BAB 3

METODOLOGI PENELITIAN DAN PERANCANGAN SISTEM

3.1 Metodologi Penelitian

Metodologi penelitian "Implementasi Algoritma Sentencepiece untuk Meningkatkan Performa Klasifikasi Naïve Bayes Classifier pada Artikel Berita" terdiri dari beberapa tahapan yaitu pengumpulan data, *text preprocessing*, pelatihan model, klasifikasi, dan evaluasi.

3.1.1 Pengumpulan Data

Data dalam bentuk artikel berita yang digunakan pada penelitian ini diambil dari portal berita Kompas Indonesia dengan menggunakan *web scraping*. Data artikel berita tersebut terdiri dari 12 kategori berita yaitu, kesehatan, makanan, edukasi, bola, homey, keuangan, otomotif, property, sains, teknologi, travel, dan lifestyle. Setiap kategori terdiri dari 1500 artikel.

3.1.2 Text Preprocessing

Tahap *text preprocessing* merupakan tahap untuk menyiapkan data artikel berita agar dapat diolah dan dimengerti oleh komputer. Tahapan *text preprocessing* terdiri dari:

1. *Case folding*

   Dalam setiap artikel berita, tidak ada konsistensi terhadap penulisan setiap huruf. Setiap huruf dapat ditulis dalam huruf besar ataupun huruf kecil. Maka dari itu, dilakukan proses *case folding* untuk mengubah setiap huruf besar menjadi huruf kecil dan setiap karakter yang bukan merupakan huruf akan dihilangkan dan dianggap sebagai delimiter.

2. *Tokenizing*

Tahapan *tokenizing* merupakan proses memecah kalimat menjadi kata tunggal. Pada penelitian ini, digunakan dua metode tokenisasi yang berbeda. Metode tokenisasi yang pertama menggunakan *library* dari *natural language toolkit* (NLTK). Untuk metode tokenisasi yang kedua menggunakan *library* dari *sentencepiece.*

3. *Filtering*

   Filtering merupakan proses memilih kata-kata penting yang dapat dilakukan dengan membuang kata-kata yang kurang penting (*stoplist*) atau mengambil kata-kata yang penting (*wordlist*). Pada penelitian ini digunakan metode *stoplist*. *Library* NLTK menyediakan kumpulan kata yang termasuk ke dalam kata-kata yang kurang penting (*stopword list*) seperti "yang", "ini", "kemudian", dan sebagainya

4. *Stemming*

   Stemming merupakan proses untuk menghilangkan imbuhan. Imbuhan dapat terletak di awal, tengah ataupun akhir suatu kata. Proses *stemming* akan mengubah kata berimbuhan tersebut kembali ke kata dasarnya.

3.1.3 Pelatihan Model

Model *sentencepiece* yang dilatih pada penelitian ini menggunkan dua tipe yaitu *byte-pair-encoding* dan *unigram language model.* Masing-masing tipe model akan dilatih dengan ukuran kosakata (*vocabulary size*) 2000, 4000, 8000, 16000, dan 32000. Jadi total model *sentencepiece* yang dilatih adalah 10.

3.1.4 Klasifikasi

Proses klasifikasi akan dijalankan menggunakan *multinomial naïve bayes* dan pembobotan kata akan menggunakan TF-IDF. Kedua hal tersebut akan

dijalankan menggunakan bantuan *library pipeline*. *Library pipeline* digunakan untuk menggabungkan pembobotan kata menggunakan TF-IDF dan algoritma klasifikasi *multinomial naïve bayes* sehingga dapat dilakukan *cross validation* dan penyetelan *hyperparameter*. *Pipeline* tersebut akan dimasukkan sebagai parameter dari *library gridsearchcv* yang digunakan untuk melakukan *exhaustive search*.

