Stock Market Prediction with Machine Learning and Deep Learning Models

Yashashvini Rachamallu
Masters Student
Department of Computer Science
Michigan State University
rachama2@msu.edu

Deeksha Mohanty
Masters Student
Department of Computer Science
Michigan State University
mohant11@msu.edu

Uta Nishii
Masters Student
Department of CMSE
Michigan State University
nishiiut@msu.edu

ABSTRACT

In response to the imperative need for precise and interpretable stock price predictions within the dynamic financial markets, this project introduces an innovative forecasting system. Leveraging custom deep learning models and ensemble techniques, the system focuses on delivering accurate predictions for a carefully curated selection of publicly traded companies, acknowledging the diverse factors influencing stock movements. The robustness of our approach is ensured through a dual data acquisition strategy: traditional downloading of historical stock price data from Yahoo Finance and web scraping of sentiment-enriched tweets data from finviz.com. The incorporation of sentiment analysis in financial news augments the model's ability to capture nuanced market sentiments. This research addresses the practical challenge of unreliable stock predictions by integrating advanced deep learning models and ensemble methods, presenting a comprehensive solution for investors and financial analysts navigating the complexities of financial markets. Notably, our methodology includes a two-step process: forecasting the next seven days' stock prices and subsequently combining this forecast with sentiment scores to determine investment viability. Evaluation of various models reveals a 63% accuracy rate when sentiment scores are excluded and a 47% accuracy rate when included. This dual-step methodology offers a nuanced and actionable tool for decision-makers, contributing significantly to the evolving landscape of predictive analytics and bridging gaps between traditional methodologies and the demands of contemporary financial ecosystems.

Keywords

Adjusted Close values, Stocks, Accuracy, LSTM, Sentiment, Forecast, RNN, GRU, Machine Learning, Ensemble, RMSE, Naive Bayes, Regression, MLP Regressor, Lasso, Ridge

1. INTRODUCTION

The project presented here stands at the forefront of financial analytics, skillfully marrying the art of finance with the sophisticated science of data analysis. At its core, it leverages a fusion of advanced machine learning and deep learning techniques to accurately predict stock market trends. This approach is not just an academic exercise; it's a practical solution to the intricate challenge of forecasting stock prices, an area where many have stumbled due to the market's inherent unpredictability and complexity.

A key aspect of this research is its innovative use of time series models, a cornerstone of data mining, to decode the patterns and movements in stock prices. This methodological choice demonstrates a deep understanding of both the theoretical and practical nuances of financial data analysis. The project doesn't shy away from the challenges in this domain. It acknowledges and addresses head-on the issues like the scarcity of tweets and continuous financial data for all stock tickers, as well as the difficulties posed by the volatile nature of the stock market. Such challenges make accurate predictions notoriously hard, especially when it comes to long-term investment strategies. However, this research rises to the occasion, offering robust solutions to these problems.

What sets this project apart is its practical approach to unreliable stock predictions. By integrating advanced deep learning models with ensemble methods, it presents a comprehensive and reliable solution for investors and financial analysts. This integration is particularly noteworthy as it enhances the accuracy

and reliability of stock predictions, a critical factor in a field riddled with uncertainties.

The methodology adopted here is both innovative and rigorous. It involves a two-step process that initially forecasts stock prices for the upcoming week and then integrates these forecasts with sentiment scores to evaluate investment viability. This dual-step approach is not just a theoretical construct; it has been empirically tested, showing a 63% accuracy rate without sentiment scores and a significant 47% accuracy rate with them. These figures are a testament to the project's empirical grounding and its relevance to real-world financial decision-making.

In summary, this research marks a significant stride in the field of predictive analytics. It not only tackles the existing challenges in stock market predictions but also innovatively combines various methodologies to offer a nuanced and actionable tool for today's financial decision-makers. By bridging traditional financial analysis methods with the demands of the contemporary financial ecosystem, this project significantly contributes to the evolving landscape of financial analytics, offering insights and tools that are both practical and scientifically robust.

