A COMPARATIVE STUDY OF SENTIMENT ANALYSIS MODELS IN FINANCIAL MARKETS

A Project Report Submitted

for the Course

MA498 Project I

by

Pranav Agarwal

Parth Maheshwari

(Roll No. 200123040 and 200123037)

to the

DEPARTMENT OF MATHEMATICS

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781039, INDIA

November 2023

CERTIFICATE

This is to certify that the work contained in this project report entitled "A Comparative Study of Sentiment Analysis Models in Financial Markets" submitted by Pranav Agarwal and Parth Maheshwari (Roll No.: 200123040 and 200123037) to the Department of Mathematics, Indian Institute of Technology Guwahati towards partial requirement of Bachelor of Technology in Mathematics and Computing has been carried out by him/her under my supervision.

It is also certified that this report is a survey work based on the references in the bibliography.

OR

It is also certified that, along with literature survey, a few new results are established/computational implementations have been carried out/simulation studies have been carried out/empirical analysis has been done by the student under the project.

Turnitin Similarity: 14%

Guwahati - 781 039 (Prof. Rajen Kumar Sinha)

November 2023 Project Supervisor

ABSTRACT

The intricate task of forecasting stock market movements necessitates continual refinement of predictive models to capture the multifaceted nature of financial markets. This study delves into the realm of financial time series forecasting, with a specific emphasis on integrating sentiment analysis data into predictive models. Recognizing the growing efficiency of deep learning architectures in this domain, our research compares various state-of-the-art methods in forecasting financial time series, augmented by sentiment analysis. We conducted extensive experiments, testing 54 different feature setups encompassing stock closing prices and sentiment scores across diverse datasets and metrics. Our approach involved employing 5 cutting-edge algorithmic schemes in two distinct case studies: one focusing on method comparison and the other on input feature setups. This report unveils the significant impact of incorporating sentiment analysis into forecasting models, particularly highlighting conditions under which this integration notably enhances model efficiency.

Contents

Chapter 1

Introduction

The impetus for this research stems from the longstanding human quest to predict future outcomes based on observed data. This endeavor, rooted deeply in our history, finds its modern expression in the field of time series forecasting. Particularly, financial time series forecasting has garnered immense interest due to its complexity and substantial implications. Our study zeroes in on this domain, exploring the utility of incorporating sentiment analysis—derived from social media platforms like Twitter—into the stock market forecasting models. We evaluate a plethora of state-of-the-art methods, including deep learning and other machine learning schemes, under various experimental setups. The incorporation of sentiment analysis, quantified through different methodologies, is investigated to discern its impact on the predictive accuracy of these models. Our extensive experimental process, involving multivariate setups with various sentiment time series, seeks to provide a comprehensive understanding of the dynamics at play in deep stock market forecasting.

This introduction sets the stage for a detailed exploration of the methodologies, experimental procedures, and the nuanced findings of this extensive study, aiming to contribute significantly to the field of financial time series forecasting.

1.1 Aim

The primary aim of this study is to conduct a comprehensive comparison of various advanced models in the context of financial time series forecasting, with a particular focus on the impact of incorporating open stock prices and sentiment analysis. We intend to systematically evaluate the performance of these models under different input parameter configurations and temporal scales. Specifically, our objectives are:

1. **Model Comparison**: To compare a range of state-of-the-art forecasting models, including both traditional machine learning and advanced deep learning architectures. This comparison is designed to assess the effectiveness of each model in predicting stock market trends.

2. **Input Parameter Analysis**: To investigate how the inclusion of open stock prices as an input parameter influences the predictive accuracy of these models. We aim to analyze and quantify the difference in forecasting performance when models are fed with data sets that include open prices versus those that do not.

3. **Impact of Sentiment Analysis**: To evaluate the effect of integrating sentiment analysis, derived from social media platforms, into the forecasting models. This involves examining how sentiment data, when combined with traditional stock market data (including and excluding open prices), alters the models' predictive capabilities.

4. **Temporal Scale Variation**: To explore how the choice of time step—ranging from short to long-term intervals—affects the accuracy and reliability of the models. This aspect is crucial for understanding the models' performance across different forecasting horizons and market conditions.

Through these objectives, this study aims to provide a nuanced understanding of the factors that influence the success of financial time series forecasting models. This includes identifying the most effective models under various conditions, understanding the impact of including open prices in the data set, and determining the added value of sentiment analysis in improving forecasting accuracy. The findings of this research will offer valuable insights for both academic researchers and practitioners in the field of financial forecasting.

Chapter 2

Experimental Procedure

In this research, we conducted a series of 540 experiments to assess the effectiveness of various state-of-the-art forecasting algorithms on stock market prediction, specifically focusing on TSLA stock data from Twitter. The experiments were designed to evaluate the impact of different combinations of input features, including and excluding open prices, across two distinct time steps.

2.1 Overview

2.1.1 Input Feature Combinations

The input features for our models were derived from a comprehensive analysis of TSLA stock data. We created 54 unique combinations of input features, where the first 27 combinations included only the closing prices of the stocks. The subsequent 27 combinations included both the opening and closing prices. Each set of input features was meticulously curated to ensure a thorough investigation of the data's predictive power.

