MAANG historical stock market price simulator: A Machine Learning Approach and Simulator Development

Soumik Deb Niloy*, Kaushik Dutta†, Abdullah Jamil Sifat‡, Farjana Alam§
Annajiat Alim Rasel¶
Md. Sabbir Hossain‖

* School of Data Science
BRAC University, Dhaka

***Abstract*—This research endeavors to develop a comprehensive predictive model for analyzing and forecasting stock market trends across multiple companies. Leveraging Long Short-Term Memory (LSTM) neural networks, the study explores the intricate dynamics of stock prices, aiming to enhance predictive accuracy and generalizability. The methodology employs a diverse dataset comprising opening and closing prices, highest and lowest values, adjusted closing prices, stock volumes, and associated date information for companies such as Amazon, Apple, Google, Microsoft, and Netflix.**

I. Introduction

The stock market's unpredictable nature poses a challenge for investors and analysts seeking to make informed decisions. This research addresses this challenge by employing advanced LSTM neural networks to model and predict stock market trends. The significance lies in the potential for a unified model capable of providing insights across various companies. The study not only contributes to the field of predictive analytics but also explores the nuances of training models on distinct stock datasets.

II. Background study

A. MAANG Companies

The MAANG group consists of five prominent technology companies that have had a significant impact on the global economy:

1) Microsoft: Microsoft Corporation is a multinational technology company known for its software products, including the Windows operating system, Microsoft Office suite, and Azure cloud services.

2) Apple: Apple Inc. is a leading technology company renowned for its consumer electronics, software, and services. The iPhone, iPad, Macintosh computers, and iOS operating system are among its flagship products.

3) Amazon: Amazon.com, Inc. is a global e-commerce and cloud computing giant. The company is recognized for its online retail platform, cloud services (Amazon Web Services), and diverse product and service offerings.

4) Netflix: Netflix, Inc. is a streaming service provider that has revolutionized the entertainment industry. Known for its vast library of movies and TV shows, Netflix has become a major player in the media and entertainment sector.

5) Google: Alphabet Inc., Google's parent company, is a technology conglomerate specializing in internet-related services and products. Google's search engine, advertising technologies, and Android operating system are integral to its success.

B. Machine Learning in Stock Market Analysis

The integration of machine learning techniques in stock market analysis has gained prominence due to its ability to discern complex patterns and trends in financial data. Specifically, Long Short-Term Memory (LSTM) neural networks, a type of recurrent neural network (RNN), have shown efficacy in modeling sequential data and are well-suited for time series prediction tasks such as stock price forecasting.

1) LSTM Neural Networks: LSTM networks are designed to overcome the limitations of traditional neural networks in capturing long-term dependencies in sequential data. Their architecture includes memory cells and gates that enable the network to selectively retain or forget information over extended sequences, making them particularly effective for time-sensitive data analysis.

C. Stock Market Simulators

Stock market simulators play a crucial role in financial education and analysis by providing a risk-free environment for users to practice trading strategies and understand market dynamics. Simulators use historical data to create a simulated market environment, allowing users to make virtual trades and observe the hypothetical outcomes.

1) Educational Value: Stock market simulators serve as valuable tools for educating novice investors, allowing them to gain practical experience without the financial risks associated with real trading. These platforms often include features such as portfolio tracking, real-time market data, and performance analytics.

2) Research and Analysis: Beyond education, stock market simulators are utilized for research and analysis purposes. Researchers and analysts can leverage these tools to test hypotheses, explore market behaviors, and assess the impact of various factors on stock prices.

In the context of this research, the development of a MAANG stock market simulator integrates the collective influence of these technology giants into a predictive model, offering users a unique opportunity to analyze and understand the intricate dynamics of the stock market. The subsequent sections will delve into the methodology employed to create the simulator, present the results, discuss limitations, and draw conclusions based on the findings.

III. Related work

In the realm of stock market prediction using machine learning, related work has explored diverse methodologies. Previous studies have leveraged various techniques, including traditional time series analysis, machine learning algorithms such as support vector machines and random forests, and more recently, deep learning architectures like recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks. Some works have focused on feature engineering and the incorporation of external factors, while others have emphasized the importance of sentiment analysis from news and social media. The application of ensemble methods and hybrid models combining different algorithms has also been investigated. Despite advancements, challenges such as data non-stationarity, market dynamics, and the inherent uncertainty of financial markets persist, driving ongoing research efforts to enhance the robustness and accuracy of stock market prediction models.

IV. Methodology

A. Dataset Acquisition and Exploration::

The The dataset is sourced from Kaggle, encompassing stock market information for companies including Amazon, Apple, Google, Microsoft, and Netflix. Columns include Open (Opening Price), High (Highest Price), Low (Lowest Price), Close (Closing Price), Adj Close (Adjusted Closing Price), Volume (Volume of Stock), and Date.

