CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter will explain about the methodology and model used for this research. The goal is to predict the stock prices using sentiment analysis of financial news headlines. The developed model will incorporate both the stock price movement and text sentiments based on financial news articles. By integrating deep learning and sentiment analysis, the model aims to predict future price movement of stocks for CIMB and MAYBANK. This chapter will discover the model architecture, data processing techniques and the steps involved in training the model.

3.2 The Framework

This research consists of seven phases which are:
Phase 1: Problem Identification
Phase 2: Data Collection
Phase 3: Data Pre-processing
Phase 4: Exploratory Data Analysis (EDA)
Phase 5: Modelling
Phase 6: Model Evaluation
Phase 7: Deployment
The details of the research framework for this study are shown in the Figure 1 below

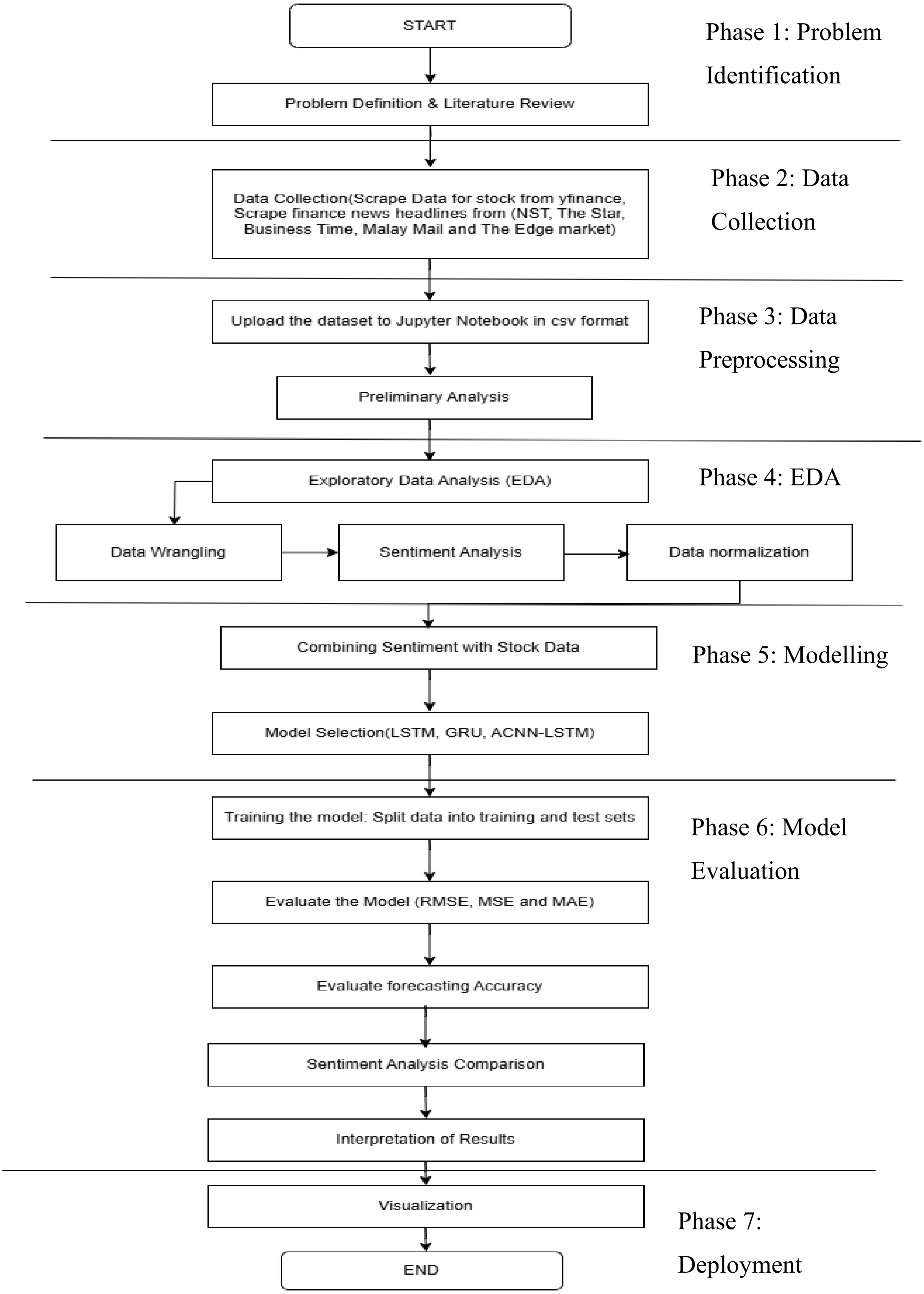

START
Problem Definition & Literature Review
Phase 1: Problem Identification
Data Collection(Scrape Data for stock from yfinance, Scrape finance news headlines from (NST, The Star, Business Time, Malay Mail and The Edge market)
Phase 2: Data Collection
Upload the dataset to Jupyter Notebook in csv format
Preliminary Analysis
Phase 3: Data Preprocessing
Exploratory Data Analysis (EDA)
Phase 4: EDA
Data Wrangling