3.1.5 Evaluasi

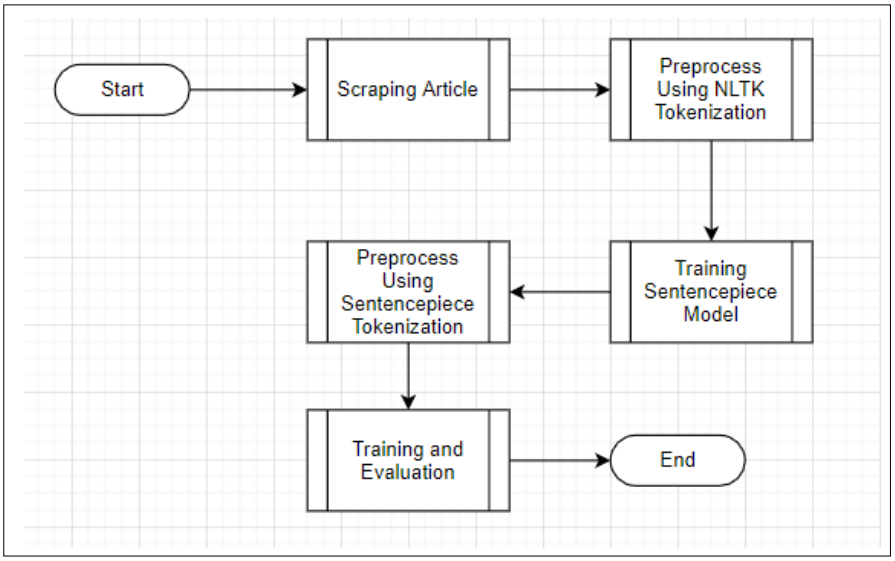
Dari setiap model yang telah melalui tahapan klasifikasi, akan menghasilkan skor terbaik yang akan dibandingkan antara satu dan yang lain. Setiap hasil juga akan memperlihatkan parameter yang digunakan untuk menghasilkan skor terbaik tersebut. Semua skor akan ditampilkan dalam bentuk grafik agar lebih mudah untuk dilihat

3.2 Perancangan Sistem

Untuk mempermudah pemahaman tentang urutan langkah proses yang dilakukan, maka digunakan *flowchart*. Terdapat 5 tahapan utama yaitu *scraping article, preprocess using NLTK tokenization, training sentencepiece model, preprocess using sentencepiece tokenization,* dan *training and evaluation*. Tahapan tersebut dapat dilihat pada Gambar 3.1.

3.2.1 Flowchart Umum

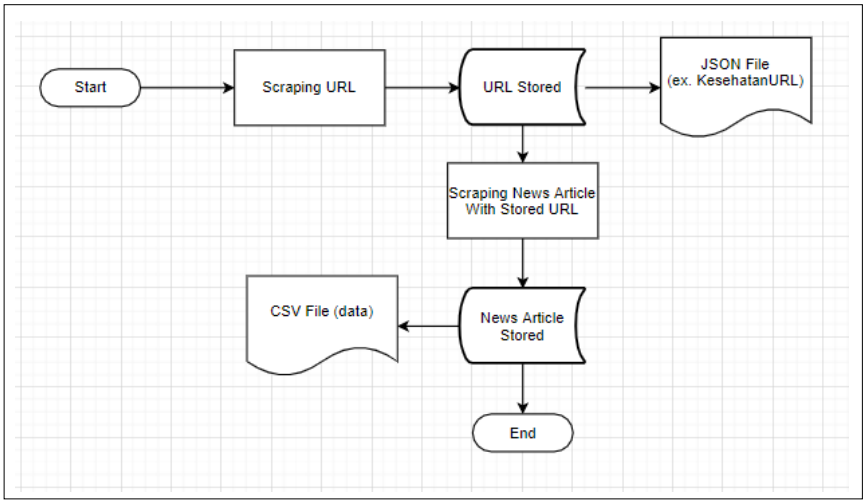
Proses yang dilakukan secara umum pada penilitian ini terbagi kedalam 5 *predefined process,* yaitu *scraping article, preprocess using* NLTK *tokenization, training sentencepiece model, preprocess using sentencepiece tokenization,* dan *training and evaluation*. Tahapan tersebut dapat dilihat pada Gambar 3.1.



