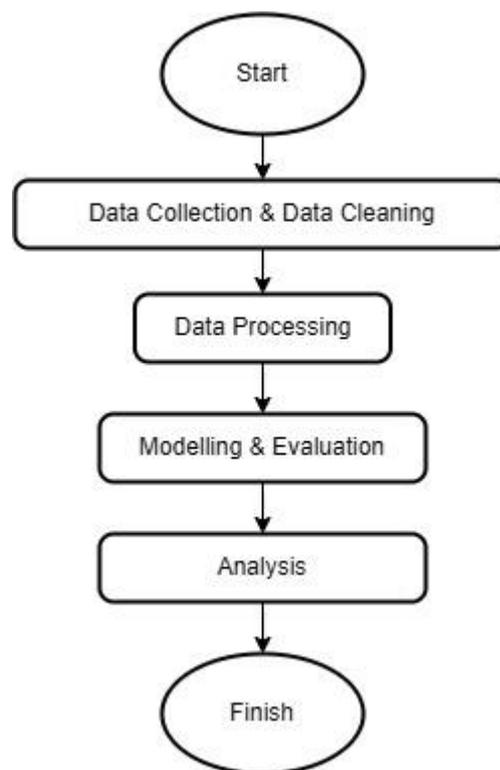


BAB III

METODOLOGI PENELITIAN

Tahapan penelitian dimulai dari proses studi literatur, data collection & data *cleaning*, data *processing*, *modelling & evaluation*, dan yang terakhir *analysis*. Tahapan penelitian dapat dilihat pada gambar 3.1.



Gambar 3.1 Tahapan penelitian

3.1 ***Data Collection & Data Cleaning***

Data yang diambil berupa *tweet* pada Desember 2022 hingga Januari 2023. Data *tweet* diambil menggunakan Teknik *scrapping* pada Python. Selanjutnya data mentah/kotor dibersihkan dan difilter untuk menghilangkan elemen yang tidak perlu (*username*, simbol, emoticon, *hashtag*, dan user yang tidak relevan).

3.2 *Data Preprocessing*

Data yang sudah dibersihkan selanjutnya dilakukan *pre-processing* data. Tahapan *pre-processing* data terdiri dari *casefolding*, *tokenizing*, *stopword removal*, dan *stemming*. Selanjutnya dilakukan *labelling* sentimen menggunakan *Lexicon InSet*. Proses ini bertujuan untuk membuat data latih/validasi dan data uji. Dataset yang sudah dilabeli sentiment dibagi menjadi data latih (80%) dan data uji (20%). Sentimen dibagi menjadi 3 jenis, *negative*, *neutral*, *positive*. Tahap selanjutnya adalah ekstraksi fitur dengan menggunakan TF-IDF dan *N-gram*, di mana fitur-fitur numerik dihasilkan dari teks *tweet* menggunakan TF-IDF dengan *N-gram* (1, 2). *N-gram* mencakup unigram (kata tunggal) dan bigram (pasangan kata), contohnya “indihome lancar”, “kenapa gangguan”.

3.3 *Modelling & Evaluation*

Dataset akan di latih menggunakan model XGBoost dengan 3 metode *hyperparameter*, yaitu *Parameter default*, *Grid Search*, dan *Randomized Search*. Setelah dilakukan pelatihan dengan 3 metode tersebut, model di evaluasi menggunakan *classification report* dan *confusion matrix*.

3.4 *Analysis*

Analisis yang dilakukan meliputi *Word Association* dan pengukuran *score* akurasi. Analisis dilakukan untuk mengidentifikasi kinerja performa tiap metode dan kata-kata yang sering muncul bersamaan dalam setiap kelas sentimen.