2.1.2 Algorithms Employed

We selected five advanced forecasting algorithms, representing the current state-of-the-art in financial time series prediction. These algorithms were chosen based on their proven effectiveness in similar contexts and their ability to handle complex data structures.

2.1.3 Time Steps

Our experiments were conducted over two different time steps to capture both short-term and long-term predictive patterns in the stock data. This approach allowed us to assess the models' performance in varying temporal contexts, providing a more holistic understanding of their capabilities.

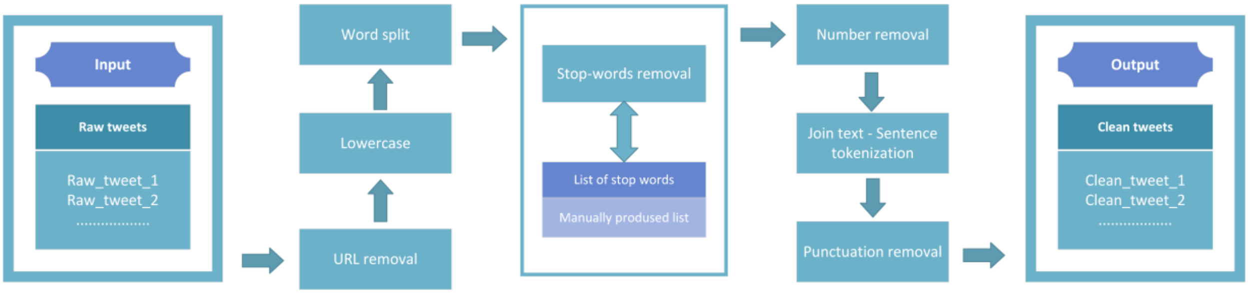
2.1.4 Data Source

The data for these experiments was sourced from Twitter, focusing on sentiment and discussion surrounding TSLA stock. The timeframe for this data collection spanned from September 30, 2021, to September 29, 2022, ensuring a comprehensive dataset that reflects a wide array of market conditions and sentiment shifts.

2.2 Data Preprocessing

In this study, stock-related data was sourced from Kaggle for a period of one year. This dataset provided the necessary stock information on a daily basis, crucial for our subsequent analysis.

The initial phase of data preparation involved removing hyperlinks and URLs from the collected tweets using the Re Python library. The textual



content was then converted to lowercase and broken down into individual words. Extraneous phrases and numerical strings, identified from a predetermined list, were eliminated. Following this, the text was restructured to its original form, and sentence-level tokenization was performed using the NLTK library. The last step of data preparation was the removal of punctuation, executed through the String module. The entire process of text data preparation is graphically represented in Figure 2 of the report.

2.3 Sentiment Analysis

In our study, sentiment analysis plays a crucial role in enhancing the predictive models for stock market forecasting. We utilized three renowned algorithms for sentiment analysis: VADER, FinBERT, and TextBlob. These algorithms were applied to Twitter data related to TSLA stock to derive sentiment scores. Furthermore, we incorporated both 7-day and 14-day rolling averages of these sentiment scores into our analysis to capture short-term and medium-term sentiment trends.

VADER (Valence Aware Dictionary and sEntiment Reasoner)

VADER is a lexicon and rule-based sentiment analysis tool specifically attuned to sentiments expressed in social media. It is unique in its ability to understand text with a mix of polarity (positive/negative) and intensity (strength of emotion). VADER is particularly effective in dealing with short texts prevalent in social media, making it an ideal choice for analyzing Twitter data. It scores texts by assigning a normalized, weighted composite score that quantifies the overall sentiment on a scale from highly negative to highly positive.

FinBERT

FinBERT is a deep learning model that has been fine-tuned specifically for the analysis of financial texts. It is based on the BERT (Bidirectional Encoder Representations from Transformers) architecture, which has revolutionized the field of natural language processing. FinBERT adapts BERT's capabilities to the financial domain, providing nuanced understanding of complex financial jargon and sentiment. This makes it exceptionally suited for analyzing tweets about stocks, as it can interpret the specific language and sentiment nuances in the financial context.

TextBlob

TextBlob is a straightforward library for processing textual data. It provides simple APIs for common natural language processing (NLP) tasks, including sentiment analysis. TextBlob's sentiment analysis module uses a combination of pattern recognition and natural language processing techniques. It is effective for basic sentiment analysis tasks and offers a good balance between simplicity and accuracy, making it suitable for analyzing large volumes of Twitter data.

Sentiment Score Integration

For each tweet, we applied these algorithms to compute sentiment scores. To add depth to our analysis, we also calculated 7-day and 14-day rolling averages of these sentiment scores. The rolling averages help in smoothing out short-term fluctuations and highlighting more sustained sentiment trends over time. This approach allows our models to not only capture the immediate sentiment reactions to market events but also understand the underlying, evolving sentiment trends that might have a more delayed impact on stock prices.

Incorporating these sentiment analysis results into our forecasting models provided a rich, multi-dimensional view of the market sentiment, enhancing the models' ability to predict stock market movements with greater accuracy.

