2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) | 978-1-6654-9400-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAECT57570.2023.10118349

A New Hybrid Model ARFIMA-LSTM Combined with News Sentiment Analysis Model for Stock Market Prediction

Satyaveer
Dept. of ESE National Institute of Electronics and Information Technology
Aurangabad , *Maharashtra, India*
satyam708435@gmail.com

Prashant patel
Dept.of ESE National Institute of Electronics and Information Technology
Aurangabad , *Maharashtra, India*
pp157943@gmail.com

Harish Chandra
Dept.of ESE National Institute of Electronics and Information Technology,
Aurangabad , *Maharashtra, India .*
harishkumar18081999@gmail.com

Prashant pal
Scientist –'B',Dept. of ESE
National Institute of Electronics and Information Technology
Aurangabad , *Maharashtra, India*
prashantpal@nielit.gov.in

Shashank kumar singh
Scientist-'B', Dept. of ESE
National Institute of Electronics and Information Technology
Aurangabad , *Maharashtra, India*
shashank@nielit.gov.in

Abstract— **we have used the advantages of both data-driven and news-driven models and came up with a novel hybrid model. This model combined the data-driven ARFIMA-LSTM with the NEWS SENTIMENT analysis model. So, in order to obtain a better and more accurate prediction at each time scale, we combined the capabilities of a data-driven model (ARFIMA-LSTM) with a news sentiments-driven model. This model outperforms the traditional models, such as SVM, Random Forest, ARIMA, KNN, GRNN, and LSTM, and it accurately forecasts the market.**

Keywords: - ARFIMA-LSTM, NLP, STOCK MARKET FORECASTING, RNN, RMSE, MSE, MAPE, AI, ANN.

I. INTRODUCTION

In this rapidly developing society, everything is digitalized, and we have steadily transitioned from traditional technologies to modern ones. Everything becomes interconnected. As a result, stock exchanges and global trade are intertwined in today's world. Trading has become incredibly popular as a potential source of income thanks to the development from physical currencies to digital currencies and digital trading platforms. As a result, more people started participating in the trading, which increased its complexity and level of competition. Because of this, conventional techniques and approaches to market forecasting are ineffective. We were continually in search of fresh, trustworthy prediction methods which can predict the market with great accuracy at each scale like short term (Intraday) to long term.

Stock market, which calls for knowledge, precise trading prediction skills, and other attributes. Fundamentally, traders must amass knowledge and expertise in their fields. People have started to think about investing in these kinds of financial industries as a result of recent advancements in technology and the stock market. Additionally, the capacity to foresee outcomes helps the trader succeed. In recent years with the evolution of social media and communication technology and the introduction of new machine learning models and AI, forecasting is become more accurate and reliable, specially, in the stock market that is our prime focus in this paper. Analyzing the stock market data has always been important to businesspeople and traders because it gives a way to select more lucrative stocks. These data must be processed using very effective machine learning models because they are vast in volume, extremely complicated, noisy, and contain some missing values. However, recent studies have demonstrated that the huge amount of publicly available internet information, including Wikipedia usage trends and fresh news from the major media, companies progress report, new government policies regarding business and public welfare have an observable effect on the daily price chart movement in the stock market[7],[8],[9]. Therefore, we can achieve more accurate results with using these two methodologies (data driven and news sentiment analysis) together. So, in this paper we have described a methodology to combine two powerful models ARFIMA-LSTM and NEWS SENTIMENT ANALYSIS models and successfully designed a hybrid model. Paper arrangement is as follows. Section II, III discusses the ARFIMA-LSTM model and its advantages. Section IV discusses the evaluation criteria. Section V, VI discusses the news sentiment analysis model. Section VII discusses the hybridization of both models. Section VIII discusses the simulation results of the hybrid model. And finally, section IX discusses the conclusion.

II. ARFIMA-LSTM MODEL

In this section we will discuss the ARFIMA and LSTM separately for better understanding.

