Utilizing ARIMA for Stock Price Prediction: Insights From Daily Trading Data

Necati Sefercioğlu
Electronics and Communication Engineering
Istanbul Technical University Faculty of Electrical and Electronics Engineering
Istanbul, Turkey
necati.sefercioglu@outlook.com

Abstract—**Stock price prediction using machine learning methods are highly valuable for traders. To create more accurate predictions, traders require many inputs like market parameters, target stock's data, economic and political inputs. State-of-the-art models are quite complex models with large-scale data inputs and many parameters. Their dependency for powerful hardware, high computation cost and long process of data collection makes these models impractical for more simpler tasks. In this paper, an implementation of auto-regressive integrated moving average (ARIMA) model for stock price prediction is presented using Amazon's daily trading data (opening price, highest price, lowest price and trading volume) in order to provide an easy-to-use method which only requires daily stock data as its input. Also, implementation of different models like SARIMA (seasonal auto-regressive integrated moving average) and LSTM networks (long short-term memory neural networks) are discussed. This paper can be used as a guide to a user-friendly implementation of machine learning methods for stock price prediction for traders.**

Index Terms— **auto-regressive integrated moving average (ARIMA) model, stock price prediction, time series forecasting.**

I. INTRODUCTION

STOCK price prediction is a very attractive subject for many people. Investors want to maximize their profits, economists and governments want to observe the market, and etc. by predicting the future price of the stocks. Yet with the impact of random noise and high volatility of the stock market, the trends are complex and hard to predict [1]. Also, stock market behavior is unpredictable and is affected by the global economy, governments and financial assumptions and many other variables which is another challenge [2].

Traditionally, forecasting in stock market is done by two methods: Fundamental Analysis and Technical Analysis. In fundamental analysis, industry efficiency, economic values, and the political environment are taken into account. In technical analysis, statistical and mathematical methods are used based on the idea that stock price's past trends affect its current trend [2], [3].

While fundamental and technical analysis continue to maintain their importance, new methods has emerged in financial markets for making investment decisions with the advances in artificial intelligence. Machine Learning (ML) based approaches for stock analysis are usually implemented under the category "time series forecasting with ML".

Stock price prediction by using machine learning methods build up on the same idea that the technical analysis uses: historical stock data -past trends and patterns- affect the current stock trend. However, instead of using the traditional technical analysis methods for understanding the relation, it uses ML methods. In the past, stock market predictions were only done by financial experts. Now with the progress of learning techniques, data scientists have also started solving prediction problems [4].

Previous works have shown that time series forecasting models with machine learning are promising for stock price prediction, having already reached short-term, medium-term, and long-term prediction capabilities for stock market index, stock price and predict the rise or fall of stock price [1]. Related work about the state-of-the-art models and highly accurate complex models will not be presented in this paper due to the scope, but it is highlighted that Shuzhen Wang provides a great scope about these models [1].

The state-of-the-art models give the best predictions. However, they can't be widely used because of the resource limitations which include: the need of very powerful hardware, large data collection and high computation costs. The hardware limitations occur as most of the state-of-the-art models are deep learning models which require powerful GPUs or AI-specialized hardware which are both quite expensive and are not widely available. The need of large data collection occurs from: 1) the more data used for training the better accuracy is reached for predictions. 2) when the number of independent variables is increased, the prediction for target variable can be made better in changing market situations as more variables are presented in the model. Finally, the high computation costs occur from the energy demand of these required powerful hardware. Also, this is not the case for just state-of-the-art models. Most of the complex models do have the same limitations as well.

In this study, an implementation of auto-regressive integrated moving average (ARIMA) model for stock price prediction is presented using Amazon.com, Inc's historical stock data. The goal of this study was to make an implementation of stock price prediction model using machine learning which doesn't have many complicated requirements so that it can be widely used, quickly implemented and easily used for different stocks. To achieve these goals, the limitations mentioned for the state-of-the-art models had to be resolved: 1) ARIMA model was chosen because it doesn't need very powerful hardware and it can be easily implemented by our daily computers. 2) Daily trading data: opening price, highest price, lowest price and trading volume are used. These data are easily accessible. Also, general stock market data are not included, only the target stock's daily trading data is included, so that data is easily obtained and there are not so many independent variables which increases the size of the data. 3) Since the models can be run on daily computers, there isn't any problem of any computational cost.