2. RELATED WORKS

The project is algorithm-oriented, and is primarily centered around the development and refinement of predictive algorithms in the financial sector. Our focus has been to meticulously craft a prediction model that not only addresses the complexities of the stock market but also harnesses the power of sentiment analysis to enhance prediction accuracy. To this end, we have employed a diverse array of models, including Long Short-Term Memory (LSTM), Autoregressive Integrated Moving Average (ARIMA), Linear Regression, Lasso Regression, Multilayer Perceptron (MLP) Regressor, Gated Recurrent Unit (GRU), and Recurrent Neural Network (RNN). Each of these models brings its unique strengths to the table, allowing us to compare and contrast their performance in a rigorous, data-driven manner.

Relevant related work we have drawn our inspirations from include “Stock Price Prediction using Sentiment Analysis and Deep Learning for Indian Markets” by Darapaneni et al, which adopts a novel approach by integrating sentiment analysis from news data with stock price prediction. This combination of financial data and sentiment analysis represents an innovative step in predictive modeling. It also utilizes data from reputable sources like Yahoo Finance and finviz.com provides a broad and comprehensive dataset. This enhances the reliability and validity of the findings. The employment of a variety of models, including LSTM, Random Forest Regressor, and potentially Multivariate LSTM, demonstrates a thorough exploration of different predictive techniques. This diversity allows for a more comprehensive analysis.

However, the method implemented in the paper faced challenges with the relevance of news data sourced from websites. The inclusion of irrelevant news could distort the results, affecting the accuracy of sentiment analysis. The method of computing daily sentiment by concatenating all news of a single day might dilute strong sentiments, leading to potentially misleading neutral overall sentiment scores. The method also encountered difficulties in effectively implementing the Multivariate LSTM model to incorporate sentiment data with close prices. This indicates a limitation in the current methodology and potential for improvement.

Another work we have drawn inspiration from is by M K Ho et al. The method involves using 70% of the data for training and optimizing the model parameters, and 30% for testing the models on an unseen dataset. This distribution ensures that the models are well-trained and also rigorously tested. The use of the ARIMA model is beneficial for making predictions using time series data. The methodology starts with the identification of models based on the behavior of autocorrelation (ACF) and partial autocorrelation (PACF) plots, followed by parameter estimation and diagnostic checking using the Ljung-Box statistic test. This systematic approach ensures a thorough and reliable model development process. Also, the implementation of Neural Networks, allows for complex processing of information and can potentially capture intricate patterns in the data. The methodological approach of processing input data through multiple nodes and layers enhances the model's ability to make nuanced predictions. The use of LSTM network, a type of recurrent neural network, is also advantageous for predicting stock prices, which are inherently sequential and time-dependent.

However, the methodology in this paper describes a general structure of neural networks but does not delve into specific architectures or configurations, which could limit the understanding of how these models are specifically applied to the stock price prediction problem. Also, LSTM models are sensitive to the quality of input data. Inaccuracies or anomalies in the data can significantly impact the model's performance. Both ARIMA and LSTM models heavily rely on historical data, which may not always be a reliable indicator of future trends, especially in volatile and unpredictable markets like stock exchanges.

3. PROBLEM STATEMENT

In the dynamic and complex world of financial markets, accurate stock price prediction remains a significant challenge, often due to the unpredictable nature of market movements and the myriad of factors influencing stock behaviors. This project seeks to address this challenge by developing an innovative forecasting system that integrates advanced machine learning and deep learning techniques. The primary objective is to create a sophisticated, data-driven tool for precise and interpretable stock price predictions, catering to the needs of investors and financial analysts in today's fast-paced financial landscape.

