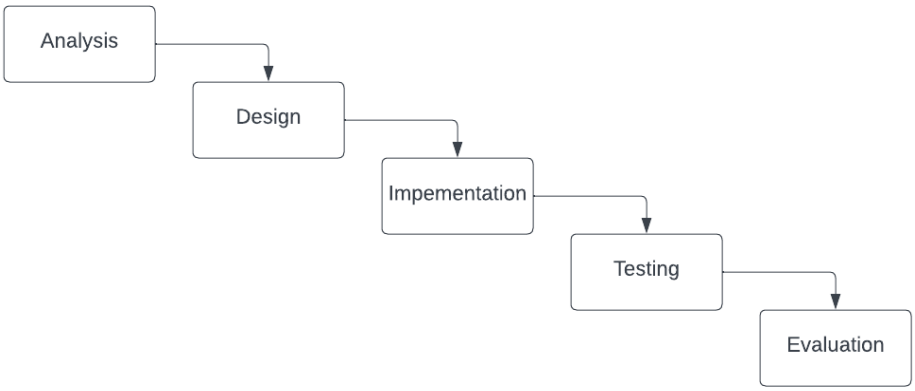
BAB III
METODE PENELITIAN

Bab ini menjelaskan langkah-langkah penelitian dalam membangun model prediksi harga saham Alfamart berbasis *deep learning* dan sentimen publik dari media sosial X. Tahapan-tahapan dalam metode ini meliputi: Desain Penelitian, Analisis Kebutuhan penelitian, Perancangan sistem prediksi harga saham, Implementasi, Pengujian, dan Evaluasi. Setiap tahapan diselesaikan terlebih dahulu sebelum melanjutkan ke tahap berikutnya.

3.1 Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperimental komparatif untuk mengevaluasi performa algoritma *deep learning* dalam prediksi harga saham. penelitian mencakup dua tahap utama: pertama, pengembangan model prediksi harga saham menggunakan tiga algoritma *deep learning* (RNN, LSTM, dan ELM) dengan data teknis saham AMRT; kedua, integrasi analisis sentimen dari media sosial X untuk meningkatkan akurasi prediksi. Penelitian ini bersifat eksploratif dan komparatif dengan tujuan mengidentifikasi model terbaik berdasarkan metrik evaluasi MAPE, RMSE, dan R2.

Dalam proses pengembangan model prediksi, penelitian ini mengadopsi pendekatan *Waterfall Methodology*, karena proses pengembangannya bersifat terstruktur dan linier, di mana setiap tahap diselesaikan secara menyeluruh sebelum berpindah ke tahap berikutnya. Pendekatan ini dimulai dari analisis kebutuhan, dilanjutkan dengan perancangan sistem, implementasi, pengujian, hingga evaluasi akhir. Alur penelitian ini digambarkan pada Gambar 3.1 berikut:



Gambar 3. 1 Alur penelitian dengan pendekatan *waterfall*

Metodologi *waterfall* dipilih agar alur penelitian lebih sistematis, memudahkan pelacakan hasil di setiap fase. Sesuai dengan karakteristik penelitian ini yang tidak memerlukan banyak iterasi ulang seperti pada pengembangan sistem berbasis *Agile methodology*.

3.2 Analisis Kebutuhan Penelitian

Tahap awal dalam penelitian ini adalah mengidentifikasi kebutuhan dan tujuan dari penelitian. Pada tahap ini, kebutuhan yang diperlukan dalam penelitian mulai dari *dataset*, algoritma yang digunakan, dan lingkungan implementasi tempat program dijalankan akan ditentukan.

3.2.1 Data

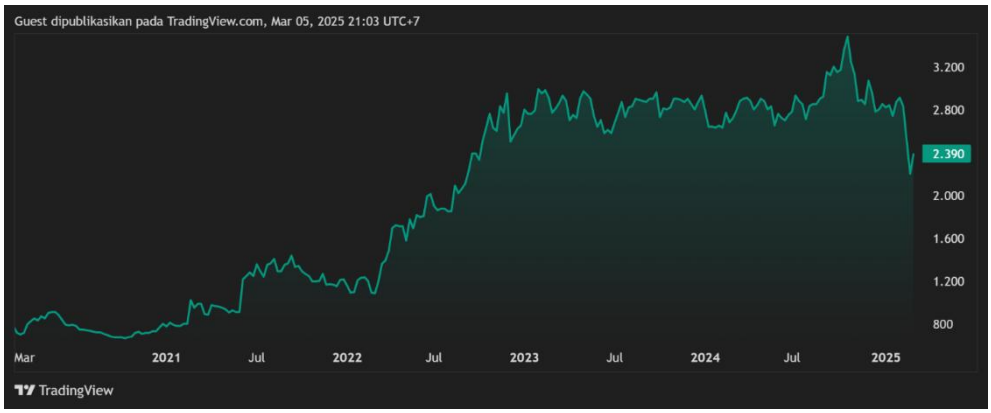
Dalam penelitian ini akan digunakan dua *dataset*, yaitu *dataset* saham AMRT dan *dataset* sentimen masyarakat Indonesia yang didapat dari media sosial X. *Dataset* saham AMRT akan digunakan sebagai basis untuk proses *training* model *deep learning*. Sedangkan *dataset* sentimen publik akan digunakan sebagai fitur tambahan yang akan diintegrasikan dengan *dataset* AMRT. Data sentimen ini akan digunakan sebagai fitur tambahan untuk proses pelatihan model prediksi *deep learning* dengan sentimen.

