Hybrid Autoregressive Integrated Moving Average and Long Short-Term Memory model for stock index prediction based on advisor recommendations

Mani PADMANABHAN

Faculty of Computer Applications, School of Social Sciences and Languages (SSL),
Vellore Institute of Technology (VIT), Vellore, India

mani.p@vit.ac.in

Abstract: As per capita incomes rise, more investors engage in high-risk, high-return stock investments. Market uncertainties significantly affect investor outcomes, making accurate stock price forecasting critically important. This study proposes a hybrid forecasting model combining Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) networks enhanced by sentiment analysis of publicly available financial advisor commentary. By extracting the emotional tone from advisor-level textual posts through the advanced natural language processing, the sentiment scores are integrated with traditional market data to improve the prediction of the National Stock Exchange (NSE) closing prices. This combined approach captures both linear market trends and nonlinear sentiment-driven fluctuations, overcoming the limitations of separate methodologies. The empirical results on the NSE indices highlight a significant relationship between the sentiment dynamics and the stock price variation. The hybrid ARIMA and LSTM model incorporating macroeconomic variables and sentiment indicators demonstrates improved forecasting accuracy, aiding investors in risk reduction and decision-making. This research advances the understanding of the behavioural influences on the financial markets and offers a practical tool for evidence-based trading strategies in the volatile emerging markets.

Keywords: Web Crawling Techniques, Stock Prediction, Recurrent Neural Network, Sentiment Analysis.

1. Introduction

Ever since its inception, stocks have been recognized as one of the most important types of virtual capital characterized by the great risk and the extreme force of influence. However, the dynamic and, often, unpredictable nature of the market continues to challenge even experienced investors. The capability to study, analyze and predict NSE equity prices has remained one of the most essential necessities in the modern research in the field of finance. Currently, there are two major paradigms that led the industry, namely the fundamental and technical analysis. The fundamental analysis involves a comprehensive evaluation of macroeconomic factors, firm-specific performance metrics, and broader industry trends. The practice is likewise intrinsically subjective, leaning mainly on the analysts' experience and their judgmental determinations regarding a stock's intrinsic value. By contrast, the technical analysis relies on the quantitative modeling and the statistical analysis of the earlier price and volume activity, a method that yields a perspective grounded more in the scientific fact, measurable directly through objective numerical data. Although the technical approaches enable a higher probability of making accurate predictions, in most cases they are not adequate when used alone. Much more so in the highly volatile modern markets, especially the dynamic indices like National Stock Exchange Fifty (NIFTY 50), the exclusive application of one and the same methodological approach often does not help to reflect the multi-dimensional nature of the movement in prices. Conventional formulations further are unsuited to the integration of the sentiment drivers of a qualitative nature, most notably being the sentiment of the investors, media influence and the general market sentiment which exert a tremendous force of influence on the short-term price movements. Within artificial environments, automated stock analysis platforms are configured and deployed to apply data-driven techniques, ranging from machine learning and natural language processing (NLP) to time-series forecasting, to emulate human decision-making processes. Consequently, they can probe market trends and forecast the evolution of the equities traded on venues such as the NSE, a task that would formerly have required employing dedicated forecasting models. The workflow fuses the collection and processing of large, structured datasets such as historical stock prices with unstructured feeds (e.g., financial news updates and social media messages).

https://doi.org/10.33436/v35i3y202507

In the proposed combination of the ARIMA and LSTM approaches the analysis is carried out fully automatically, concurrently detecting linear market trends and nonlinear sentiment-driven shifts. By utilizing a sentiment analysis pipeline grounded in the natural language processing results, the mood indicators of the investors are extracted from the textual inputs and subsequently incorporated into the predictive ensemble. Consequently, the methodology overcomes the long-standing limitations of the conventional temporal modelling and sentiment analysis in isolation by uniting them into a single cohesive framework, independent of or stacked directly on the quantitative time series techniques and the qualitative sentiment analytics Nelson, Pereira & de Oliveira (2016). The model allows a greater predictive accuracy through the incorporation of understanding of the matter of cognitive biases that are involved in the judgment of the human mind, reduces the exposure of the investors to risks and helps in the identification of more viable entry and exit points in the financial markets. Figure 1 provides an own overview of how manual and automated stock-analysis methods are shown:

