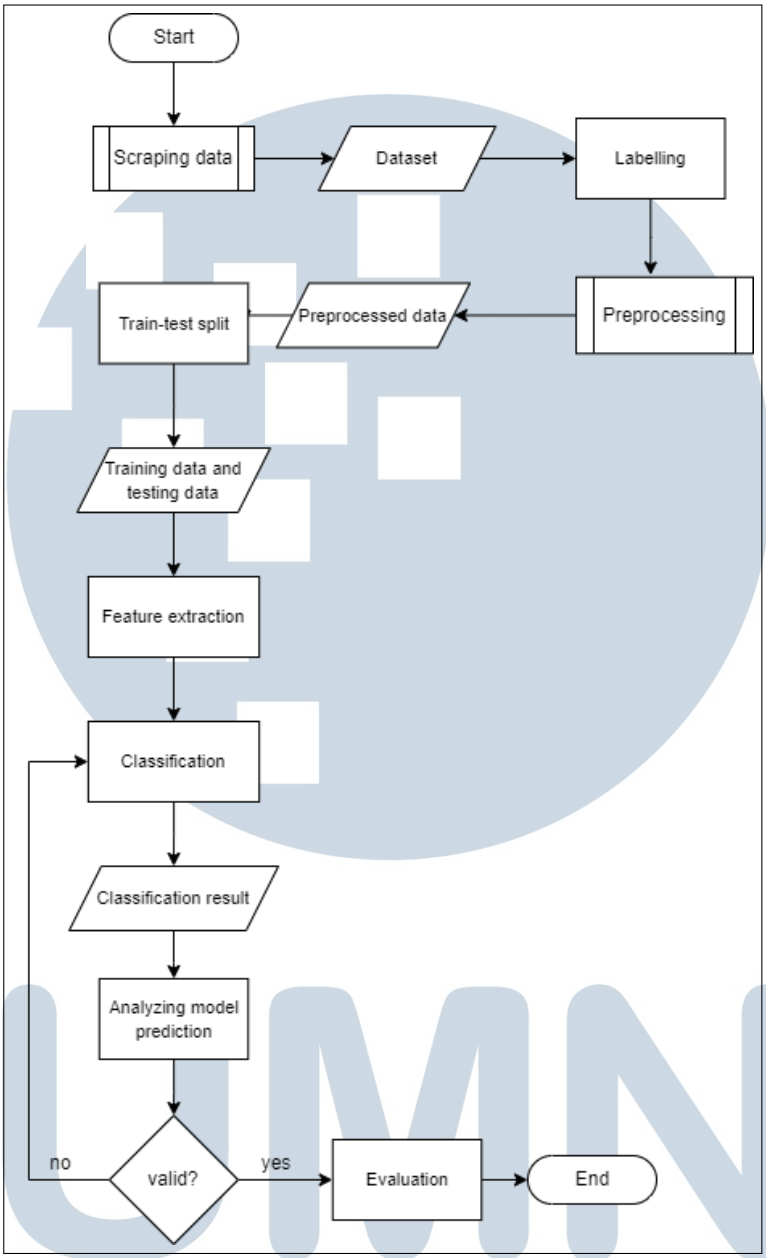
BAB 3
METODOLOGI PENELITIAN

3.1 Metodologi Penelitian

Metode penelitian ini memprediksi hasil yang dapat menyatakan apakah obrolan tersebut merupakan obrolan yang mengandung konten SARA atau tidak. Data yang digunakan untuk penelitian adalah data yang diambil dari grup Discord. Tahapan penelitian terdiri dari Tahap Pengumpulan Data, Tahap *Preprocessing*, Tahap *Text Transformation*, Tahap Klasifikasi, dan Tahap Evaluasi.
Berikut tahapan penelitian yang dirangkai dalam flowchart pada Gambar 3.1.