3.2.1.1 *Dataset* Saham

Dataset saham yang akan digunakan dalam penelitian ini adalah *dataset* AMRT.csv yang didapat dari Github. *Dataset* ini dipilih karena diambil dan

diolah dari *website* PT Bursa Efek Indonesia. Semua data yang ada dalam *dataset* adalah milik PT Bursa Efek Indonesia. Penggunaan *dataset* ini harus mengacu pada syarat penggunaan yang dimiliki oleh PT Bursa Efek Indonesia. *Dataset* ini memiliki data yang sangat lengkap untuk digunakan sebagai fitur untuk proses *training* pada model *deep learning* untuk memprediksi harga saham AMRT. *Dataset* ini juga terbukti sesuai dengan data pergerakan harga saham dari *platform trading* terkemuka bernama Tradeview. Perbandingan visualisasi pergerakan harga saham dari data AMRT.csv dan Tradeview dapat dilihat pada Gambar 3.2 dan Gambar 3.3 berikut ini:

Gambar 3. 2 Visualisasi pergerakan harga saham AMRT dari *dataset* AMRT.csv



Gambar 3. 3 Pergerakan harga saham AMRT dari platform Tradeview

Dataset AMRT.csv memiliki data pergerakan saham dari tanggal 29 Juli 2019 sampai 21 Februari 2025. Namun dalam penelitian ini, data yang akan

digunakan dalam proses *training* adalah data dari 20 Januari 2024 sampai 20 Januari 2025.

Dengan pertimbangan tersebut, maka *dataset* ini akan digunakan sebagai basis untuk *training* model *deep learning* yang akan digunakan untuk prediksi harga saham AMRT pada penelitian ini.

3.2.1.2 *Dataset* Sentimen

Pada penelitian ini, *dataset* sentimen akan dikumpulkan menggunakan teknik *crawling* terhadap media sosial X. Proses tersebut dilakukan dengan bantuan skrip Tweet Harvest yang dapat mengumpulkan data *tweet* dari media sosial X secara otomatis. Data yang dikumpulkan adalah *tweet* pengguna X di Indonesia yang menyebutkan kata "alfamart" di dalamnya. Data yang telah dikumpulkan akan disimpan dalam format CSV di penyimpanan lokal perangkat. Data yang dikumpulkan menyesuaikan data *training* dari AMRT.csv yaitu *tweet* dari tanggal 20 Januari 2024 sampai 20 Januari 2025.

3.2.2 Algoritma yang Akan Digunakan

Penelitian ini akan menggunakan algoritma yang berbeda dalam proses analisis sentimen dan prediksi harga saham. Dalam proses analisis sentimen, tiga algoritma akan diuji terhadap performanya terhadap data sampel sentimen yang diambil secara acak dari *dataset* sentimen dan diberi label sentimen. Tiga algoritma tersebut adalah pendekatan Lexicon, NN, dan BERT. Setelah algoritma yang menghasilkan model klasifikasi sentimen terbaik didapat, model tersebut akan digunakan untuk melakukan klasifikasi sentimen pada keseluruhan *dataset* sentimen yang hasilnya akan digunakan sebagai fitur tambahan dalam pembuatan model prediksi harga saham dengan data sentimen.

Sementara itu, tiga algoritma *deep learning* akan digunakan dalam pembuatan model prediksi harga saham. Algoritma yang akan digunakan adalah algoritma RNN, LSTM, dan ELM. Ketiga algoritma tersebut dipilih karena telah terbukti dapat memprediksi pergerakan harga saham pada beberapa penelitian yang telah dilakukan.

3.2.3 Lingkungan Implementasi

Dalam melakukan penelitian ini, dibutuhkan lingkungan yang berupa perangkat keras dan perangkat lunak yang dapat menjalankan program untuk melakukan proses pengumpulan data, klasifikasi sentimen, pelatihan, pengujian, dan evaluasi model prediksi. Perangkat keras yang akan digunakan dalam penelitian ini adalah laptop Victus 15 dengan spesifikasi *processor* Ryzen 5 8645HS, GPU RTX 4050 dengan VRAM 6GB dan RAM 16GB. Perangkat ini sudah cukup kuat untuk menjalankan perangkat lunak dan program yang akan digunakan dalam penelitian ini.

Selain perangkat keras yang memadai, beberapa perangkat lunak juga harus disiapkan untuk menjalankan kode program untuk melaksanakan penelitian ini. Penelitian ini menggunakan bahasa pemrograman python dalam proses EDA, *Preprocessing* data, pembuatan model klasifikasi sentimen, dan pembuatan model prediksi harga saham. Sedangkan proses pengumpulan data sentimen dilakukan menggunakan skrip TweetHarvest yang dijalankan menggunakan Node.js. Versi Python yang digunakan adalah versi 3.11. Versi ini dipilih karena mendukung penggunaan *library* TensorFlow dan PyTorch yang akan digunakan dalam proses pelatihan model *deep learning*. Sedangkan versi Node.js yang digunakan adalah versi 22.

Google Colab akan digunakan untuk menjalankan program *crawling* data sentimen. Google Colab digunakan dalam proses *crawling* untuk menghindari hal yang dapat mengganggu proses pengumpulan data sentimen yang cukup memakan waktu. Karena Google Colab dijalankan secara daring pada server milik Google, maka program tidak akan langsung berhenti jika perangkat keras atau jaringan internet mati. Data yang telah dikumpulkan akan langsung disimpan di Google Drive secara otomatis. Selain itu, program Microsoft Excel 2021 digunakan untuk proses konversi label sentimen menjadi data numerik dan proses integrasi data sentimen terhadap data saham.

Kemudian, program untuk pembuatan model prediksi dilakukan pada program VSC (Visual Studio Code). program ini dipilih karena mendukung untuk menulis dan mengedit program dalam bahasa Python. VSC juga dipilih karena mendukung penggunaan Jupyter Notebook yang akan digunakan untuk proses pembuatan model prediksi mulai dari *preprocessing* data hingga pelatihan dan evaluasi model.