The core of our research involves a dual data acquisition strategy: traditional historical stock price data sourced from Yahoo Finance and sentiment-enriched tweets data obtained through web scraping from finviz.com. This approach allows for a comprehensive analysis that considers both numerical stock data and qualitative market sentiments. Our methodology includes the development and implementation of custom deep learning models and ensemble techniques, focusing on a two-step process: first forecasting the next seven days' stock prices and then integrating these forecasts with sentiment scores to evaluate investment viability.

Our data mining tasks span various domains, including time series analysis for stock price prediction, sentiment analysis to gauge market mood, and classification techniques to categorize investment recommendations. The project extends beyond mere prediction; it aims to develop a user-friendly investment recommendation system that translates complex algorithms into actionable insights for real-world financial decision-making. This system is designed to be intuitive and practical, allowing users to input stock names and dates to receive tailored investment advice based on our predictive models.

In summary, by blending traditional predictive models with modern sentiment analysis and creating a user-friendly investment recommendation tool, we are pushing the boundaries of what's possible in financial forecasting, offering a sophisticated, data-driven solution for today's dynamic financial markets.

4. METHODOLOGY

Our methodology is robust and multifaceted. We train our best-performing model on seven meticulously curated datasets, each offering a unique perspective on stock market behaviors. This approach allows us to predict the next seven days' adjusted closing values of stocks—a task we undertake with two distinct datasets: one with normal data and another with hybrid data, which includes appended sentiment scores. This dual dataset approach provides a comprehensive view, enabling us to analyze stock trends both in the absence and presence of market sentiment.

Further refining our predictive capabilities, we implement Logistic Regression on both the Normal and Hybrid Data. This step is crucial as it helps us to understand the impact of sentiment scores on our predictions, offering a deeper insight into how public perception and market news can influence stock movements.

One of the most practical aspects of our project is the development of an Investment Recommendation system. This system is designed to be user-friendly and highly practical. Users can input a stock name and date, and the system, leveraging our customized predictive algorithms, will provide a recommendation. This feature transforms our complex algorithms into a tangible tool that can be utilized by investors, traders, and financial analysts to make informed investment decisions.

The methodology of our report delineates a comprehensive approach to predictive analytics in the financial domain, focusing on data collection, preprocessing, sentiment analysis, visualization, and

predictive modeling. Each step is designed to ensure accuracy and reliability in our predictions and investment recommendations.

4.1 Data Collection

At the core of our methodology is a meticulous data collection process utilizing Yahoo Finance and finviz.com as our primary sources, chosen for their reliability and extensive information coverage. These sources provided a blend of historical financial data and real-time news data. Our analysis focuses on a diverse group of companies, including Hertz Global Holdings, Chevron Corporation, Barclays PLC, Toyota, Amazon, Walmart, and Google, offering a broad perspective on market trends across various sectors.

During the aforementioned phase, we diligently curated a dataset spanning from January 1, 2015, to November 30, 2023, incorporating historical data for the aforementioned seven companies. Subsequently, we extended our data collection efforts to capture social sentiments directed towards these companies, particularly in the form of tweets, associating them with their respective posting dates. In essence, our dataset presents a comprehensive representation of both the historical performance of these companies and the concurrent social sentiments expressed through online platforms. This integrated approach enhances our ability to discern patterns and interactions between financial indicators and public perceptions.

