CHAPTER 4

INITIAL FINDINGS

4.1 Introduction

This chapter will discuss about the accuracy of the model to predict stock market prices for Malaysian banking (CIMB and Maybank) based on sentiment analysis of news headline using machine learning. This chapter begin with the step about data collection, preprocessing, exploratory data analysis (EDA), developing model and implementing model using machine learning techniques. The machine learning used in this project are LSTM, GRU and ACNN-LSTM. Based on the results of machine learning, it was found that the ACNN-LSTM model techniques had highest percentage of accuracy compared to others model. The details of the results and analysis are presented in this chapter.

4.2 Data Collection

To collect news article related to CIMB and Maybank, the data collection process was carried out using the multi-threaded Selenium-based scraping framework. The news headlines data was retrieved from five major websites which are Malay Mail, The Edge Market, Business Today, New Straits Time and The Star. Stock price data was obtain using the Yahoo Finance (yfinance) for both CIMB and Maybank. After the data was collected, the data was saved in CSV file format to facilitate for further analysis.

The total data collected are display below:

Bank	Total Articles Scraped	Sources
CIMB	14 369	Malay Mail, The Edge Market, Business Today,
Maybank	17 853	

		New Straits Time, The Star

Table 4-1 Dataset of News Headline

Each record includes the news headline, published date and category/section.

	headline	date	category
1	99 Speedmart's Extended Hours Could Enhance Groups SSSG: CIMB	June 9, 2025	NEWS
2	CIMB Thai Shares Suspended Over Public Float Rules	June 6, 2025	NEWS
3	CIMB Remains Concerned On 7-Eleven's Long Term Outlook	June 6, 2025	NEWS
4	CIMB Has Tools To Navigate Tariff Headwinds	June 3, 2025	NEWS
5	Today's Shares: CIMB Slips 1.15% Despite Analysts' Optimism Post-1QFY25 Results	June 3, 2025	MARKETS

Figure 4-1News Headline data

Yahoo Finance (via yfinace library) was used to scraped historical stock data between 2019-01-01 and 2025-12-31

Bank	Total Daily Records	Ticker
CIMB	1581	1023.KL
Maybank		1155.KL

Table 4-2 Yahoo Finance Dataset

All the data are compress in RNDPM file and will be upload to google colab for further analysis.

4.3 Data Preprocessing

Data preprocessing was conducted before sentiment analysis and correlation with financial stock data.

4.3.1 Date Formatting and Standardization

- News Headline

  The date format for news portal is vary due to the different data structure of news websites. Some entries contained delimiters such as pipes (|) or symbols (@) which are used to separate date and time elements. To standardise this, it

was split into separate Date and Time columns. The Date values were then reformatted into ISO format (YYYY-MM-DD) for consistency across all the datasets.

- Stock Market Data

  Yahoo finance date column contains both time and date data. The date and time string were further split into Date and Time column. The Date was formatted but the time column is removed because it was redundant data.

4.3.2 Column Reduction

- News Headline

  The Time column was removed from news articles datasets. It is because the Time column was not relevant in this study.

- Stock Market Data

  Yahoo Finance contain Dividends and Stock Split columns. Both columns are not needed for this analysis thus it was dropped during preprocessing.

4.3.3 Category Filtering

To improve the quality of sentiment analysis, only financial related articles were kept. Other than that, are discard to reduce noise from the data. For example, in The Edge Markets, categories such as ‘Market close,’ ‘Hot stock’ and ‘Market Open’ were retained while ‘Aviation’ and ‘Branded’ were removed. The same filter was also applied to others news dataset. Additionally, the section column in Malay Mail was renamed to category. It is to maintain the consistency across sources.

4.3.4 Dataset Merging, Re-Indexing and Null Handling

After all the individual files were cleaned, they were grouped by bank (CIMB and Maybank). The group was then merged using pandas.concat() into two unified datasets. Then, the merge datasets only retained the record between January 1, 2019 and June 13, 2025. This is to ensure the alignment with the available stock data for the comprehensive analysis. Any null values in the Date and Category column were dropped.