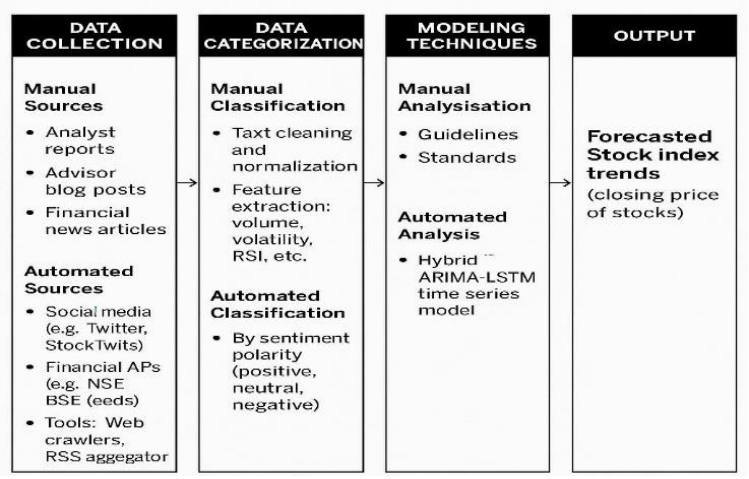

Figure 1. Manual Vs Automated Stock Analysis

In the NSE market, the analysis relies on time-series data for the NIFTY 50 and Sensex indices, and, with the results in hand, the Python-based hybrid ARIMA and LSTM model fuses ARIMA tasked with extracting the linear component of the historical price trajectories with a LSTM neural network that handles the non-linear residuals and intricate temporal dynamics. Expert analytics substantiates that the model founded on this NSE depositor integrated approach outperforms the standalone techniques and delivers a markedly higher level of performance than the purely technical techniques that ignore sentiment. In particular, the evidence shows that the model aids investors in controlling their downside risk without sacrificing significant returns, which in turn makes this model extremely valuable for decision-making in the Indian capital markets.

2. Related research work

ARIMA models are quite useful when the short-term and linear trends are involved but are not suited to the complex dynamical behavior. Conversely, LSTM networks prevalently perform well with modelling non-linearity and long-term dependencies and thus they become very suitable in the process of extracting residual patterns on financial time series. Having combined both of these approaches, the current model aims to differentiate macro-level patterns and micro-level fluctuations Chen et al. (2021). In addition, the sentiment features that can be extracted based on news sources and contents of the social media further enhance the contextual understanding of price dynamics, particularly when they arise due to incidences of higher volatility or the occurrence of market beating events. Among the earlier ones of the hybrid methods, a hybrid approach by Zhang (2003) integrates ARIMA and a feed-forward neural network. The present study shows that the ARIMA performs better and is overall consistent in its performance mainly due to its ability to capture linear patterns in a forecast, whereas the neural networks offer additional input on non-

linear patterns. Further studies used more advanced deep-learning models, which led to the use of LSTM networks to identify long-term dependencies and other characteristics of sequences in the stock data (Fischer & Krauss, 2018). The current paper thus attempts not only to integrate ARIMA and LSTM further under the same roof, but also to demonstrate the synergetic benefits of using the two in tackling non-linearities that cannot be efficiently handled by the traditional models Chen et al. (2016). In particular, the hybrid stock forecasting model involves the use of ARIMA to fit the trend component, and LSTM forecasts and predicts the elaborate residual variation.When contrasted with the Machine learning models, namely the ARIMA and LSTM model, the integrated model produced fewer forecasting errors.

2.1. Hybrid forecasting approaches in financial analysis

Combined linear and non-linear forces predicting the trend in the stock market is one of the most challenging activities. Traditional linear models like the Autoregressive Integrated Moving Average (ARIMA) are effective in capturing consistent patterns and trends within time-series data. Table 1 presents own research the key findings from the earlier study and highlights their relevance to the proposed framework.

