

BAB I

PENDAHULUAN

A. Latar Belakang

Pasar saham adalah instrumen investasi yang sangat dinamis dan kompleks, di mana harga terbentuk dari interaksi berbagai faktor: kinerja fundamental perusahaan, arah kebijakan moneter dan fiskal, ekspektasi investor, serta peristiwa geopolitik dan makroekonomi global [1]. Sebagai contoh, perubahan suku bunga bank sentral atau berita krisis kesehatan dapat memicu reaksi berantai pada harga saham secara instan. Tingkat fluktuasi ekstrem seperti yang terjadi selama pandemi COVID-19 menunjukkan bahwa metode statistik linier tradisional sering gagal merespons perubahan pola secara cepat. Oleh karena itu, dibutuhkan pendekatan yang mampu menangkap dinamika non-linear, efek memori jangka panjang, serta memperhitungkan faktor-faktor psikologis dan sentimen publik.

Model *Autoregressive Integrated Moving Average* (ARIMA) pertama kali diperkenalkan dalam [1] karena kemampuannya memisahkan komponen *autoregresif* (AR), *integrasi* (I), dan *moving average* (MA) untuk meramalkan tren dan musiman. Namun, asumsi kestabilan pola statistik dan hubungan linier dalam model ARIMA menjadi kelemahan kritis saat menghadapi pola kompleks atau lonjakan tiba-tiba, terutama dalam data finansial yang sarat ketidakpastian [2]. Keterbatasan inilah yang membuka ruang bagi pendekatan seperti *Long Short-Term Memory* (LSTM) yang dirancang khusus untuk menangkap hubungan non-linier dan ketergantungan jangka panjang.

Long Short-Term Memory (LSTM) pertama kali diperkenalkan dalam [2] untuk memperluas kemampuan *Recurrent Neural Network* (RNN) dengan menambahkan “gate” yang mengatur aliran informasi, sehingga mengatasi vanishing gradient dan mempertahankan konteks jangka panjang. Penelitian [3] menunjukkan bahwa LSTM dapat menurunkan MSE hingga 20% dibandingkan

ARIMA, dan kombinasi stacked autoencoder dan LSTM terbukti memberikan akurasi lebih stabil pada berbagai skenario pasar [4]. Studi [5] memperkuat validitas penggunaan LSTM pada prediksi deret waktu non-linear. Namun, sebagian besar penelitian tersebut belum mengeksplorasi peran informasi sentimen pasar, maupun kinerja model dalam menghadapi periode fluktuasi ekstrem seperti pandemi COVID-19.

Dalam beberapa penelitian terakhir, data sentimen dari media sosial semakin diakui sebagai variabel penting yang semula bersifat eksogen, namun kini terbukti memiliki pengaruh signifikan terhadap pembentukan harga saham. Penelitian [6] memanfaatkan pola emosional *tweet* untuk memprediksi *Dow Jones Industrial Average* (DJIA) dengan akurasi hingga 87%, sedangkan penelitian [7] menggabungkan fitur sentimen dengan algoritma *XGBoost* untuk meningkatkan nilai R^2 hingga 8%. Walaupun demikian, masih terbatas penelitian yang mengintegrasikan sentimen secara langsung ke dalam arsitektur LSTM dan membandingkannya secara paralel dengan ARIMA, terutama untuk menganalisis dampak sentimen terhadap prediksi pada tiga fase pasar: sebelum, selama, dan setelah pandemi COVID-19.

B. Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, terdapat beberapa permasalahan yang dapat dirumuskan dalam penelitian ini:

1. Bagaimana menggabungkan data sentimen dari media sosial dengan data saham untuk melihat dampak sentimen terhadap prediksi harga saham?.
2. Seberapa baik kinerja model LSTM dibandingkan dengan *baseline model* ARIMA dalam hal akurasi dan efektivitas untuk prediksi harga saham?.