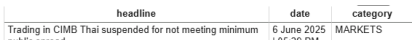
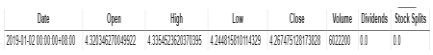
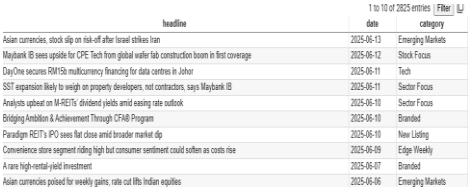
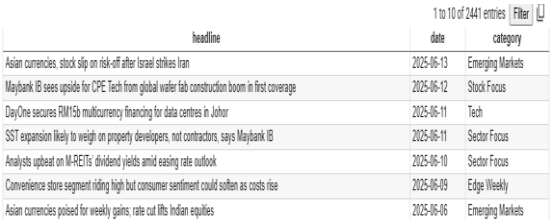
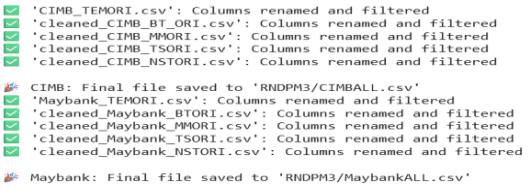
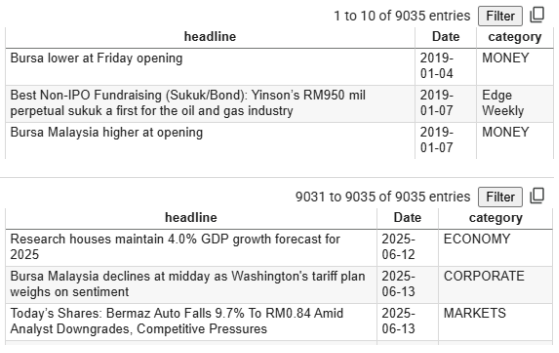
Preprocessing step	**Before preprocess**	**After preprocess**
Reformat date and drop column		
Reformate date and drop column		
Category Filtering		
Merging		
Re-Indexing		

Table 4-3 Preprocessing Input and Output of Dataset

4.4 Exploratory Data Analysis (EDA)

Data preprocessing was conducted before sentiment analysis and correlation with financial stock data.

4.4.1 Data Cleaning

Data cleaning is one of the most important steps when dealing with web-scraped data. It is due to the raw data of news headlines are usually have inconsistent formatting, contain missing values, noise, and a redundant field. By cleaning the data, the quality of sentiment analysis and statical modelling are expected to improved.

All character were converted to lowercase to ensure the consistency. It also will help in reducing the lexical duplication. The digits, punctuation and other non-alphabetic characters are eliminated by using the regular expressions. Extra spaces and tabs were removed from the text. NLTK was deployed to removed the English stop words and retain only the meaningful words. Any missing headline values were replaced with empty string. This will prevent the processing errors during analysis. This process was applied to CIMB and Maybank dataset. To reduce noise and redundancy, only three columns are kept ensuring the efficiency of analysis and modelling

1 to 10 of 9035 entries Filter

headline	Date	category
bursa lower friday opening	2019-01-04	MONEY
best nonipo fundraising sukukbond yinsons rm mil perpetual sukuk first oil gas industry	2019-01-07	Edge Weekly
bursa malaysia higher opening	2019-01-07	MONEY
bursa malaysia remains higher midmorning	2019-01-07	MONEY
bursa malaysia remains positive territory midday	2019-01-07	MONEY
bursa malaysia remains higher midafternoon	2019-01-07	MONEY
best privatisation timely sweet exit oldtown shareholders	2019-01-08	Edge Weekly
best share placement khazanah sells rm mil cimb block tight discount	2019-01-08	Edge Weekly
disappointments year year full disappointments bursa	2019-01-08	Edge Weekly
bursa marginally higher opening	2019-01-08	MONEY

Figure 4-2 Cleaned dataset of News headline

4.1.1 Sentiment Analysis

Sentiment analysis is performed using VADER (Valence Aware Dictionary and Sentiment Reasoner) tool. The tool is capable to quantify the emotional tone of the news headline and it effective for domain-specific analysis like financial news. Each headline was passed through VADER's sentiment analyser. The analyser will categorise the score into positive, neutral and negatives. Table below show the compound score that used in the classification of the sentiment label

Compound Score	Sentiment label
>0.05	Positive
<-0.05	Negative
Else	Neutral

Table 4-4 Sentiment Analysis Compound score

The analysis is applied to CIMB and Maybank headlines dataset.