Gambar 3.1. Tahapan penelitian

3.1.1 Scraping Data

Scraping Data merupakan tahap pengumpulan data (*data collection*) yang akan digunakan sebagai data latih dan data uji. Data diambil dari 5 grup Discord dengan pengambilan 100 data pada setiap grup, sehingga total data yang diambil adalah 500 data. Data diambil dengan menggunakan teknik *scraping* menggunakan website Export Comments https://exportcomments.com/. Data yang diperoleh dari

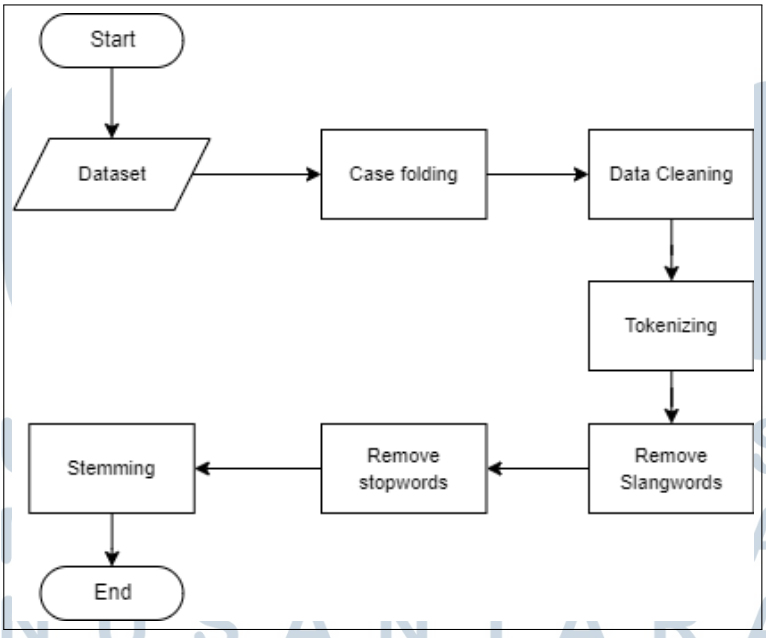
hasil *scraping* menggunakan website Export Comments berupa file dengan format csv.

3.1.2 Labeling

Labeling merupakan proses pelabelan data yang akan dibedakan berdasarkan kategori. Proses pelabelan data menggunakan dua kategori, yaitu `sara` dan `nsara`. Proses *labeling* dilakukan secara manual oleh peneliti, kalimat yang dikategorikan sebagai `sara` merupakan kalimat yang merupakan sindiran/pelecehan yang mengatasnamakan Suku, Agama, Ras, dan Antargolongan berdasarkan pengertian SARA pada jurnal "Comparison of SARA Issues Sentiment between Online News Portal and Social Media Towards the 2019 Election" [22].

3.1.3 Tahap Preprocessing

Preprocessing merupakan tahap memproses data agar data yang didapatkan lebih terstruktur dan mempermudah proses perhitungan. Gambar 3.2 menampilkan tahapan *preprocessing* yang terdiri dari tahap *case folding*, *data cleaning*, *tokenizing*, *slangword removal*, *stopwords removal*, dan *stemming*

