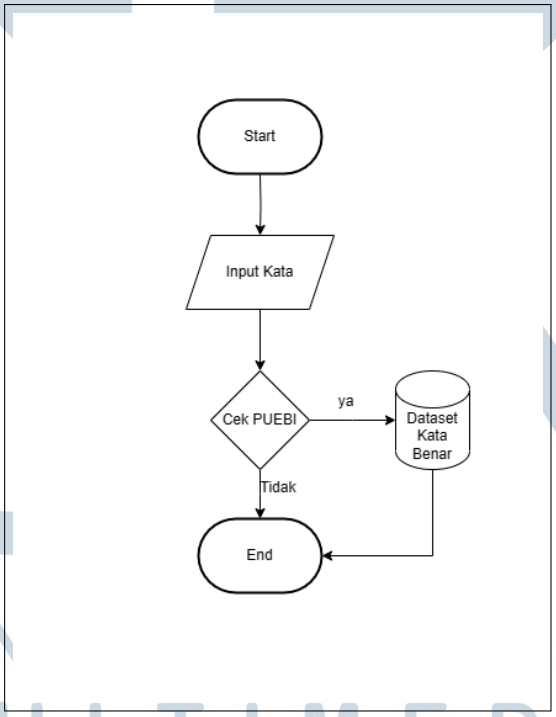
BAB 3
METODOLOGI PENELITIAN

Pada bagian ini akan dijabarkan tentang tahapan-tahapan yang dilakukan dalam penggunaan metode *Multinomial Naive Bayes* untuk mendeteksi kesalahan tik pada artikel berita Tribunnews.

3.1 Pengumpulan Data

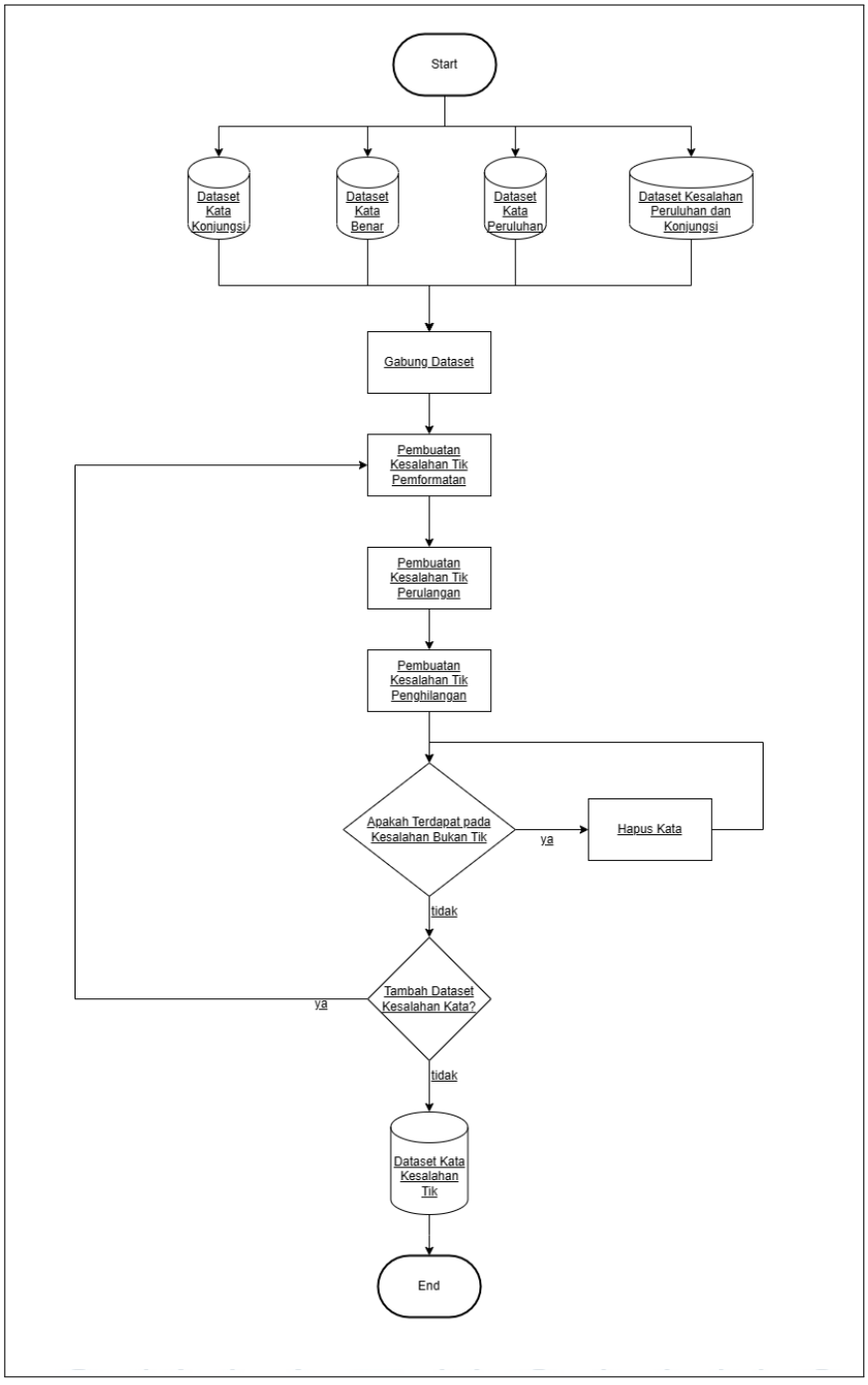
Pada tahapan ini akan dikumpulkan sejumlah data, data *training* yang berupa Kata berlabel benar atau terdapat dalam PUEBI dan kata kesalahan tik. Data *test* yang diambil melalui cara *web crawling* secara manual, mencari dataset di kaggle atau website penyedia dataset lainnya, dan mengisikan dataset secara manual melalui *excel* ataupun *text editor*.