3.3 Perancangan dan Implementasi Sistem Prediksi Harga Saham

Perancangan sistem mencakup seluruh rancangan arsitektur sistem, data, dan alur kerja penelitian. Tahap ini mencakup proses pengumpulan data, pengolahan data teks, klasifikasi sentimen dan integrasi data, EDA, *preprocessing*, dan *training* model prediksi.

3.3.1 Pengumpulan Data

Tahap awal dari proses pembangunan model prediksi adalah mengumpulkan data yang dibutuhkan untuk proses pembangunan model. Data yang dibutuhkan dalam penelitian ini adalah data teknis saham AMRT, dan data sentimen yang didapat dari media sosial X.

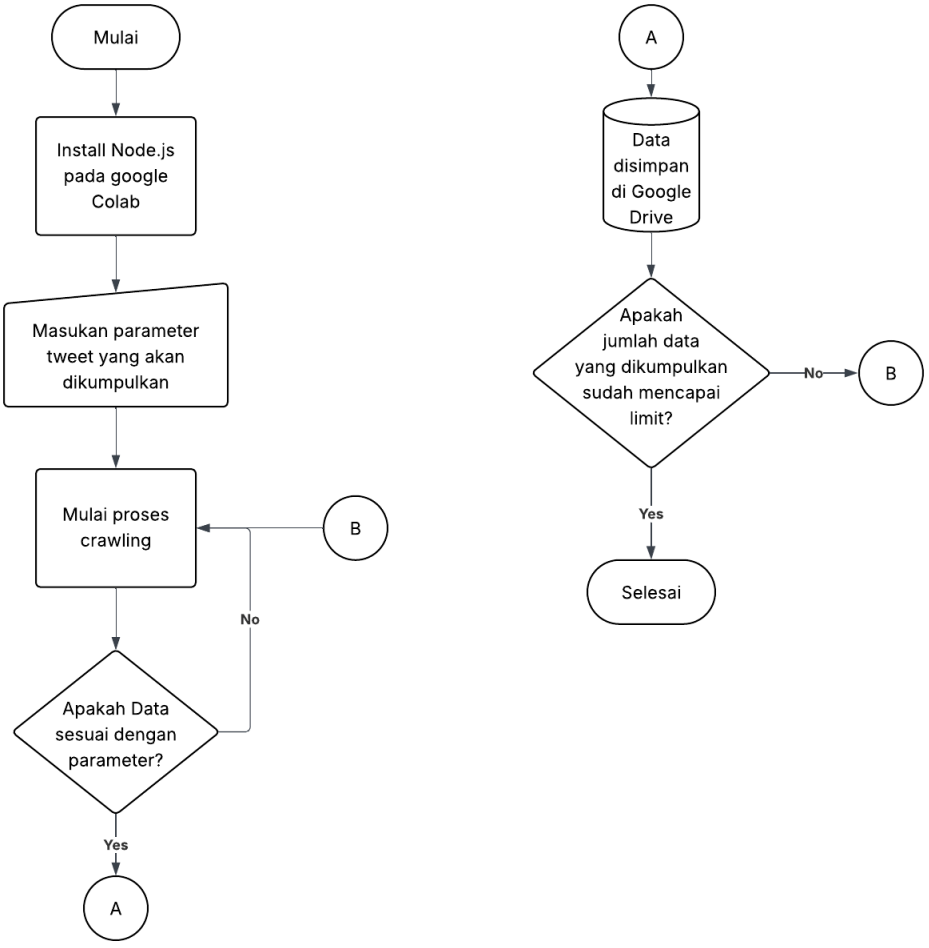
1. Pengumpulan Data Saham

Data saham yang akan digunakan pada penelitian ini adalah data pergerakan harga saham AMRT. *Dataset* yang akan digunakan pada penelitian ini diunduh dari repositori Github (https://github.com/wildangunawan/Dataset-Saham-IDX/blob/master/Saham/Semua/AMRT.csv) yang diunggah oleh seorang pengguna dengan nama Wildan Gunawan.

2. Pengumpulan Data Sentimen

Pengumpulan data sentimen pada penelitian ini akan dilakukan dengan teknik *crawling* (pengumpulan) *tweet* dari X berdasarkan kriteria pencarian tertentu. Proses ini dibantu oleh skrip Tweet Harvest yang dibuat oleh seorang pengguna Github dengan nama Helmi Satria. Tweetharvest secara teknis tidak menggunakan API X resmi, melainkan meniru pencarian *tweet* secara

langsung melalui *browser* menggunakan *headless browser* yaitu Playwright. Kode program untuk melakukan proses *crawling* ini dijalankan pada platform Google Colab. Proses *crawling* data *tweet* dijelaskap pada Gambar 3.4.



Gambar 3. 4 *Flowchart* Proses *Crawling* Data *Tweet*

Proses *crawling* akan diawali dengan memasang Node.js versi 22 agar Google Colab dapat menjalankan skrip TweetHarvest. Sebelum proses *crawling* dimulai, parameter yang dibutuhkan oleh TweetHarvest untuk melakukan *crawling* harus ditetapkan terlebih dahulu. Parameter yang harus disiapkan adalah token akses akun X, nama *file* dengan format CSV, kata kunci pencarian yaitu “alfamart”, batasan tanggal *tweet* yaitu dari 20 Januari 2024 sampai 21 Februari 2025, dan batas *tweet* yang akan disimpan dalam satu kali proses *crawling*.

