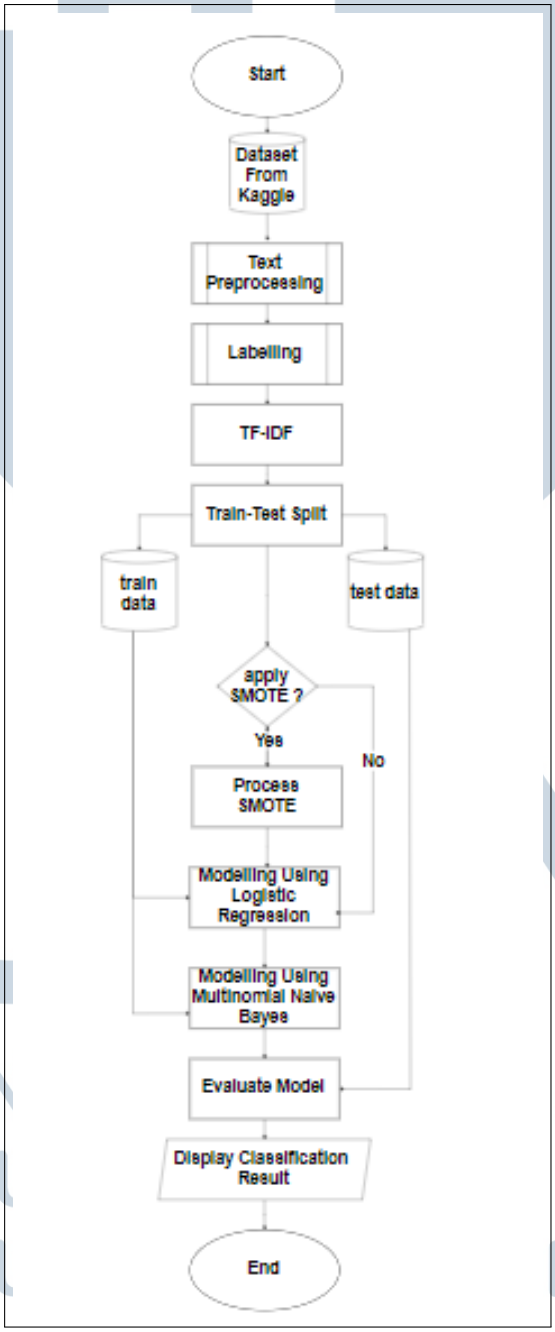
BAB 3
METODOLOGI PENELITIAN

3.1 Gambaran Umum Penelitian



Gambar 3.1. Gambaran Umum Penelitian

Seperti yang ditunjukkan pada gambar 3.1., pertama data berupa teks yang di *download* dari Kaggle akan di *preprocessing* untuk menghilangkan *noise* yang tidak diperlukan sehingga dapat dimodel dengan baik. Setelah itu data akan dilabel menjadi 2 label yaitu positif dan negatif. Setelah itu data akan dilakukan pembobotan dengan TF-IDF *Vectorizer*. Setelah itu akan dilakukan pemisahan antara data latih dan data uji. Setelah itu, data latih bisa dilakukan SMOTE terlebih dahulu untuk menyeimbangkan jumlah data dengan label positif dan negatif di dalam kumpulan data latih. Jika tidak maka data latih akan langsung di model dengan *Logistic Regression* dan *Multinomial Naive Bayes*. Setelah dimodel, data uji akan dibandingkan dengan hasil klasifikasi model untuk mengevaluasi seberapa bagus model tersebut.

3.2 Perancangan Sistem

Tahap-tahap yang dilalui pada perancangan sistem adalah sebagai berikut.

3.2.1 Preprocessing

Preprocessing adalah pemrosesan data yang dilakukan sebelum membuat model agar hasil pembuatan model menjadi lebih baik. Tahap-tahap yang dilalui pada *preprocessing* adalah sebagai berikut.

A Handling Duplicates

Pada tahap ini, baris data yang muncul lebih dari 1 dalam tabel akan dihapus. Setelah itu, akan di cek total baris data di dalam tabel setelah data duplikat dihapus. Tujuannya agar data yang dibuat model tidak berlebihan dan hasil klasifikasi menjadi lebih baik.

B Remove Empty Values

Pada tahap ini, baris baris dalam tabel yang mengandung data dengan nilai yang kosong akan dihapus. Setelah itu akan di cek lagi jumlah baris di dalam tabel setelah data dengan nilai yang kosong dihapus. Hal ini bertujuan agar tidak ada data yang tidak memiliki nilai.