Table 1. Related research papers

Paper	**Key Insights**	**Relevance to Proposed Study**
Xiao et al. (2022)	Demonstrates the benefit of ARIMA + LSTM hybrids	Validates the proposed model architecture
Halder (2022)	Uses FinBERT to integrate news sentiment	Mirrors the sentiment integration approach
Jiang (2024)	Offers comparative results with ARIMA	Helps benchmark the hybrid method
Kasture & Shirsath (2024)	Demonstrates hybrid sentiment models in India	Provides relevance in local Indian context
Lim et al. (2023)	Validates decomposition of linear/non-linear series	Supports methodological adaptability

It has been analyzed in terms of related studies asserting that ARIMA was a productive model in the data representation of linear relationships, while the neural network component captured nonlinear residuals, resulting in enhanced forecasting performance. Following this work, Fischer and Krauss used LSTM networks, a variation of the recurrent neural network (RNN), to model temporal dependencies in financial data. Their study revealed that the LSTM models performed far better than the conventional models as they managed to learn about long-range sequential patterns of stock market data (Fischer & Krauss, 2018). Bao, Yue & Rao (2017) used the deep learning framework that stacked the autoencoders with the LSTM in order to emphasize the opportunities of large-scale neural architecture on financial forecasting. Following up on this research, Livieris, Pintelas, E. & Pintelas, P. (2020) developed a model of CNNLSTM to predict gold prices, which had once again confirmed the importance of deep learning to grasp complicated market behaviors. A direct comparison analysis of the ARIMA against the LSTM and the hybrid model showed that the combination of the ARIMA and the LSTM model outperformed the individual component models significantly, particularly when it came to contemporarily estimating the trend and short-term fluctuations (Siami-Namini, Tavakoli & Siami Namin, 2018). The study by Gite et al. (2021) constitutes a systematic review of the existing approaches to stock-market prediction, and points out the fact that there is increasingly more interest in the hybrid approaches in which this problem can be solved through integrating the accuracy of the statistical models with the flexibility of deep learning. These works confirm the soundness of the hybrid ARIMA and LSTM framework as an effective instrument to improve the accuracy and reliability of the stock-market forecasts.

2.2. Sentiment analysis, and hybrid modeling techniques

The use of the sentiment analysis in forecasting a stock-market has taken a more significant position in the recent scholarship. According to Chen et al. (2022), there are convincing results that the sentiment indicators based on financial news have a high temporal correlation with the market mood and investor perception which are some of the most telling antecedents of the forthcoming price changes and supported these findings with additional evidence of the short-long term predictive model diversities where the performance of the short-term predictive models has once again been established as higher since they include the historical prices and technical indicators as structured variables and the text sentiment as unstructured data. Consequently, the sentiment analysis proved to be an alternative to the conventional quantitative instruments, one of the most vital modalities used to forecast the financial markets where emotion-based data was a valuable supplement to the currently applied modelling (D'Angelo & Palmieri, 2021). Given that capital markets are highly dependent on the mood of the investors, the mood of the press and the wider community, which is always reflected by the phrasing in any given news story, the analyst report or social media update, the tone of the sentiment is extracted by natural language processing and artificial-intelligence-based machine-learning algorithms simply scoring of optimism. The empirical findings consistently demonstrate that a positive mood is linked to a rise in the price whereas a negative mood is an indication of a price decline in the near future (Brown & Cliff, 2004). Forecasting models that incorporate sentiment measures and historical price and volume information have the effect of capturing a broader swath of market forces. When the hybrid model is used, sentiment scores will be another input that enhance the model performance, at least detecting any extreme price changes or high volatility that cannot be easily attributed to the previous prices. This kind of analysis is quite relevant to the emerging economies like India's, where investor psychology plays a bigger part; the sentiment analysis would offer hands-on knowledge, a sharper prediction and would enhance real-time decisions of the traders as well as the analysts Deng et al. (2013). In the context of emerging markets such as India, where indices like Nifty 50 and Sensex are highly responsive to both macroeconomic events and investor sentiment, the inclusion of sentiment features is particularly valuable. These hybrid models enable a multidimensional view, capturing not only the numerical patterns but also the qualitative drivers of market behavior.