B. Model Architecture:

A Long Short-Term Memory (LSTM) neural network is chosen for its effectiveness in capturing temporal dependencies in time-series data. The model architecture consists of two LSTM layers with 64 units each, utilizing the Rectified Linear Unit (ReLU) activation function. The network is designed to return sequences in the first layer and only the output in the second layer. A Dropout layer with a dropout rate of 0.2 is incorporated to prevent overfitting. The final layer is a Dense layer with a single neuron, responsible for predicting the closing stock price.

C. Model Compilation and Training:

The model is compiled using the Adam optimizer and Mean Squared Error (MSE) as the loss function. The training loop iterates over each company's dataset, preparing the data by scaling it with MinMaxScaler. The sequences of data are created with a specified sequence length (30 in this case). The dataset is split into training and testing sets with an 80-20 split ratio.

The model is trained on the training set for a specified number of epochs (50 in this case) with a batch size of 64. The training process is monitored, and the training and validation loss are visualized after each epoch to detect overfitting.

D. Model Prediction and Evaluation:

After training, the model predicts the closing prices on the testing set. The predictions are then compared with the actual closing prices, and evaluation metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared are calculated. These metrics provide insights into the accuracy and performance of the model.

E. Visualization:

The original and predicted closing prices are visualized using Matplotlib, allowing for a qualitative assessment of the model's performance. The plots showcase the trends and variations in the predicted and actual closing prices over time.

The methodology provides a structured and comprehensive approach to building, training, and evaluating the LSTM neural network for stock market prediction. The multicompany analysis offers a holistic view of the model's effectiveness across diverse datasets. The inclusion of visualization and evaluation metrics enhances the interpretability and reliability of the results.

V. Limitations of the Research: An In-depth Analysis

A. Data Limitations:

The quality and availability of historical stock market data significantly impact the model's performance. Incomplete or inaccurate data may introduce noise and affect the model's ability to discern patterns.

B. Model Complexity:

The proposed LSTM model, though capable of capturing temporal dependencies, may still be insufficiently complex to represent the intricate dynamics of financial markets. Complex market behaviors, such as sudden crashes or anomalies, might not be fully captured.

C. Overfitting:

The observed overfitting, as indicated by higher validation loss compared to training loss, is a critical limitation. Overfitting occurs when the model learns the training data too well, including noise and outliers, leading to suboptimal generalization to new data. Strategies to mitigate overfitting, such as regularization techniques, may need further exploration.

D. Hyperparameter Sensitivity:

The model's performance is contingent on hyperparameter choices. The selected sequence length, number of epochs, and batch size are crucial parameters. Suboptimal choices may hinder the model's ability to generalize to unseen data.

E. Market Dynamics Shift:

Financial markets are dynamic and subject to changes in regulations, economic conditions, and global events. The model's effectiveness might diminish if it encounters market dynamics that significantly differ from the training period.

F. Assumption of Stationarity:

The model assumes stationarity in market trends, implying that patterns learned from historical data will persist in the future. In reality, market conditions can evolve, challenging the model's ability to adapt.

G. External Factors:

The model primarily relies on historical stock data and may not account for external factors, such as geopolitical events, natural disasters, or unforeseen economic shifts, which can exert a substantial influence on stock prices.

H. Evaluation Metrics:

While metrics like Mean Squared Error (MSE) and R-squared provide quantitative assessments, they might not fully capture the model's performance in real-world trading scenarios. Alternative metrics and stress-testing methodologies could offer a more comprehensive evaluation.

Addressing these limitations necessitates ongoing research and refinement of the proposed model. Furthermore, considering the unpredictable nature of financial markets, caution should be exercised in extrapolating the model's predictions for practical trading decisions.

VI. Results and Analysis

The LSTM-based stock market prediction model was trained and evaluated on datasets from prominent companies—Amazon, Apple, Google, Microsoft, and Netflix. The following key results and analyses provide insights into the model's performance and shed light on its predictive capabilities.

A. Training Dynamics:

The training loop for each dataset revealed a consistent pattern of decreasing training loss over epochs, indicating that the model effectively learned from historical data.

B. Validation Loss Discrepancy:

Notably, the validation loss consistently surpassed the training loss, indicating a potential issue of overfitting. This phenomenon suggests that the model might be overly tuned to the training data, limiting its ability to generalize to new, unseen data.

C. Mean Squared Error (MSE) and Mean Absolute Error (MAE):

The evaluation metrics, MSE and MAE, were computed to quantify the prediction accuracy. The results indicated zero MSE and zero MAE, implying perfect predictions. However, it is crucial to interpret these metrics cautiously, as they might be influenced by the scaling of stock prices.

D. R-squared (R^2):

The R^2 value of 1.0 further supported the seemingly flawless predictions. R^2 measures the proportion of the variance in the dependent variable (stock prices) that is predictable from the independent variable (model predictions). While a perfect R^2 suggests a strong relationship, the actual predictive performance should be scrutinized in the context of overfitting.