II. BACKGROUND RESEARCH

ARIMA is one of the most popular methods for time series prediction. It is a linear regression model that is used to track linear trends in stationary time series data where future values are generated from linear functions observed in the past [5].

SARIMA (seasonal auto-regressive integrated moving average) model is an extension of the ARIMA model which includes seasonality components. It is used to make better predictions when the trends show seasonality. SARIMA can create better predictions than ARIMA model when the data shows seasonal trends where data patterns repeat at regular intervals over time.

LSTM (long short-term memory) networks, are a special type of RNNs (recurrent neural networks). They are highly successful at handling sequences of data especially when important patterns and trends are spread out over time, which makes them a great choice in time series forecasting problems. For example, Sisodia et al. proposed an LSTM model that had an accuracy of 83.88% on price prediction, using ten years of historical data for the NIFTY 50 index of India's National Stock Exchange (NSE) [6].

LSTM networks are expected to give better results than SARIMA and ARIMA models. However, their implementation is very complex compared with SARIMA and ARIMA models. Moreover, sometimes vanishing gradients problem occur and LSTM model may have difficulty learning effectively. Just like ARIMA, SARIMA's implementation is easier when compared with other more complex models. However, when the seasonality's period increases, computation of SARIMA gets harder. The data's nature is very important when using SARIMA. For instance, if daily data is used and periods are monthly, this may not cause any problem in computation part, but if periods are yearly for the same data, computation might become very hard. ARIMA model doesn't have these problems due to its less complex nature, but its learning capability is smaller accordingly.

III. METHODOLOGY

A. Dataset

Amazon.com, Inc.'s daily stock price (AMZN stock) data is obtained from Kaggle. The dataset has Apache 2.0 license. Columns include the following: Date, Open, High, Low, Close, Adj Close and Volume representing date of the data, opening stock price of the day, highest stock price of the day, lowest stock price of the day, closing stock price of the day, adjusted closing price of the day and trading volume of the day. Data is recorded according to business days. Weekends and holidays are not included in the dataset. There are no missing values in the dataset. Data recording start from 15.05.1997 and ends with 05.12.2023.

It is observed that "Close" column is exactly equal to "Adj Close" column. Therefore, "Adj Close" column is deleted in data preprocessing in order to reduce the dimensionality of the data.

TABLE I
FORMAT OF DATASET AFTER PREPROCESSING WITH SOME EXAMPLE ROWS

Date	Open	High	Low	Close	Volume
2023-01-03	85.459999	86.959999	84.209999	85.820000	76706000
2023-01-04	86.550003	86.980003	83.360001	85.139999	68885100
2023-01-05	85.330002	85.419998	83.070000	83.120003	67930800
2023-01-06	83.029999	86.400002	81.430000	86.080002	83303400

Also, the closing price history of the AMZN is plotted in order the visualize this data set and obtain a better understanding of the data.

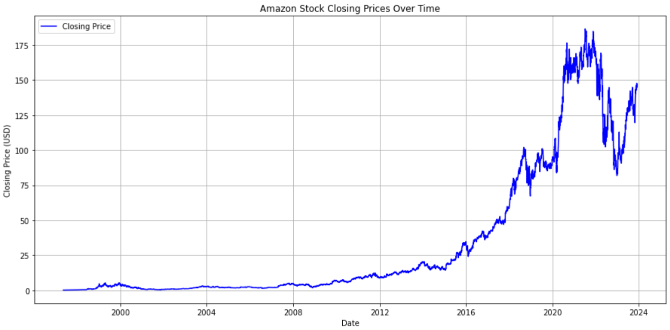

Fig. 1. AMZN Closing Price History

B. Model Determination

In order the find the best model that suits this paper's goal, LSTM networks and more complex models are eliminated from the start as their implementation difficulties inhibit their widespread usage.

Next, to decide between ARIMA and SARIMA models, time series decomposition is applied according to closing price variable as it is the target variable.