C Casefolding

Pada Tahap ini, semua huruf besar yang ada pada data dirubah menjadi huruf kecil. Data yang sudah di dilakukan *casefolding* dapat dilihat pada tabel 3.1.

Tabel 3.1. Perbandingan Data Original dan yang Sudah di *Casefolding*

Original Data	Casefolded
Good price, works flawless in my Samsung S4! Normal SanDisk quality! That is why I go with this brand and only this brand..	good price, works flawless in my samsung s4! normal sandisk quality! that is why i go with this brand and only this brand..

D Data Cleaning

Pada tahap ini, komponen-komponen penyusun data yang tidak dibutuhkan seperti tanda baca dan *link* URL akan dihapus. Data yang sudah di dilakukan *cleaning* dapat dilihat pada tabel 3.2.

Tabel 3.2. Perbandingan Data Original dan yang Sudah di *Cleaning*

Original Data	Cleaned
Good price, works flawless in my Samsung S4! Normal SanDisk quality! That is why I go with this brand and only this brand..	good price works flawless in my samsung normal sandisk quality that is why go with this brand and only this brand

E Tokenization

Pada Tahap ini, kalimat-kalimat pada setiap baris data yang menjadi variabel independen (kalimat-kalimat ulasan produk) akan dipecah menjadi kata-kata yang menyusun setiap kalimat tersebut dan dipisahkan dengan tanda koma (dalam bentuk *list*). Data yang sudah di dilakukan *tokenization* dapat dilihat pada tabel 3.3.

Tabel 3.3. Perbandingan Data Original dan yang Sudah Dilakukan *Tokenizing*

Original Data	Tokenized
Good price, works flawless in my Samsung S4! Normal SanDisk quality! That is why I go with this brand and only this brand..	[’good’, ’price’, ’works’, ’flawless’, ’in’, ’my’, ’samsung’, ’normal’, ’sandisk’, ’quality’, ’that’, ’is’, ’why’, ’go’, ’with’, ’this’, ’brand’, ’and’, ’only’, ’this’, ’brand’]

F Stopword Removal

Pada Tahap ini, semua kata-kata yang terdapat pada *stoplist* bahasa inggris yaitu kata-kata yang tidak memiliki arti yang penting seperti ”*but*”, ”*as*” dan ”*are*” akan dihapus. Data yang sudah dilakukan *stopword removal* dapat dilihat pada tabel 3.4.

Tabel 3.4. Perbandingan Data Original dan yang Sudah Dilakukan *Stopword Removal*

Original Data	Stopword Removed
Good price, works flawless in my Samsung S4! Normal SanDisk quality! That is why I go with this brand and only this brand..	[’good’, ’price’, ’works’, ’flawless’, ’samsung’, ’normal’, ’sandisk’, ’quality’, ’go’, ’brand’, ’brand’]

G Lemmatization

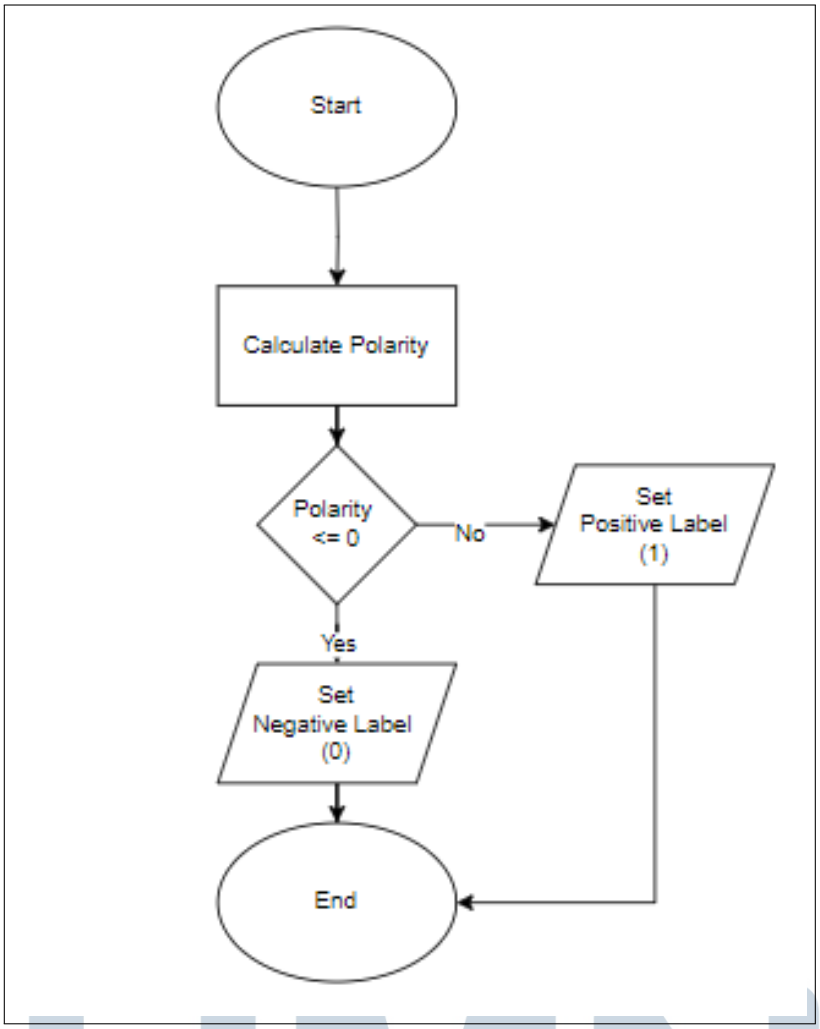
Pada Tahap ini, kata-kata dirubah menjadi kata dasar dengan mengelompokkan kata-kata tersebut berdasarkan jenis katanya atau bisa disebut juga *lemma*. *Lemma* dari sebuah kata ada 4 jenis yaitu *adjective* (kata sifat), *verb* (kata kerja), *noun* (kata benda) dan *adverb* (kata keterangan). Setelah dilakukan *lemmatization*, kata-kata digabung kembali menjadi 1 kalimat dalam bentuk *string*. Data yang sudah di dilakukan *lemmatization* dapat dilihat pada tabel 3.5.