1 to 10 of 9035 entries Filter

headline	Date	category	Sentiment
bursa lower friday opening	2019-01-04	MONEY	-0.296
best nonipo fundraising sukukbond yinsons rm mil perpetual sukuk first oil gas industry	2019-01-07	Edge Weekly	0.6369
bursa malaysia higher opening	2019-01-07	MONEY	0.0
bursa malaysia remains higher midmorning	2019-01-07	MONEY	0.0
bursa malaysia remains positive territory midday	2019-01-07	MONEY	0.5574
bursa malaysia remains higher midafternoon	2019-01-07	MONEY	0.0
best privatisation timely sweet exit oldtown shareholders	2019-01-08	Edge Weekly	0.802
best share placement khazanah sells rm mil cimb block tight discount	2019-01-08	Edge Weekly	0.5423
disappointments year year full disappointments bursa	2019-01-08	Edge Weekly	-0.7184
bursa marginally higher opening	2019-01-08	MONEY	0.0

Figure 4-3 Sentiment Analysis of News Headline

Daily sentiment scores are calculated by averaging all news headlines for each day before combining it with stock prices data. The total of positive, negative, and neutral headlines is also calculated.

1 to 10 of 1871 entries Filter

Date	Open	High	Low	Close	Volume	Dividends	Stock Splits	Avg_Sentiment	Max_Sentiment	Min_Sentiment	Headline_Count	Positive_Headlines	Negative_Headlines	Price_Up
2019-01-02	4.320346270049922	4.33545236203704	4.244815810114329	4.267475128173828	6022200.0	0.0	0.0							0
2019-01-03	4.2901340860756845	4.320346270049922	4.244815810114329	4.267475128173828	5198100.0	0.0	0.0							0
2019-01-04	4.252368927001953	4.282581102200455	4.244815701123511	4.252368927001953	3506600.0	0.0	0.0	-0.296	-0.296	-0.296	1.0	0.0	1.0	1
2019-01-07	4.282581348314875	4.312793531179086	4.259922027587891	4.259922027587891	10358800.0	0.0	0.0	0.23886	0.6369	0.0	5.0	2.0	0.0	1
2019-01-08	4.282581220160947	4.312793404135183	4.259921902101448	4.267475128173828	11516600.0	0.0	0.0	0.0086555555555555	0.802	-0.7184	9.0	3.0	4.0	0
2019-01-09	4.282581220160947	4.297687312148065	4.259921902101448	4.267475128173828	12079200.0	0.0	0.0	0.2361833333333333	0.802	0.0	6.0	3.0	0.0	0
2019-01-10	4.275027994088567	4.297687312148065	4.244815810114329	4.267475128173828	8114200.0	0.0	0.0	0.2553	0.5106	0.0	2.0	1.0	0.0	0
2019-01-11	4.275027994088567	4.2901340860756845	4.244815810114329	4.267475128173828	6385100.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0
2019-01-14	4.267475128173828	4.2901340860756845	4.244815810114329	4.267475128173828	4324900.0	0.0	0.0	0.0636	0.4215	-0.4404	6.0	3.0	2.0	1
2019-01-15	4.267474286663064	4.305239329105084	4.252368197654742	4.282580375671387	9463200.0	0.0	0.0	0.23091	0.6369	-0.296	10.0	5.0	1.0	0

Figure 4-4 Daily Sentiment Dataset

The daily sentiment analysis data was then merged with daily stock prices data for CIMB and Maybank using the date as a key. The missing stock value were forward filled while missing sentiment value are set to zero. The clean dataset for each bank then containing both sentiment and financial information and can be used for further analysis.