Jika semua parameter sudah ditetapkan, skrip TweetHarvest dapat dijalankan. TweetHarvest akan meniru proses pencarian *tweet* secara langsung pada browser menggunakan akun yang token aksesnya digunakan. Semua *tweet* yang muncul dari proses pencarian akan disimpan ke folder berbentuk CSV pada Google Drive. Namun, satu akun X hanya dapat menyimpan sebanyak 600 *tweet* dalam sehari. Hal ini disebabkan batasan terhadap proses *crawling* yang ditetapkan oleh Elon Musk. Oleh karena itu, proses pengumpulan data sentimen pada penelitian ini akan dilakukan secara bertahap menggunakan beberapa akun media sosial X hingga data selama satu tahun didapat. Setelah data sentimen selama satu tahun didapat, data yang tersimpan pada Google Drive akan diunduh dan digabungkan menjadi satu *dataset* sentimen.

3.3.2 Pengolahan Data Teks

Sebelum dilakukan klasifikasi sentimen, data teks yang diperoleh dari media sosial X perlu melalui tahap *preprocessing* atau pengolahan awal. Tujuan dari tahapan ini adalah untuk membersihkan, menyederhanakan, dan menormalisasi teks agar data teks dapat diproses secara lebih akurat oleh algoritma klasifikasi yang akan digunakan. tahapan pengolahan teks yang dilakukan dalam penelitian ini adalah sebagai berikut:

- *Text Cleaning*

  Proses ini dilakukan untuk membersihkan data *tweet* dari tanda baca, simbol, angka, emoji, *hashtag*, *mention*, spasi ganda, dan karakter kosong. Proses ini akan dibantu dengan library re (regular expression), emoji (untuk menghapus emoji), dan string (untuk karakter khusus).
- *case folding*

  Proses ini akan mengubah menjadi huruf kecil. Hal ini dilakukan agar kata-kata seperti "Alfamart", "alfamart", dan "ALFAMART" dianggap sama. Proses ini dilakukan dengan menggunakan syntax text.lower()
- *Stemming*

Stemming bertujuan mengubah kata ke bentuk dasarnya (*root word*). Dalam penelitian ini digunakan *stemming* Bahasa Indonesia dari *library* Sastrawi yang efektif dalam menangani imbuhan dan bentuk turunan kata.

- *Vektorisasi*

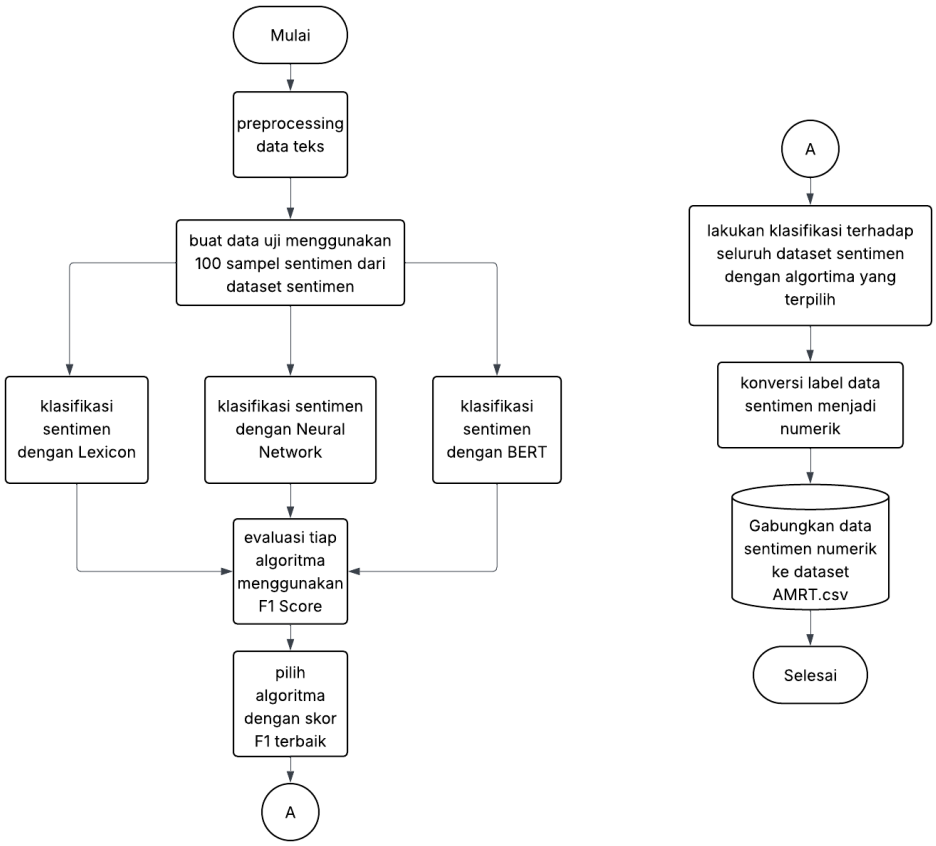
  Vektorisasi mengubah teks menjadi representasi numerik yang dapat diproses oleh model klasifikasi sentimen. Untuk algoritma *Neural Network*, setiap teks yang telah melalui *stemming* akan diubah menjadi vektor menggunakan TF-IDF. Sementara itu, model BERT tidak memerlukan *stemming* atau TF-IDF karena sudah menggunakan *tokenizer* internal dan representasi vektor kontekstual dari setiap kata.

Setelah data teks dibersihkan, data akan difilter dan data *tweet* dengan nama pengguna “alfamart”, “alfa”, dan “alfakarir” akan dihapus dari *dataset*. Hal ini dilakukan karena data *tweet marketing* dan pengumuman resmi tidak termasuk data sentimen publik.