3. Sentiment analysis of financial advisor data sets

The present study is based on the web-crawling method to assemble a set of financial documents that are scraped out of the social-media pages of the certified financial advisors. The data body of the problem provides a profound case study of linguistic and topical factors that constitute the foundation of the contemporary communication in the field of finance. Web crawling collects unstructured financial text, which FinBERT then analyzes to quantify sentiment, creating an integrated pipeline for the AI-based financial analysis and informed decision-making.

3.1. Web crawling techniques

The use of manual data acquisition procedures, such as regular forms and direct data entry, is still prevalent as the procedures are straightforward, inexpensive, and have little requirement of set up. Nevertheless, these techniques have a tendency to be characterized by duplication or inconsistency and can also be met with copyright issues. They also rely mostly on the level of skills of the involved workforces. Also, it is common that the manual processes face challenges when facing complicated information is implied, because much of the input will have to be laboriously messaged into pre-set categories, a problematic time-consuming process Hu et al. (2021). Automated systems are more efficient and scalable by contrast. An example can be the web crawling technique, which is completely automated and reduces the amount of labor since bulk data can be gathered with little human involvement, making the process much faster. Ontology-based, semi-automated methods are an incremental step and add a greater degree of structure and semantic clarity. These have assisted in automatically structuring unstructured information which

provides a compromise in between the manual and fully automated systems. However, the automated and semi-automated procedures are likewise restricted. The ontology- based systems should be monitored and reviewed continually to be able to evolve according to the changes in data structures, syntax, and contents. Ontologies could require new development over time as the sets change. Also, such systems are not always completely flexible and can include partial automation with the possible loss in accuracy. Web crawlers can be divided into four classes, depending on their goals and limitations: general-purpose, focused, incremental and real-time (or stream) crawlers. The general-purpose crawlers like Googlebot are configured to crawl the whole web by randomly following hyperlinks. Focused crawlers use one or more heuristics or classifiers to limit traversal to pages relevant to a set of topic or keywords and increase the relevance and efficiency. Incremental crawlers crawl efficiently to update the data crawled before and hence the elimination of redundancy can be done by updating the previously crawled data. Streaming crawlers or Real-Time crawlers can be used to scrape data at high frequencies, characteristics especially useful in cases where one wants to track breaking-news. Web crawling requires a careful design to manage challenges such as robots.txt compliance, duplicate content detection, polite crawling (delay between requests), and handling JavaScript-heavy or dynamic web pages. Recent advancements incorporate machine learning and natural language processing into crawling pipelines to enhance the quality of data selection and content extraction. Table 2 explains the details of the web crawling techniques based on own research.

Table 2. Web crawling techniques

Technique	**Objective**	**Advantages**	**Limitations**	**Use Case**
Incremental Crawling	Update previously crawled data	Data freshness; avoids re-downloading	Complexity in change detection	News/article updates; e-commerce price checks
Real-Time Crawling	Live data extraction (streaming)	Low latency; up-to-date information	Resource-intensive; difficult rate-limiting	Social media monitoring; stock sentiment
Adaptive Crawling	Learn and adjust based on crawl results	Intelligent navigation; improves over time	Requires initial training; overhead	Web mining with feedback learning
API-Based Crawling	Retrieve data via public/ private APIs	Structured data; fewer errors	Limited access; rate limits; no full-page context	Twitter API, Financial APIs (Yahoo, NSE)

The Named Entity Recognition (NER) filters only advisor quotes or opinions. Integrate Google News API or RSS feeds to automate real-time crawling. Apply Topic Modeling (LDA) to group advisor sentiments by financial sectors. Store results in time-series format for further use in prediction models. The algorithm in question implements the text (preprocessing), sentiment (classification), and web crawling algorithms, in order to create a structured time stamped sentiment dataset of the market, using qualitative, open sources (blogs, forums, financial commentary websites, etc.). The final product is to generate data that will be utilizable in carrying out empirical studies, predictive analytics, or real-time support in the financial arena. The major component of the data acquisition module is the looping-based web crawling mechanism. Each cycle produces a query string created by adding key words to a base URL (e.g. via Google Search API, search queries on specific sites or via RSS). Here are the advisor blogs, and stock outlook, market prediction, and interest rates are the keywords that are used in the search. After the posts have been downloaded, the FinBERT sentiment model is implemented. The model predicts a positive on the text; inflation peaking in Q3 is positive for equities with a score of 0.76, and a date of 2025-07-10. Positive trends on a given day may be revealed by aggregated daily averages over all posts as a result of rosy expectations.