11 to 20 of 1871 entries Filter

Date	Open	High	Low	Close	Volume	Avg_Sentiment	Max_Sentiment	Min_Sentiment	Headline_Count	Positive_Headlines	Negative_Headlines	Price_Up	No_News
2019-01-16	4.282581220160947	4.282581220160947	4.259921902101448	4.267475128173828	5718700.0	-0.008125	0.6486	-0.4404	8.0	2.0	3.0	0	0
2019-01-17	4.267475128173828	4.267475128173828	4.244815810114329	4.267475128173828	6605800.0	0.0683	0.2732	0.0	4.0	1.0	0.0	1	0
2019-01-18	4.267474286663064	4.290133240096762	4.252368197654742	4.282580375671387	5980300.0	0.1307	0.5574	-0.296	2.0	1.0	1.0	0	0
2019-01-20	4.267474286663064	4.290133240096762	4.252368197654742	4.282580375671387	5980300.0	0.2732	0.2732	0.2732	1.0	1.0	0.0	0	0
2019-01-21	4.267474286663064	4.290133240096762	4.252368197654742	4.282580375671387	5980300.0	0.2263	0.2263	0.2263	1.0	1.0	0.0	0	0
2019-01-22	4.320346446776484	4.35811185836792	4.282581395342706	4.35811185836792	16396400.0	0.1103857142857142	0.2732	0.0	7.0	3.0	0.0	0	0
2019-01-23	4.320346025316835	4.343005342092759	4.275027751922607	4.275027751922607	12430100.0	0.0793799999999999	0.7003	-0.5994	5.0	2.0	1.0	0	0
2019-01-24	4.2901340860756845	4.312793404135183	4.244815810114329	4.267475128173828	13410700.0	0.1646333333333333	0.4939	0.0	3.0	1.0	0.0	1	0
2019-01-25	4.297687178932736	4.297687178932736	4.267474995894985	4.290133963094482	11523500.0	0.0	0.0	0.0	1.0	0.0	0.0	0	0
2019-01-28	4.305239934185425	4.305239934185425	4.259921660791198	4.275027751922607	10608400.0	-0.2047333333333333	0.0	-0.3182	3.0	0.0	2.0	1	0

Figure 4-5 Sentiment Analysis with Stock Market

4.1.2 Descriptive Statistics

After preprocessing, each dataset contains 14 columns and 1871 rows for CIMB and 1876 rows for Maybank. The dataset ranges from 2 January 2019 to 13 June 2025 and has no null value. The date is also aligned for both sentiment and stock data.

- Stock price distribution

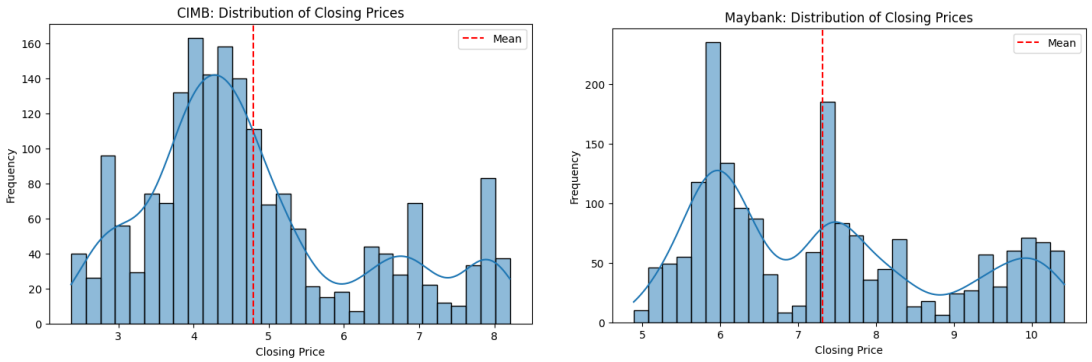


Figure 4-6 Stock price distribution chart for CIMB and Maybank

Based on the graph, CIMB showed a unimodal distribution centered around RM 4.25 with lower volatility. Maybank on the other hand displayed a wider spread and it peaked around RM 5.85-RM 6.20 reflecting a broader trading range