3.3.3 Klasifikasi Sentimen dan Integrasi Data

Klasifikasi sentimen dalam penelitian ini dibagi menjadi tiga kelas, yaitu positif, netral, dan negatif. Pemilahan ke dalam tiga kelas ini dipilih agar klasifikasi sentimen tidak hanya berfokus pada kutub emosional (positif atau negatif) saja, tetapi juga mampu menangkap teks-teks yang tidak relevan atau ambigu terhadap perusahaan Alfamart. Sentimen netral digunakan sebagai penanda untuk teks yang secara isi tidak menunjukkan opini emosional yang kuat atau tidak secara langsung merujuk pada aktivitas bisnis Alfamart. Hal ini penting karena dalam proses *crawling* data dari media sosial X, tidak semua *mention* terhadap kata "alfamart" mengandung opini tentang perusahaan, melainkan bisa berupa konten umum, promosi, atau obrolan tanpa relevansi terhadap citra perusahaan. Dengan demikian, klasifikasi tiga kelas ini bertujuan untuk menyaring *noise* dan hanya mempertahankan opini publik yang benar-benar memiliki potensi memengaruhi persepsi pasar terhadap saham AMRT.

Untuk mengklasifikasikan sentimen dari data teks tersebut, tiga pendekatan algoritma yang berbeda akan digunakan. Pendekatan tersebut adalah klasifikasi sentimen berbasis Lexicon, *Neural Network* sederhana, dan model *pre-trained* berbasis BERT. Proses klasifikasi sentimen menggunakan ketiga algoritma yang dipilih dan proses integrasinya terhadap *dataset* saham dijelaskan sebagai pada gambar 3.5.



Gambar 3. 5 *Flowchart* proses klasifikasi dan Integrasi data Sentimen

1. Klasifikasi Sentimen Berbasis Lexicon

Pendekatan ini menggunakan kamus Lexicon yang berisi daftar kata-kata dalam Bahasa Indonesia yang telah diberi skor sentimen. Kamus lexicon diperoleh dari repositori Github sumber terbuka yang sesuai dengan konteks Bahasa Indonesia. Pada kamus Lexicon ini berisi 3.609 kata positif dan 6.609 kata negatif. Setiap kata yang berada pada kategori positif akan diberi skor 1, dan kata yang berada pada kategori negatif akan diberi skor -1.

Setiap teks diuji secara langsung terhadap kamus ini, dan total skor dihitung berdasarkan kata-kata yang ditemukan. Jika skor total melebihi ambang tertentu, maka sentimen ditentukan sebagai:

- Positif jika skor total > 0
- Netral jika skor total = 0
- Negatif jika skor total < 0

2. Klasifikasi Sentimen Menggunakan *Neural Network*

Model *Neural Network* akan dikembangkan secara sederhana dan dilatih secara terbatas untuk klasifikasi tiga kelas (positif, netral, negatif). Untuk pelatihan, digunakan 500 data pertama dari *dataset* yang telah diberi label sentimen secara manual oleh peneliti.

Langkah-langkah proses pelatihan model adalah sebagai berikut:

- Pengolahan data teks
- Pemisahan data latih dan validasi
- Pelatihan model *Neural Network*
- Evaluasi model menggunakan data uji yang sama dengan algoritma lainnya

3. Klasifikasi Sentimen Menggunakan BERT

Pendekatan ketiga akan menggunakan model *pretrained transformer* berbahasa Indonesia yaitu w11wo/indonesian-roberta-base-sentiment-classifier. Model ini merupakan hasil pelatihan sebelumnya pada data sosial media berbahasa Indonesia dan telah tersedia di *platform* HuggingFace.

Model ini langsung digunakan untuk mengklasifikasikan 100 data uji tanpa perlu pelatihan ulang. Proses inferensi dilakukan pada data uji yang sama untuk menjamin kesetaraan evaluasi.

4. Evaluasi dan Pemilihan Model Terbaik

Ketiga pendekatan di atas dievaluasi berdasarkan skor F1 terhadap 100 data uji yang diambil acak dari *dataset* sentimen yang telah dilabeli secara

manual. Evaluasi dilakukan untuk masing-masing kelas (positif, netral, negatif) dan rata-rata F1-score digunakan untuk menentukan model terbaik.

Model dengan skor F1 tertinggi kemudian digunakan untuk mengklasifikasikan seluruh *dataset* sentimen yang telah dikumpulkan. Dengan demikian, setiap baris data sentimen akan mendapatkan label sentimen seperti *positive*, *neutral*, atau *negative*.

5. Konversi dan Agregasi Data Sentimen

Jika seluruh data sudah diklasifikasikan, label sentimen akan dikonversikan menjadi data numerik menggunakan fungsi IF pada program Microsoft Excel. Pada proses ini, data dengan label sentimen akan dikonversikan menjadi seperti ini: *negative* = -1, *neutral* = 0, *positive* = 1.

Jika data label sentimen telah dikonversi menjadi data numerik, data sentimen harian diperoleh dengan menjumlahkan nilai-nilai sentimen per tanggal menggunakan fitur PivotTable.

6. Integrasi Data Sentimen ke *Dataset* Saham

Langkah akhir dari proses ini adalah mengintegrasikan data sentimen harian ke dalam *dataset* harga saham AMRT.csv. Proses integrasi akan dilakukan menggunakan fungsi VLOOKUP pada Microsoft Excel, dengan mencocokkan kolom tanggal pada data sentimen harian dan data harga saham. Hasilnya adalah *dataset* saham AMRT yang telah memiliki fitur tambahan berupa sentimen harian, yang akan digunakan sebagai salah satu fitur *input* pada model prediksi.