Sentiment Analysis
Data normalization
Combining Sentiment with Stock Data
Model Selection(LSTM, GRU, ACNN-LSTM)
Phase 5: Modelling
Training the model: Split data into training and test sets
Evaluate the Model (RMSE, MSE and MAE)
Evaluate forecasting Accuracy
Sentiment Analysis Comparison
Interpretation of Results
Phase 6: Model Evaluation
Visualization
END
Phase 7: Deployment

Figure 3-1Research framework of Predicting stock market using sentiment analysis

3.3 Problem Identification

The primary goal of this research is to increase a system studying version that leverages sentiment evaluation of financial news headline to are expecting stock market movement. By using deep learning techniques including LSTM, GRU and ACNN-LSTM, this looks at objectives to capture the relationship among the sentiment in the information and inventory charge trends.

But there are several demanding situations need to be solved to ensure the version accuracy.

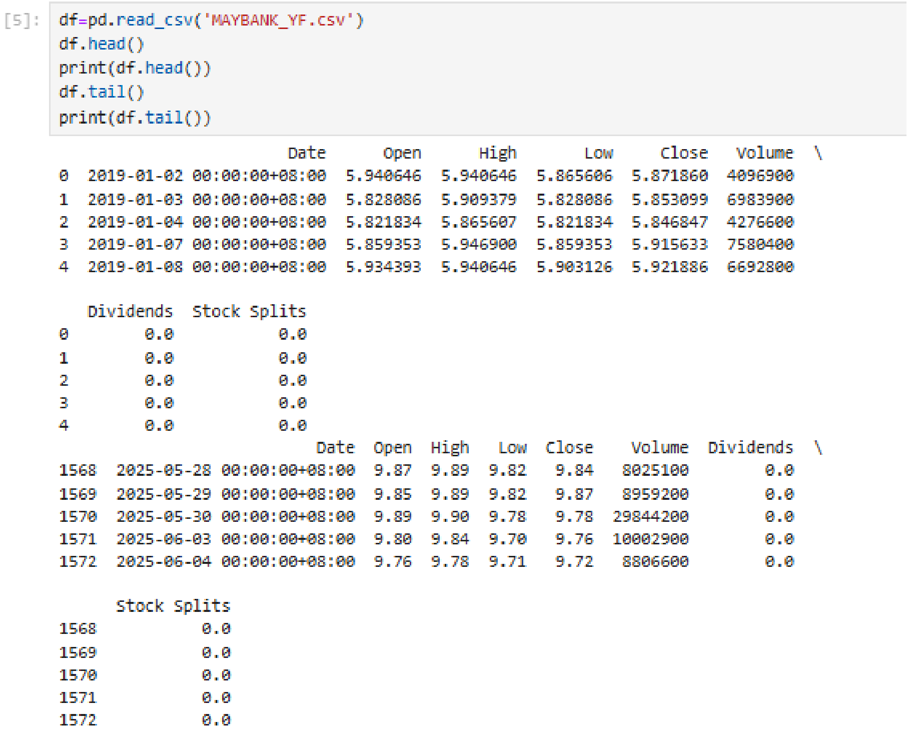
1. Making sure the facts great and consistency is the main problem as the dataset want to be complete. For these studies, the dataset will include stock fees and economic information headlines related to Maybank and CIMB from 2019 to 2025. The dataset must not have missing value and must properly handle. Only financial headlines and topic related to it are chosen to ensure the relevancy of the dataset.