4.2 Text Preprocessing and Sentiment Analysis

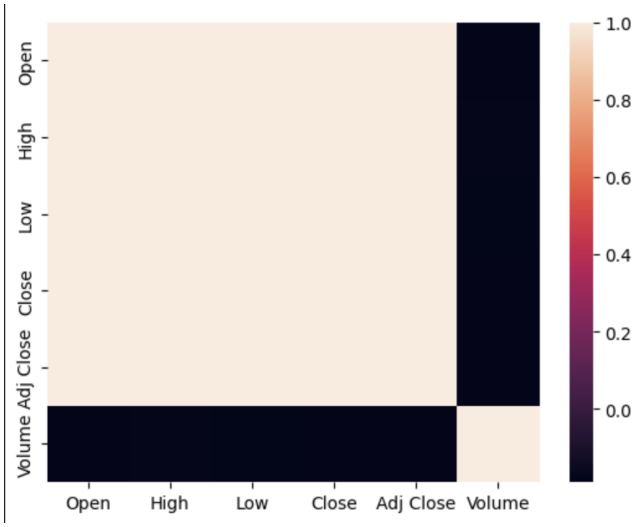
A critical component of our methodology is the preprocessing of text data and sentiment analysis. Here, we address missing values and clean the data to ensure its quality. The Vader Sentiment Analyzer is used and predicted the positive, negative, compound and neutral score of the text.. This step is crucial for interpreting the sentiment conveyed in news articles and social media posts, which can significantly impact stock prices.

4.3 Visualizations

Visualization plays a vital role in our methodology, aiding in the interpretation and understanding of complex data sets.

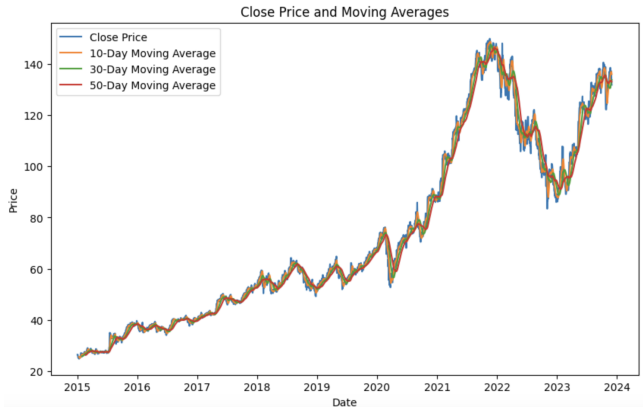
4.3.1 Heat Map

Our heat map visualization demonstrates a high correlation among all attributes, with most values nearing 0.99. This indicates a strong interrelationship between the various data points we've analyzed.



4.3.2 Moving Averages Chart

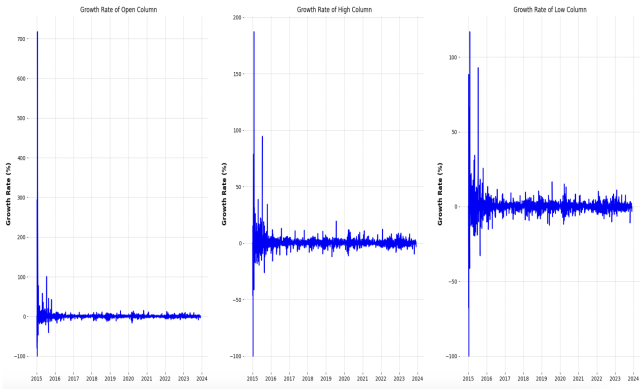
We utilize moving averages of different lengths (such as 50-day, 100-day, 200-day) to compare short-term trends against long-term trends. Observing the convergence or divergence of these averages helps indicate potential market changes.



4.3.3 Growth Rate Chart

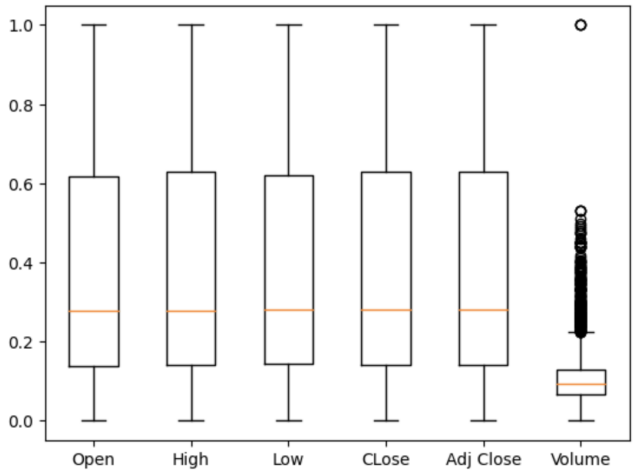
The growth rate charts are employed to identify trends in stock prices. A consistently positive growth rate

signals an uptrend, whereas a negative rate suggests a downtrend.