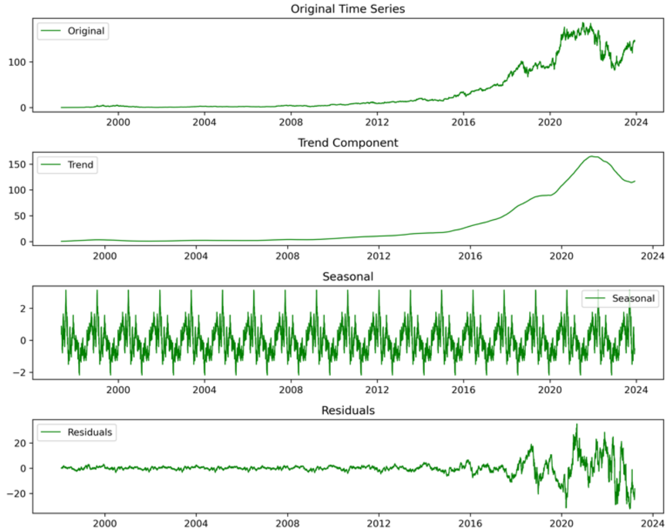

Fig. 2. AMZN Closing Price History Time Series Decomposition

From the time series decomposition, it can be easily seen that this dataset shows a significant seasonality. When the plot showing the seasonal component of the time series is analyzed, it is observed that this data has a seasonality with periods somewhere around 1.5 to 2 years. Considering that data is indexed daily according to business days (around 250 days a year), periods correspond to around 375 to 500 days. This is a significantly high value for the seasonality parameter of the SARIMA model. As discussed before, computation of SARIMA model is not possible in daily computers -or takes too much time- with these high seasonality parameters as initial trials also showed. In the mentioned trials, SARIMA model was tried to be implemented using the target variable closing price. When the seasonality parameter was set around the 250 (a business year) or around 375, the computation failed. When the seasonality parameter was set around 20 (a business month) computation succeeded, yet the seasonal part of SARIMA didn’t affect the model as desired since the period was too small to obtain the seasonal trends and it acted like ARIMA. Therefore, as a result of computability limitations experienced from the large periodicity of seasonality, ARIMA model is chosen over SARIMA model for this study, having already eliminated the more complex models.

ARIMA is a single variable model which predicts the future outputs of the target variable using the past trends of this target variable. Yet in the dataset, the columns open, high, low and volume is all assumed to affect the close (closing price) and we are limited to a single variable as a result of the selection of ARIMA model.

Due to the nature of the stock market, all opening price, highest price, lowest price and closing price are usually within a close range in daily basis, where not many exceptions are observed like jumps between these values or extreme values. It is important to note that these exceptions occur, but they are both rare, and they are not effective in long term trends. When these values are plotted in very short periods, the difference between their trends can be seen. However, when these trends are plotted in longer periods, their plots overlap and single trend is seen. Therefore, when target value of closing price is chosen for the ARIMA model, we do not lose any trend relation by discarding opening price, highest price and lowest price due to the nature of ARIMA where past trends are used to predict future values and the nature of these variables which have overlapping trends except in very small periods. The open, high, low and close values in table 1 can be reviewed to gain a better understanding about this situation which will show that these values are indeed in a very close range.

The real issue arises in including the effect of trading volume to the closing prices. The trading volume has a very different trend compared to the closing price. Figure 3 shows the plot of trading volume history, which should be compared with Figure 1 where historical closing prices are plotted in order to observe the difference.

