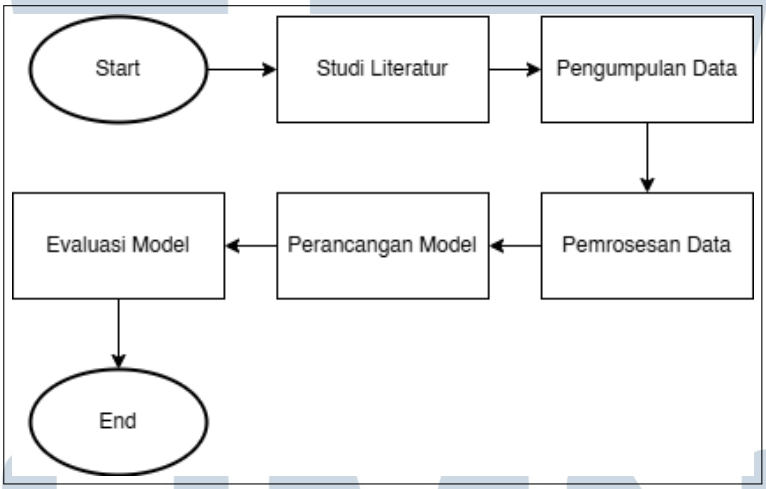
BAB 3
METODOLOGI PENELITIAN

Bab ini menjelaskan metodologi penelitian yang digunakan untuk menganalisis ujaran kebencian di platform X. Penelitian ini memanfaatkan algoritma *Convolutional Neural Network* (CNN) untuk melakukan klasifikasi ujaran kebencian dan Word2Vec untuk melakukan *word embedding*. Metodologi diterapkan dengan cara yang terorganisir di setiap fase penelitian, dimulai dari studi literatur, pengumpulan data, pemrosesan data, perancangan model, dan evaluasi model. Tahapan-tahapan tersebut dapat dilihat pada Gambar 3.1



Gambar 3.1. Alur metodologi penelitian

3.1 Studi Literatur

Pada langkah ini, studi literatur dilakukan untuk mengumpulkan informasi dari berbagai studi yang membahas tentang topik yang relevan, seperti ujaran kebencian dan algoritma *machine learning* yang digunakan untuk melakukan klasifikasi. Penelitian ini menelaah sejumlah jurnal ilmiah dan artikel dengan periode publikasi dari tahun 2018 sampai 2024, yang berasal dari sumber nasional serta internasional. Studi literatur dilakukan dengan tujuan sebagai landasan dalam memahami topik penelitian yang dilakukan. Pemahaman yang didapatkan pada langkah ini mendukung pengembangan penelitian. Berdasarkan hasil studi literatur yang telah dilakukan, ditetapkan penggunaan Convolutional Neural Network (CNN) sebagai model utama dalam proses klasifikasi. Untuk representasi fitur dari data teks, digunakan metode word embedding dengan

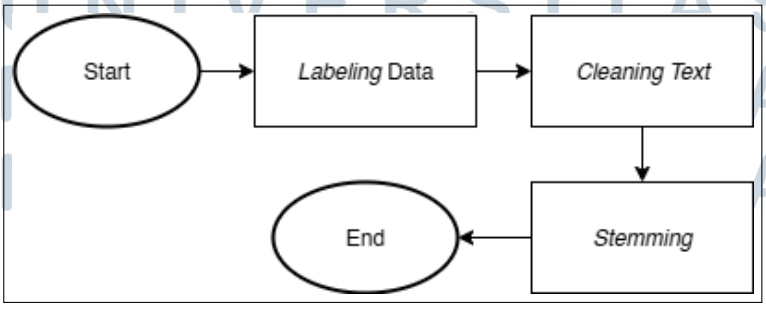
pendekatan Word2Vec, yang mampu menangkap hubungan semantik antar kata secara kontekstual. Proses pemilihan parameter terbaik (hyperparameter tuning) dilakukan menggunakan Optuna, sebuah framework optimasi yang efisien dan berbasis pendekatan Bayesian. Selanjutnya, untuk mengevaluasi kinerja model yang dibangun, digunakan sejumlah metrik evaluasi yang mencakup accuracy, precision, recall, f1-score, serta confusion matrix, guna memberikan gambaran menyeluruh terkait performa klasifikasi yang dihasilkan.

