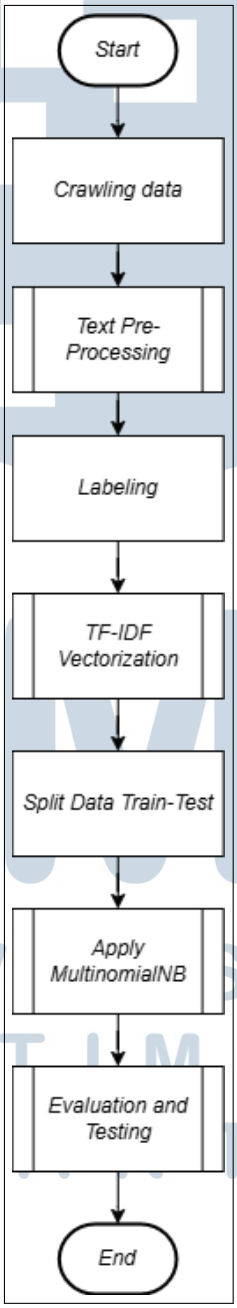
BAB 3
METODOLOGI PENELITIAN

3.1 Gambaran Umum Alur Penelitian

Bab ini berisikan keseluruhan rancangan jalannya setiap langkah penelitian, Gambar 3.1 merupakan alur penilitian yang disusun dalam bentuk *flowchart*.



Gambar 3.1. *Flowchart* utama

3.2 Metodologi Penelitian

Dalam penelitian ini dilakukan implementasi algoritma *Multinomial Naive Bayes* untuk analisis sentimen komentar YouTube terhadap video renovasi musik Eka Gustiwana, dengan beberapa metode penelitian sebagai berikut:

3.2.1 Studi Literatur

Studi literatur dilakukan untuk memperoleh informasi yang relevan guna mendukung penelitian ini. Fokus utama meliputi pengambilan data komentar dari YouTube menggunakan YouTube API, pemahaman algoritma *Multinomial Naive Bayes* untuk klasifikasi teks, metode TF-IDF untuk ekstraksi fitur, serta *confusion matrix* sebagai alat evaluasi performa. Selain itu, metode SMOTE digunakan untuk mengatasi ketidakseimbangan data dengan melakukan oversampling pada kelas minoritas agar model memiliki performa yang lebih baik. Teknik validasi silang *K-Fold* juga diterapkan untuk menguji kestabilan dan generalisasi model secara menyeluruh pada data yang terbagi menjadi beberapa subset. Referensi teori diperoleh dari berbagai sumber seperti penelitian terdahulu, dokumentasi resmi, video pembelajaran, dan jurnal online yang membahas analisis sentimen terhadap komentar di media sosial.

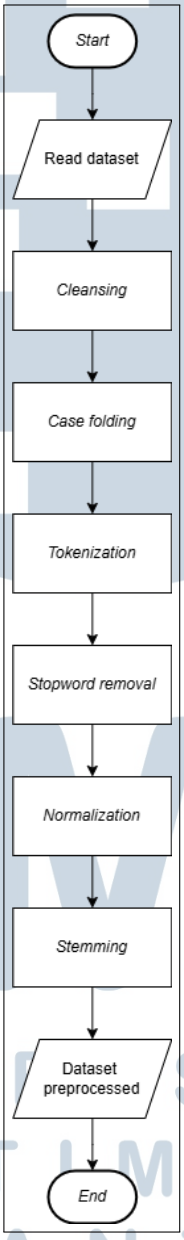
3.2.2 Crawling Data

Pengumpulan data dilakukan dengan menarik data *comment* video di Youtube menggunakan Google API. Video dipilih dari *channel* Eka Gustiwana yang berjudul ”RENOVASI MUSIK — TURU NIGHT (Flashlight by Jessie J)”. Data *comment* yang terambil adalah sebanyak 1000 data yang dipilih dari hasil proses *crawling*.

3.2.3 Text-Preprocessing

Gambar 3.2 menunjukkan tahapan dalam proses *text preprocessing*. Pada tahap ini, data diproses lebih lanjut melalui serangkaian tahapan pembersihan, meliputi *cleansing*, *case folding*, *tokenization*, *stopword removal*, *normalization*, dan *stemming*. Langkah *cleansing* dilakukan untuk menghapus angka dan tanda baca yang tidak relevan. Selanjutnya, huruf-huruf dalam teks diseragamkan menjadi huruf kecil melalui *case folding*. Proses *tokenization* bertujuan memisahkan kalimat menjadi kata-kata individu, sedangkan *stopword removal* menghilangkan kata-kata

umum yang tidak memiliki makna penting. Setelah itu, *normalization* memperbaiki penulisan kata tidak baku menjadi bentuk standar. Terakhir, *stemming* digunakan untuk mengubah kata-kata menjadi bentuk dasarnya. Hasil akhir dari keseluruhan proses ini adalah fitur kata hasil *stemming* dan label kelas yang siap digunakan dalam pemodelan data.



