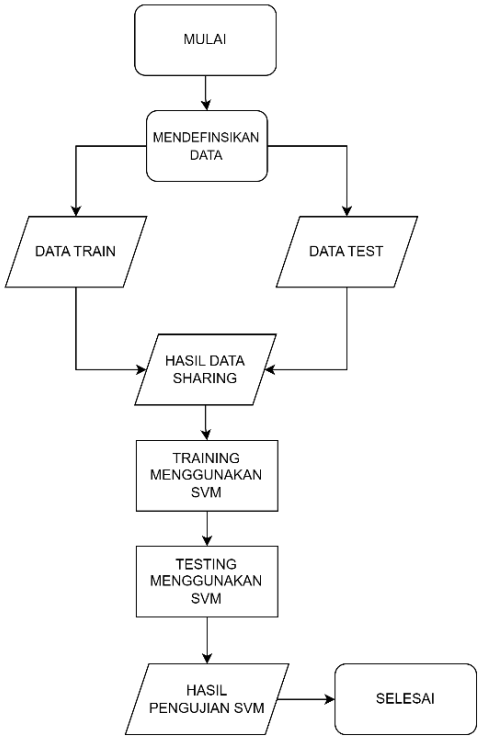
$tf_{i,j}$ adalah banyaknya kata i pada dokumen j

N adalah banyaknya dokumen

df_i adalah banyaknya dokumen yang mengandung kata ke-i

3.2.4. Klasifikasi Menggunakan *Support Vector Machines*

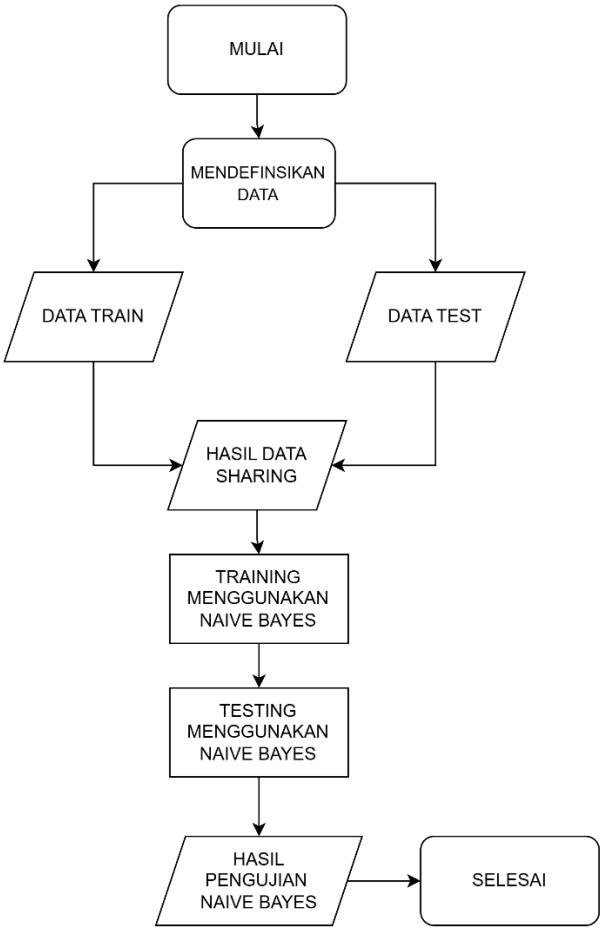
Tahapan selanjutnya setelah mendapatkan hasil *preprocessing, Lexicon Based*, dapat dilakukan klasifikasi dengan metode *Support Vector Machine* untuk melihat tingkat akurasi, presisi, recall dan F-1 score dari model SVM. SVM adalah metode klasifikasi yang dilakukan dengan cara mencari *term* atau kondisi serupa dengan parameter tertentu agar dapat menentukan kategori akhirnya. Kelebihan menggunakan metode ini adalah dapat menghasilkan model klasifikasi yang baik dengan formulasi yang jelas dan hanya sedikit parameter yang diatur. Selain itu, metode ini mudah dalam penerapannya karena penentuan SVM dapat ditentukan menggunakan QR (*Quadratic Programming*) dan memiliki kemampuan generalisasi yang tinggi. Pada Metode ini langkah-langkah proses klasifikasi dapat dilihat pada Gambar 3.9.



Gambar 3.9 Alur Penelitian Model SVM

3.2.5. Klasifikasi Menggunakan *Naïve Bayes*

Pada tahap klasifikasi *Naïve Bayes*, data *tweet* yang telah melalui tahapan *preprocessing*, klasifikasi *Lexicon* dan pembobotan akan diproses untuk melihat tingkat akurasi, presisi, recall dan F-1 *Score* nya. Dalam proses pengklasifikasian, keuntungan dari metode *Naïve Bayes* adalah data pelatihan yang dibutuhkan dalam jumlah kecil untuk menentukan estimasi parameter. Perbedaan yang paling mendasar antara teorema Bayes dengan metode lainnya adalah parameter Bayes dianggap menjadi variabel random, sedangkan dalam statistik klasik, parameter tidak dapat diketahui. Hubungan antara peluang bersyarat dari kejadian H dan X disebut teorema Bayes. Gambar 3.10 berikut adalah alur penelitian dari model *Naïve Bayes*.



Gambar 3.10 Alur Penelitian Model *Naïve Bayes*

3.3. Evaluasi Sistem

Evaluasi dilaukan dengan menggunakan *confusion matrix*. Pada penelitian ini, terdapat 3 label kelas pada *Confusion Matrix* , yaitu: kelas positif, kelas negatif, dan kelas netral, dapat dilihat pada Tabel 3.2 untuk contoh *confusion matrix* 3 kelas dan penentuan *variable* kelas benar dan salah.