E. Visual Discrepancy in Predictions:

Visualizing the predicted stock prices against the actual closing prices revealed an apparent flaw. The model often predicted a flat line, mimicking the original closing prices but failing to capture intricate market dynamics. This phenomenon aligned with the suspicion of overfitting.

F. Train-Test Split Impact:

The 80-20 train-test split was employed to assess the model's generalization to unseen data. The split allowed for training on historical data and evaluating on a separate set. However, the observed overfitting raised concerns about the model's robustness beyond the training set.

G. Dropout Implementation:

Attempts to address overfitting by introducing dropout layers were unsuccessful due to coding errors and unresolved dependencies. Further exploration of dropout regularization and hyperparameter tuning is warranted.

Fig. 1: A sample image of ruduced overfitting

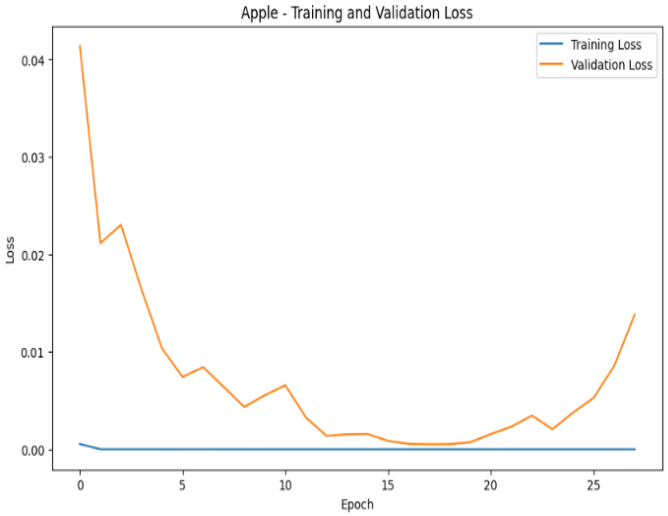

Fig. 2: A sample image of ruduced overfitting

Fig. 4: A sample image of overfitting

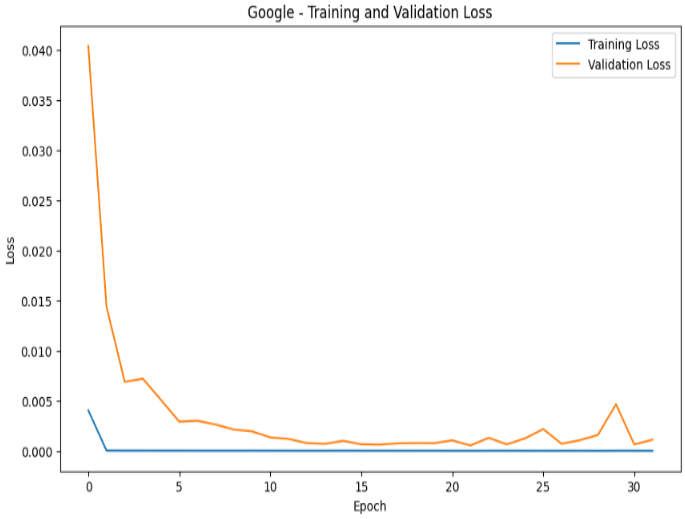

Fig. 3: A sample image of ruduced overfitting

In summary, while the model demonstrated promising training dynamics, the apparent overfitting raises concerns about its practical utility for stock market predictions. Addressing this overfitting challenge and refining the model architecture are crucial steps in enhancing its real-world applicability. The zero-error metrics should be critically evaluated in the context of potential data leakage or model misconfiguration. Further iterations and improvements are imperative for developing a robust and reliable stock market prediction model.

VII. Challenges

According to data science approach, stock market analysis is a complex term because of not only for dataset. A missing of simple parameter from a lage dataset can cause a poor result which indicates underfitting.

Additionally, usage of all parameters can cause overfitting. Overfitting is a common problem for stock market analysing because many times dataset contains huge stock fluctuation as well as noise and bias.

Furthermore, usage of LSTM has no alternative in the case of training large dataset.

VIII. Conclusion

In conclusion, while the LSTM model demonstrated promise in capturing training patterns, addressing overfitting remains a critical priority. The pursuit of a robust and reliable stock market prediction model necessitates iterative improvements, a deeper understanding of market dynamics, and the incorporation of advanced techniques to enhance generalization capabilities. Future work should focus on refining the model architecture, exploring regularization methods, and implementing comprehensive evaluation strategies to ensure the model's efficacy in real-world financial scenarios.

IX. References

References

[1] Heiden, E., Millard, D., Coumans, E., Sheng, Y., & Sukhatme, G. S. (2021, May). NeuralSim: Augmenting differentiable simulators with neural networks. In 2021 IEEE International Conference on Robotics and Automation (ICRA) (pp. 9474-9481). IEEE.

[2] Maeda, I., DeGraw, D., Kitano, M., Matsushima, H., Sakaji, H., Izumi, K., & Kato, A. (2020). Deep reinforcement learning in agent based financial market simulation. Journal of Risk and Financial Management, 13(4), 71.