2.4 Algorithms

The study employed five advanced algorithms for stock market prediction, leveraging the capabilities of the tsAI library in Python. Below is a table summarizing each algorithm and its full name:

tsAI Library

The tsAI library, available in Python, is a versatile and powerful toolkit for time series analysis with deep learning. It is designed to facilitate the implementation and evaluation of various time series models, especially in

Algorithm	
TCN	Temporal Convolutional Network
GRU_FCN	Gated Recurrent Unit - Fully Convolutional Network
XCM	Explainable Convolutional Multivariate Model
LSTM	Long Short-Term Memory
LSTM_FCN	Long Short-Term Memory - Fully Convolutional Network

Table 2.1: List of Algorithms Used

complex forecasting tasks like stock market prediction. tsAI offers a user-friendly interface and integrates seamlessly with other Python-based data analysis frameworks. This library was instrumental in our study for implementing the aforementioned algorithms efficiently and effectively.

Data Splitting

In our experimental setup, we used a data splitting ratio of 15:85 for test to train sets. This split was chosen to provide a substantial amount of data for training the models, ensuring their robustness and accuracy, while still retaining a meaningful portion for testing and validating the models' predictive capabilities.

2.5 Metrics

In this study, we employed two widely recognized metrics for evaluating the performance of our predictive models: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These metrics are pivotal in quantifying the accuracy of our models in forecasting stock prices.

Mean Absolute Error (MAE)

MAE is a measure of the difference between the predicted values and the actual values. It is calculated as the average of the absolute errors between the predictions and the actual observations, giving us a straightforward indication of the average error magnitude. The formula for MAE is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.1}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations. A lower MAE value indicates a model with higher accuracy and precision in its predictions.

Root Mean Squared Error (RMSE)

RMSE is a quadratic scoring rule that measures the average magnitude of the error. It is the square root of the average of squared differences between prediction and actual observation. The formula for RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.2}$$

where y_i and $\hat{y}_i$ are the actual and predicted values, respectively, and n is the number of observations. RMSE gives a relatively high weight to large errors, making it particularly useful when large errors are particularly undesirable. A lower RMSE value indicates a model with higher accuracy.

Both MAE and RMSE are critical in assessing the performance of our forecasting models. They provide insights into different aspects of the prediction error, contributing to a comprehensive understanding of the models' accuracy.

2.6 Experiment Execution

Each of the 540 experiments involved training the selected algorithms on the designated input feature combinations and time steps. The performance of each model was then evaluated based on its predictive accuracy, with particular emphasis on its ability to forecast stock market trends effectively.

2.6.1 Analysis Methodology

The results of these experiments were systematically analyzed to draw meaningful conclusions about:

1. The impact of including open stock prices in the predictive models.

2. The efficacy of different algorithms in handling various input feature combinations and time steps.

3. The overall capability of these models to leverage sentiment analysis for enhanced stock market prediction.

This rigorous experimental procedure was designed to provide a comprehensive and nuanced understanding of the dynamics involved in financial time series forecasting, particularly in the context of integrating sentiment analysis and varying input features.

Chapter 3

Results

3.1 Key Findings

Our study's results yield several key findings regarding the impact of timestep, the influence of open prices in input data, a comparison across different sentiment analysis tools, and the variability in model performance.

Impact of Timestep

- For all sentiment analysis tools, we observed that the average MAE and RMSE are generally higher with a timestep of 14 compared to 7. This indicates that predictions become less accurate as the prediction window increases. For example, in the "FinBERT_TS14_with open price" configuration, the mean MAE is significantly higher (32.62) compared to the "FinBERT_TS7_with open price" configuration (23.71).

Influence of Open Price

- The inclusion of the open price in the input data shows a variable impact on the error metrics. In some instances, such as "TextBlob_TS7", the inclusion of open price slightly reduces the mean MAE and RMSE, suggesting improved accuracy. Conversely, in other cases like "Vader_TS14", including open price results in a slight increase in errors. This inconsistency indicates that the relevance of open prices might vary depending on the specific model and data characteristics.

Comparison Across Sentiment Analysis Tools

- Among the sentiment analysis tools, FinBERT tends to exhibit higher error metrics (both MAE and RMSE) in longer timesteps (14) compared to TextBlob and Vader. This might indicate potential overfitting or reduced effectiveness over longer prediction intervals. On the other hand, TextBlob and Vader display more consistency across different timesteps and with the inclusion of open prices. Notably, Vader shows lower average errors in most configurations, suggesting better overall performance for the dataset used.

Outliers and Variability

- We noted significant variability in the maximum and minimum values of MAE and RMSE across all configurations, pointing to the fluctuating performance of the models. For instance, "FinBERT_7_TS14_without open price" recorded a high maximum MAE (51.43), indicating occasional substantial prediction errors. This variability highlights the need for a cautious approach when interpreting the models' predictive capabilities.

3.2 Comparative Analysis of Models

This analysis compares the performance of five different models used in stock market predictions: GRU_FCN, LSTM_FCN, LSTM, TCN, and XCM. The performance is evaluated in terms of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), with a lower value indicating better predictive accuracy.