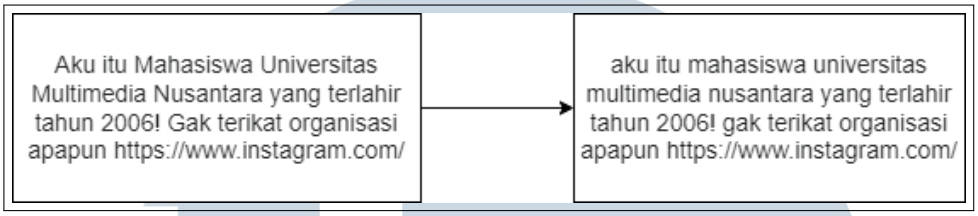


Gambar 3.2. Proses *preprocessing*

1. ***Case Folding***

   Case folding adalah proses mengubah seluruh kalimat yang menggunakan

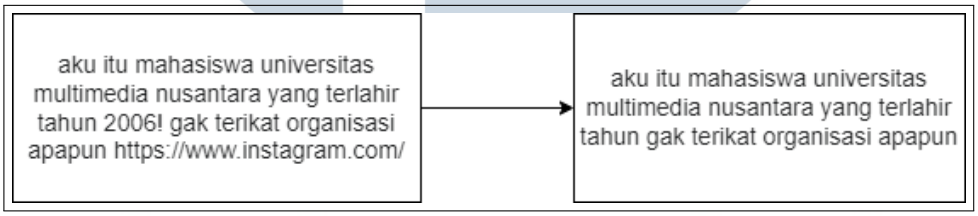
huruf kapital menjadi huruf kecil. Berikut Gambar 3.3 menampilkan contoh proses *case folding*:



Gambar 3.3. Proses *case folding*

2. ***Data Cleaning***

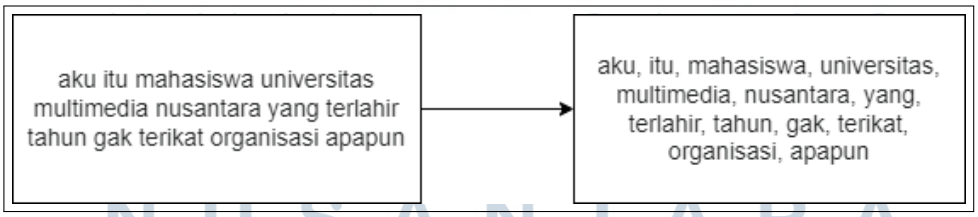
   Pemrosesan data berupa *data cleaning*, yaitu pembersihan data dengan menghapus angka, *hyperlink*, tanda baca atau *punctuation*, serta menghapus *whitespace leading* dan *trailing* dengan menggunakan fungsi *regex substraction*. Proses *cleaning data* menggunakan library NLTK atau *Natural Language Tool Kit*. Berikut Gambar 3.4 menampilkan contoh proses *cleaning*:



Gambar 3.4. Proses *cleaning*

3. ***Tokenizing***

   Tokenizing adalah proses memisahkan kalimat menjadi bentuk kata berdasarkan tiap kata penyusunnya. Berikut Gambar 3.5 menampilkan contoh proses *tokenizing*:

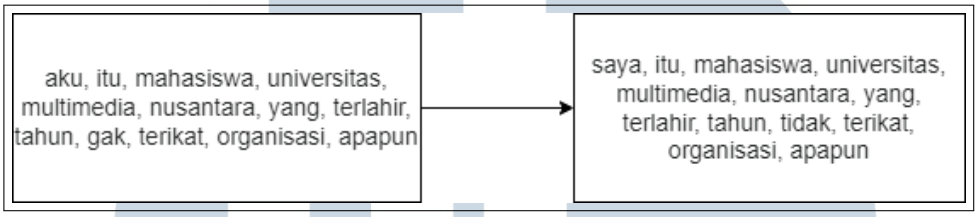


Gambar 3.5. Proses *tokenizing*

4. ***Slangwords Removal***

   Slangwords removal adalah proses menghilangkan kata slang serta mengu-

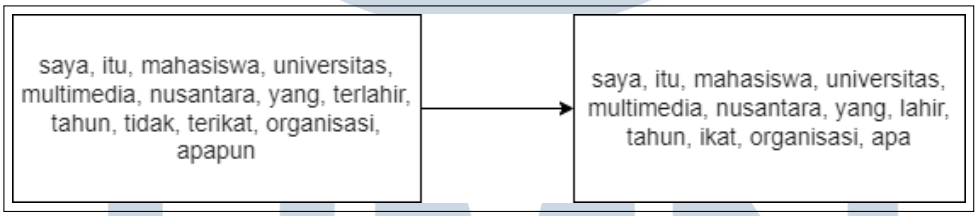
bah kata-kata yang menggunakan bahasa tidak baku menjadi bahasa baku. Proses *slangwords removal* menggunakan *slangwords dictionary* dalam format JSON yang dikumpulkan secara manual. Berikut contoh Gambar 3.6 menampilkan contoh proses *Slangwords removal*:



