Chapter 3

Research Methodology

3.1 Introduction

This chapter describes the approach used in this research work in detail especially the process of combining sentiment analysis and time series forecasting to predict stock market. This is because to recognize the interrelation between the structured and unstructured data types in financial analytics is challenging given the nature of the data, and by leveraging natural language processing and deep learning approaches in this study, it shall effectively solve this research problem. The practical aspect of the work also meets these objectives since the proceeding from data collection to deployment is best practices, thus offering a blueprint for more investigations.

3.2 Research Framework

To address the research objectives efficiently and with an adequate level of comprehensiveness, the research frameworks follow a clear data science life cycle. It is composed of interdependent subprocesses and each of them results in the production of a predictive model. The design includes problem formulation and data collection and initial assessment, cleaning and exploring, feature engineering and modeling phases. This analysis involves the combination of two techniques proves to be the main working model of the methodology. Assessment and implementation follow the created and improved solutions accompanied by continuous monitoring to make the solutions Roi-oriented.

Prominent in this framework is the ingration of quantitative (Numeric) and Qualitative (Text) data. Text analysis applied to the data corresponds to the attitude of the public and media, while time series analysis utilizes historical {price behavior}. These insights are then integrated into the hybrid model used in understanding market dynamics. Besides, this framework can not only solve the imminent research concern but also present a potential blueprint for future research on financial analytics.

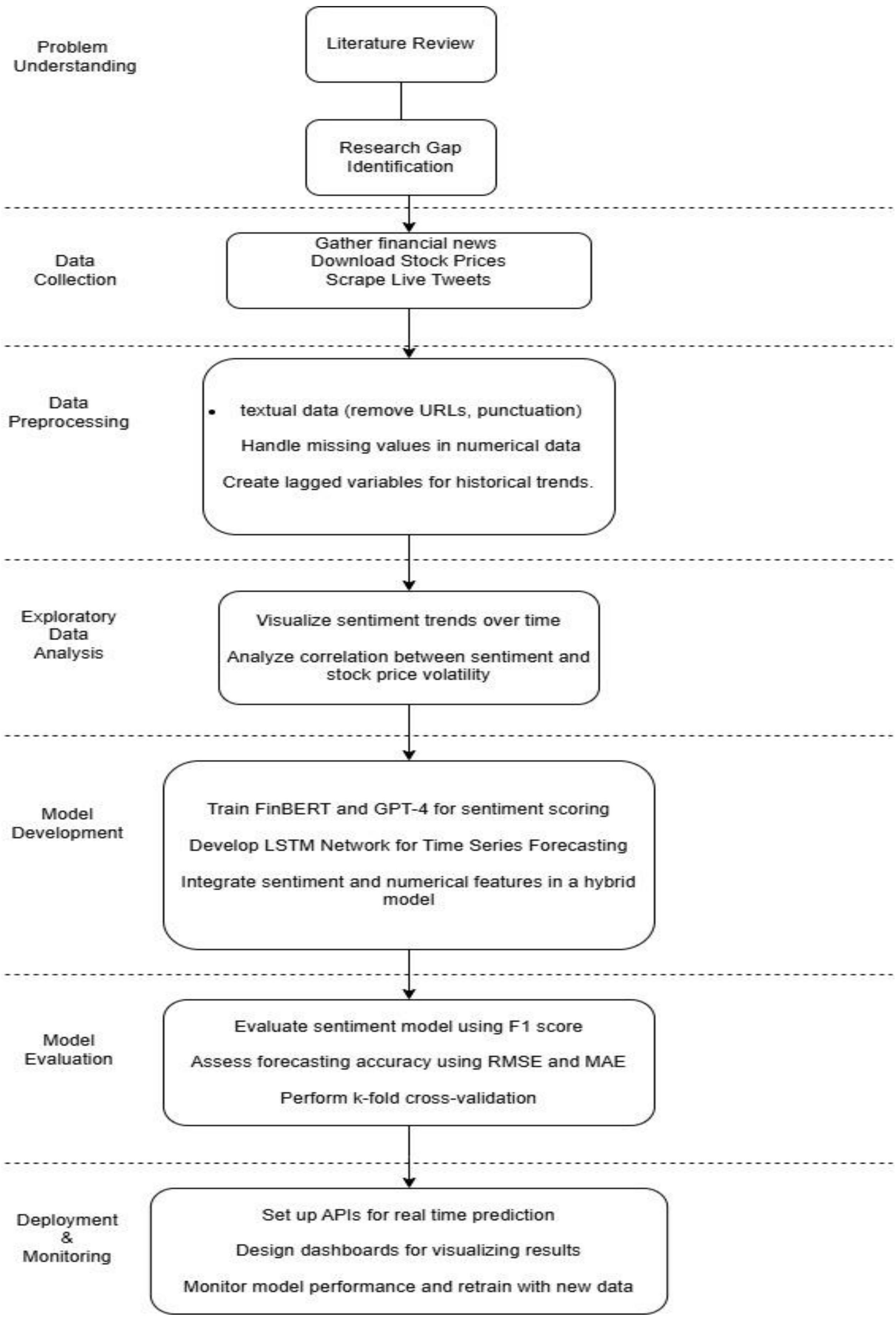
Problem Understanding
Literature Review
Research Gap Identification
Data Collection
Gather financial news
Download Stock Prices
Scrape Live Tweets
Data Preprocessing
• textual data (remove URLs, punctuation)
Handle missing values in numerical data
Create lagged variables for historical trends.
Exploratory Data Analysis
Visualize sentiment trends over time
Analyze correlation between sentiment and stock price volatility
Model Development
Train FinBERT and GPT-4 for sentiment scoring
Develop LSTM Network for Time Series Forecasting
Integrate sentiment and numerical features in a hybrid model
Model Evaluation
Evaluate sentiment model using F1 score
Assess forecasting accuracy using RMSE and MAE
Perform k-fold cross-validation
Deployment & Monitoring
Set up APIs for real time prediction
Design dashboards for visualizing results
Monitor model performance and retrain with new data