Findings

MAE Mean Comparison

- The model TCN shows the lowest mean MAE (28.23), suggesting it has, on average, the most accurate predictions.
- The models LSTM_FCN and GRU_FCN have nearly identical mean MAE values (28.52), indicating similar levels of average accuracy.
- The model XCM has a slightly higher mean MAE (28.67), followed closely by LSTM (28.70).

RMSE Mean Comparison

- Similar to MAE, TCN exhibits the lowest mean RMSE (35.01), which implies better performance in terms of prediction error magnitude.
- The models LSTM, XCM, LSTM_FCN, and GRU_FCN show closely related mean RMSE values (ranging from 35.33 to 35.53), indicating comparable prediction quality.

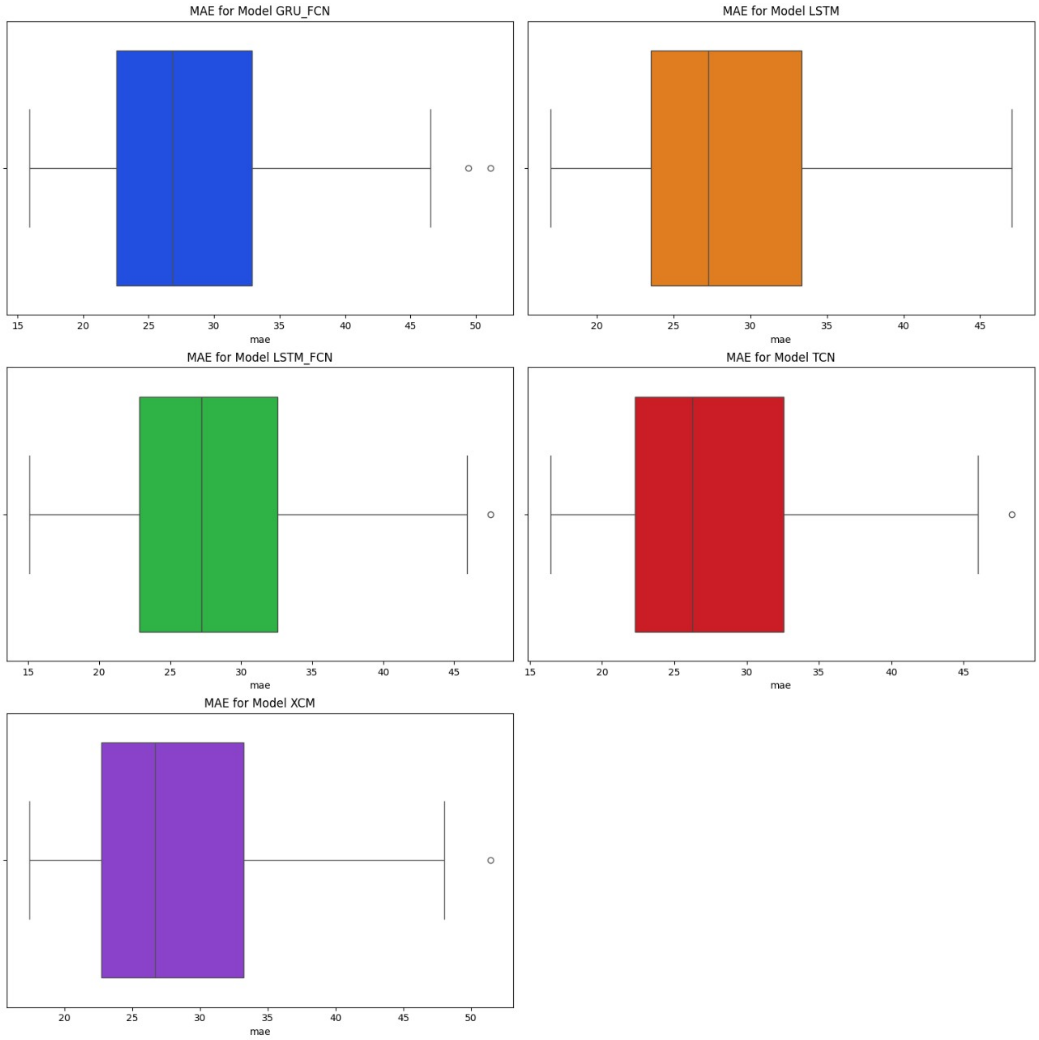


Figure 3.1: MAE Box Plots

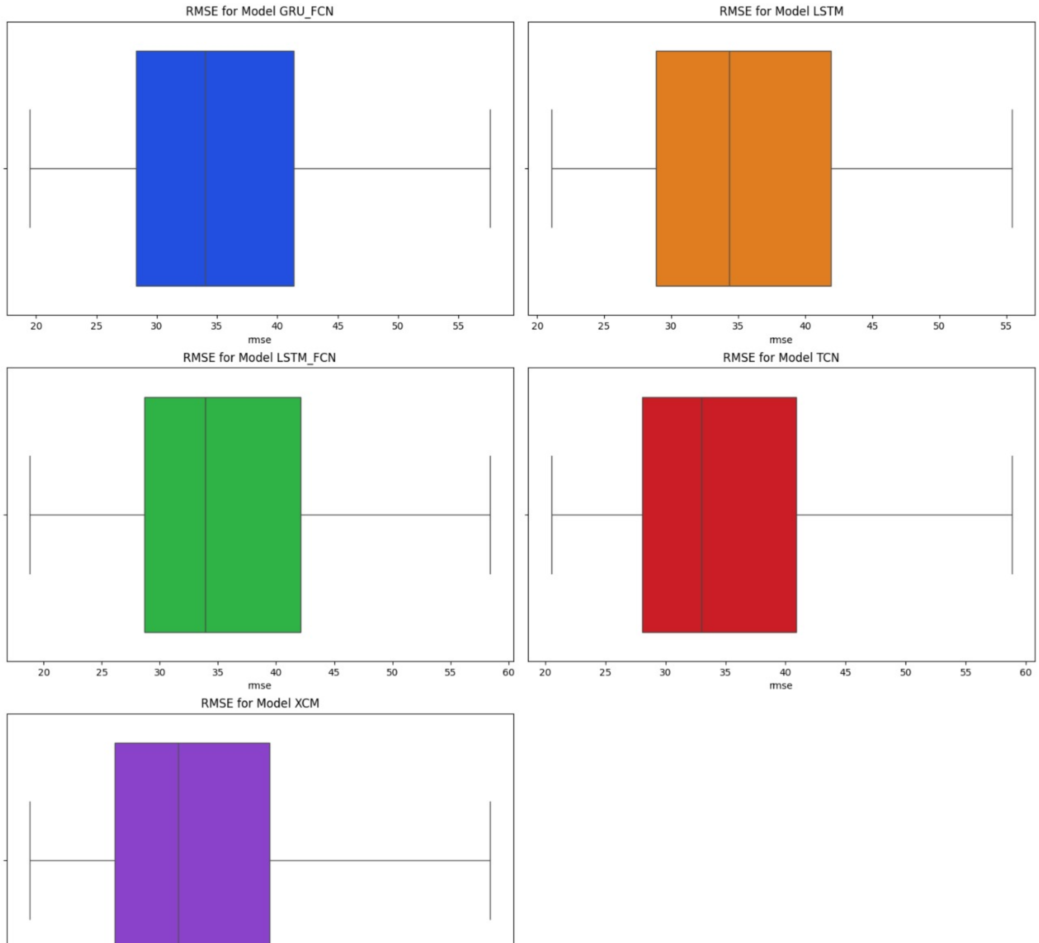


Figure 3.2: RMSE Box-Plots

Variability (Standard Deviation) Analysis

- LSTM exhibits the lowest standard deviation in both MAE (7.07) and RMSE (8.14), suggesting consistent performance.
- The model XCM shows the highest standard deviation in both metrics, indicating greater variability in its predictions.

Range Analysis (Min-Max)

- TCN demonstrates a wide range in both MAE (16.44 - 48.32) and RMSE (20.50 - 58.82), suggesting fluctuating prediction accuracy.