- Correlation heatmaps

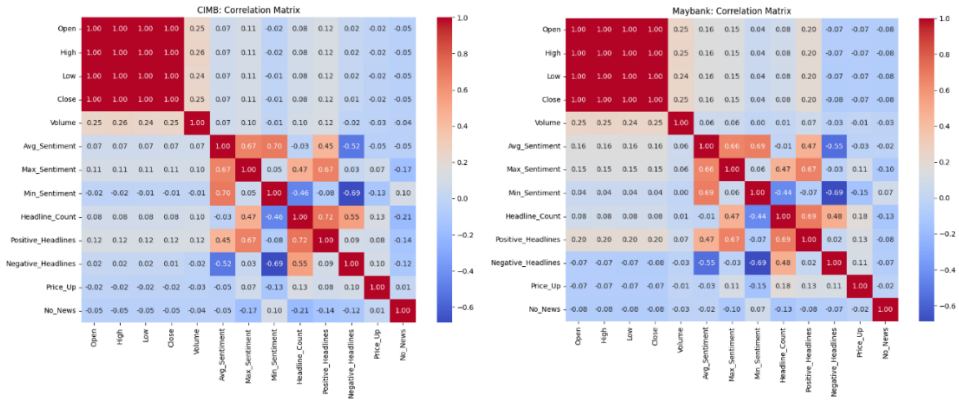


Figure 4-7 Correlation heatmaps for CIMB and Maybank

The heatmap indicated that sentiment indicators such as Avg_Sentiment, Positive_Headlines, Negatives_Headlines had weak positive correlations with price variables. This indicated that sentiment of the news headlines is not directly influences the price stock market.

- Time Series for Price vs Sentimet

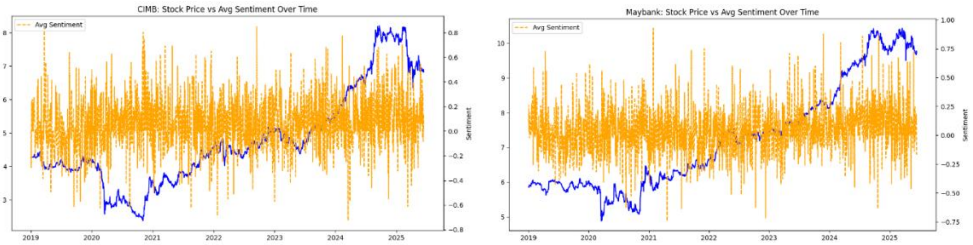


Figure 4-8 Time Series Chart for Price vs Sentimet for CIMB and Maybank

Based on the time series plot, CIMB showed no strong pattern between sentiment and Price. Maybank on the other hand has increases in average sentiment with the upward trends. Maybank showed strong evidence of the relation between sentiment and prices.

- Sentiment Score Distribution

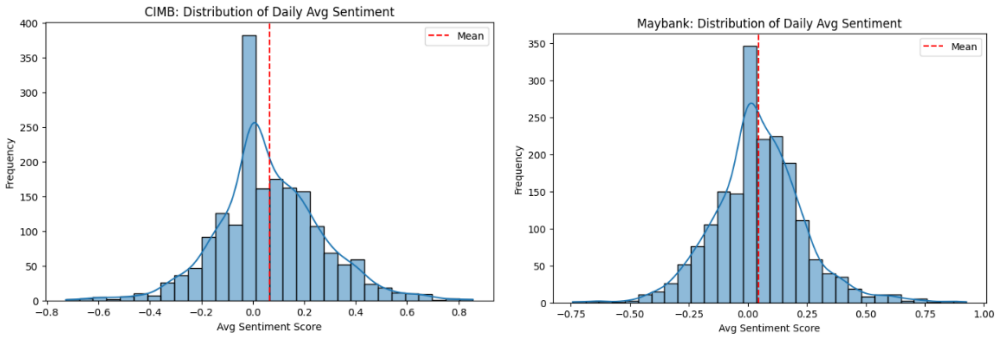


Figure 4-9 Sentiment Score Distribution for CIMB and Maybank

The distribution graph showed that CIMB sentiment distribution was mostly around 0 indicated most of the news are natural. Maybank slightly skewed to the right proving there are more positive sentiment in its news coverage.