3.2 Pengumpulan Data

Data yang digunakan untuk penelitian ini didapat dari pengguna platform X dengan cara menggabungkan data sebelumnya dan melakukan *scrapping* melalui Twitter Search API. Proses pengambilan data melibatkan kata kunci yang telah ditentukan sebagai contoh ujaran kebencian serta bahasa yang kasar, berdasarkan hasil penelitian serta literatur yang sudah ada. Data yang telah terkumpul dilakukan anotasi lebih lanjut terhadap subjek, kategori, dan tingkat keparahan ujaran kebencian. Anotasi dibagi menjadi dua kategori, yaitu ujaran kebencian dan *abusive language*. Proses anotasi melibatkan 30 orang anotator dari berbagai latar belakang untuk memastikan objektivitas dan keakuratan data.

3.3 Pemrosesan Data

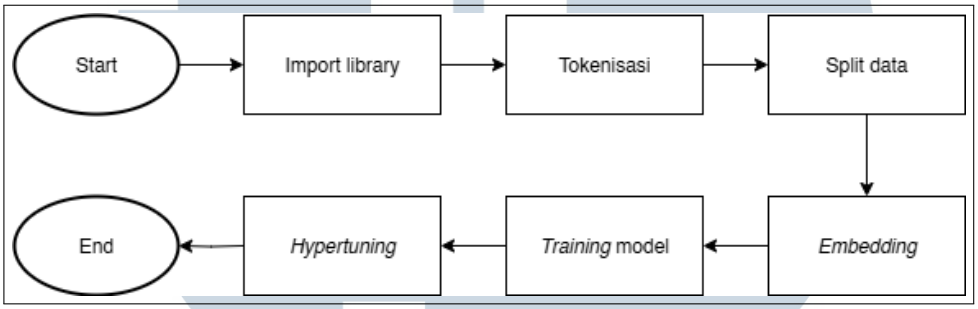
Pada tahap ini, dilakukan prapemrosesan data dengan mencakup beberapa bagian seperti menghapus data duplikat, menangani *missing values*, membersihkan teks dengan menghapus URL, hashtag, unicode, simbol atau tanda baca, kata dan karakter berulang, spasi berlebih, emoji, mengubah *slang word*, dan *stemming*. Tahapan ini bertujuan untuk mempersiapkan data yang relevan dengan kebutuhan model nantinya. Detail dari tahapan pemrosesan data dapat dilihat pada Gambar 3.2



Gambar 3.2. Alur pemrosesan data

3.4 Perancangan Model

Setelah data telah diproses, langkah berikutnya adalah merancang model CNN. Perancangan dilakukan dengan tujuan untuk menghasilkan model yang dapat dengan baik mendeteksi ujaran kebencian pada platform X. Pada fase pelatihan, langkah-langkah yang perlu dilakukan meliputi: *import library*, tokenisasi, *split* data, *embeding*, *training* model, *hypertuning*, dan evaluasi. Untuk lebih jelasnya, rangkaian alur perancangan model akan dijelaskan dengan Gambar 3.3



Gambar 3.3. Alur perancangan model

Langkah-langkah perancangan model yang ditunjukkan pada Gambar 3.3 dapat dijelaskan sebagai berikut:

- *Import library*: *install* dan memasukkan *library* yang dibutuhkan untuk perancangan model. *Library* yang digunakan pada penelitian ini yaitu pandas, tensorflow, gensim, dan Optuna.
- Tokenisasi: mengubah data teks yang bersifat simbolik menjadi format yang terstruktur dan dapat dikenali secara numerik oleh model. tokenisasi berperan sebagai jembatan yang mengonversi teks ke dalam bentuk representasi yang dapat diolah lebih lanjut, seperti vektor angka atau urutan indeks. Representasi ini memungkinkan model untuk mengenali pola dan makna dari suatu ujaran.
- *Split* data: pemisahan data menjadi data latih, validasi, dan uji. Data latih digunakan untuk melatih model agar dapat mempelajari pola dan hubungan dari data. Data validasi digunakan untuk menguji model selama pelatihan untuk menghindari *overfitting*. Sedangkan data uji digunakan untuk mengukur performa akhir model secara independen dan objektif.