- LSTM_FCN and XCM have the highest maximum MAE (47.56 and 51.43 respectively) and RMSE (58.44 and 59.76 respectively), indicating occasional large prediction errors.

3.3 Top Performers

3.3.1 With Open Price

This section presents the top five performing configurations involving sentiment analysis tools with open prices. The performance is evaluated based on the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics, with lower values indicating better predictive accuracy.

1. Vader_TS7_with open price

- **MAE Mean**: 22.431
- **RMSE Mean**: 27.730

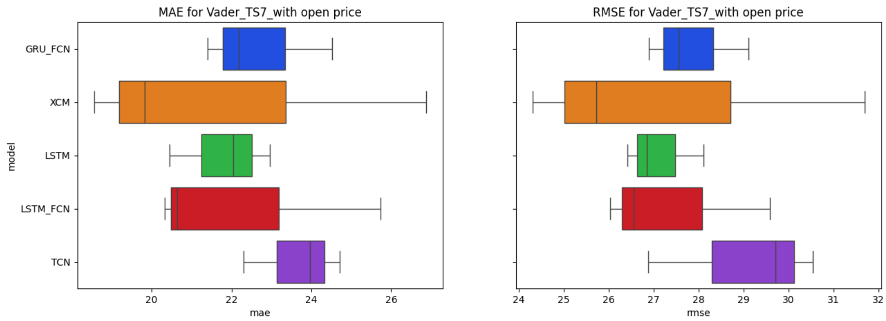


Figure 3.3: Vader TS7 with open price BoxPlots

This configuration showcases the best performance among the analyzed setups, with the lowest MAE and RMSE means, indicating high accuracy in predictions.

2. TextBlob_TS7_with open price

- **MAE Mean**: 23.377
- **RMSE Mean**: 28.816

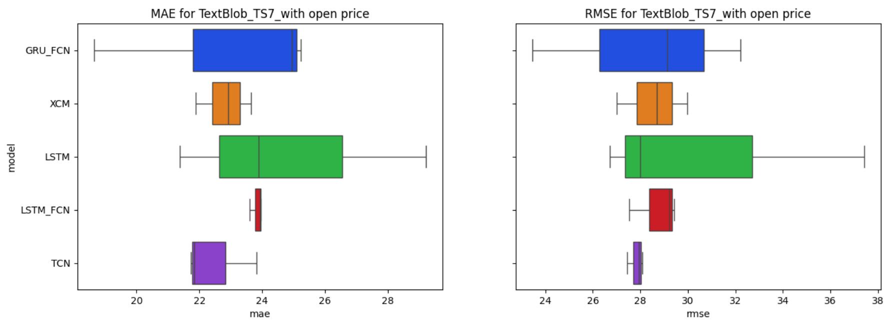


Figure 3.4: TextBlob TS7 with open price BoxPlots

TextBlob, when analyzing data with a 7-day timestep and including open prices, demonstrates strong predictive capabilities, as reflected in its MAE and RMSE values.