Gambar 3.6. Proses *slangword removal*

5. ***Stopwords Removal***

   Stopwords removal, yaitu mengurangi jumlah *token* dengan menghapus kata yang sering muncul namun kurang memiliki makna. *Stopwords removal* dilakukan dengan menggunakan library NLTK atau Natural Language Tool Kit. Berikut Gambar 3.7 menampilkan contoh proses *stopwords removal*:



Gambar 3.7. Proses *stopwords removal*

6. ***Stemming***

   Stemming merupakan proses untuk mengubah kata yang memiliki imbuhan menjadi kata dasar, dengan cara menghilangkan imbuhan yang terdapat dalam kata tersebut. Proses stemming dilakukan dengan menggunakan library Sastrawi. Berikut Gambar 3.8 menampilkan contoh proses stemming:

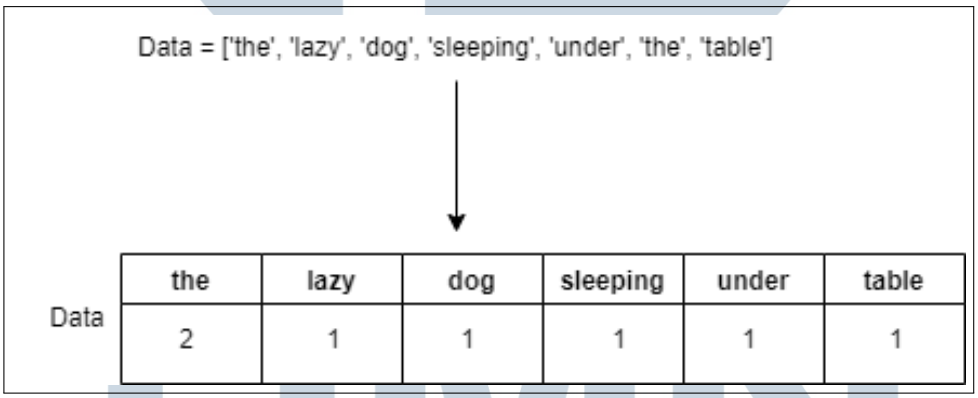
Gambar 3.8. Proses *stemming*

3.1.4 Train-test Split

Train-test split merupakan proses pembagian data. Data yang telah melewati tahap *preprocessing* akan dilakukan pembagian menjadi data latih dan data uji. Data latih digunakan untuk melatih algoritma dalam mengklasifikasikan model. Sedangkan data uji digunakan untuk menguji performa dari model yang telah dilatih. Pembagian data latih adalah 70 persen dan data uji adalah 30 persen.

3.1.5 Feature Extraction

Tahap ekstraksi fitur menggunakan algoritma CountVectorizer. Tahap ini dilakukan perubahan dari fitur teks menjadi sebuah representasi vector untuk dilakukan perhitungan pada tahap klasifikasi dengan cara mengubah data yang berupa *token-token* kata menjadi *vector*. Gambar 3.9 merupakan contoh proses ekstraksi fitur dengan CountVectorizer.



Gambar 3.9. Proses *countvectorizer*

3.1.6 Classification

Proses klasifikasi dengan algoritma Naïve Bayes Classifier menggunakan library sklearn serta mengimpor library bernama MultinomialNB yang terdapat dalam bahasa pemrograman python.

3.1.7 Analyzing Model Prediction

Setelah melalui proses klasifikasi, dilakukan tahap prediksi terhadap model dengan menulis fungsi untuk memprediksi suatu kata yang termasuk ke dalam kat-

egori SARA atau bukan. Apabila hasil prediksi valid, maka penelitian akan lanjut ke tahap selanjutnya, yaitu evaluasi.

3.1.8 Evaluation

Tahap ini dilakukan untuk mengevaluasi baik atau buruknya metode klasifikasi yang telah digunakan. Tahap evaluasi diukur dengan menggunakan *Confusion Matrix*, yaitu evaluasi berdasarkan hasil *accuracy*, *precision*, *recall*, dan *F1-Score*.