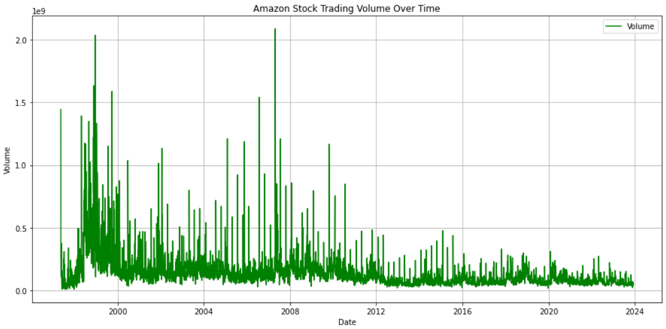

Fig. 3. AMZN Trading Volume History

Since the aim is to predict the future closing prices, closing price is selected as the target variable and the other variables are discarded. While this doesn’t create any problem for the open, high and low variables as discussed before, it creates an important problem for volume variable as we can’t see the effect of trading volume while making new predictions. This is a very significant limitation as prices and volumes express information about an asset’s intrinsic value to entities when an equilibrium state is present, and trading volume also reflects the differing expectations about the future performance of an asset among investors [7]. This limitation will be discussed further with resolution strategies in the later parts of the paper.

In order to see the short-term and medium-term results of the ARIMA model for stock price prediction, the 2023 data of the stock starting from the 01.01.2023 and ending with the 05.12.2023 will not be used for training process of ARIMA. Instead, the trained model will forecast the stock prices for these dates -prediction for nearly a whole year- and these predictions will be compared with the actual prices of the 2023, which was not included in the training part.

The model was tested with different lookback periods in order to obtain the best results. Results are compared using RMSE (Root Mean Square Error) as the error metric. RMSE is chosen because its unit is the same with the target variable, USD (United States Dollar) in this context, to be specific.

Also, 95%, 75% and 50% confidence intervals are calculated and plotted alongside the results to obtain a better understanding about the performance of the ARIMA model.

The optimal parameters of ARIMA model is calculated with trial and error method with the initial guess calculated from auto_arima() function in pmdarima.arima library in python. Also, ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) are calculated and plotted in order to be make more informed guesses in trial-and-error part.

C. ARIMA Model

A process is an ARIMA*(p,d,q)* process if it is stationary and if for every *t*

$$\Phi_p(B)(1-B)^d X_t = \theta_0 + \Theta_q(B)a_t \tag{1}$$

with an AR operator

$$\Phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^p \tag{2}$$

and a MA operator

$$\Theta_q(B) = 1 - \theta_1 B - \cdots - \theta_q B^q \tag{3}$$

where ϕ_m represents the m-th AR coefficient, θ_n the nth MA coefficient, X_t stock price, a_t the error, and B backward shift operator, and d is a nonnegative integer When *d=0*, model reduces to an ARMA*(p,q)* model [8].

IV. RESULTS

With many different trial and errors for the ARIMA model with different ARIMA parameters (p, d, q) with an initial guess of (0,1,0) coming from auto_arima() function, and with many different trials for the lookback periods for model training, following results are obtained:

1) Results are categorized under 3 categories in which: 1) prediction breaks the current trend in upwards direction and predicted prices show an increasing trend. 2) prediction breaks the current trend and stays steady without an increase or decrease staying linear in a constant price, which is equal to the price at the starting date of the prediction. 3) prediction breaks the current trend in downwards direction and predicted prices show a decreasing trend.

2) Best results for the prediction of the 2023 AMZN closing prices are obtained with a lookback period of 3 years. When the training data starts at 01.01.2020 and end at 30.12.2022, it gave the best results for the prediction of next year in each previously defined case.

3) Best results for predictions of each case came when *p=2* and *q=2*. The cases occurred with by the variation of d. When *d=0* it showed the increasing trend, when *d=1* it showed the linear trend, when *d=2* it showed the decreasing trend. The p parameter represents the number of lag observations included in the model. Getting best results when p=2 shows that increasing the number of past data points for future predictions made a positive impact for the results with this lookback period. Therefore, it can be said that trends rely heavily in older inputs when 3-year lookback period is used. The q parameter represents the number of lagged forecast errors that should be included in the model. Getting best results when q=2 shows that increasing the number of past forecast errors for future predictions made a positive impact for the results with this lookback period. Therefore, it can be said that trends rely heavily in older inputs' errors when 3-year lookback period is used. The d parameter represents number of times that the raw observations are differenced. Differencing is done to make the time series stationary. With the increasing number of differencing, the trends went from going upwards, to steady, and finally to decreasing.