Tabel 3.5. Perbandingan Data Original dan yang Sudah Dilakukan *Lemmatization*

Original Data	Lemmatized
Good price, works flawless in my Samsung S4! Normal SanDisk quality! That is why I go with this brand and only this brand..	good price work flawless samsung normal sandisk quality go brand brand

3.2.2 Labelling

Proses *labelling* data ditunjukkan pada gambar 3.2.



Gambar 3.2. *Labelling*

Pertama, data yang sudah di *preprocessing* akan dihitung nilai polaritas dari setiap baris data. Polaritas menggambarkan seberapa positif atau seberapa negatif opini seseorang. Jika polaritasnya kurang dari atau sama dengan 0 maka data akan dikategorikan sebagai sentimen berlabel negatif sementara jika polaritasnya lebih dari 0 maka data dapat dikatergorikan sebagai sentimen berlabel positif.

3.2.3 TF-IDF

TF-IDF adalah salah satu metode ekstraksi fitur. TF-IDF dihitung dengan mengalikan nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). Nilai TF adalah frekuensi seberapa sering sebuah kata muncul dalam sebuah dokumen. Nilai IDF adalah frekuensi banyaknya jumlah dokumen dimana sebuah kata itu muncul. Nilai TF-IDF sebuah kata akan tinggi jika sebuah kata sering muncul dalam satu dokumen tapi jarang muncul di dokumen lain.

3.2.4 Modelling dan Evaluasi

Ada 2 skenario pembuatan model yaitu pembuatan model dengan SMOTE dan tanpa SMOTE.

A Modelling Tanpa SMOTE

Pada pembuatan model tanpa SMOTE, pertama data akan dilakukan *train-test split* dengan 2 rasio data latih dan data uji yaitu 80:20 dan 70:30. Lalu data latih akan dilakukan pemodelan dengan *Logistic Regression* dan *Multinomial Naive Bayes*. Setelah itu data uji akan dilakukan evaluasi dengan membandingkan dengan hasil klasifikasi model untuk mengetahui seberapa bagus hasil kerja model. Hasil kerja model akan diukur menggunakan *confusion matrix*. Metrik-metrik yang akan diukur untuk menguji kinerja model adalah akurasi, presisi, *recall* dan *f1-score*.

B Modelling dengan SMOTE

Pada pembuatan model dengan SMOTE, pertama data akan dilakukan *train-test split* dengan 2 rasio data latih dan data uji yaitu 80:20 dan 70:30. Setelah itu, data latih akan di *resample* menggunakan SMOTE untuk menyeimbangkan jumlah data berlabel positif dan negatif pada kumpulan data latih. Lalu data latih akan dilakukan pemodelan dengan *Logistic Regression* dan *Multinomial Naive Bayes*. Setelah itu data uji akan dilakukan evaluasi dengan membandingkan dengan hasil klasifikasi model untuk mengetahui seberapa bagus hasil kerja model. Hasil kerja model akan diukur menggunakan *confusion matrix*. Metrik-metrik yang akan diukur untuk menguji kinerja model adalah akurasi, presisi, *recall* dan *f1-score*.