A. ARFIMA

Auto regressive fractional integrated moving average is referred to as ARFIMA [1]. ARFIMA is focused on long-memory operations. As we know that long-memory processes are stable processes with a gradual decay of the autocorrelation function. The ARMA (autoregressive moving-average) is nested within the ARFIMA model, which offers a sparse parameterization of long-memory processes. Short-memory processes can be fractionally integrated to produce long-memory processes. That is what the ARFIMA model does.

The ARFIMA model filters linear tendencies effectively and then the output data from ARFIMA is immediately supplied to the LSTM model, which produces the final result, making ARFIMA a useful filtering tool for LSTM models.

B. LSTM (LONG- SHORT TERM MEMORY)

It is a type of recurrent neural network. Therefore, it is crucial to comprehend neural networks and recurrent neural networks before moving on to LSTM.

NEURAL NETWORKS:

Inspire by biological neural networks, an artificial neural network is a layered arrangement of interconnected neurons. We can do sophisticated operations on data thanks to combinations of different algorithms, not just one.

C. RECURRENT NEURAL NETWORKS (RNN)

An exclusive class of neural networks known as RNN was developed to handle the short-term input. The neural network uses loops to process input in line with the internal state that RNN neurons have established as a cell state or memory [1]. RNNs have recurrent ‘tanh' layers that give them the ability to store information. But not for a long time, which is why LSTM models are required.

Recognizing long-term correlations in data requires the use of a special type of recurrent neural network called an LSTM. This is made possible by the model's recurring module, which consists of four levels that interact with one another. Following is the given figure to better understand how the RNN and LSTM model operates:

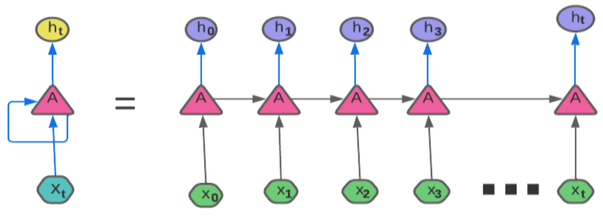

Fig 1: Diagram of RNN Neural Network

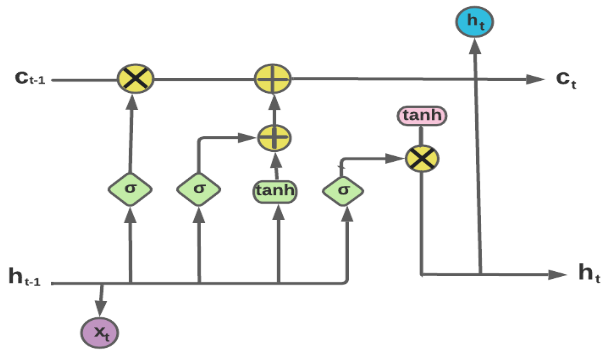

Fig 2: Structure of Hybrid LSTM Neural Network

III. ADVANTAGES OF HYBRID ARFIMA-LSTM MODEL

- Being a mix of both ARFIMA and LSTM, it is more effective than the prior LSTM models and has good prediction accuracy.
- This model greatly minimizes the over fitting issue and volatility.
- It is appropriate for market forecasting on any time scale.

IV. EVALUATION BENCHMARK FOR PERFORMANCE

Mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE) are used to assess how well the proposed nonlinear combination model performs

A). MEAN ABSOLUTE ERROR

$$= \frac{1}{N}\sum_{i=1}^{N}|y_t - \hat{y}_t| \qquad (1)$$

B). ROOT MEAN SQUARE ERROR

$$= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_t - \hat{y}_t)^2} \qquad (2)$$

C). MEAN ABSOLUTE PERCENTAGE ERROR

$$= \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_t - \hat{y}_t}{y_t}\right| 100\% \qquad (3)$$

V. MARKET EXPERT CREATED DICTIONARY BASED NEWS SENTIMENT ANALYSIS MODEL

News is important in the realm of investing because it gives investors information, they need to make judgments on the stock markets. People's emotions and ideas can be shaped and influenced by it, which affects their choice to purchase or sell in markets. Analysis of social media like twitter, BBC WORLD (News channel), text analysis makes it possible to ascertain the subjectivity and viewpoint of the text. News emotion has a substantial impact in forecasting daily stock returns in the stock market, claims World Bank researcher Samuel P. Fraiberger [7].