- Headline Volume vs Stock Price

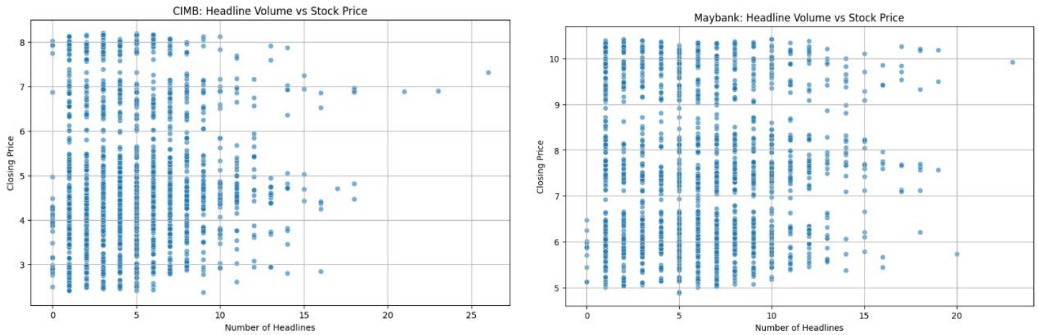


Figure 4-10 Headline Volume vs Stock Price for CIMB and Maybank

Scatterplots between headline count and close price showed that there is no linear trend between these two variables. However major headline volume spikes in Maybank are followed with stock movement activity. This validate that headline count is one of the contributors to market relevant events.

- Sentiment and the closing price

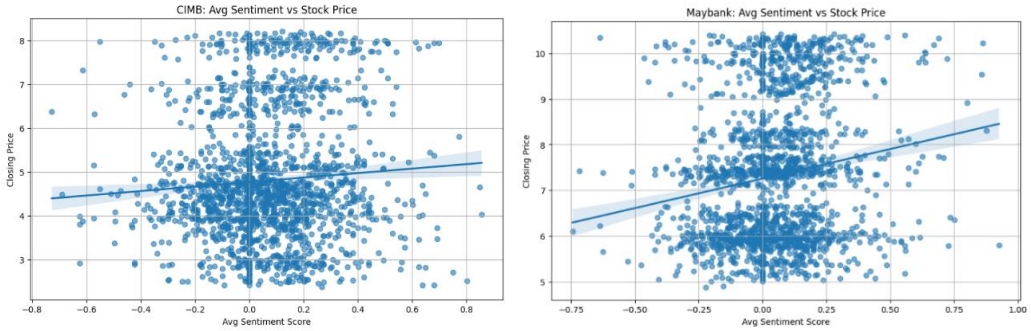


Figure 4-11 Sentiment and the closing price for CIMB and Maybank

The graph revealed that the relationship between Avg_Sentiment and Close price are flat and weak for CIMB. For Maybank, the slope was inclined a bit indicated that higher sentiment related with stock prices market.

- Feature Importance Summary for Stock Price Prediction

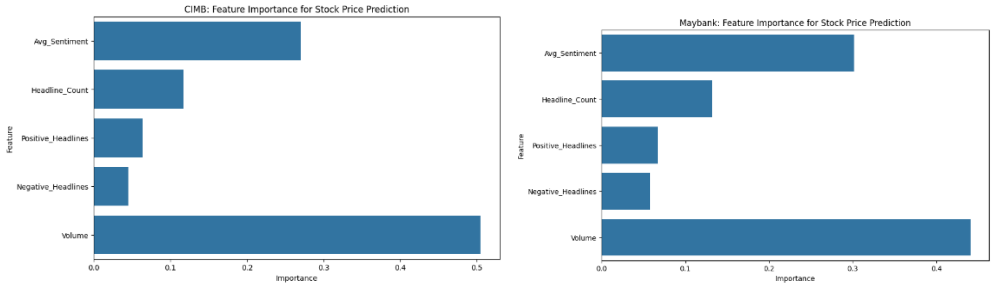


Figure 4-12 Feature Importance Summary for Stock Price Prediction

The bar charts above summarize which features were most important in predicting the stock prices of CIMB and Maybank. Volume had the highest importance indicating that trading activity for both banks affects price movements. The Average Sentiment Score is the second. It means that news sentiment has a noticeable but smaller impact.

4.5 Model Selection, Training & Evaluation

This study compared three deep-learning model to predict the next-day closing price prediction of stock market. A sliding-window of 5 days formed input sequences. 80% of each series used for training while another 20% for testing. To prevent over-

fitting, all model is then trained using Adam optimiser, MSE loss, mini batch=32 and early-stopping=10

Model	Key idea
LSTM	To capture lone-range temporal patterns
GRU	Fewer parameter
ACNN-LSTM	Extract short-term patterns with 1-D convolutions, then modelling the long-range context

Table 4-5 Core Idea of Model Selection

Evaluation matric used in this project are Mean Squared Error (MSE), square-root (RMSE) and Mean Absolute Error (MAE). Table below showed the result of each model.