4.3.4 Box Plot Analysis

Our box plot analysis reveals the presence or absence of outliers in our data attributes. Notably, the only attribute with outliers is Volume, which could be indicative of significant trading activities or market events.



4.3.5 Predictive Modeling and Classification

Our methodology's core is the development and refinement of predictive models. We utilize a range of machine learning models and neural networks, along with ARIMA for time-series analysis. The models are trained on seven distinct datasets to forecast the next adjusted close values. We differentiate between 'Normal Data' and 'Hybrid Data', the latter including appended sentiment scores.

For classification modeling, we perform logistic regression on both Normal and Hybrid Data sets. This step is crucial for understanding the impact of sentiment scores on the stock predictions.

The culmination of our methodology is an investment recommendation system. This system takes as input a stock name and date, and using the trained models, provides a recommendation to invest or not. This system is designed to be user-friendly and accessible, transforming complex data and analytics into actionable insights.

In summary, our methodology is a multi-faceted approach combining rigorous data collection, advanced text preprocessing, sophisticated visualizations, and state-of-the-art predictive modeling. This approach ensures that our investment recommendation system is not only accurate but also grounded in a thorough analysis of financial data and market sentiments.

5. EXPERIMENTAL EVALUATION

In this section, we detail our experimental framework, focusing on the dataset characteristics pre- and post-preprocessing, including data instances and attributes. We then describe the evaluation measures used, and the software tools employed. This concise overview aims to ensure clarity and ease the replication of our study

5.1 Experimental Setup

5.1.1 Dataset

In the preprocessing phase of our experiment, we initially identified missing data points in our dataset, which primarily consisted of historical data from seven companies over eight years. These gaps were attributed to public holidays and weekends when the stock market was closed. To address this, we employed a forward-filling method, where missing values were assumed to be identical to the preceding or following day's data, ensuring no days were left without data. Further analysis, particularly through visualization techniques, led us to the conclusion that the 'Adjusted Close' (ADJclose) price should be our primary predictive attribute for investment recommendations. This decision was based on the unique characteristic of ADJclose being influenced by all other attributes, both preceding and subsequent, a feature not shared by other attributes in the dataset.

Utilizing this insight, we created seven distinct datasets, one for each of the companies under study.

We added a new 'invest' attribute to our dataset, based on 'Adjusted Close' (AdjClose) values for short-term investment analysis. If the AdjClose value increases from one day to the next, we label 'invest' as 'yes', suggesting a favorable investment. If it decreases, we label it as 'no', indicating a less favorable investment opportunity. This approach provides a straightforward method for assessing daily investment potential.

Additionally, we developed a dataset featuring a forecast for the next seven days, utilizing the most effective model, which will be elaborated upon in subsequent sections of this paper. In our endeavor to enrich the data, we created two distinct datasets: one incorporating sentiment scores and another without them. We then merged the 7-day forecast data with the sentiment scores, aligning them based on the date attribute. This integration resulted in a dataset we named 'forecast_sentiment'. The other dataset, simply containing the 7-day forecast, was named 'forecast'.

5.1.2 Evaluation Measures and Software Details

In this study, we employed distinct evaluation metrics to assess the performance of our models, tailored to the specific nature of the tasks - classification and forecasting.

For the classification aspect of our analysis, we utilized 'Accuracy' as the primary metric. Accuracy is a straightforward measure, calculating the proportion of correctly predicted instances against the total number of instances. This metric is particularly useful in evaluating the overall effectiveness of our classification model in correctly identifying categories or classes in our dataset.