2. The model must accurately capture the sentiment (positive, neutral and negatives) based on the financial news headlines. The use of sentiment analysis models such as VADER is important to make sure the financial news is properly interpreted. To train the model, the sentiment scores need to be combine with historical stock price data.

3. It is critical to account for the dynamic nature of the stock market. This because the stock prices movement are influenced by the geopolitical events, economic policies, and global market sentiments even it is not usually address in news headlines. The model designed must be able to adapt to market evolves also must consider the external factors.

3.4 Data Collection and Preprocessing

There are two datasets used in this study which are the stock price datasets and financial news headline. The stock prices datasets were extracted from CIMB and Maybank, two main blue-chip companies in banking sector in Bursa Malaysia. Both are chosen because of the financial performance and reputation. The yfinance, a python library tools that allow to access Yahoo Finance's stock data is used to fetch the stock price data. The datasets include daily stock prices open, high, low, close prices, and volume for the years 2019 to 2025. The data consist of 35384 records for both CIMB and MAYBANK



```
[5]: df=pd.read_csv('MAYBANK_YF.csv')
     df.head()
     print(df.head())
     df.tail()
     print(df.tail())
```

```
                        Date      Open      High       Low     Close   Volume  \
0  2019-01-02 00:00:00+08:00  5.940646  5.940646  5.865606  5.871860  4096900
1  2019-01-03 00:00:00+08:00  5.828086  5.909379  5.828086  5.853099  6983900
2  2019-01-04 00:00:00+08:00  5.821834  5.865607  5.821834  5.846847  4276600
3  2019-01-07 00:00:00+08:00  5.859353  5.946900  5.859353  5.915633  7580400
4  2019-01-08 00:00:00+08:00  5.934393  5.940646  5.903126  5.921886  6692800

   Dividends  Stock Splits
0        0.0           0.0
1        0.0           0.0
2        0.0           0.0
3        0.0           0.0
4        0.0           0.0
                           Date  Open  High   Low  Close    Volume  Dividends  \
1568  2025-05-28 00:00:00+08:00  9.87  9.89  9.82   9.84   8025100        0.0
1569  2025-05-29 00:00:00+08:00  9.85  9.89  9.82   9.87   8959200        0.0
1570  2025-05-30 00:00:00+08:00  9.89  9.90  9.78   9.78  29844200        0.0
1571  2025-06-03 00:00:00+08:00  9.80  9.84  9.70   9.76  10002900        0.0
1572  2025-06-04 00:00:00+08:00  9.76  9.78  9.71   9.72   8806600        0.0

      Stock Splits
1568           0.0
1569           0.0
1570           0.0
1571           0.0
1572           0.0
```