Gambar 3.1 Flowchart Umum

3.2.2 Flowchart Scraping Article

Pada tahapan *scraping article*, dilakukan proses pengumpulan URL dari artikel berita yang disimpan dalam bentuk JSON dan URL tersebut digunakan untuk mengumpulkan artikel berita. Artikel berita tersebut kemudian disimpan dalam bentuk CSV. Tahapan *scraping article* dapat dilihat pada Gambar 3.2.

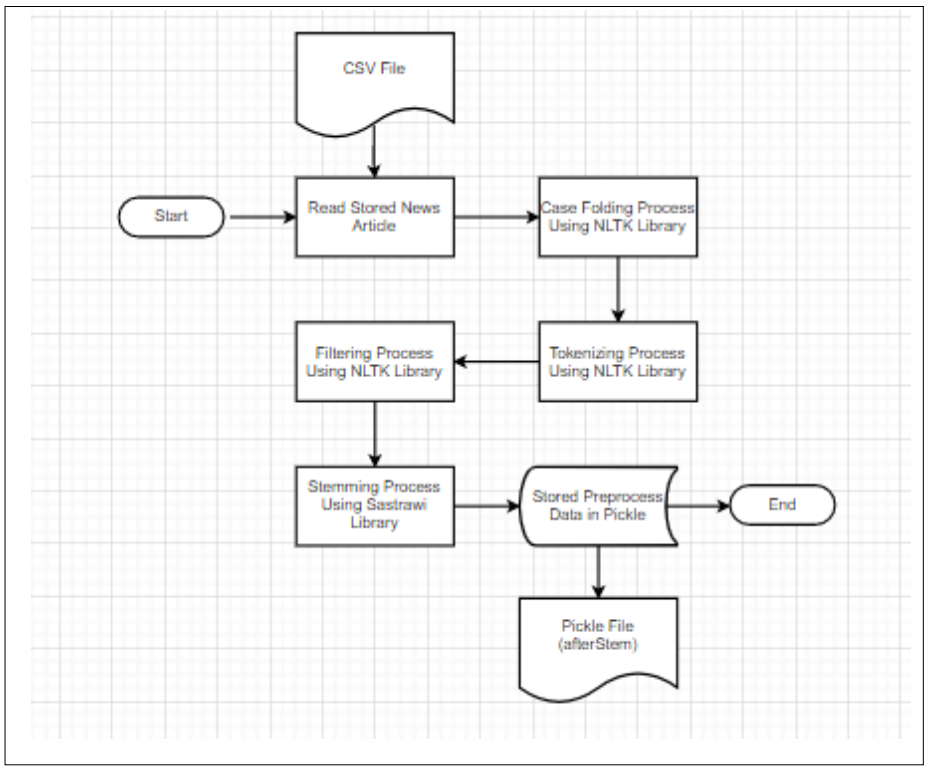


Gambar 3.2 Tahapan Scraping Article

3.2.3 Flowchart Preprocess Using NLTK Tokenization

Berdasarkan Gambar 3.1, tahapan setelah *scraping article* adalah *preprocess using* NLTK *tokenization*. Pada tahapan ini, data yang telah

