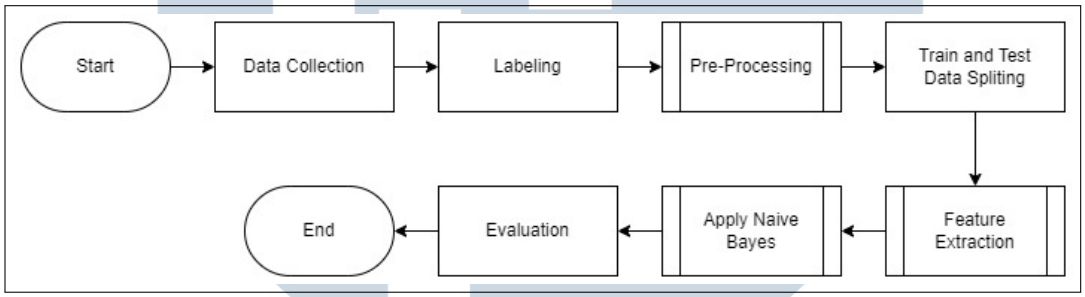
BAB 3
METODOLOGI PENELITIAN

3.1 Metodologi Penelitian

Dalam pelaksanaan penelitian analisis sentimen pengguna media sosial twitter terhadap minat membayar pajak menggunakan algoritma Naive Bayes akan menggunakan metodologi penelitian yang digambarkan menggunakan *flowchart* seperti yang dapat dilihat sebagai berikut:



Gambar 3.1. *Flowchart* Metodologi Penelitian

3.1.1 Data Collection

Tahapan awal yang dilakukan pada penelitian kali ini adalah *data collection* atau pengumpulan data. Pengumpulan data akan dilakukan dengan cara *crawl data* atau pengumpulan data secara otomatis yang bertujuan untuk mendapatkan data mentah atau *raw data* yang nantinya akan digunakan dalam penelitian. Pada tahapan ini dilakukan pengumpulan data dari sosial media twitter yang mengandung kata kunci yang telah ditetapkan, yaitu ”stop bayar pajak”, ”berhenti bayar pajak”, ”stop pajak”, ”tetap bayar pajak”, dan ”tetep bayar pajak” yang diunggah sejak tanggal 21 Februari 2023 sampai dengan 3 April 2023. Kata kunci ”Stop Bayar Pajak” dan ”Setop Bayar Pajak” yang digunakan untuk proses pengumpulan data diambil berdasarkan viralnya #StopBayarPajak yang dibuat secara terang-terangan oleh sebagian masyarakat Indonesia yang dipicu oleh kasus yang melibatkan sang pegawai pajak [37]. Sedangkan untuk kata kunci lainnya diambil dengan tujuan sebagai penyeimbang data dan membuat pemodelan yang nantinya akan digunakan dapat menganalisa sentimen kata yang lebih bervariatif. Pengumpulan data dilakukan secara otomatis menggunakan dua jenis *scraping tools*. Data yang sudah dikumpulkan secara otomatis tersebut kemudian digabungkan dan disimpan

menjadi sebuah *dataset* dalam format *comma seperated values* (CSV) sebagai bentuk *raw data*.

3.1.2 labeling

Setelah mendapatkan data pada tahapan *crawling data*, proses dilanjutkan dengan tahapan memberikan label atau *labeling* yang dilakukan dengan tujuan agar model dapat mempelajari data tersebut. Tahapan *labeling* dapat dilakukan dengan dua cara. Cara pertama yang dapat dilakukan adalah dengan *automatic labeling* dengan bantuan *library* pada *python* yang mewajibkan data diterjemah ke dalam Bahasa Inggris terlebih dahulu karena *library* tersebut hanya dapat membaca data yang Berbahasa Inggris. Cara kedua adalah dengan *manual labeling* yang dilakukan dengan meminta bantuan dari beberapa orang yang kiranya dapat memberikan label secara manual terhadap data. Pada penelitian ini tahapan *labeling* dilakukan dengan menggunakan cara *manual labeling*. Hal ini disebabkan banyaknya data yang salah diterjemah pada saat menggunakan *automatic labeling* sehingga menyebabkan pelabelan menjadi rancu.