C. Batasan Masalah

Untuk menjaga fokus penelitian dan menghindari pembahasan yang terlalu luas, penelitian ini dibatasi pada ruang lingkup tertentu agar hasil yang diperoleh lebih terarah. Adapun batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Penelitian ini menggunakan data historis harga penutupan saham *Apple Inc.* (AAPL) periode 2017-2024 dari Yahoo Finance.
2. Penelitian ini hanya membandingkan performa model LSTM dengan *baseline model* ARIMA
3. Penelitian ini menggunakan data Twitter dengan analisis sentimen VADER. Nilai netral (0) digunakan jika tidak ada tweet.
4. Evaluasi kinerja model dalam penelitian ini hanya dilakukan dengan menggunakan metrik seperti *Mean Squared Error* (MSE), *Mean Absolute Error* (MAE), koefisien determinasi (R^2), dan *Mean Absolute Percentage Error* (MAPE).

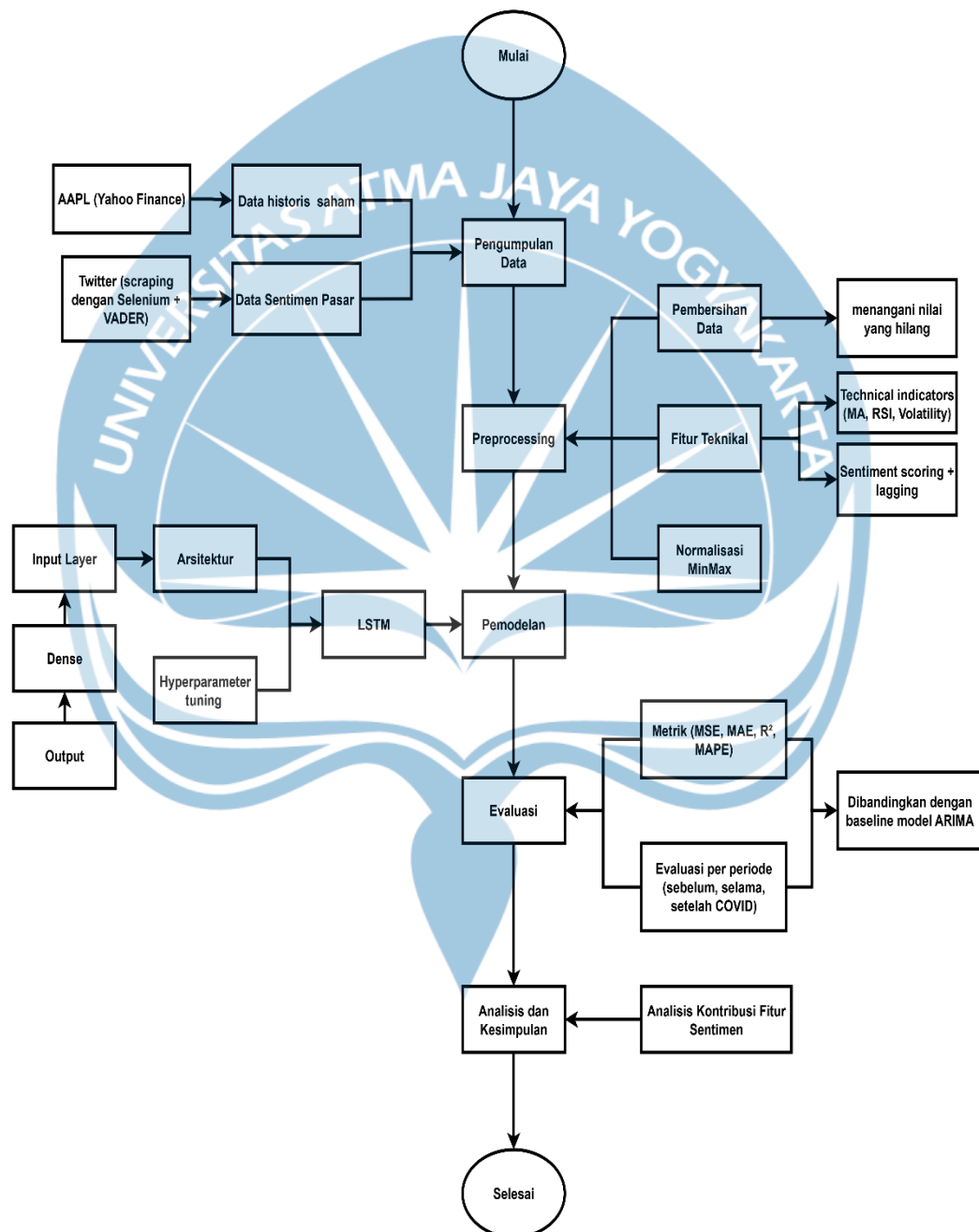
D. Tujuan Penelitian

Penelitian ini bertujuan untuk menjawab rumusan masalah yang telah disampaikan dan memberikan gambaran mengenai pengaruh data sentimen terhadap prediksi harga saham. Tujuan yang ingin dicapai yaitu:

1. Menerapkan metode penggabungan data sentimen dari media sosial dengan data saham untuk melihat dampak sentimen terhadap prediksi harga saham.
2. Membandingkan kinerja model LSTM dengan *baseline model* ARIMA dalam hal akurasi dan efektivitas prediksi harga saham.

E. Metode Penelitian

Penelitian ini dilakukan dalam beberapa tahap seperti tampak pada Gambar 1.1.



Gambar 1.1. Alur Metode Penelitian

Beberapa langkah – langkah yang dilakukan yaitu:

1. Pengumpulan Data