The last procedure involves consulting a financial sentiment lexicon that determines the keywords and sentiment scores. To every word of the cleaned text, the lexicon identifies its value

in the summation of the sentiment reading. Figure 2 - Flowchart illustrating the own web crawling process for financial advisor data.

- A score is retrieved from the lexicon.
- A cumulative or average sentiment score is computed.
- The text is labeled based on score thresholds (e.g., score > 0.5 → Positive).

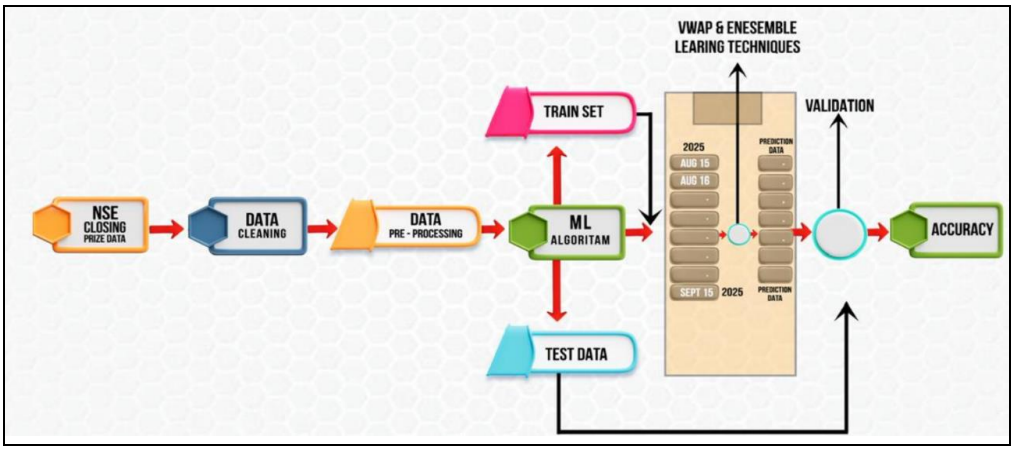

Figure 2. Flow diagram for web crawling of financial advisor

The framework regards the extraction of a financial sentiment from web-based advisory sources. The structured web crawling accompanied by the NLP-based text preprocessing and sentiment classification outputs a stable set of sentiment-rich data which can be used in the in-depth financial analysis. The modular design makes the framework highly adaptable, that is, it can support many kinds of applications, languages, and types of analytical details, which in the sentiment-aware financial research systems renders the framework a central inventory. Frameworks: The sentiment-analysis pipeline output was a labeled set of advisors generated commentary, and text in processed and cleaned form, with sentiment labels or scores assigned. All of the advisory statements were classified as either positive, neutral, or negative by synthesizing financial sentiment lexicons, and machine-learning methods. The example in stating a comment on the market volatility which indicated a score of -0.45 is regarded as negative but one which was cautious but stable was labeled as neutral. Such sentiment scores may be standardized between -1 and +1, indicating not only the polarity (whether it is a strong opinion in the market, or not) but also its strength. The temporal aggregation provides a cumulative perspective on the sentiment patterns by summarizing data collected over time. The qualitative achievements of sentiment stability in perceptions of market translate to pertinent explanations of how the financial advisors approach the market. These understandings can be directly incorporated into predictive models that seek to forecast the stock market as is the case in this study proposal.