A. Best Result For Case 1: Upward Trend

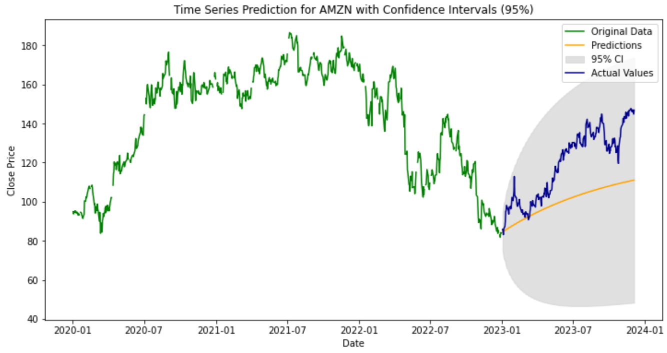

Fig. 4. AMZN Stock Price Prediction For Case 1 Against Whole Lookback Period

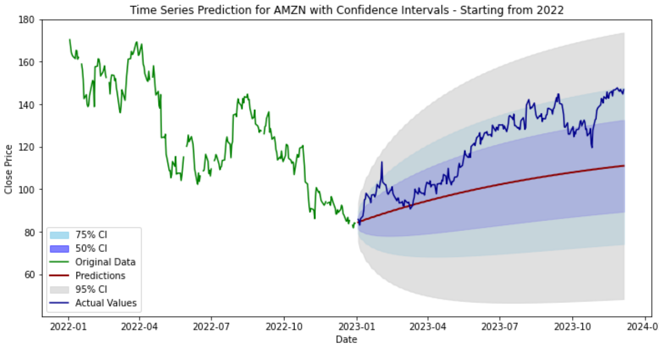

Fig. 5. AMZN Trading Price Prediction For Case 1 Between 01.01.2022 and 05.12.2023

RMSE for ARIMA(2,0,2) = 22.012

Under 1st category which represents the predictions with upward trends, best result is obtained by parameters (2,0,2). RMSE is equal to 22.012 $. Just like the actual 2023 data, predictions show an increasing trend, yet they are below the actual prices. Whole 2023 AMZN prices are actually inside the 95% confidence interval and most of the predictions are inside the 75% confidence interval. Also, 50% confidence interval also includes a significant part of the actual prices. Considering the simplicity of the model, this is actually a very good result.

B. Best Result For Case 2: Steady Trend

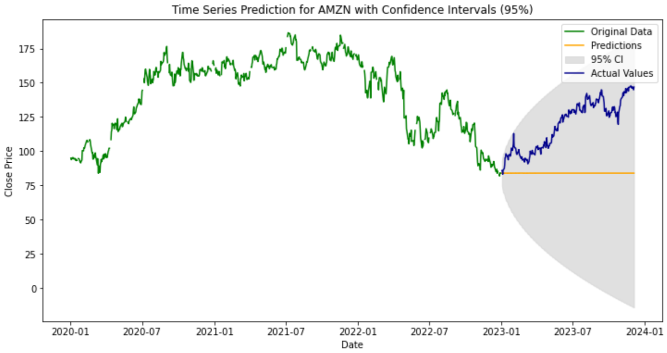

Fig. 6. AMZN Stock Price Prediction For Case 2 Against Whole Lookback Period

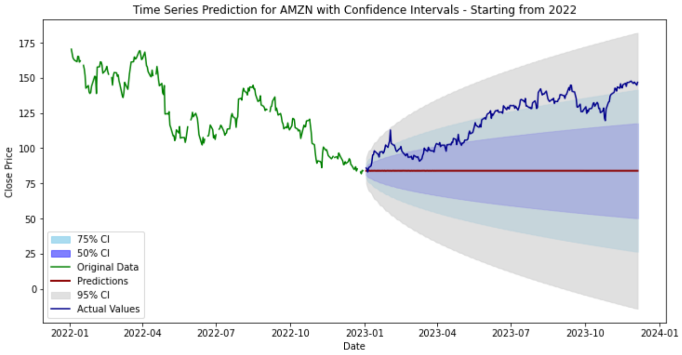

Fig. 7. AMZN Trading Price Prediction For Case 2 Between 01.01.2022 and 05.12.2023