Penelitian ini mengumpulkan dua jenis data utama: data historis saham AAPL (2018–2024) dari Yahoo Finance menggunakan library *yfinance*, dan data sentimen pasar dari Twitter melalui teknik scraping berbasis Selenium dengan modul *Twitter_Scraper* serta analisis VADER [6], [7]. Data saham mencakup harga penutupan harian (*Close*), sementara data Twitter difokuskan pada tweet berbahasa Inggris yang mengandung kata kunci terkait AAPL dan pasar saham seperti “*stock*”, “*price*”, “*bullish/bearish*”, serta istilah spesifik Apple seperti “*iPhone*” atau “*Tim Cook*”. Proses scraping dilakukan dengan autentikasi akun dan dilengkapi filter berbasis daftar kata kunci positif, negatif, dan netral untuk menjaga relevansi, dengan batas maksimal 100 tweet per hari [8]. Data mentah hasil scraping kemudian dianalisis menggunakan algoritma VADER untuk menghasilkan skor sentimen harian, dan disimpan dalam format CSV guna menjamin reproduktibilitas penelitian.

2. Preprocessing

Tahap preprocessing diawali dengan penanganan *missing values* menggunakan forward-fill dan *interpolasi* linier untuk menjamin kelengkapan data deret waktu [9]. Selanjutnya dilakukan ekstraksi fitur teknikal meliputi *Moving Average* (MA-5/MA-20), *Relative Strength Index* (RSI-14), return harian, volatilitas 5-hari, momentum 5-hari, serta penerapan metode lagging dengan membuat fitur *lag* harga 1-5 hari sebelumnya untuk menangkap ketergantungan temporal [10]. Transformasi sentimen Twitter menjadi skor numerik menggunakan VADER [7] dilakukan dengan menerapkan *lag* 1 hari (*Sentiment_Lag1*) untuk mencerminkan dampak tertunda sentimen pada harga saham. Metode lagging ini secara khusus menciptakan fitur historis (y_{t-1} hingga y_{t-5}) yang merepresentasikan nilai harga masa lalu sebagai prediktor kunci. Seluruh fitur kemudian dinormalisasi ke rentang [0,1] menggunakan Min-Max Scaler untuk mencegah dominasi fitur berskala besar [11], termasuk fitur interaksi tambahan seperti rasio MA5/MA20 dan RSI/MA5 yang meningkatkan

kemampuan model menangkap dinamika pasar [10].