4. Experiment design

The securities market in India mainly works due to the existence of two mega exchanges namely the National Stock Exchange (NSE) and the Bombay Stock Exchange (BSE). The BSE Sensex and NSE Nifty 50 are their main indices which are commonly known as benchmark indices used in the measurement of the overall performance of the Indian equity market. In this work, both indices will be employed in order to track and predict the market trends. A group of sentiment indicators lies at the core of the forecasting model using the activity of individual, institutional and foreign investors. This paper presents a hybrid forecasting system based on linear and nonlinear models, the ARIMA and LSTM) nets in order to increase the prediction accuracy. Cumulatively, the models are supposed to give a better prediction of the trends of Sensex and Nifty indices. Table 3 presents the raw crawled advisor content along with the corresponding sentiment scores, generated through our experimental framework.

Table 3. Raw Crawled Advisor Content and Sentiment Score

Date	Extracted Advisor Text	Cleaned Text	Sentiment Label	Sentiment Score
25-06-2025	"According to Rakesh Mehta, the market is likely to remain volatile in the near term due to global cues..."	"Market likely remains volatile near term due global cues"	Negative	-0.45
24-05-2025	"Sonia Kapoor suggests investors remain cautious and focus on defensive sectors amid ongoing rate hikes."	"Investors remain cautious focus defensive sectors ongoing rate hikes"	Neutral	0.10

Sensitivity measures the model's ability to correctly detect positive cases, while specificity evaluates its accuracy in identifying negatives. Positive predictive value (PPV) indicates the likelihood that the predicted positives are true positives. The F1-score balances precision (PPV) and sensitivity, providing a single measure of a model's reliability, especially in imbalanced datasets. Collectively, these metrics offer a comprehensive assessment of the classification performance beyond the overall accuracy. The algorithm does post-processing processes after classification and the foremost such process is accumulation of sentiment by date so that researchers can calculate the daily average sentiment scores and follow the market sentiment over the time trend and through temporal planes.

4.1. Label generation

Binary class labels are used in order to forecast future trends. $L_{ij} \in \{0, 1\}$ are created on the basis of the directional rise of the values of stock indices. A value of 1 is provided when the price goes up relative to the last time step otherwise 0. This constitutes the supervised learning ground truth. The opinion of the various investor types is represented by using text data, and the sentiment indicators are determined in the text by means of textual data. Net capital flow data Y_i is collected, representing the buying or selling pressure in the market. These features capture the market psychology and behavioral trends. The full dataset is constructed by combining price data X_i, sentiment indicators S_{ij}, and capital flow values Y_i. The background of time-series modeling and the activity of classification are implemented on the present multimodal data set. The raw text was subject after the gathering stage to preprocessing preliminaries including normalization and cleaning exercise of conventional corpus level undertakings. Step of preprocessing is a very important step towards ensuring the effectiveness and accuracy of the sentiment classification. Preprocessing: The first step undertaken by the system during the preprocessing step is to parse-out embedded scripts and styles, Web page HTML tags, leaving no processing overhead spent parsing meaningless tags. Then, alien characters, smileys, and excessive whitespaces undergo a strain removing to enhance robustness of the labeling process. A process is that the strings retrieved are tokenized into individually detected words, which are normalized by all being in lower cases. At this intermediate step, optional lexical compression steps may be undertaken, were, in the present example, the words can be compressed, by using stemming or lemmatization schemes, to their semantically equivalent roots. Collectively, the measures normalize the textual inputs and render the homogeneity of the corpus to derive a uniform sentiment analysis. The central part of the system is the Sentiment classification module that attaches a sentiment label to each of the preprocessed documents. Two significant methodologies become possible. The dictionary approach entails the utilization of a set financial lexicon, words that hold sentiments. In the lexicon, the algorithm returns frequency and the polarity data in each token on every document and each document has a total sentiment score by adding the scores of all the detected polarity to give the score of the entire document. An example of this can be seen whereby occurrence of the word

bullish like, recovery, or buy would render a positive contribution and repeated instances like recession, crash, or decline can assume a negative weight. Otherwise, the machine-learning approach implies utilizing already trained models, which are supposed to identify contextual innuendoes as a typical example of a financial text. The text that passed through the preprocessor is then passed through the model chosen and the output only gives one label (positive, negative or neutral) and this is accompanied with the confidence level in Table 4. Comparative metrics of different models.