3.3.4 *Exploratory Data Analysis* (EDA)

Exploratory Data Analysis atau EDA adalah fondasi dari proses *preprocessing*. Melalui proses ini, nilai-nilai yang hilang, *outlier*, serta hubungan antar variabel yang relevan dalam konteks prediksi harga saham dapat diidentifikasi. Dalam penelitian ini, EDA akan dilakukan terhadap data historis harga saham AMRT.csv yang telah ditambahkan data sentimen harian. Tahapan EDA mencakup beberapa aktivitas utama sebagai berikut:

A. Statistik Deskriptif

Statistik deskriptif akan digunakan untuk memberikan gambaran umum terhadap distribusi data. Untuk data saham pada penelitian ini, dihitung nilai minimum, maksimum, rata-rata (*mean*), standar deviasi, dan kuartil. Hal ini membantu peneliti dalam memahami sebaran harga saham selama periode pengamatan.

B. Visualisasi Data Deret Waktu

Visualisasi data dalam bentuk grafik deret waktu akan dibuat untuk memantau pergerakan harga saham dari waktu ke waktu. Grafik ini memungkinkan peneliti untuk mengidentifikasi tren umum, pola musiman (*seasonal pattern*), maupun volatilitas harga.

C. Korelasi Antar Variabel

Analisis korelasi dilakukan untuk mengetahui sejauh mana hubungan antara fitur-fitur numerik terhadap fitur target. Koefisien korelasi Pearson akan digunakan dalam analisis ini. Hasil uji korelasi setiap fitur terhadap target akan ditampilkan menggunakan *heatmap*. Pemetaan hasil uji korelasi menggunakan *heatmap* ini dapat membantu peneliti dalam memilih kombinasi fitur yang akan digunakan dalam proses pelatihan model prediksi.

D. Pendeteksian *Missing Value* dan *Outlier*

Pemeriksaan *missing value* dilakukan untuk memastikan kelengkapan data. Jika ditemukan data kosong, maka akan dilakukan penanganan dengan pengisian menggunakan nilai rata-rata. Sementara itu, Pendeteksian *outlier* dilakukan menggunakan visualisasi *Boxplot*.

3.3.5 *Preprocessing*

Preprocessing merupakan tahap penting yang dilakukan sebelum proses pelatihan model dimulai. Tujuannya adalah untuk mempersiapkan data agar dapat diolah secara optimal oleh algoritma *deep learning*. *Preprocessing*

dilakukan terhadap data AMRT.csv yang telah terintegrasi terhadap dengan *dataset* sentimen dan telah melalui proses EDA. Proses *preprocessing dataset* AMRT.csv yang akan dilakukan pada penelitian ini adalah sebagai berikut:

A. Penanganan *Missing Value*

Jika *Missing Value* ditemukan pada *dataset*, maka penyebab terjadinya *Missing Value* tersebut akan diidentifikasi. Terdapat kemungkinan di mana data harga saham pada tanggal tertentu kosong karena hari libur pasar saham. Jika data tersebut ditemukan maka data pada baris tersebut akan dihapus. Jika terdapat data sentimen yang kosong karena tidak adanya aktivitas media sosial pada hari tertentu, maka nilai kosong pada data sentimen akan diisi dengan angka 0 sebagai representasi dari sentimen netral.

B. Penanganan *Outlier*

Outlier dapat terdeteksi pada volume perdagangan dan jumlah *tweet* harian. Jika data *outlier* tersebut terlihat tidak masuk akal, maka data tersebut akan dihapus, namun jika data *outlier* muncul akibat berita besar atau rumor pasar, maka data akan dipertahankan dalam analisis karena relevan secara kontekstual

C. Transformasi Data Waktu

Kolom tanggal pada *dataset* akan dikonversi ke dalam format *datetime* dan digunakan untuk menyortir data agar urutan baris mencerminkan urutan kronologis. Meskipun pembagian data tidak berdasarkan waktu, pengurutan tetap dilakukan untuk menjaga konsistensi data.

D. Normalisasi Data

Untuk memastikan kestabilan pelatihan model dan mempercepat konvergensi, fitur numerik dinormalisasi ke dalam rentang 0 hingga 1 menggunakan teknik *Min-Max Scaling*.

Dengan tahapan *preprocessing* ini, data yang semula berasal dari dua sumber berbeda telah diolah dan disiapkan dalam format yang sesuai untuk

digunakan dalam pelatihan dan evaluasi model prediksi harga saham berbasis *deep learning*.

3.3.6 Training

Tahap *training* atau pelatihan merupakan proses inti dalam pembangunan model prediksi harga saham. Pada tahap ini, data yang telah melalui *preprocessing* digunakan untuk melatih model agar mampu mempelajari pola dari fitur-fitur *input* terhadap nilai target yang akan diprediksi, yaitu harga penutupan saham.