In our forecasting domain, we used 'Root Mean Square Error' (RMSE) and 'R-squared' (R^2) as evaluation metrics. RMSE measures the average errors between predicted and actual values, emphasizing larger errors for a more robust model assessment. R^2, or the coefficient of determination, complements this by quantifying how well the model predicts unseen samples, offering a broader view of the model's predictive accuracy and overall effectiveness.

The primary software tool employed for our data analysis and processing was Visual Studio Code. The hardware used for running our experiments featured an Intel CPU with a clock speed of 2.0 GHz. This hardware configuration provided the necessary computational power and speed to efficiently handle the data processing and analysis tasks, ensuring minimal delays and optimal performance throughout our research.

5.2 Experimental Results

The below sub-sections details the results obtained on experimentation.

5.2.1 Best Model Analysis

To identify the best-performing model for our study, we initially focused on a single dataset from Google, implementing various models to determine their efficacy based on the Root Mean Square Error (RMSE) and 'R-squared' (R^2).

The R-squared values derived from several regression models, as illustrated in Figure [1], led us to conclude that Linear Regression, Ridge Regression, Lasso Regression, and MLP emerged as the top-performing models. This conclusion was based on their superior R-squared values, indicating their effectiveness in our analysis.

In addition to traditional machine learning predictors, our study also incorporated deep learning models, specifically Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNN), and Gated Recurrent Units (GRU). For these models, we employed Adam's Optimizer and set Mean Squared Error (MSE) as the loss function. The training was conducted over 200 epochs, ensuring ample learning and optimization of the models' parameters.

The Figure [2] and Figure [3] contains the forecasting results of all the mentioned models on the google dataset. Each model forecasted the stock for the following days. As illustrated clearly in the figures we can see which model is forecasting accurately.

This approach allowed for a straightforward comparison of each model's forecasting accuracy under consistent conditions. Each model was rigorously implemented and assessed using the Google dataset, focusing on R^2 score, which signifies higher precision in forecasting. The effectiveness of each model has been written in the Table [1]. This comprehensive evaluation led us to identify the most accurate model, which then formed the basis of further analysis and application in our research.

Models	R^2 score
Linear Regression	0.91
Lasso Regression	0.91
Ridge Regression	0.91
MLP	0.91
RNN	0.94
LSTM	0.94
GRU	0.97

Table [1]

The data presented in our tables and figures unequivocally demonstrate the superior performance of deep learning methods over traditional machine learning (ML) approaches, with Gated Recurrent Units (GRU) emerging as the top performer. However, due to significant hardware constraints, utilizing GRU for the entire dataset is impractical. On the other hand, Long Short-Term Memory (LSTM) networks, while outperforming ML methods, show comparable effectiveness to GRU, especially when considering the R-squared scores and corresponding plots. This makes LSTM a more feasible choice for our analysis, balancing performance with hardware limitations.

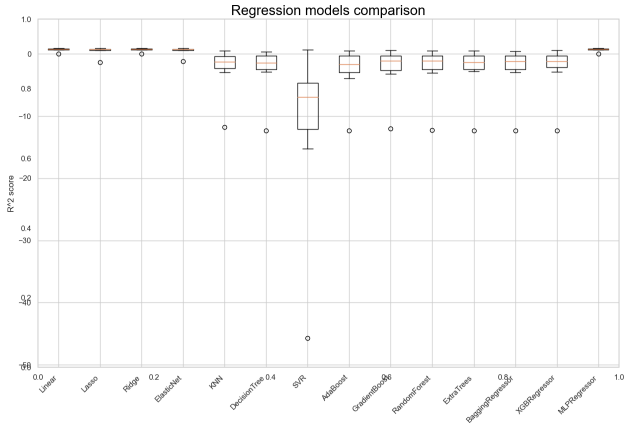


Figure [1]

Figure [2]

Figure [3]

5.2.2 Classification

We applied logistic regression to build two classification models, one for each of the 'forecast_sentiment' and 'forecast' datasets. The model using the 'forecast_sentiment' dataset achieved an accuracy of 47%, while the model built on the 'forecast' dataset resulted in a higher accuracy of 63%.