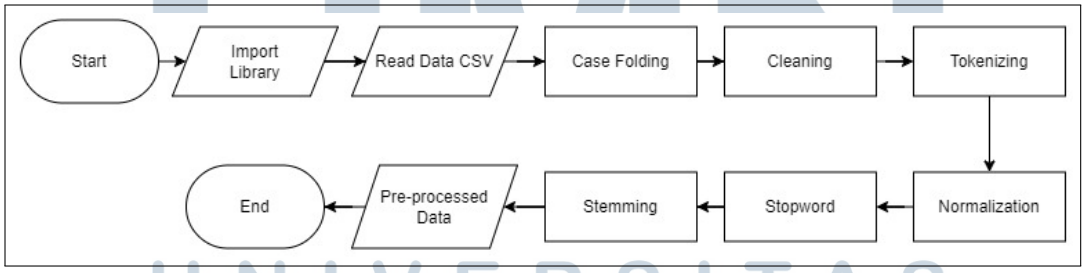
Manual labeling dilakukan dengan cara memahami makna dari data *tweet* yang telah dikumpulkan dan kemudian diberikan label negatif, netral, atau positif sesuai dengan pemahaman yang didapatkan. *Manual labeling* dilakukan oleh tiga orang, yaitu kedua orang tua penulis yang sudah menjadi wajib pajak dan satu kolega penulis yang belum menjadi wajib pajak. Dari ketiga pemberi label tersebut akan diambil label dengan perbandingan yang paling banyak. Dalam proses pemberian label, ketiga pemberi label akan diberikan petunjuk pengisian label yang didapatkan setelah melakukan wawancara melalui pesan singkat bersama dengan Ibu Jufika Martalina, S.Hum., M.Hum. selaku mantan dosen bahasa indonesia yang mengajar di StiKes Mercubakti Jaya Padang. Setelah wawancara dan disimpulkan, didapati hasil petunjuk pengisian label seperti yang dapat dilihat pada Tabel 3.1.

Tabel 3.1. Petunjuk Pengisian Label

Nomor	Petunjuk
1	**Baca dan pahami** data tweet yang berada pada kolom ”tweet” kemudian berikan label apakah tweet tersebut termasuk label ***Positive / Neutral/ Negative***.
2	***Positive:*** berupa saran maupun dukungan untuk tetap membayar pajak dan tidak setuju terhadap gerakan yang memboikot pajak.
3	***Neutral:*** berupa pertanyaan maupun berita.
4	***Negative:*** berupa hasutan atau ajakan maupun bentuk anarkis terhadap gerakan yang memboikot pajak.

3.1.3 Pre-processing

Kemudian penelitian dilanjutkan dengan tahapan *pre-processing*. *Pre-processing* dilakukan dengan tujuan untuk merubah data yang sebelumnya tidak terstruktur menjadi terstruktur atau rapih. Dalam penelitian ini terdapat beberapa tahapan *pre-processing* yang digunakan, dimulai dengan membaca data. Kemudian dilanjutkan dengan *case folding*, *cleaning*, *tokenizing*, *normalization*, *stopword*, *steaming*, dan kemudian didapatkan data yang telah dilakukan *pre-processing*. Adapun tahapan-tahapan *pre-processing* tersebut dapat dilihat dalam *flowchart pre-processing* pada Gambar 3.2 berikut.



Gambar 3.2. *Flowchart Pre-processing*

1. Case Folding

   Case folding merupakan proses merubah seluruh huruf besar atau *uppercase* yang terdapat data teks menjadi huruf kecil atau *lowercase*.

2. Cleaning

Cleaning merupakan proses membesihkan data dengan menghapus tulisan yang tidak diperlukan, seperti *link*, *emoticon*, *symbol*, *punctuation*, dan *username*.

3. Tokenizing

   Tokenizing merupakan suatu teknik yang dilakukan guna memecah kalimat menjadi kata-kata.

4. Normalization

   Normalization dilakukan guna memperbaiki kata-kata yang ditulis secara singkat maupun kata-kata tidak baku menjadi kata-kata yang baku.

5. Stopword

   Stopword digunakan dengan tujuan untuk menghapus kata yang tidak memiliki arti atau makna yang biasanya masuk kedalam kategori kata penghubung.