3. FinBERT_TS7_with open price

- **MAE Mean**: 23.708
- **RMSE Mean**: 29.080

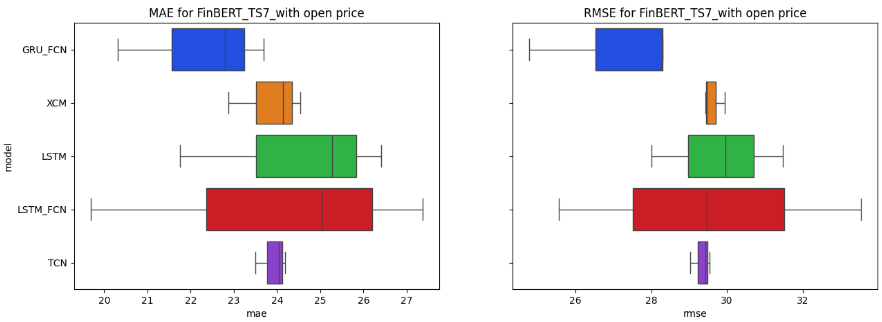


Figure 3.5: FinBERT TS7 with open price BoxPlots

FinBERT's performance with a 7-day timestep and open prices included ranks third, showing slightly higher error metrics compared to Vader and TextBlob in the same setup.

4. Vader_14_TS7_with open price

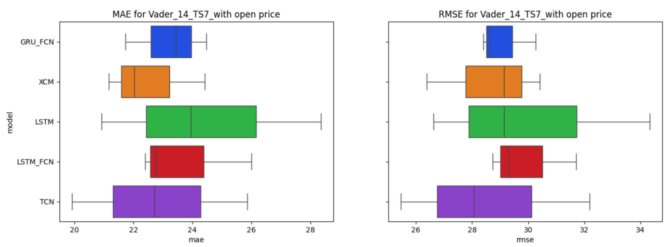


Figure 3.6: Vader14 TS7 with open price BoxPlots

- **MAE Mean**: 23.340
- **RMSE Mean**: 29.250

In this configuration, Vader's performance with a 14-day rolling average and open prices indicates a minimal increase in error metrics compared to its 7-day counterpart.

5. FinBERT_14_TS7_with open price

- **MAE Mean**: 24.020
- **RMSE Mean**: 29.641

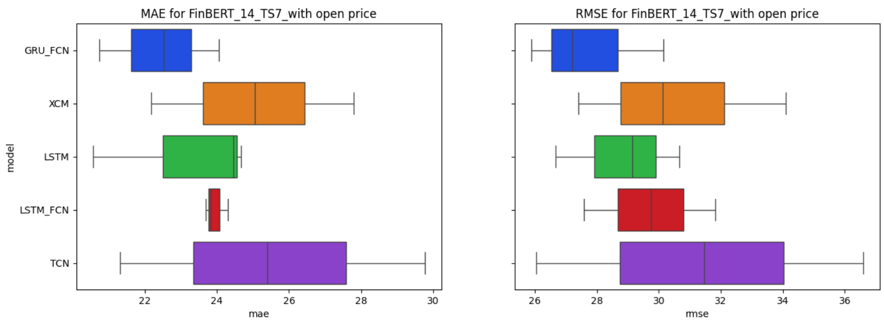


Figure 3.7: FinBERT 14 TS7 with open price BoxPlots

The FinBERT model with a 14-day rolling average and open prices shows higher error metrics, suggesting a slight decrease in accuracy compared to its 7-day timestep setup.

These results highlight the nuanced impact of including open prices in the models and the effectiveness of different sentiment analysis tools over varying timesteps.

3.3.2 Without Open Price

This section presents the top five performing configurations involving sentiment analysis tools without open prices. The performance is evaluated based on the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics, with lower values indicating better predictive accuracy.

1. Vader_7_TS7_without open price

- **MAE Mean**: 21.910
- **RMSE Mean**: 26.947

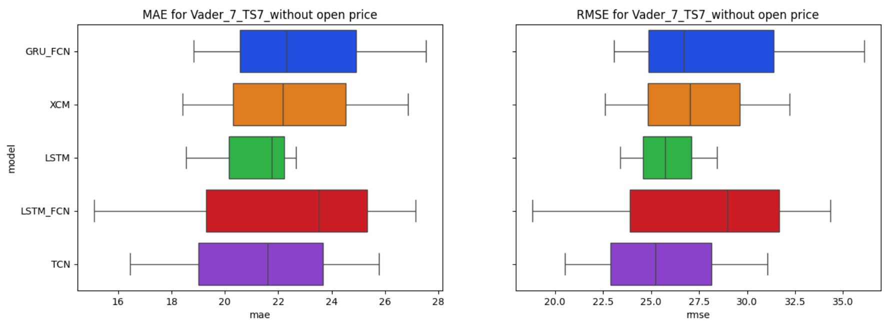


Figure 3.8: Vader7 TS7 without open price BoxPlots

This configuration exhibits the best performance among the setups without open prices, with the lowest MAE and RMSE means, indicating high accuracy in predictions.