Gambar 3.1. Flowchart Pengumpulan Data Manual

Pada Gambar 3.1 adalah alur pengumpulan kata yang dilakukan secara manual. Pada mulanya terdapat input kata yang ditemukan dalam artikel yang diperiksa apakah penulisannya merujuk ke PUEBI atau tidak dan jika terdapat kata yang termasuk PUEBI maka akan dimasukkan kedalam dataset awal. Proses pengecekan tiap kata yang dimasukkan dilakukan secara manual dengan

membandingkan kata yang dimasukkan dengan penulisan PUEBI. Dalam proses ini sumber utama dataset tambahan untuk kata dikumpulkan dari artikel berita Tribunnews. Dataset kata benar selain ditambahkan secara manual juga didapatkan dari penelitian lain seperti dataset kata konjungsi dan kata peluluhan. Dataset konjugsi berisikan kata penghubung sedangkan dataset peluluhan berisikan kata yang sudah luluh atau kata yang mengalami perubahan bentuk saat diberikan imbuhan, contohnya kata ”tanam” sebagai kata dasar berubah menjadi ”menanam” dimana ”t” meluluh menjadi ”n”.



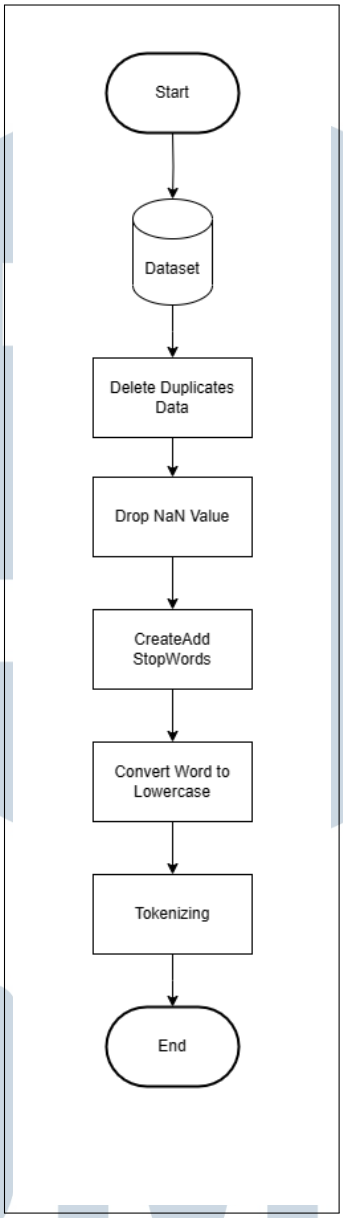
Gambar 3.2. Flowchart Pembuatan Data Kata Kesalahan Tik

Pada gambar 3.2 memperlihatkan alur pembuatan kata kesalahan tik. Pengumpulan dataset untuk kata kesalahan tik dibuat secara otomatis menggunakan kode pemrograman. Proses pembuatan dataset ini berdasarkan dengan landasan teorikesalahan tik dan dibuat setidaknya satu buah dataset kata kesalahan setiap jenis kesalahan tik yang ada. Data set yang didapatkan digabungkan menjadi

satu dataset besar yang berisi kata benar dan dari dataset ini melewati 4 proses pembuatan kata dimana satu proses membuat satu dataset kesalahan kata dari dataset kata benar. Setelah kata kesalahan tik dibuat dilakukan pengecekan apakah ada kata kesalahan tik yang termasuk bentuk kesalahan penulisan lainnya seperti kesalahan kata peluluhan dan konjungsi, bila ada kata tersebut akan dihapus dari dataset yang dibuat dan dilakukan pengecekan ulang untuk memastikan kata sudah terhapus. Proses ini juga memungkinkan untuk membuat data kata salah berkali-kali untuk memperbesar ukuran dataset yang digunakan.