A Natural Language Processing (NLP) method for determining the Sentiment of the text's extremes is news sentiment analysis or opinion mining (positive, negative, or even neutral) [7],[8]. In order to determine polarity, machine learning architectures such as Support Vector Machines, Boosting and Bagging algorithms, and Random Forests give sentiment ratings to the categories within words in statements. [8].A variety of decision trees are used to create the Random Forests machine learning algorithm, which produces precise and reliable predictions [3]. While employing the data bagging approach to train, to enhance efficiency, this technique additionally adds unpredictability to the data.

In this work, sentiment analysis is used to assess the positivity or negativity of many datasets of daily aggregated news headlines. The positive class made reference to the news stories that led to an increase in the stock price the next day, while the negative class made reference to a fall in the stock price.

VI. IMPLEMENTATION OF THE MODEL

In this study, we retrieved thorough intraday (60-minute gap) pricing information for every stock from

Investing.com, while the news pieces were retrieved from One of the best and most reliable sources of stock-specific news in India is Moneycontrol.com.

A. DATA SOURCES AND PREPROCESSING

To conduct the sentiment analysis, a dataset of individual news stories from the previous six months about the stocks that make up the Nifty Pharma Index is used. A web scraper was originally developed to handle the format of the article links that needed to be deleted. The moneycontrol.com news API was then used to obtain the news articles. With the help of the Python Beautiful Soup package, the complete text of the articles was downloaded from the specified URLs. Only items from the preceding six months for each company were saved by the scraper's design. Additionally, the information includes quarterly progress findings that are crucial for analysis.

B. DATA TRANSFORMATION

The text corpus is converted into numerical vectors using a Python module dubbed "pattern." The primary text of each news article was taken and converted into n-grams [7]. For the sentiment analysis, unigrams (words with a single token) are employed. The context of the text was not conveyed by the unigrams, though. As an illustration, the word "decline/fall" is normally negative word, yet the phrase "costs decreased" is beneficial to the company. The text corpus was eventually used to generate bigrams (word token counts of two) and trigrams (word token counts of three). These made it easier to comprehend the polarity of the words themselves as well as the surrounding sentences, which made it possible to capture the context much more accurately. Following that, we manually constructed a glossary with terms that had particular meanings related to stock market firms. Each definition in the dictionary was assigned one of three sentiment categories: positive (Buy), negative (Sell), or neutral (Hold). After checking the created n-grams against the dictionary, the polarity of every corresponding word was either positive or negative. Depending on whether a match was discovered. The emotion ratings were then created based on the frequency of the positive and negative terms, which were then counted. For instance, the emotion would be +5 if there were five positive words. Neutral words have no bearing on the result. The scores were cross-validated against the stock prices to understand how each new article affected the scores. To evaluate the effectiveness of the sentiment analysis model, a simple portfolio strategy was developed. The decision to "buy," "sell," or "hold" would be made based on the score (do nothing). The stock will be purchased following the release of the news if the aggregate news score is positive and higher than a set threshold, and vice versa. Below is a schematic diagram that depicts the entire process.