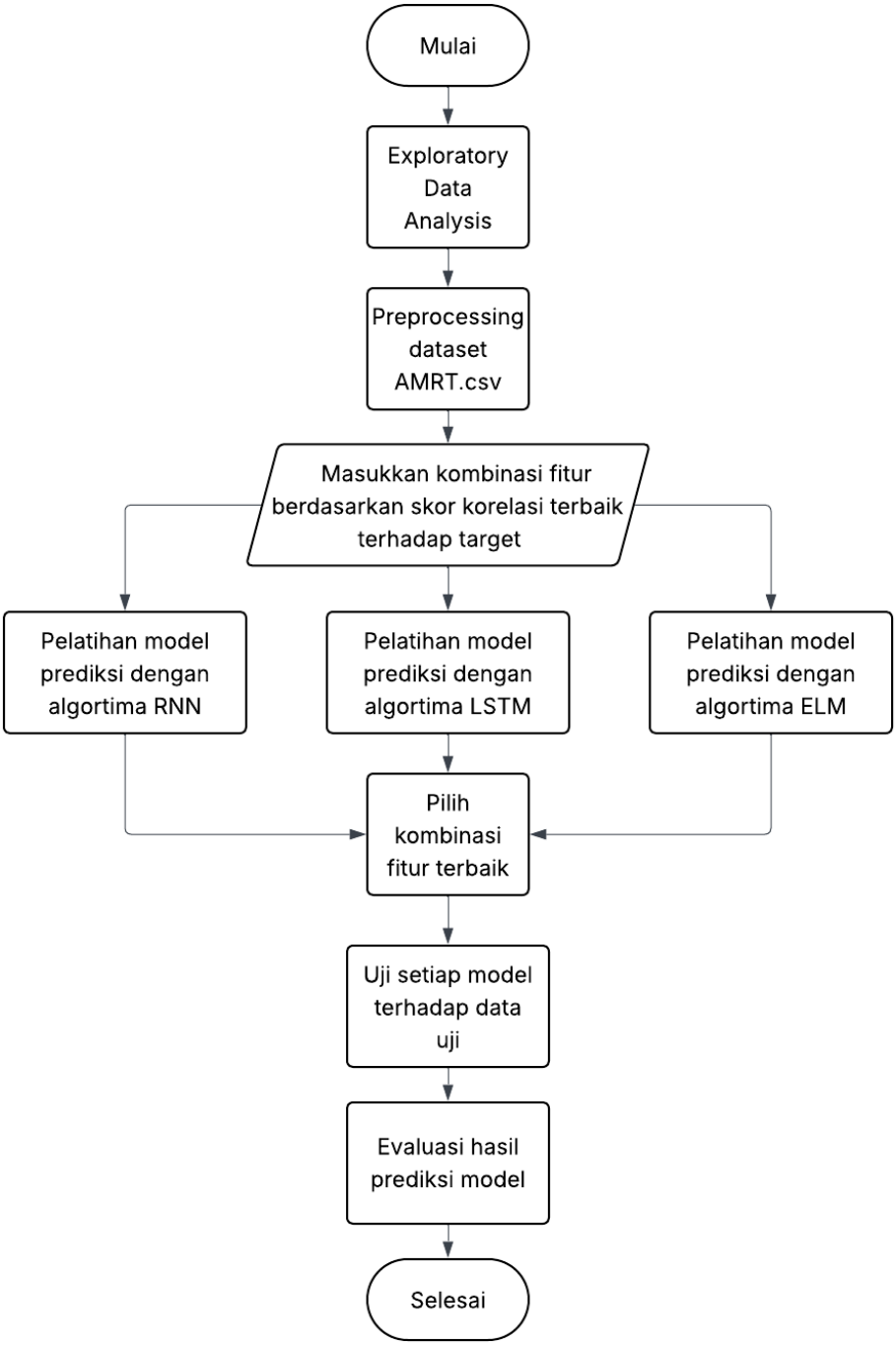
Penelitian ini akan dilakukan dengan melakukan eksperimen pelatihan model dengan tiga jenis algoritma yang berbeda, yaitu algoritma RNN, LSTM, dan ELM. Masing-masing algoritma akan diuji dalam dua skenario berbeda, yakni tanpa menggunakan data sentimen harian dan dengan menggunakan data sentimen harian. Detail proses pelatihan dijelaskan sebagai berikut:

A. Rentang Waktu *Dataset*

Data yang digunakan untuk pelatihan diambil dari *dataset* AMRT.csv, yang mencakup harga saham harian PT Sumber Alfaria Trijaya Tbk (AMRT). Rentang data yang digunakan dalam proses *training* adalah data dari tanggal 20 Januari 2024 sampai 20 Januari 2025. *Dataset* ini digunakan secara konsisten pada seluruh algoritma dan skenario untuk menjaga kesetaraan pembandingan.

B. Proses Pelatihan Model Prediksi Tanpa Sentimen

Pada tahap awal, setiap algoritma akan dilatih dan diuji dengan berbagai kombinasi fitur tanpa menyertakan fitur sentimen harian. Tujuan dari tahap ini adalah untuk menemukan kombinasi fitur terbaik yang menghasilkan model dengan rata-rata nilai R^2 tertinggi berdasarkan evaluasi K-Fold *Cross Validation*. Proses ini digambarkan pada gambar 3.6.



Gambar 3. 6 Proses pembuatan model prediksi tanpa sentimen

Seperti yang terlihat pada gambar 3.5, fitur yang akan digunakan dalam proses ini adalah fitur yang memiliki nilai korelasi tinggi yang telah didapat dari proses uji korelasi pada proses EDA. Setiap kombinasi fitur akan diuji