RMSE for ARIMA(2,1,2) = 39.531

Under 2nd category which represents the predictions with steady trends, best result is obtained by parameters (2,1,2). RMSE is equal to 39.531 $. Unlike the actual 2023 data, predictions show a steady trend and they are below the actual prices. Whole 2023 AMZN prices are actually still inside the 95% confidence interval and still most of the predictions are inside the 75% confidence interval. Yet, 50% confidence interval fails to include a significant part of the actual prices. Considering the simplicity of the model, and difference of the steady prediction trend from the actual upward, this is actually a quite good result from the perspective of confidence intervals.

C. Best Result For Case 3: Downward Trend

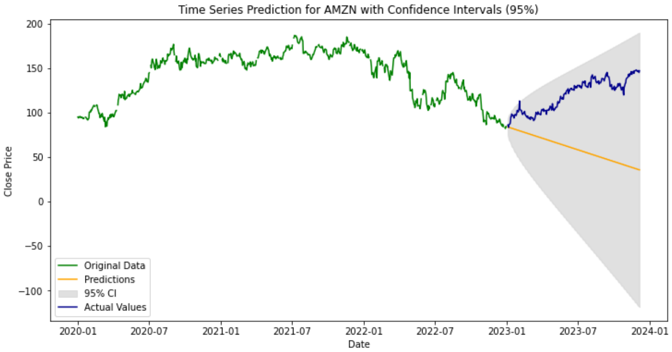

Fig. 8. AMZN Stock Price Prediction For Case 3 Against Whole Lookback Period

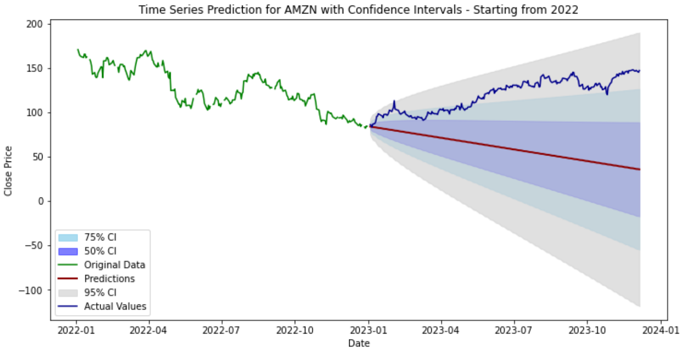

Fig. 9. AMZN Trading Price Prediction For Case 3 Between 01.01.2022 and 05.12.2023

RMSE for ARIMA(2,2,2) = 67.359

Under 3rd category which represents the predictions with downward trends, best result is obtained by parameters (2,2,2). RMSE is equal to 67.359 $. Unlike the actual 2023 data, predictions show a downward trend and they are way below the actual prices. Nearly whole 2023 AMZN prices are still inside the 95% confidence interval but there is a small number of outliers. This time very small number of the predictions are inside the 75% confidence interval and, 50% confidence nearly fails to include any part of the actual prices except very few instances. Model fails to produce an acceptable result.

D. Short-Term Results and Medium-Term Results

(Since Case 3 failed to produce an acceptable result, only Case 1 and Case 2 will be evaluated in this part)

In both cases, short-term predictions are more accurate than the medium-term predictions. It is seen that the models predict the near future better than the following dates. In 1st case, predictions are very close the real prices and even the 50% confidence interval contains most of the actual prices in first half of the year. Then, the gap between predictions and actual values start to increase. In 2nd case, predictions are close the real prices and the 75% confidence interval contains significant part of the actual prices in first 4-5 months of the year. Then, the gap between predictions and actual values start to increase. Considering the different trends between the actual prices and predictions in 2nd case, this result is surprisingly good.

V. CONCLUSION

The main goal of this study was to serve as a create an easy-to-use implementation of a machine learning model for stock price prediction which can be used widespread. Especially depending on the results from case 1, this is reached. The ARIMA selection as the ML model enables this implementation to be done in personal computers without experiencing any hardware limitation. Only daily closing prices for a stock is needed, which makes the computation easy and significantly simplifies the data collection part since stock price history data is widely available for most stocks. The model is not company-specific so after the first implementation, it can be reused for different stocks by using its data for training. These makes this method very attractive to users as they can make an easy and fast prediction from their data, without any additional computation cost as the model is comfortably ran on local devices.