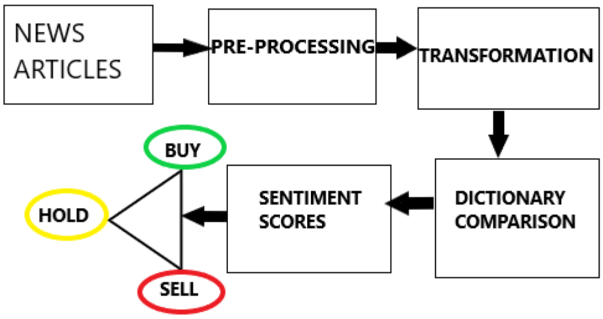

Fig. 3: Diagram of DICTIONARY BASED NEWS SENTIMENT MODEL

VII. HYBRIDIZATION OF THE TWO MODELS (ARFIMA-LSTM AND NEWS SENTIMENTS ANALYSIS MODEL):

For hybridization of two model the following figure illustrates the procedure:

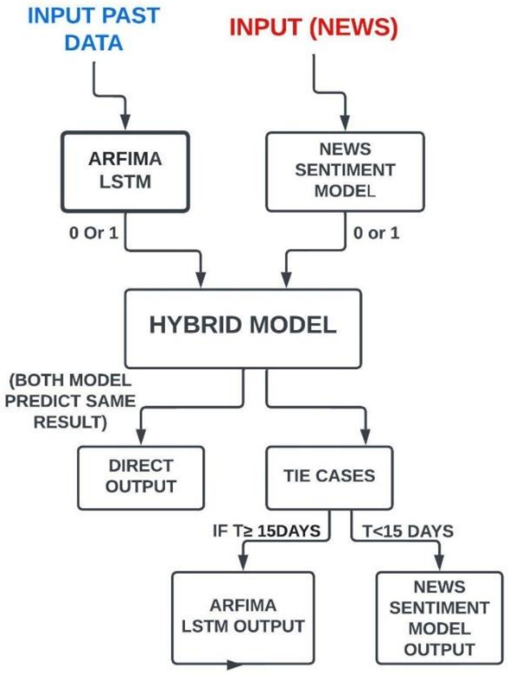

Fig.4: Implementation of Hybrid model

First model (ARFIMA-LSTM) is predicting the market using past data while second model (NEWS SENTIMENTS ANALYSIS) predict the market using news sentiments. Each sub model's output falls into one of the two established groups, Positive Market (BUY, denoted by a 1) and Negative Market (SELL, represented by 0). The prediction for each occurrence provided by two sub models is sent as an input to the hybrid model by design. The following are the criteria on which a hybrid model bases its outcome:

A. If two sub models predicts the market positive than the output of hybrid model is positive means buying the stock is the viable option and a user can place the buying order with calculated stop loss and calculated profit booking.

B. If two sub models predict the market negative than the output of hybrid model is negative and selling is a viable option with proper stop loss and profit booking.

C. In case of a tie, means one model predict the market positive and another model predict the market negative than the time duration for which the prediction is maded will act as a tie breaker as follows:

Let prediction time is t than if t<15 days than we will go with the decision given by sentiment analysis model because news are the decisive factor for the price movement for short duration (for less than 15 days). Additionally, if the sentiment score-based model advises buying a stock, the prediction is only deemed accurate if the stock increases by more than 0.5%. Consequently, a 0.5% threshold has been established for both purchase and sell choices. The "1%" condition is set for neutral or hold decisions, which indicates that if the stock moves in either direction for more than 1%, the prediction is wrong.
3)If t>=15 days than we will go with the decision given by ARFIMA-LSTM model because this model predicts very well for the long duration by using past data as an input with RMSE accuracy is about 80%, which is comparable to traditional predicting rivals.