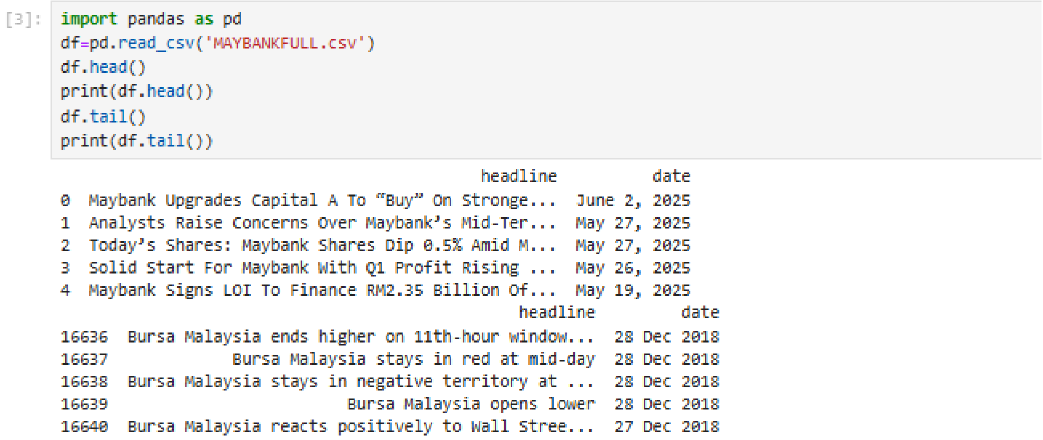
Figure 3-2 The code and information first 5 rows and last 5 rows of Maybank stock data

Same process was conducted for different dataset. Below are summarized of the data

Dataset Name	Source	Format	Number of rows
CIMB_YF	yfinance	.csv	1581
MAYBANK_YF	yfinance	.csv	1581

Table 3-1 Dataset for Stock Market

Second, Selenium is used as scrapping tools for financial news headlines from several finance news websites such as New Straits Times, The Star, Malay Mail, The Edge Market and Business Today. The Selenium can handle web pages that require user interaction like clicking next button or scrolling to load more content. Besides, the use of BeautifulSoup is limited for the websites that use JavaScript to load headlines and content. There is total 32222 headlines scrapped from this website related to CIMB, Maybank, business, and finance.



```
[3]: import pandas as pd
     df=pd.read_csv('MAYBANKFULL.csv')
     df.head()
     print(df.head())
     df.tail()
     print(df.tail())
```

```
                                         headline          date
0  Maybank Upgrades Capital A To “Buy” On Stronge...  June 2, 2025
1  Analysts Raise Concerns Over Maybank’s Mid-Ter...  May 27, 2025
2  Today’s Shares: Maybank Shares Dip 0.5% Amid M...  May 27, 2025
3  Solid Start For Maybank With Q1 Profit Rising ...  May 26, 2025
4  Maybank Signs LOI To Finance RM2.35 Billion Of...  May 19, 2025
                                             headline         date
16636  Bursa Malaysia ends higher on 11th-hour window...  28 Dec 2018
16637              Bursa Malaysia stays in red at mid-day  28 Dec 2018
16638  Bursa Malaysia stays in negative territory at ...  28 Dec 2018
16639                        Bursa Malaysia opens lower  28 Dec 2018
16640  Bursa Malaysia reacts positively to Wall Stree...  27 Dec 2018
```

Figure 3-3 The code and Information of first 5 rows and last 5 rows of Maybank news headlines

All the data uploaded in Jupyter Notebook for further analysis. Initial analysis dataset is performed to check the data structure, types, and missing value. Any duplicated data are handled properly.

Dataset Name	Source
CIMB_MMORI	https://www.malaymail.com/
CIMB_NSTORI	https://www.nst.com.my/business
CIMB_TSORI	https://www.thestar.com.my/
CIMB_BTORI	https://www.businesstimes.com.sg/
CIMB_TEMORI	https://theedgemalaysia.com/
CIMBALL	Merged all the dataset of CIMB headlines
MAYBANK_MMORI	https://www.malaymail.com/
MAYBANK_NSTORI	https://www.nst.com.my/business
MAYBANK_TS	https://www.thestar.com.my/
MAYBANK_BT	https://www.businesstimes.com.sg/

MAYBANK_TEM	https://theedgemalaysia.com/
MAYBANKFULL	Merged all the dataset of Maybank headlines

Table 3-2 Dataset for news headlines

3.5 Data Cleaning and Sentiment Analysis

The data cleaning process is proceeded after the dataset was uploaded and analysed. This step to ensure the quality of the datasets before used for machine learning training.