However, there are 2 main limitations of this model:

1) Users still have to predict whether the trend will increase, stay still, or decrease after the prediction date. This means that model is not able to detect which way that the trend is going to move, but it will give good prediction when it knows whether predicted trend will break the current trend in upwards direction, or downward direction, or stay steady around current price.
2) Trading Volume is not included in predictions

A. Comments About First Limitation: Trend Direction

With more basic models, it is hard to guess the direction of the trends as they are dependent to many variables ranging from the economical parameters to social events, and more complex models are needed for predictions that include deciding which direction trend is going to change.

In the scope of complex models, Shi et al. presented DeepClue model which visually interprets text-based deep learning models in predicting stock price movements using deep learning, and they deployed the model to predict S&P 500 stocks using mainstream financial news and firm-specific tweets [9]. DeepClue model concentrates more on the effects of social variables on predictions so it is expected find the trend breaking points from social input when the change of trend is not expected from technical variables. Also, there are some state-of-the-art models with very accurate predictions relying solely on the technical data, when there is no disturbance from social inputs. There are also hybrid-models which includes both parts.

As the proposed ARIMA implementation can't predict the direction of trend all by itself, a question may arise about the target audience of this model. This model targets the people who are familiar with stock market with a goal of seeing how the stock might move for different kind of trends -upwards, downwards, stable- or with a goal of making price predictions for exact dates, having already decided which way the trend will probably go from their own fundamental and technical analysis.

B. Comments About Second Limitation: Trading Volume Problem

Trading volume is an important variable for the stock market and it greatly influences the market dynamics and the stock prices, as previously discussed. Therefore, its effect should also be analyzed while predicting the stock prices. However, ARIMA is a single input model. In order to solve this issue while still using ARIMA, we need a new single variable which represents the correlation of closing prices and trading volumes. For this purpose, dimensionality reduction can be used, which is a powerful method to obtain the compact low-dimensional representation of the observed high-dimensional data [10]. In this case, it can be applied to “closing price” and “trading volume” columns to obtain a single variable that can be later fed into ARIMA model. However, it is important to note that, it might not be easy to do dimensionality reduction on just 2 variables, so adding other desired economic parameters might make the process easier, and also more accurate.

REFERENCES

[1] Wang, S., *A Stock Price Prediction Method Based on BiLSTM and Improved Transformer.* IEEE Xplore, 2023.

[2] Pandey, S., *Forecasting of the Stock Market Price using LSTM- CNN Model with Various Representations of Collected Dataset.* IEEE Xplore, 2023.

[3] Biswas, M., et al., *Predicting Stock Market Price: A Logical Strategy using Deep Learning.* IEEE Xplore, 2021.

[4] Nabipour, M., et al., *Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data;a Comparative Analysis.* IEEE Xplore, 2020.

[5] Wang, Y. and Y. Guo, *Forecasting Method of Stock Market Volatility in Time Series Data Based on Mixed Model of ARIMA and XGBoost.* IEEE Xplore, 2020.

[6] Singh, P., et al., *Harnessing a Hybrid CNN-LSTM Model for Portfolio Performance: A Case Study on Stock Selection and Optimization.* IEEE Xplore, 2023.

[7] Zhanhui, C. and Y. Xin, *Heterogeneous Beliefs, Trading Volume, and Seemingly Emotional Stock Market Behavior.* IEEE Xplore, 2007.

[8] Mauludiyanto, A., et al., *ARIMA Modeling of Tropical Rain Attenuation on a Short 28-GHz Terrestrial Link.* IEEE Xplore, 2010.

[9] Shi, L., et al., *DeepClue: Visual Interpretation of Text-Based Deep Stock Prediction.* IEEE Xplore, 2019.

[10] Liu, Z., et al., *Sparse Low-Rank Preserving Projection for Dimensionality Reduction.* IEEE Xplore, 2019.