3.2 Preprocessing Data

Pada tahap ini data-data hasil dari *web crawling* secara manual, dataset dari direktori dan dataset *excel* akan digabungkan menjadi satu buah dataset. Data train akan dilakukan beberapa proses *text processing* untuk mengubah data train menjadi data yang siap untuk diolah. Berikut adalah *flowchart* dari langkah *Preprocessing*. Gambar 3.3 memperlihatkan alur proses *preprocessing*.

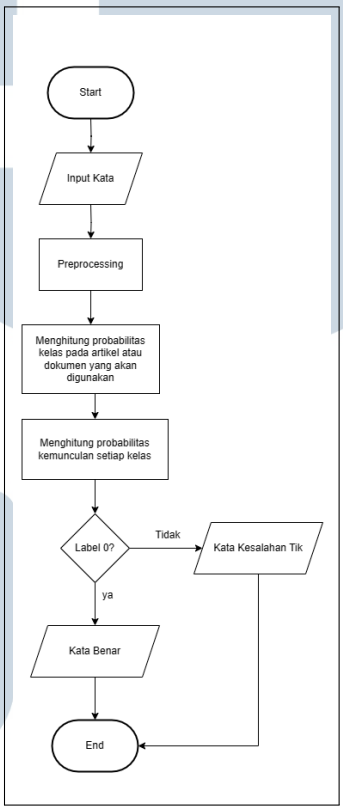


Gambar 3.3. Flowchart Preprocessing

Pada tahap awal terdapat 3 buah dataset yakni 1 dataset hasil Web Crawling dan 2 dataset yang digunakan pada penelitian sebelumnya tentang kata konjungsi dan peluluhan. Ketiga dataset ini digabungkan dan diberikan nilai label *correct* sedangkan untuk data kata yang mengalami kesalahan tik akan di buat berdasarkan macam-macam kesalahan tik kata dengan label *incorrect*. Dataset kemudian digabungkan menjadi satu dataset besar yang berisi kata kesalahan tik dan kata benar. Proses preprocessing akan menghapus angka dan semua tanda baca kecuali tanda (”-”) baik pada dataset maupun artikel. Data juga akan dibuat dalam bentuk token.

3.3 Pembangunan Model

Pada tahap ini akan menerapkan algoritma *Multinomial Naive Bayes* pada data hasil *preprocessing*. Nantinya keluaran dari tahap ini adalah sebuah model pembelajaran mesin yang sudah dapat mengklasifikasikan kata berdasarkan input yang berupa sebuah dokumen berita. Pada pembuatan model digunakan keseluruhan data dari dataset untuk dijadikan data *train*, hal ini dilakukan untuk memaksimalkan informasi kata yang diterima model karena model yang dibuat sangat bergantung pada dataset sehingga apabila ada kata yang tidak terdapat dalam model maka model tidak dapat mendeteksi kata tersebut.



Gambar 3.4. Flowchart Model *Multinomial Naive Bayes*

3.4 Evaluasi dan Testing

Setelah model dibangun maka akan dilakukan evaluasi untuk mengukur tingkat akurasi dari model yang telah dibuat dan juga untuk mengukur akurasi hasil deteksi maka akan dilihat menggunakan *f1-Scores*. Selain itu evaluasi juga akan dilakukan dengan menggunakan artikel. Evaluasi dengan artikel berita dimulai dengan membersihkan artikel dari nama orang atau kata bernilai unik lainnya,

namun dapat juga secara langsung memasukkan artikel tetapi keluaran dari artikel yang memiliki kata unik dilabeli kata *incorrect* sebagai tanda bahwa kata bernilai unik termasuk kesalahan kata tik. Artikel yang masuk program akan dihapus angka. tanda baca kecuali tanda baca (”-”) dan diubah menjadi token untuk setiap kata pada artikel. Token kata ini digunakan oleh model untuk menentukan apakah suatu kata termasuk kesalahan tik atau tidak dan juga proses evaluasi ini dilakukan per-artikel. Alur proses pembuatan model dari algoritma *Multinomial Naive Bayes* dapat dilihat pada gambar 3.4 .

3.5 Spesifikasi Sistem

Dalam pembuatan aplikasi, digunakan spesifikasi hardware sebagai berikut:

- CPU: Intel Core i7-7700HQ
- RAM: 24 GB
- OS: Windows 10

Perangkat lunak yang digunakan adalah sebagai berikut:

- Python 3.9.13
- Anaconda 2021.0.4