2. FinBERT_14_TS7_without open price

- **MAE Mean**: 22.274
- **RMSE Mean**: 28.971

In this setup, FinBERT with a 14-day rolling average, excluding open prices, shows relatively low error metrics, ranking second in performance.

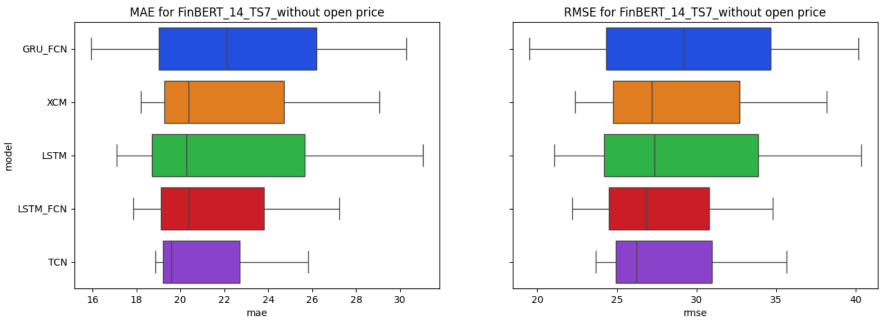


Figure 3.9: FinBERT14 TS7 without open price BoxPlots

3. TextBlob_TS7_without open price

- **MAE Mean**: 22.449
- **RMSE Mean**: 27.718

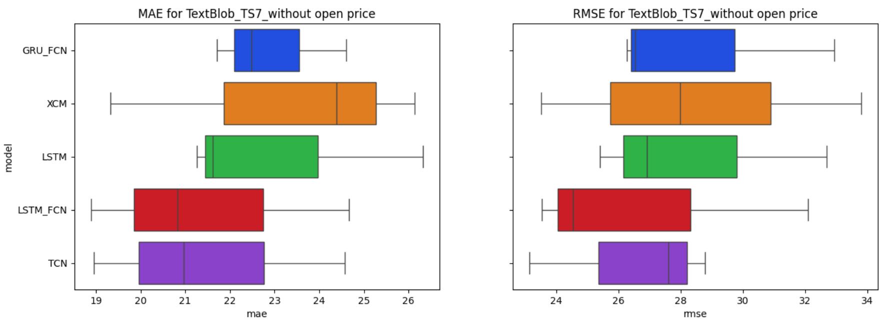


Figure 3.10: TextBlob TS7 without open price BoxPlots

TextBlob with a 7-day timestep without open prices demonstrates strong predictive capabilities, as reflected in its MAE and RMSE values.

4. Vader_14_TS7_without open price

- **MAE Mean**: 22.857
- **RMSE Mean**: 29.047

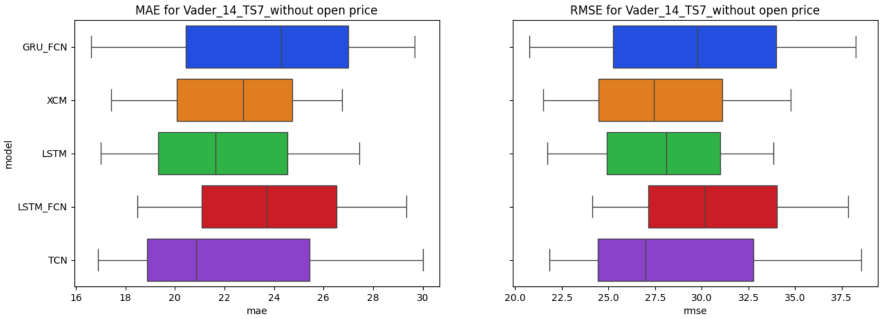


Figure 3.11: Vader14 TS7 without open price BoxPlots

This configuration of Vader with a 14-day rolling average, excluding open prices, shows slightly higher error metrics compared to its 7-day counterpart.

5. FinBERT_TS7_without open price

- **MAE Mean**: 23.658
- **RMSE Mean**: 29.872

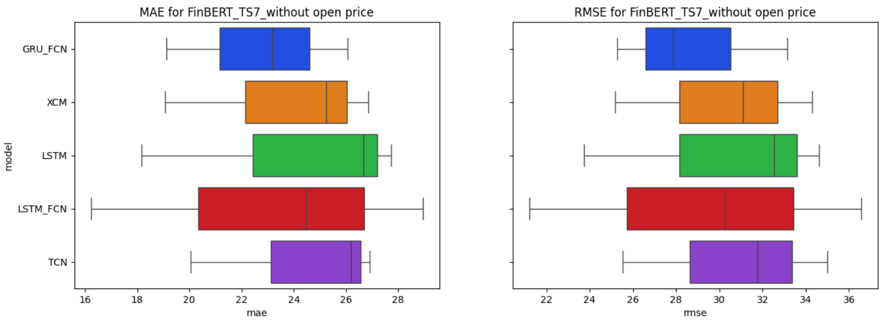


Figure 3.12: FinBERT TS7 without open price BoxPlots

The FinBERT model with a 7-day timestep, excluding open prices, exhibits higher error metrics, suggesting a decrease in accuracy compared to its 14-day rolling average setup.