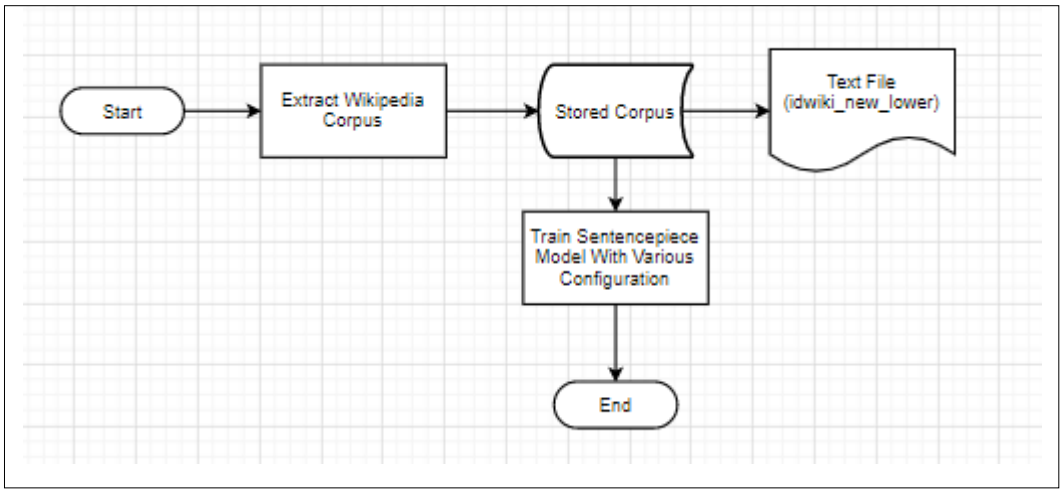
dikumpulkan akan melalui proses seperti *case folding, tokenization, filtering,* dan *stemming*. Setelah proses *stemming* selesai, hasilnya disimpan dalam bentuk *pickle.* Tahapan ini dapat dilihat pada Gambar 3.3.



Gambar 3.3 Tahapan Preprocess Using NLTK Tokenization

3.2.4 Flowchart Training Sentencepiece Model

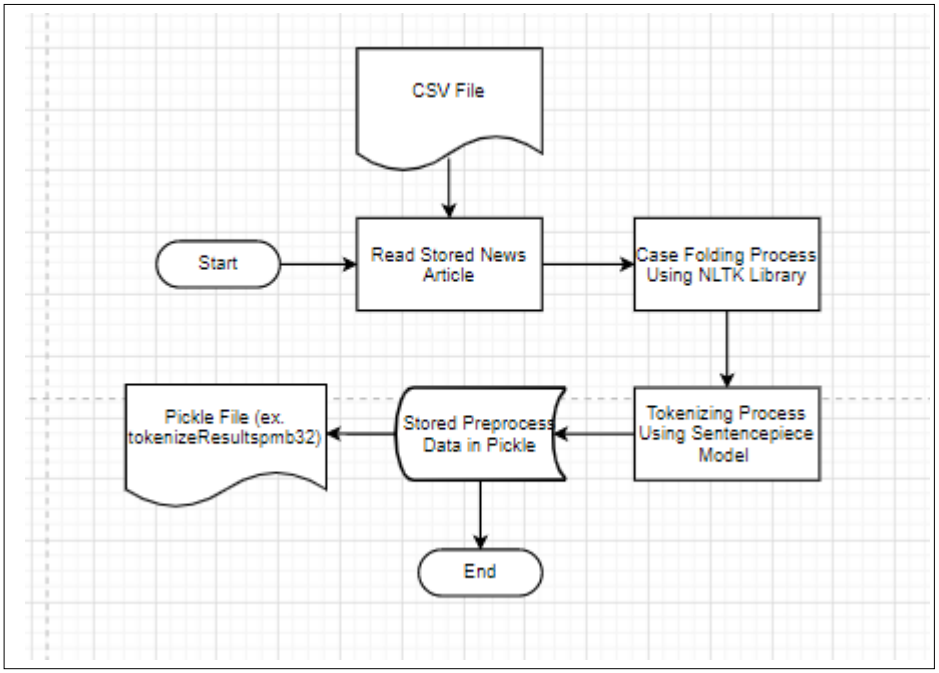
Berdasarkan Gambar 3.1, tahapan berikutnya adalah *training sentencepiece model.* Data yang digunakan untuk melatih model *sentecepiece* diambil dari Wikipedia dan dilakukan ekstraksi menjadi teks. Kemudian dilakukan pelatihan model *sentencepiece* dengan beberapa konfigurasi seperti *model type* dan *vocabulary size. Model type* terdiri dari *unigram language model* dan *byte-pair-encoding*. Untuk *vocabulary size* bervariasi dari 2000, 4000, 8000, 16000 dan 32000. Tahapan ini dapat dilihat pada Gambar 3.4.