5.3 Discussion

Upon analyzing the results through the plots, it became evident that the forecasts for the next seven days using the best model were quite accurate and consistent. Regarding the classification models, the accuracies achieved were not as high as anticipated. This lower performance can be primarily attributed to the lack of comprehensive sentiment data, which limited the models' ability to make more accurate predictions.

6. CONCLUSIONS

This project introduces a novel forecasting system for precise stock price predictions, combining deep learning models and ensemble techniques. Leveraging historical stock data from Yahoo Finance and sentiment-enriched tweets from finviz.com, the approach addresses the challenges of market unpredictability. The two-step methodology achieves a 63% accuracy rate without sentiment scores and 47% with sentiment analysis, offering a nuanced tool for decision-makers. The project's integration of advanced models and sentiment analysis distinguishes it in the financial analytics landscape, providing practical solutions for investors and analysts. The methodology involves robust data collection, innovative visualizations, and a user-friendly investment recommendation system. Experimental evaluation highlights the effectiveness of LSTM and GRU models, with logistic regression achieving accuracies of 47% and 63% for datasets with and without sentiment scores. Overall, the project contributes significantly to predictive analytics, bridging gaps between traditional methodologies and contemporary financial demands.

7. FUTURE WORKS

Expanding our dataset is essential for enhancing model accuracy and capturing a broader range of market trends. A larger, more diverse dataset will allow for a deeper analysis of complex market behaviors, leading to more robust and comprehensive models.Implementing our best-performing model, GRU, with GPU support is a priority. GPUs offer increased computational power, enabling faster and more efficient processing. This upgrade will allow for more sophisticated modeling and potentially more accurate predictions.

Transitioning to long-term market trend analysis will provide a more detailed understanding of market dynamics. By analyzing extended historical data, our models can uncover deeper patterns and offer insights into long-term investment strategies and market shifts.

8. REFERENCES

[1] Darapaneni, Narayana & Paduri, Anwesh & Sharma, Himank & Manjrekar, Milind & Hindlekar, Nutan & Bhagat, Pranali & Aiyer, Usha & Agarwal, Yogesh. (2022). Stock Price Prediction using Sentiment Analysis and Deep Learning for Indian Markets.

[2] M K Ho et al 2021 J. Phys.: Conf. Ser. 1988 012041

[3] Sayah, F. (n.d.). Stock Market Analysis & Prediction Using LSTM. Kaggle. Retrieved 2023, from https://www.kaggle.com/code/faressayah/stock-market-analysis-prediction-using-lstm

[4] Adusumilli, R. (2019, December 26). Predicting Stock Prices Using a Keras LSTM Model. Towards Data Science. https://towardsdatascience.com/predicting-stock-prices-using-a-keras-lstm-model-4225457f0233

[5] Stock Closing Price Prediction using Machine Learning Techniques, Procedia Computer Science, Volume 167, 2020, Pages 599-606, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.326

[6] Hiransha M, Gopalakrishnan E.A., Vijay Krishna Menon, Soman K.P., NSE Stock Market Prediction Using Deep-Learning Models, Procedia Computer Science, Volume 132, 2018, Pages 1351-1362, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.05.050.

[7] Jadhav, K. (2021). Stock Market Prediction Web App using Machine Learning And Sentiment Analysis [GitHub repository]. GitHub. https://github.com/kaushikjadhav01/Stock-Market-Prediction-Web-App-using-Machine-Learning-And-Sentiment-Analysis/tree/master

[8] Yahyaoui, Y. (2023, October). Time Series Analysis and Stock Market Prediction. Kaggle. https://www.kaggle.com/code/yassineyahyaouii/time-series-analysis-and-stock-market-prediction

[9] Project Code can be found at : Project_Code