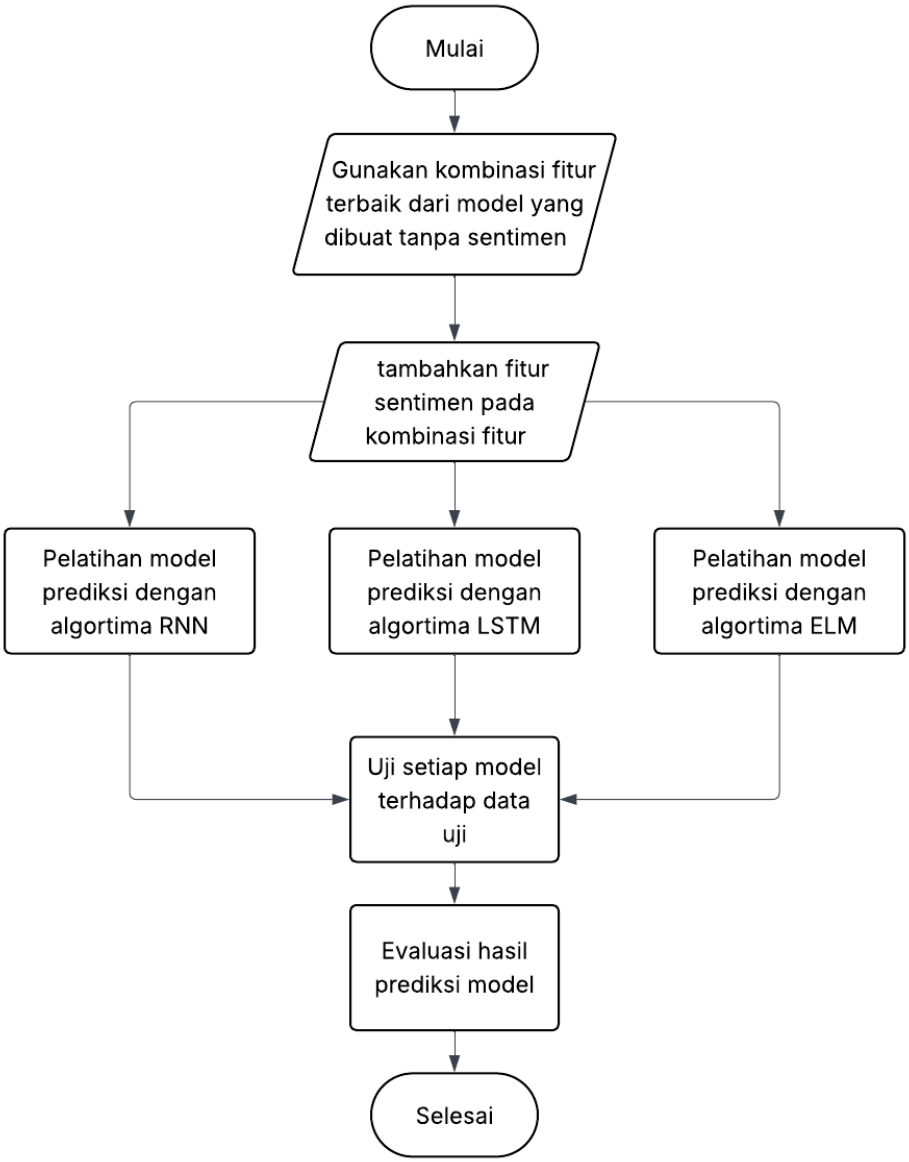
menggunakan K-Fold *Cross Validation*, dan nilai R^2 rata-rata pada data uji dari setiap *fold* akan dicatat. Model dengan kombinasi fitur dan parameter yang menghasilkan nilai R^2 tertinggi akan dipilih sebagai model terbaik tanpa data sentimen untuk masing-masing algoritma.

C. Komparasi dan Seleksi Model

Jika seluruh proses pelatihan telah selesai dilakukan, hasil dari setiap model yang dihasilkan dari setiap kombinasi fitur pada ketiga algoritma yang telah disebutkan akan dibandingkan. Model yang menghasilkan nilai rata-rata R^2 tertinggi dari pengujian K-Fold *cross validation* akan dipilih sebagai perwakilan dari algoritma tersebut. Kombinasi yang digunakan dalam melatih model tersebut juga akan digunakan pada proses pelatihan model prediksi degan data sentimen.

D. Proses Pelatihan Model Prediksi Dengan Sentimen

Jika kombinasi fitur terbaik telah diperoleh, maka proses pelatihan ulang pada model akan dilakukan dengan menggunakan kombinasi fitur terbaik yang sama, namun kali ini ditambahkan fitur tambahan yaitu data sentimen harian. Proses pelatihan yang sama dengan proses pelatihan model tanpa data sentimen akan digunakan kembali pada proses pelatihan model ini. Dengan menambahkan fitur sentimen, penelitian ini bertujuan untuk mengamati apakah keberadaan data opini publik dari media sosial X dapat meningkatkan performa model prediksi harga saham, ditinjau dari nilai evaluasi rata-rata R^2. Proses pelatihan model prediksi dengan sentimen dapat dilihat pada gambar 3.7.



Gambar 3. 7 Proses pembuatan model prediksi dengan integrasi data sentimen

3.4 Pengujian

Pada proses ini, setiap model yang sudah dipilih dari tiap algoritma akan diuji terhadap data *test.* data *test* adalah data dari *dataset* AMRT.csv dari tanggal 21 Januari 2025 hingga 21 Februari 2025.

Hasil prediksi dari model yang diuji akan dibandingkan dengan data aktual. Perbandingan ini dilakukan menggunakan *linechart*. Selain itu, hasil

prediksi model akan dievaluasi menggunakan metrik evaluasi yang sudah ditentukan.

3.5 Evaluasi

Proses ini akan mengevaluasi hasil penelitian dan menghasilkan kesimpulan serta saran dari penelitian yang telah dilakukan. Pada proses ini, data hasil evaluasi model dari masing-masing algoritma yang dipilih akan dibandingkan. Model dibandingkan berdasarkan skor yang dihasilkan metrik evaluasi MAPE, RMSE, dan R^2 dari hasil prediksi terhadap data *test*. Hasil komparasi tersebut akan menjadi acuan analisis dampak integrasi data sentimen masyarakat Indonesia terhadap *dataset* AMRT.csv terhadap hasil prediksi harga saham AMRT menggunakan algoritma *deep learning*. Apakah integrasi data sentimen ini berdampak positif dan dapat menaikkan akurasi model, atau malah berdampak negatif dan mengurangi hasil akurasi model

Hasil evaluasi ini akan menjadi dasar dalam pengambilan kesimpulan terhadap efektivitas integrasi data sentimen ke dalam sistem prediksi harga saham. Nilai yang didapat dari metrik evaluasi juga akan digunakan untuk komparasi performa model tanpa sentimen dan dengan sentimen. Hasil komparasi tersebut akan menjadi kesimpulan dari penelitian ini, apakah integrasi data sentimen terhadap *dataset* saham AMRT dapat menambah akurasi model prediksi