These results underscore the nuanced impact of excluding open prices in the models and highlight the effectiveness of different sentiment analysis tools over varying timesteps.

Chapter 4

Conclusion

This research has provided critical insights into the performance of various models used in sentiment analysis and stock market predictions. The study focused on evaluating models like GRU_FCN, LSTM_FCN, LSTM, TCN, and XCM, and drew comparisons based on Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) metrics.

4.1 Model Performance Analysis

- **Model Consistency:** LSTM demonstrated the greatest consistency in prediction accuracy, indicated by the lowest standard deviation in both MAE and RMSE.

- **Best Average Performance:** TCN showed the best average performance based on mean values of MAE and RMSE. However, it also exhibited considerable variability, suggesting fluctuating prediction accuracy.

- **Error Magnitude:** Higher maximum values in MAE and RMSE for

LSTM_FCN and XCM suggest these models, while performing comparably on average, may experience larger errors in specific instances.

4.2 General Conclusions

- **Timestep Sensitivity:** The accuracy of predictions generally decreases with an increase in the prediction window. This trend underscores the importance of selecting appropriate timesteps for predictive models in the context of stock market analysis.
- **Influence of Open Price:** The impact of including open price data in the models varies, suggesting that its relevance may depend on specific market conditions or the inherent characteristics of the respective model.
- **Comparative Performance:** Vader exhibited relatively consistent performance across various configurations, suggesting its robustness. FinBERT, however, showed higher sensitivity to longer timesteps.
- **Variability and Extremes:** Observed variability in error metrics, particularly the occasional high errors, necessitates further investigation into model consistency and reliability.

These comprehensive findings contribute to a deeper understanding of the dynamics in stock market prediction models and can guide future model selection and development, enhancing the predictive accuracy and reliability of such models in financial market analysis.

4.3 Recommendations for Future Work

Further research could explore more advanced data preprocessing techniques or additional features that might enhance model accuracy and consistency. Particularly, investigating the reasons behind high variability in certain model configurations and exploring ways to reduce these extremes would be beneficial.

4.3.1 Feature Engineering and Selection

Future research can explore more sophisticated feature engineering techniques to extract valuable information from the data. Investigate the relevance of additional features such as trading volume, news sentiment, technical indicators, and economic data. Implement feature selection algorithms to identify the most influential features for prediction.

4.3.2 Time Series Forecasting

Investigate advanced time series forecasting models beyond the models used in the current analysis (TCN, GRU_FCN, XCM, LSTM, LSTM_FCN). Experiment with hybrid models that combine traditional time series models with deep learning models for improved forecasting accuracy.

4.3.3 Ensemble Models

Explore ensemble modeling techniques to combine the predictions of multiple models. Techniques like stacking, bagging, and boosting can enhance overall prediction accuracy.

4.3.4 Cross-Asset Analysis

Extend the analysis to include multiple assets (stocks, commodities, currencies) and investigate the relationships and correlations between different asset classes. Develop models for multi-asset portfolio optimization and risk management.

4.3.5 Real-Time Predictions

Explore methods for making real-time predictions and assess the feasibility of deploying predictive models in live trading environments. Investigate the challenges and opportunities of low-latency prediction systems.

4.3.6 News Impact Analysis

Incorporate real-time news data and assess its impact on stock price movements. Develop models that can detect and react to breaking news events in real-time.

4.3.7 High-Frequency Trading

Adapt predictive models for high-frequency trading scenarios and explore the challenges and opportunities in making predictions at sub-second intervals.

In conclusion, while each model and configuration has its strengths and limitations, this analysis provides a foundation for more targeted and effective approaches in the field of stock market prediction using sentiment analysis tools.

Chapter 5

Bibliography

1. Liapis, C.M.; Karanikola, A.; Kotsiantis, S. "Investigating Deep Stock Market Forecasting with Sentiment Analysis." *Entropy* 2023, 25, 219. https://doi.org/10.3390/e25020219.

2. Bird, S. "NLTK: The Natural Language Toolkit." arXiv 2004, arXiv:cs.CL/0205028.

3. Bird, S.; Klein, E.; Loper, E. "Natural Language Processing with Python." Packt Publishing Ltd.: Birmingham, UK, 2009.

4. Hutto, C.J.; Gilbert, E. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." In Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, MI, USA, 1–4 June 2014.

5. Araci, D. "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models." arXiv 2019, arXiv:abs/1908.10063.

6. timeseriesAI. Timeseriesai/Tsai: Time Series Timeseries Deep Learning Machine Learning Pytorch FASTAI: State-of-the-Art Deep Learning Library for Time Series and Sequences in Pytorch. GitHub