The unnecessary columns, news headlines and row are removed during the data wrangling. The headlines are cleaned by removing the special characters, numbers, and punctuation. The stock price data is then aligned with the news headline based on their timestamps, ensuring that each headline corresponds to the correct stock price data. The numbers of frequent words used in news and stock datasets are identify during EDA. Besides, it helps to understand trends, patterns, and common terms in the financial headlines.

News headlines are analysed using natural language toolkit (NLTK) sentiment VADER. It is a prebuilt tools used to extract sentiment from text data. The VADER model calculates the polarity value and categorized into Positive (>0.1), Negative (<-0.1) and Neutral (>-0.1, <0.1). The SentimentIntensityAnalyzer from VADER is applied to calculate the sentiment score for each headline. The score is then used to identify the sentiment type and to predict the stock price movements.

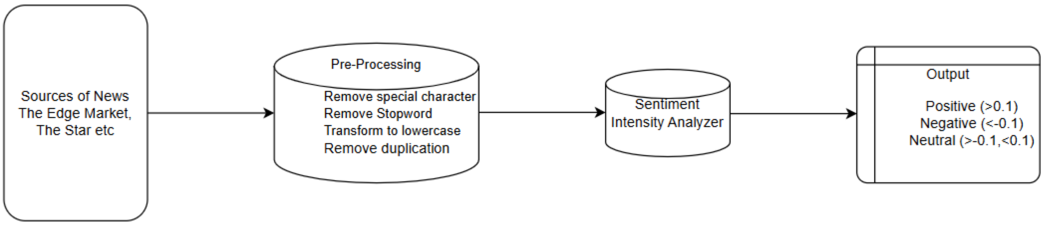


Figure 3-4 Sentiment Analysis Process FlowFigure

The data normalization step begin after the sentiment scores are determined. Stock prices data need to be normalized to ensure the value ranges (between 0 and 1) is not biased before use them for machine learning.

3.6 Model Selection and Evaluation

The stock price data is then combined with the cleaned and processed sentiment score dataset. The datasets are used for model training. The sentiment values are aligned according to the date of events.
Three deep learning model are selected for this study:

- LSTM (Long Short-Term Memory)
  A LSTM model is selected to provide prediction for the stock prices. Its ability to remember long-term sequence data is better compare to other machine learning algorithm thus it is suitable for time-series forecasting. As stock prices data are influenced by the long period of historical data, the application of LSTM algorithm can identify the long-range relationship. The stock prices datasets will be divided to 80% for training and the remaining 20% for testing. The model will be using in the Keras package in Python.

- GRU (Gated Recurrent Unit)
  GRU model will be used for comparison. The model is capable to learn long-term dependencies even with simpler architecture. It also has fast computational speed. The parameter used for this model are fewer compare to LSTM which is more efficient to train a large dataset. GRU is expected to perform like LSTM due to the simpler model result to quicker iterations. The same training data split will be used. However, the hyperparameters will be adjusted for optimization.

- ACNN-LSTM (Attention Convolutional Neural Network-LSTM):
  This hybrid model also will deploy in this project. The model combines the CNN ability in extracting text features from news headlines with LSTM power to learn temporal patterns of stock market movement. The

attention mechanism allows to focus on the important parts of the news. It will improve the prediction accuracy. This approach is useful when working with textual (new headlines) and sequential data (stock prices).

The data is split into training and test set. The training set is used to teach the models the relationship between sentiment and stock prices. The test set is used to evaluate the model capability adapting to new data. There are a lot of external factors to consider to ensure the model optimization.

The Root Mean squared Error (RMSE) is used to evaluate the efficiency of the model to predict the stock prices. This model is commonly used in time series prediction. RMSE will calculate the differences between the observed prices and the predicted values. The smaller the RMSE, the closer the prediction to the actual value. Therefore, the RMSE will be use to assist in predicting prices of stock market. The dropout layer is added to each LSTM layer to avoid overfitting. Mean Squared Error (MSE) and Mean Absolute Error (MAE) are also use to assist the evaluation for model performances for broader view of model performances in predicting stock market