6. Stemming

   Stemming merupakan tahapan yang dilakukan untuk merubah data yang memiliki imbuhan menjadi kata dasar.

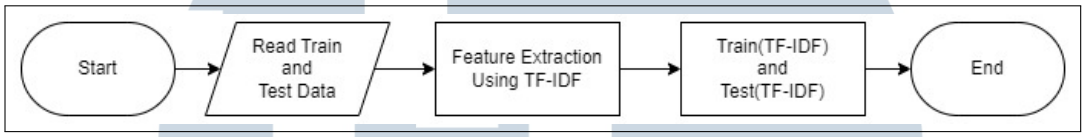
3.1.4 Train and Test Data Splitting

Setelah merapihkan data yang dilakukan pada tahapan *pre-processing* dan pemberian label pada tahapan *labeling*, penelitian dilanjutkan dengan tahapan *train and test data splitting*. Pada tahapan ini data yang telah terstruktur atau rapih akan dibagi menjadi data latih (*train*) dan data uji (*test*). Pembagian data dilakukan ke dalam tiga skenario perbandingan, yaitu 80:20, 70:30, dan 60:40.

3.1.5 Feature Extraction

Feature extraction atau ekstraksi fitur merupakan sebuah proses pengurangan dimensi di mana akan mereduksi kumpulan awal data mentah menjadi sekumpulan data yang lebih mudah dikelola untuk diproses. Model *machine learning* condong tidak dapat memahami data yang bukan numerik, sehingga data yang bukan merupakan data numerik tersebut perlu diubah kedalam bentuk numerik [36]. *Feature extraction* dilakukan menggunakan TF-IDF dengan

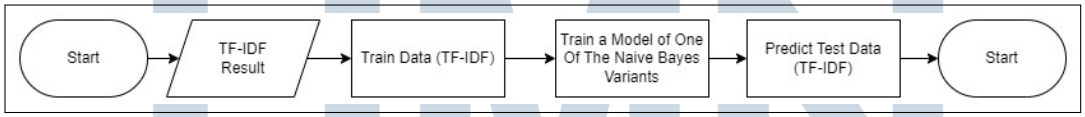
bantuan salah satu *library* python, yaitu scikit-learn. Proses *feature extraction* dalam penelitian kali ini dimulai dengan membaca data yang telah dibagi menjadi data latih (*Train*) dan data uji (*Test*) yang kemudian dilanjutkan dengan mengektrasi fitur menggunakan TF-IDF dari data tersebut. Setelah *feature extraction* dilakukan, akan didapati data latih (*Train*) dan data uji (*Test*) yang telah diektrasi fitur. Adapun proses *feature extraction* dalam bentuk *flowchart* dapat dilihat pada Gambar 3.3.



Gambar 3.3. *Flowchart Feature Extraction*

3.1.6 Apply Naive Bayes

Penelitian dilanjutkan dengan pengaplikasian model klasifikasi. Setelah mendapatkan data latih (*Train*) dan data uji (*Test*) yang telah dilakukan ektrasi fitur menggunakan TF-IDF. Data latih (*Train*) yang telah dilakukan ektraksi fitur tersebut akan digunakan untuk melatih model salah satu varian Naive Bayes. Kemudian hasil dari model yang telah diklasifikasi tersebut akan melakukan prediksi terhadap data uji (*Test*) yang telah dilakukan TF-IDF. Adapun *flowchart* dari proses tahapan *Apply Naive Bayes* dapat dilihat pada Gambar3.4



Gambar 3.4. *Flowchart Apply Naive Bayes*

3.1.7 Evaluation

Tahapan terakhir dalam penelitian kali ini adalah evaluasi atau *evaluation*. *Evaluation*. Pada penelitian ini, *evaluation* dilakukan menggunakan confusion matrix terhadap ketiga varian Naive Bayes dalam seluruh skenario perbandingan menggunakan tabel yang terdiri dari positif, netral, dan negatif. Tahapan *evaluation* menggunakan confusion matrix dilakukan untuk mendapatkan skor *accuracy*, *precission*, *recall*, serta *F1-score* guna mengetahui performa dari ketiga varian Naive Bayes dalam tiga skenario perbandingan data yang berbeda.