- *Embedding*: membuat representasi vektor kata memiliki makna serupa, bukan angka token biasa. *Embedding* juga berperan untuk membantu model mengenali kesamaan dan perbedaan antar kata dalam konteks ujaran kebencian, serta meningkatkan kemampuan generalisasi model dalam mengidentifikasi variasi bahasa yang mengandung unsur ofensif.
- *Training* model: melakukan pelatihan model CNN menggunakan data yang telah di-*preprocessing*. *Training* dilakukan sebagai pengujian apakah model dan dataset yang digunakan bisa bekerja dengan baik.
- *Hypertuning Parameter*: memberikan sejumlah kombinasi parameter untuk diuji. *Hypertuning* diperlukan untuk menemukan kombinasi nilai hiperparameter terbaik agar model dapat mencapai performa yang maksimal berdasarkan metrik evaluasi tertentu.

3.4.1 Hypertuning

Hypertuning parameter atau penyetelan hiperparameter merupakan proses penting dalam pengembangan model pembelajaran mesin, terutama untuk meningkatkan kinerja model secara optimal [43]. Hiperparameter adalah parameter yang tidak dipelajari secara langsung dari data pelatihan, melainkan ditentukan sebelum proses pelatihan dimulai, seperti *learning rate*, jumlah neuron, jumlah *layer*, *batch size*, dan lain sebagainya. Pemilihan nilai hiperparameter yang tepat sangat menentukan kemampuan model dalam melakukan generalisasi terhadap data yang belum pernah dilihat sebelumnya. Oleh karena itu, proses *hypertuning* diperlukan untuk menemukan kombinasi nilai hiperparameter terbaik agar model dapat mencapai performa yang maksimal berdasarkan metrik evaluasi tertentu.

Pada penelitian ini, Optuna digunakan sebagai salah satu kerangka kerja untuk melakukan *hypertuning* secara efisien dan fleksibel. Optuna memiliki keunggulan dibandingkan dengan metode pencarian hiperparameter konvensional seperti *random search*. Salah satu kelebihan utama Optuna adalah kemampuannya dalam melakukan optimisasi berbasis pendekatan *Bayesian optimization* melalui teknik *Tree-structured Parzen Estimator* (TPE). Teknik ini memungkinkan proses pencarian dilakukan secara lebih efisien karena memperkirakan distribusi nilai hiperparameter terbaik berdasarkan iterasi sebelumnya.

Optuna menunjukkan efisiensi yang lebih baik dalam proses tuning dan mampu menemukan konfigurasi model yang optimal pada berbagai jenis tugas

pembelajaran mesin. Penelitian yang dilakukan oleh Akiba (2019) dalam publikasi berjudul ”Optuna: *A Next-generation Hyperparameter Optimization Framework*” menunjukkan bahwa penggunaan Optuna dapat menghasilkan performa model yang lebih tinggi dalam waktu yang lebih singkat dibandingkan dengan metode pencarian lainnya [44]. Dengan demikian, penggunaan Optuna sebagai alat bantu dalam proses hypertuning dapat memberikan kontribusi signifikan terhadap efektivitas dan efisiensi pengembangan model prediktif.

3.5 Evaluasi Model

Pada tahap evaluasi, model yang telah dilatih diuji menggunakan data validasi untuk mengukur performanya dalam mendeteksi ujaran kebencian pada platform X. Metode evaluasi yang digunakan adalah *accuracy*, *precision*, *recall*, dan *F1-score*. *Confusion matrix* juga digunakan untuk menganalisis kesalahan klasifikasi dan memahami distribusi prediksi model terhadap kelas yang benar.