Tabel 3.2 Confusion Matrix 3x3

	Prediksi			
Aktual		A	B	C
	A	*True A*	*False AB*	*False AC*
	B	*False BA*	*True B*	*False BC*
	C	*False CA*	*False CB*	*True C*

1. *Accuracy*

Variabel yang digunakan untuk menggambarkan seberapa akurat model dalam mengklasifikasikan data dengan benar.

2. *Precision*

Variabel yang digunakan untuk menggambarkan akurasi antara data yang diminta dengan hasil prediksi yang diberikan oleh model.

3. *Recall*

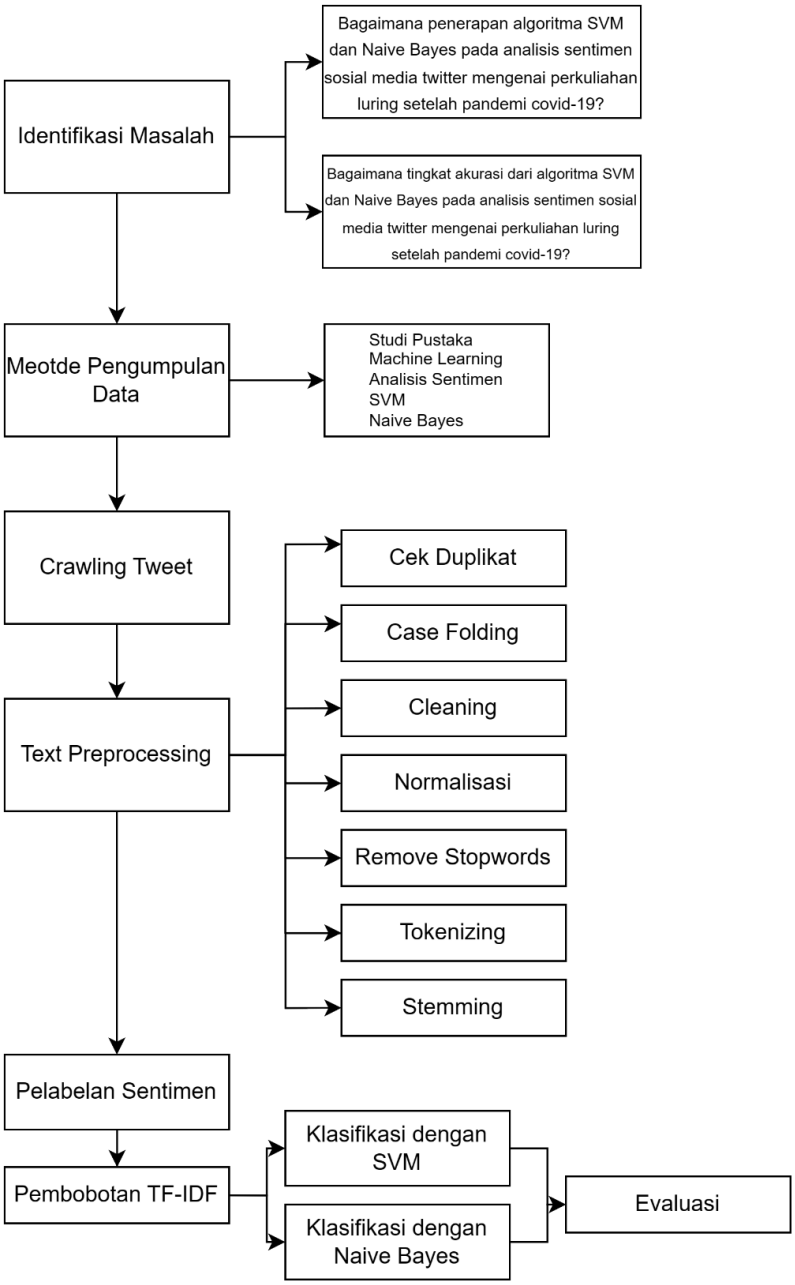
Variabel yang digunakan untuk menggambarkan keberhasilan model dalam menemukan kembali sebuah informasi.

4. *F Measure (F1 Score)*

Variabel yang digunakan untuk menggambarkan perbandingan rata-rata *precision* dan *recall* yang dibobotkan. *Accuracy* dapat digunakan sebagai acuan performansi algoritma jika dataset memiliki jumlah data *False Negatif* dan *False Positif* yang sangat mendekati (*symmetric*). Namun jika jumlahnya tidak mendekati, maka sebaiknya menggunakan F1 *Score* sebagai acuan.

Setelah selesai menentukan variabel-variabel yang telah ditentukan, akan terdapat 4 macam hasil yaitu akurasi, presisi, *recall*, dan *F-1 Score*. Peneliti akan menghitung hasil metode analisis sentimen tersebut berdasarkan setiap variabel di atas, untuk menentukan hasil akurasi yang paling tinggi.

3.4. Diagram Alur Penelitian



Gambar 3.11 Diagram Alur Penelitian

Berdasarkan langkah–langkah tahapan penelitian pada Gambar 3.11 dijelaskan bahwa:

1. Tahapan petama diawali dengan menentukan rumusan masalah dalam penelitian ini.
2. Tahapan kedua melakukan studi pustaka terkait materi penelitian.

3. Tahapan ketiga melakukan pengumpulan data dengan *crawling* data *tweet*.
4. Tahap keempat melakukan *text preprocessing* untuk mengumpulkan data dan memproses data yang telah dikumpulkan.
5. Tahap kelima melakukan pemberian label data dan pembobotan TF-IDF.
6. Tahapan keenam pengujian evaluasi menggunakan *confusion matrix* dari pengujian algoritma SVM dan *Naïve Bayes*.
7. Tahapan ketujuh peneliti mengambil kesimpulan dan saran dari penelitian yang telah dilakukan.