Figure 3.1: Overall Research Framework

3.3 Problem Understanding

In the context of analyzing the stock market as an object of the investor's focus, implying fluctuations in its dynamics, these factors are an obvious fact. In traditional conjugations primarily linear and static relationships are employed to model these influences, which does not make much sense in particularly non-linear and dynamic fields when textual data such as, financial news or social media posts are included into the prospective analysis. The idea of this study is to fill this gap by integrating raw price numbers with sentiment values estimated from text messages. The integrated strategy allows having a deeper understanding of the behavior in the market and makes forecast more accurate and exact.

Here, the performance evaluation of this project is defined in terms of metrics that consider both sentiment classification and predictive accuracy. The following are the performance indicators for sentiment analysis model: Precision, recall and F1 scores Forcasting performance is measured using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Comparison with other models includes with ARIMA models, as well as stand alone LSTM and other LSTM variants with an aim of proving that the proposed hybrid architecture outperforms the other models.

3.4 Data Collection

Continuing the research framework, the next step is Phase 2: Data Collection. Determining the data source is a crucial step in the sentiment analysis process. In this project, it is essential to ensure that the review data used is original and obtained directly from a trusted source, which in this case is the players themselves, without any modifications or manipulation. Once the data source is determined, the next step is to proceed with data collection. During the data collection phase, the relevance of the data to the project objectives was also considered to ensure that the analysis provided results aligned with the project's goals.

3.4.1 Data Source

The data source for the analysis process was taken from three primary sources. The stock market dataset has 15 NYSE-listed companies stock price for a period of 20 years. The dataset also has features such as Open, High, Low, Close and Volume. For sentiment analysis, the textual information is derived from financial news articles, talking about the same companies over the same time period. In fact, these are the sources of sentiment analysis in the text of these data, which tells us the perception of market about it.

3.4.2 Data Collection Method

The historical stock prices were collected using Python libraries like “pandas” and “yfinance”. The “yfinance” library was used to fetch the stock data. Then the fetched dataset was stored in a CSV file for further processing. The “pandas” library provided data handling, transformation, and storage. For acquiring financial news data, news articles and headlines were scraped through web scraping techniques elaborately in Python through “BeautifulSoup” and “Requests” libraries. The web scraping process deals with fetching the news from financial online news sources, extracting relevant HTML components, and saving entries for sentiment analysis. Later the textual data were processed through transformers library to apply FinBERT model that scores the sentiment of the news articles as well.

3.5 Data Preprocessing

Data preprocessing is a crucial step aimed at preparing raw data for further analysis. Raw data collected through web scraping often contains incomplete data (e.g., missing values or empty data), inconsistencies, and information that is irrelevant to the study's objectives. These issues can reduce the accuracy of the analysis. The data preprocessing process ensures that the dataset is clean and consistent before moving on to the analysis phase.