VIII. SIMULATION AND RESULT

Now, we will see the results for ARFIMA-LSTM model with dictionary- based news sentiment model that we have simulated for SENSEX STOCK MARKET index. We have compared the performance with other available models on the basis of MAE, RMSE and MAPE. The results are shown below:

Table 1. Prediction performance on the basis of MAE, RMSE, and MAP

SR. NO	MODEL NAME	ROOT MEAN SQUARE	MEAN ABSOLUTE ERROR	MEAN PERCENTAGE ERROR
1	ARFIMA	0.2720	0.13560	0.1650
2	ARFIMA-LSTM	0.0542	0.02696	0.0030
3	GENERALIZED REGRESSION NEURAL NETWORK (GRNN)	0.0635	0.03156	0.0116
4	ARIMA	0.3134	0.15700	0.1899
5	ARFIMA-LSTM WITH NEWS SENTIMENT **(Proposed Model)**	0.0532	0.02680	0.0015

IX. CONCLUSION

The simulation results that we have seen in the results section table1 allow us to draw the conclusion that our model, which combines ARFIMA-LSTM with NEWS SENTIMENTS ANALYSIS outperforms the existing models like SVM, GRNN, and KNN etc. Using two sub models of the hybrid model, the model was able to anticipate the market with an accuracy of up to 97%. By assessing the model on a larger data set of various equities available in the NSE, BSE, and international stock markets, the research can be furthered. Additionally, additional social media channels could be included to boost the model's public perception component. Future changes could allow for more flexibility in the Positive and Negative categories because right now, even a small deviation from the previous day's closing price signals can spoil an entire results or prediction. A neutral class may also exist. Finally, we can state that there is always a chance to develop a prediction model that is more accurate because the stock market is a very complex puzzle that cannot always be predicted with 100% accuracy because it depends on the collective judgment of thousands of human minds and numerous machine learning algorithms, some of which are even able to predict the results of other ML algorithms.

REFERENCES

[1] A. H. Bukhari, M. A. Z. Raja, M. Sulaiman, S. Islam, M. Shoaib and P. Kumam, "Fractional Neuro-Sequential ARFIMA-LSTM for Financial Market Forecasting," in IEEE Access, vol. 8, pp. 71326-71338, 2020, doi: 10.1109/ACCESS.2020.2985763.

[2] M. Usmani, M. Ebrahim, S. H. Adil and K. Raza, "Predicting Market Performance with Hybrid Model," 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), 2018, pp. 1-4, doi: 10.1109/ICEEST.2018.8643327.

[3] S. Vazirani, A. Sharma and P. Sharma, "Analysis of various machine learning algorithm and hybrid model for stock market prediction using python," 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), 2020, pp. 203-207, doi: 10.1109/ICSTCEE49637.2020.9276859.

[4] V. Rajput and S. Bobde, "Stock market prediction using hybrid approach," 2016 International Conference on Computing, Communication and Automation (ICCCA), 2016, pp. 82-86, doi: 10.1109/CCAA.2016.7813694.

[5] Y. Wang and Y. Guo, "Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost," in China Communications, vol. 17, no. 3, pp. 205-221, March 2020, doi: 10.23919/JCC.2020.03.017.

[6] S. S. Alotaibi, "Ensemble Technique With Optimal Feature Selection for Saudi Stock Market Prediction: A Novel Hybrid Red Deer-Grey Algorithm," in IEEE Access, vol. 9, pp. 64929-64944, 2021, doi: 10.1109/ACCESS.2021.3073507.

[7] D. Shah, H. Isah and F. Zulkernine, "Predicting the Effects of News Sentiments on the Stock Market," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 4705-4708,
doi: 10.1109/BigData.2018.8621884.

[8] S. Sridhar and S. Sanagavarapu, "Analysis of the Effect of News Sentiment on Stock Market Prices through Event Embedding," 2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS), 2021, pp. 147-150, doi: 10.15439/2021F79.

[9] Z. Wang, S. -B. Ho and Z. Lin, "Stock Market Prediction Analysis by Incorporating Social and News Opinion and Sentiment," 2018 IEEE International Conference on Data Mining Workshops (ICDMW), 2018, pp. 1375-1380, doi: 10.1109/ICDMW.2018.00195.