Bank	Model	MSE	RMSE	MAE
CIMB	LSTM	0.441	0.664	0.601
	GRU	0.547	0.740	0.665
	ACNN-LSTM	0.060	0.244	0.187
Maybank	LSTM	1.009	1.005	0.969
	GRU	0.522	0.723	0.689
	ACNN-LSTM	0.054	0.232	0.177

Table 4-6 Evaluation of Model

From the table, it can be concluded that ACNN-LSTM model produces the lowest value compare to another model.

To visually compare prediction performance, six-line graphs were generated, each showing the actual closing price vs. predicted price from the test set

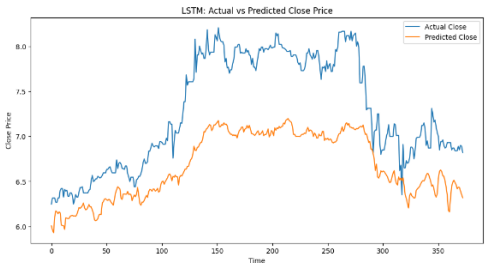
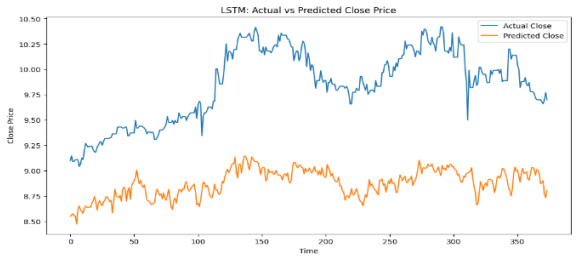
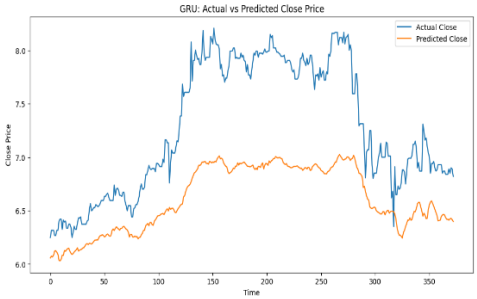
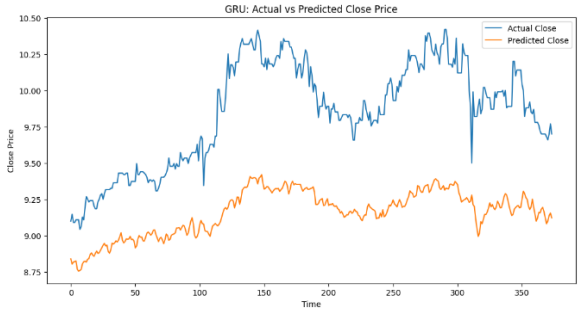
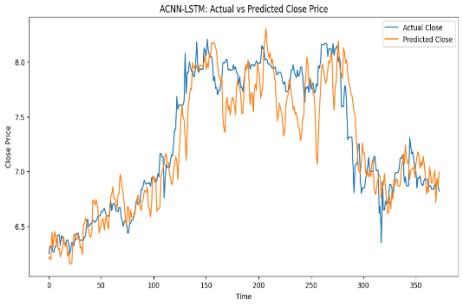
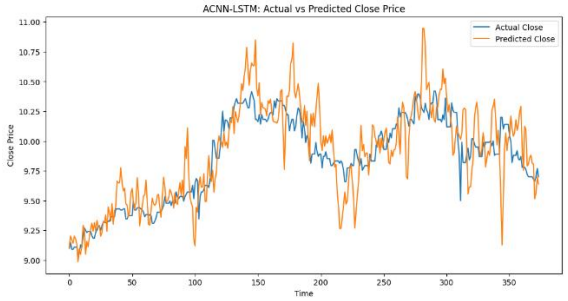
	CIMB	MAYBANK
LSTM		
GRU		
ACNN-LSTM		

Table 4-7 Line Graph for Deep Learning Model

From the table, the convolutional front-end in ACNN LSTM proved to capture short-horizontal patterns. After that, the dense layers of the model will learn longer-term dependencies thus contribute to the lowest prediction error for both CIMB and Maybank. For Maybank GRU outperformed LSTM but for CIMB, LSTM is superior compare to GRU. This show that parameter efficiency is not enough without feature extraction. Overall, ACNN-LSTM is the best model for forecasting and trading-strategy simulations for both banks