Gambar 3.4 Tahapan Training Sentencepiece Model

3.2.5 Flowchart Preprocess Using Sentencepiece Tokenization

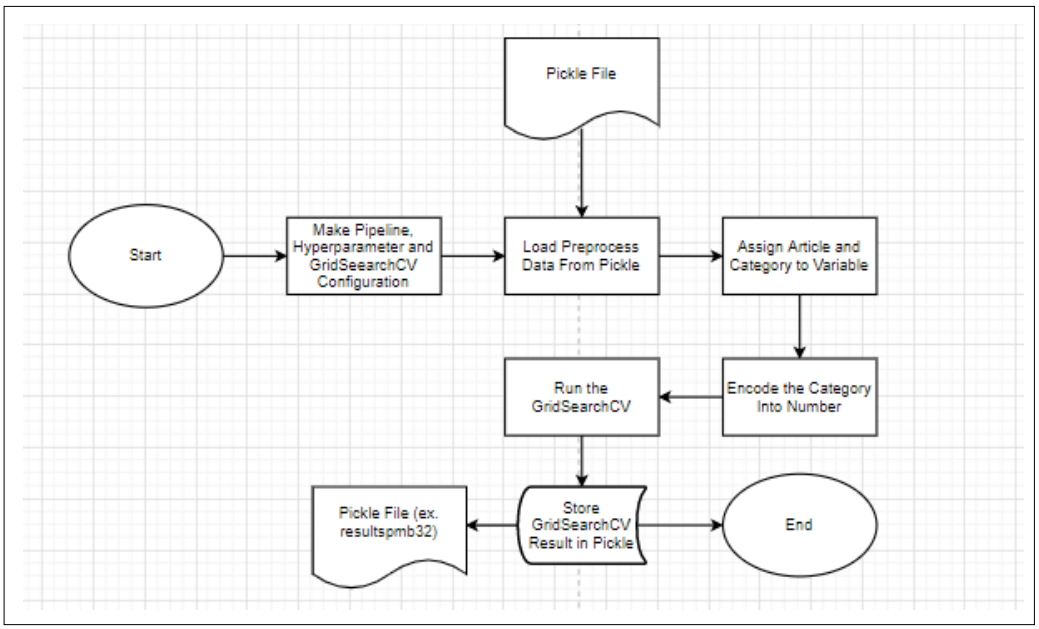
Setelah tahapan *training sentencepiece model,* tahapan berikutnya adalah *preprocess using sentecepiece tokenization*. Tahapan ini terdiri dari *case folding* dan *tokenization* menggunakan model *sentencepiece*. Setelah proses tokenisasi selesai, hasilnya disimpan dalam bentuk *pickle*. *Flowchart* tahapan ini dapat dilihat pada Gambar 3.5.



Gambar 3.5 Tahapan Preprocess Using Sentencepiece Tokenization

3.2.6 Flowchart Training and Evaluation

Berdasarkan Gambar 3.1, proses terakhir adalah *training and evaluation*. Pada tahapan ini dilakukan proses pelatihan model klasifikasi dan evaluasi. Dibuat sebuah *pipeline*, *hyperparameter*, dan *gridsearch*. Kemudian hasil *preprocess* akan digunakan untuk pelatihan model klasifikasi. Kemudian hasilnya disimpan dalam bentuk *pickle*. Tahapan ini dapat dilihat pada Gambar 3.6.



Gambar 3.6 Tahapan Training and Evaluation