3.5.1 Data Cleaning

The raw data undergoes a few preprocessing procedures before the final analysis. The subsequent processes are as follows. For the Stock Price Data, the missing values are dealt using forward-filling techniques in pandas. Duplicate records are identified and removed since they are deemed irrelevant to the entire data integrity. Statistical methods are also adopted to detect anomalies for adjusting extreme price fluctuations. For the News Sentiment Data, the textual data was preprocessed through nltk and re libraries. This involved removal of punctuation marks, special characters, and stop words. Tokenization and lemmatization are performed using “nltk.WordNetLemmatizer” to standardize the words and improve the accuracy of sentiment analysis.

3.5.2 Data Transformation

Once the initial dataset is cleaned, added transformations are done with the proper-python libraries. For Stock Price Data, dates have been standardized using “pandas.to-datetime()” in order to conveniently conduct the time-series analyses since they are consistent in format. The percentage changes were calculated using “pct_change()” with respect to stock prices which are normalized for variations in stock price. For news Sentiment Data, numerical sentiment scores were obtained from the transformation of sentiment labels using the FinBERT model gained through the transformers library. This enables correlation analysis, between sentiment and stock prices.

3.5.3 Feature Engineering

For improving the predictive power, some new features are added to the dataset.

Sentiment Aggregation: This is achieved by aggregating sentiment scores over different time windows with the help of daily, weekly, and monthly rolling of sentiment scores using the pandas rolling function. This helps in bringing the effect of sentiment aggregation across time.

Technical Indicators: Some stock market indicators, including Moving Averages, Relative Strength Index (RSI), Bollinger Bands, among others, are calculated using the ta (technical analysis) library to help predictive modeling.

Lag Features: Past values of stock price and sentiment are included as lag features using shift() in pandas in an attempt to introduce historical dependence into the model.

3.6 Exploratory Data Analysis (EDA)

Exploratory Data Analysis helps users to get a better understanding of many aspects of the given dataset. Where descriptive analyses involve computation of summary statistics such as mean, median, standard deviation and correlation coefficients, then statistical techniques are employed. Heat maps represent the coefficient between the sentiment scores and stock price swing, while the time series plots show series and trends of data over time.

For instance, the inspection of the heatmap of the correlation coefficient uncovers the nature of the correlation between general or sector-specific sentiment coming from the globe's financial media and daily stock price changes that facilitates feature selection. Wherein the stock price trends per security are plotted against the overall sentiment as a function of time through a time-series plot.

3.7 Model Development

In the proposed hybrid model structure, the obtained sentiment analysis informs the time-series forecasting. The sentiment analysis is first done using FinBERT, a transformer model fine-tuned for the financial text, and GPT-4 which has text understanding ability. These models therefore generate sentiment scores that will in turn act as inputs for the time series model.

The forecasting component is based on using Long Short-Term Memory (LSTM) networks that allow considering sequential dependencies. These aspects are in this hybrid architecture, while input layers work with numerical and textual data, hidden LSTM layers for working on sequence data, functional layers dealing with sentiment-derived data. Parameter optimization is conducted with grid search with regards to such features as learning rate and batch size and the number of LSTM units to optimize the performance of model.

3.8 Model Evaluation

Selecting a performance measure for the hybrid model and for the individual and composite components is given careful consideration. It is classified based on accuracy on how well it categorizes sentiments then using parameters like precision, recall and F1 score. Moreover, for evaluating the time- series forecasting part of the model Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), which provide the measure of prediction error.

Validation methods include k – fold cross validation to check the model's stability and compared with conventional model such as SARIMA and only LSTM model. On the basis of this comparison, the enhancement resulting from the use of sentiment analysis and enhanced forecasting are vehicled.

3.9 Deployment and Monitoring

The last of them is to apply the extracted hybrid model for the practical use with some organization or project. APIs imply data exchange in the real-time environment, thus, the model can be updated continuously. An explorative dashboard is created for visual display of predictions, sentiments, and other KPI's, in a manner that allows for easy tracking via.

Testing and observation allow the model to reach its best consistently. Assessments of the prediction accuracy are done periodically and the algorithm is updated with new data after some time due to performance decline. Triggers of notification and alerts are included here to draw user's attention when there supposed to be fundamental changes in the market behavior.

3.5 Conclusion

This chapter outlined the methodology employed to conduct the project. Beginning with Phase 1, the research identified gaps in existing studies through a literature review, as discussed in Chapter 2. Next, the methodology moved to data collection and data preprocessing phases to ensure clean, structured and meaningful datasets. In addition, feature engineering was carried out to enhance the analytical potential of the datasets. The later steps are mentioned to give an overview of the holistic approach that is taken to materialize the project.