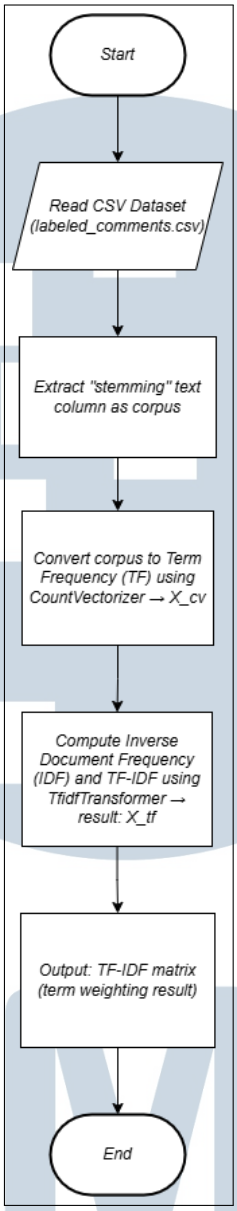
Gambar 3.2. *Flowchart Text Pre-Processing*

3.2.4 Labeling

Pelabelan sentimen pada data dilakukan menggunakan pendekatan *lexicon-based* dengan memanfaatkan dua daftar kata, yaitu `positive.tsv` dan `negative.tsv`, yang masing-masing berisi kosakata bernuansa positif dan negatif. Setiap komentar pada kolom `stemming` dianalisis dengan menghitung jumlah kata positif dan negatif. Jika jumlah kata positif lebih banyak, maka komentar diberi label *positive*; sebaliknya, jika kata negatif lebih dominan, maka komentar diberi label *negative*. Komentar yang tidak mengandung kata dari kedua daftar tidak diberi label dan dihapus dari data. Hasil akhir pelabelan ini disimpan dalam berkas `labeled_comments.csv` untuk digunakan pada tahap pelatihan model klasifikasi.

3.2.5 TF-IDF

Setelah korpus teks dihasilkan dari kolom hasil *stemming*, korpus tersebut diubah menjadi representasi numerik menggunakan *CountVectorizer* untuk menghitung *Term Frequency* (TF), yaitu frekuensi kemunculan setiap kata dalam dokumen. Matriks TF ini kemudian diproses menggunakan *TfidfTransformer* untuk menghitung bobot *Term Frequency-Inverse Document Frequency* (TF-IDF), yang mengalikan frekuensi kata dengan bobot kebalikan dari frekuensi kata tersebut pada seluruh dokumen. Proses ini menghasilkan matriks TF-IDF yang merepresentasikan bobot pentingnya setiap *term* pada masing-masing dokumen, sehingga dapat digunakan untuk analisis lebih lanjut dalam pengolahan teks.

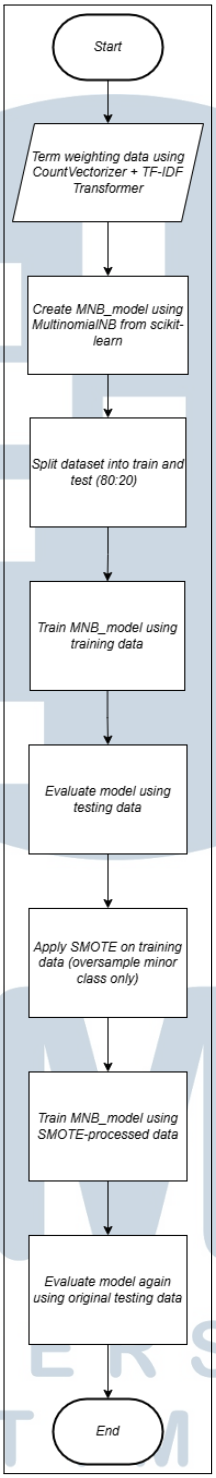


Gambar 3.3. *Flowchart* TF-IDF

3.2.6 Multinomial Naive Bayes

Alur proses dalam penelitian ini digambarkan pada Gambar3.4. Proses dimulai dengan melakukan transformasi data teks ke bentuk representasi numerik menggunakan *CountVectorizer* dan *TF-IDF Transformer* untuk memberikan bobot pada setiap kata. Setelah itu, model klasifikasi dibangun menggunakan algoritma *Multinomial Naive Bayes* (MNB) yang disediakan oleh pustaka *scikit-learn*. Dataset kemudian dibagi menjadi data latih dan data uji dengan rasio 80:20. Model pertama dilatih menggunakan data latih awal dan dievaluasi terhadap data uji untuk

mendapatkan performa awal. Untuk mengatasi permasalahan ketidakseimbangan kelas dalam data, metode SMOTE diterapkan pada data latih untuk melakukan *oversampling* terhadap kelas minoritas. Setelah proses SMOTE, model MNB dilatih kembali menggunakan data latih yang telah diseimbangkan. Evaluasi ulang dilakukan terhadap model ini menggunakan data uji asli (tanpa modifikasi) untuk membandingkan hasil performa model sebelum dan sesudah dilakukan *oversampling*. Tahapan ini bertujuan untuk mengetahui efektivitas penerapan SMOTE dalam meningkatkan kinerja klasifikasi terhadap data teks yang memiliki distribusi kelas tidak seimbang.



Gambar 3.4. *Flowchart Multinomial Naive Bayes*

3.2.7 Evaluation and Testing

Evaluasi model klasifikasi sentimen dilakukan dalam dua skenario utama untuk mengukur kinerja algoritma *Multinomial Naive Bayes* dengan representasi fitur TF-IDF, yaitu menggunakan pembagian data latih dan uji (*train-test split*) dan metode *Stratified K-Fold Cross Validation*. Pada skenario pertama, data dibagi dalam rasio 80:20 dan model diuji baik tanpa penyeimbangan data maupun setelah penerapan SMOTE pada data latih untuk menangani ketidakseimbangan kelas. Skenario kedua menggunakan *K-Fold Cross Validation* untuk memberikan evaluasi yang lebih stabil dan representatif, dengan SMOTE tetap diterapkan hanya pada data latih di setiap fold. Evaluasi dilakukan menggunakan metrik akurasi, *precision*, *recall*, *f1-score*, serta *confusion matrix*.