CHAPTER 1

INTRODUCTION

In a world where the average annual fixed deposit interest rate is dropping drastically, people with conservative financial management observe literal evaporation of their wealth. So, in order to make profits, more and more people are inclined towards investment in stocks with high returns and high liquidity. To make profit, the accurate prediction of market movement is very crucial. Accurate predictions can significantly heighten decision-making processes, decrease investiture risks and step up profitability. But the complexness and unpredictability of the stock market make it difficult for researchers and investors to predict market movement. Traditionally, statistical techniques such as time series analysis and regression models have been employed to calculate stock prices. However, these methods often fall unforesightful in identifying intricate patterns within stock data, especially when influenced by extraneous variables such as market sentiment and financial news. In today's interconnected world, people's notion toward any product can change at any time. Within the blink of an eye, there will be innumerous opinions & discussions on social media. Financial articles will start publishing articles. All of these combined will inevitably fluctuate the stock price. So, predicting stock market movements without integrating market sentiment is not possible.

In recent years, advancements in machine learning (ML) and Deep Learning (DL) have provided many helpful tools to tackle this problem. Natural Language Processing (NLP), a subfield of artificial intelligence (AI), has been performing sentiment analysis successfully. Basically, sentiment analysis is the process of extracting insights from large volumes of unstructured textual information. In sentiment analysis, the emotional tone or opinion (positive, negative, neutral) available in financial news, social media posts or analyst reports is analyzed. This comes very handy to predict the probable fluctuation in any stock price. After sentiment analysis, the extracted sentiment features are combined with historical stock prices to make future predictions. But this whole process has to go through different obstacles that hurt accuracy.

This project uses advanced NLP models like FinBERT & GPT-4 to perform sentiment analysis on textual data collected from multiple sources. It then uses an advanced Deep Learning model, LSTM (Long Short Term Memory) network to make predictions from the sentiment scores and numerical data. By taking a hybrid approach, this project promises to enhance prediction accuracy.

1.1 Problem Background:

Sentiment analysis lends great help to incorporate public sentiment in predicting stock market movements by converting unstructured textual data into quantifiable sentiment scores. Then these scores are incorporated into predictive models to provide a more holistic take on the factors that influence stock prices. Advanced Natural Language Processing models like FinBERT (Finance Bidirectional Encoder Representations from Transformers) and GPT-4 (Generative Pre-trained Transformer) have improved sentiment analysis. These models can classify sentiment e.g. positive, negative, neutral. By doing so, they can predict how much sentiment may impact stock prices. But according to, classical Machine Learning models like logistic regression, when tuned properly, display more effectiveness. Five key metrics, i.e., Accuracy, Precision, Recall, F1 Score, and ROC AUC, were used to assess the performance.

It is found from recent research that, FinBERT is a financial domain-centric model and it has very high potential in understanding financial terms and concepts with remarkable precision. However, it is resource-intensive in nature. That's why it is likely to hurt computational efficiency. GPT-4 is a versatile language model. It has great abilities to create and realize human-like text. That's why it is perfect choice for processing unstructured news data. But the exploration of its predetermined and heuristic approach may restrict its precision in particular financial situations. On the other hand, logistic Regression is computationally efficient. It produces dependable results when it is applied precisely and performs better than both FinBERT and GPT-4 across most metrics in spite of their advanced

text analysis capabilities. However, logistic regression faces limitation in handling stock data for a number of reasons like non-linear relationships, high-dimensional data, feature interactions, sensitivity to outliers, and multicollinearity.

This project addresses this dilemma & offers a hybrid approach by combining state-of-the-art NLP models like FinBERT & GPT-4 with LSTM network. FinBERT & GPT-4 show prominence in extracting insights from complicated data. LSTM networks are good at handling sequential data, capturing long term dependencies and robust to noise. These characteristics make them an ideal choice for prediction of stock prices.

1.2 Problem Statement:

The primary problem this project addresses originates from the dilemma of choosing between traditional Machine Learning model like logistic regression and advanced NLP algorithms like FinBERT & GPT-4. Logistic regression is computationally efficient and simple. With proper adjustment, it outperforms advanced DL algorithms. But when it comes to handling complicated data patterns, it lacks proficiency. Logistic regression cannot handle non-linear data much efficiently but the relationship between stock prices and predictors is often non-linear. It also faces problems in handling high-dimensional stock market data with numerous features as it might struggle with feature selection and regularization. Handling outliers also becomes an issue for Logistic regression. On the other hand, FinBERT & GPT-4 are highly resource intensive. These models are also computationally heavy. But these models are very strong in understanding financial terminology and human generated textual data which eventually make them suitable for sentiment analysis. To solve this dilemma, this project takes a hybrid approach. It combines these advanced NLP models with LSTM networks. LSTM networks can tackle the challenges caused by the limitations of logistic regression. Thus, this project promises to provide a model with more accuracy.

1.3 Research Goals:

The goal of this project is to enhance the prediction accuracy of stock prices by combining state-of-the-art NLP models with advanced DL algorithm LSTM. The NLP models will do sentiment analysis on data collected from various sources & generate sentiment scores. LSTM will perform time series analysis on the sentiment scores and historical stock prices to accurately predict stock market movements.

1.4 Research Objectives:

- To enhance the comprehensiveness of prediction by collecting and preprocessing sentiment data from multiple sources

- To implement FinBERT & GPT-4 for detailed sentiment analysis and develop LSTM networks for improving prediction accuracy

- To rigorously assess the model's performance with a number of evaluation metrics such as Accuracy, Precision, Recall, F1 score etc

1.5 Scope:

This project combines sentiment analysis of financial news and social media posts and performs time series forecasting to make effective stock price predictions. This covers data collection, preprocessing and model development using LSTM networks and the evaluation of predictive performances. However, the research does not include real-time data analysis. It doesn't perform a full-fledged market analysis and doesn't recommend any particular stock.

To enhance the accuracy of stock price predictions, this project comprises the integration between sentiment analysis and time series forecasting. This involves the collection of historical stock prices and trading volumes, financial indicators, as well as sentiment data from financial news, social networks, and -analyst reports-from several trusted sources. It also includes extensive preprocessing steps like cleaning, normalizing, and tokenizing textual data as a measure for data-quality improvement and compatibility. Advanced Natural Language Processing (NLP) models such as FinBERT and GPT-4 will be employed in feature extraction and sentiment analysis. High-level Long Short-Term Memory (LSTM) networks will be developed to synthesize the sentiment features with the traditional financial indicators and will follow extensive training and a thorough hyperparameter tuning as well as evaluation based on Accuracy, Precision, Recall, F1 Score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The results of the model are further validated using cross-validation and backtesting in a bid to enhance the robustness of the model.

Model creation and analysis of historical data only exclude real-time data processing and live stock price forecasts. This does not aim to be an all-encompassing analysis of market situations to include geopolitical events or macroeconomic conditions. There will not be specific stock recommendations or investment advice covered by the project, nor will it include predictions made for other financial instruments such as bonds, commodities, or derivatives. Also, more sophisticated machine learning architectures like deep reinforcement learning have been excluded from this study. The analysis is limited by the availability and quality of historical data, and the study is constrained in terms of model training and evaluation because of the machine resources available.

1.6 Expected Contribution:

- Introduction of a novel approach combining advanced NLP models and LSTM networks
- Improvement of stock price prediction accuracy
- Providing valuable insights to the investors and mitigate the risk factor
- Developing a robust framework in financial analysis sector