3. Pemodelan

Proses pemodelan diawali dengan implementasi arsitektur LSTM *single-layer* dengan 128-256 unit, mengikuti pendekatan dalam [3] yang terbukti efektif untuk data finansial. Optimasi *hyperparameter* dilakukan menggunakan *Bayesian Optimization* dengan 30 *trial* untuk menentukan kombinasi optimal unit LSTM, neuron pada *dense layer*, dan *learning rate* [3], [11]. Penelitian menerapkan determinisme ketat untuk menjamin hasil reproduisibel [12]. Model dasar kemudian mengalami *fine-tuning* spesifik periode (sebelum, selama, dan setelah COVID-19) dengan teknik *weight cloning* dan *optimizer state transfer* agar model dapat beradaptasi dengan konteks temporal tanpa kehilangan pengetahuan awal [5]. Setiap model periode diinisialisasi dengan bobot model utama, kemudian dilatih ulang pada subset data masing-masing fase pandemi.

4. Evaluasi

Evaluasi kinerja model dilakukan menggunakan empat metrik kuantitatif: MSE (memberikan penalti kuadratik pada deviasi besar), MAE (*robust* terhadap outlier), R^2 (mengkuantifikasi proporsi variansi), dan MAPE (interpretasi kesalahan relatif) [13], [14]. Proses ini dilaksanakan secara *in-sample* (seluruh periode 2018-2024) dan *out-of-sample* pada tiga subperiode COVID (sebelum, selama, setelah) mengikuti kerangka penelitian dalam [5]. Hasilnya kemudian dibandingkan dengan *baseline model* ARIMA untuk mengukur keunggulan adaptif LSTM dalam menangani fluktuasi pasar yang berbeda, dengan fokus analisis pada disparitas kinerja (khususnya MAPE dan R^2) di tiap fase pandemi.

5. Analisis Kontribusi Fitur Sentimen

Analisis ini dilakukan untuk mengukur kontribusi variabel sentimen pasar dengan membandingkan dua skenario paralel: model lengkap yang mencakup fitur sentimen dan model tanpa fitur sentimen [12]. Proses ini menggunakan arsitektur identik dan prosedur *fine-tuning* yang sama untuk setiap periode COVID-19, di

mana selisih performa (terutama pada metrik MAPE dan R^2) dikuantifikasi untuk mengevaluasi dampak eksogen sentimen terhadap akurasi prediksi. Seluruh hasil termasuk prediksi, metrik evaluasi, dan visualisasi diekspor dalam format terstruktur (CSV/JSON) untuk mendukung analisis komparatif lebih lanjut dan menjamin transparansi penelitian.



F. Sistematika Penulisan

Penulisan laporan tugas akhir ini terdiri atas enam bab dengan urutan sebagai berikut:

1. BAB I Pendahuluan

Bab ini menguraikan latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, metode penelitian, dan sistematika penulisan.

2. BAB II Tinjauan Pustaka

Bab ini berisi analisis komparatif penelitian terdahulu terkait prediksi harga saham, model tradisional dan modern, serta pemanfaatan data sentimen, dilengkapi tabel gap penelitian.

3. BAB III Landasan Teori

Bab ini menjelaskan dasar teoretis model ARIMA, arsitektur LSTM, teknik sinkronisasi data (*reindex*, *resample*, *interpolasi*), fitur teknikal (MA, RSI, volatilitas), dan metrik evaluasi (MSE, MAE, MAPE, R^2).

4. BAB IV Dataset dan Pemrosesan Data

Bab ini memaparkan deskripsi dataset (harga saham AAPL dan sentimen Twitter), tahap pra-pemrosesan (penanganan *missing values*, *lagging* sentimen), *feature engineering* (fitur teknikal dan interaksi), normalisasi, serta pembagian data per periode COVID-19.

5. BAB V Implementasi Model dan Pengujian Sistem

Bab ini mendetailkan implementasi model LSTM (*hyperparameter tuning*, *stateful transfer learning*), perbandingan kinerja model, evaluasi konsistensi output, dan analisis kontribusi fitur sentimen.

6. BAB VI Penutup

Bab ini menyajikan kesimpulan temuan penelitian dan saran pengembangan lebih lanjut.