Table 4. Evaluation Metrics for Various Models

Model	Sensitivity	Specificity	Positive Predictive value
Logistic Regression	0.950	0.947	86.11%
Decision Tree	1.00	0.956	75.36%
Naïve Bayes	1.00	0.992	92.56%

In terms of classification both results, regardless of the strategy, contain: the original text, the sentiment score or label and the time stamp in the common format.

4.2. Pseudocode ARIMA and LSTM for stock index forecasting

The pseudocode for ARIMA in stock index forecasting outlines the step-by-step process of preparing the data, identifying the model parameters, and fitting the time series to capture its linear patterns.

```
Pre: Price index data: Xi for i ∈ {1, 2, ..., n}
   Pre: Index types: j ∈ {Sensex, Nifty}
   Pre: Sentiment indicators: Sij from 3 investor types
   Pre: Net capital flow data: Yi
   Post : Predicted trend: Lij
   Post : Final price prediction: X̂i
   Post : Evaluation metrics: HR, HR1, HR2, AR, ARa, MDD, Sharpe Ratio, MDD%
1. BEGIN
2. Label Generation:
3. FOR each i in {1, 2, ..., n}:
4. FOR each j in {Sensex, Nifty}:
5. IF price at time i+1 > price at time i:
6. Lij ← 1
7. ELSE:
8. Lij ← 0
9. Collect Sentiment Data:
10. FOR each i in {1, 2, ..., n}:
11. Sij ← sentiment scores from 3 investor groups
12. Yi ← net capital inflow/outflow
13. Construct Dataset:
14. D ← {(Xi, Yi, Sij, Lij) for all i, j}
15. Create Sliding Window Data Splits:
16. Divide D into training and testing sets:
17. Dtrain1, Dtrain2, Dtrain3, Dtrain4
18. Dtest1, Dtest2, Dtest3, Dtest4
19. Feature Selection:
20. Apply correlation filtering on sentiment features
21. Apply PCA to reduce dimensionality
22. Retain key features: Sij_selected
23. ARIMA Decomposition:
24. FOR each time series Xi:
25. Fit ARIMA model → X̂i_ARIMA
26. Calculate residuals: Ri ← Xi - X̂i_ARIMA
27. ARIMA Model Training:
28. Train ARIMA on historical Xi
29. Output: residuals Ri
30. Prepare Input for LSTM:
31. Input_LSTM ← concatenate(Ri, Sij_selected, Yi)
32. Train LSTM Model:
33. LSTM_input ← Input_LSTM
34. LSTM_output ← predict future trend L̂ij
35. Train LSTM using LSTM_input → L̂ij
36. Hyperparameter Tuning:
37. Use Grid Search or Random Search on LSTM
38. Optimize: learning rate, batch size, epochs, etc.
39. Final Prediction:
40. FOR each i:
41. R̂i_LSTM ← LSTM.predict(Ri, Sij_selected, Yi)
42. X̂i ← X̂i_ARIMA + R̂i_LSTM
43. Evaluate Prediction Accuracy:
44. Compute:
45. HR ← hit rate
46. HR1 ← positive trend accuracy
47. HR2 ← negative trend accuracy
48. Evaluate Financial Performance:
49. Compute:
50. AR ← Annualized Return
51. ARa ← Average Return
52. MDD ← Maximum Drawdown
53. Sharpe Ratio ← risk-adjusted return
54. MDD% ← (MDD / peak value) * 100
55. END
```

4.3. ARIMA modeling and residual extraction

The ARIMA model is trained on the historical price index series X_i to capture linear temporal pat- terns. The difference between the predicted ARIMA output $X\hat{}$ARIMA and the actual value produces residuals R_i, which are assumed to contain nonlinear components not captured by ARIMA.

$$X_i = \hat{X}_i^{ARIMA} + R_i$$

The extracted ARIMA residuals R_i, along with the selected sentiment features and capital flow values, are used as input to train a Long Short-Term Memory (LSTM) model. The LSTM is designed to capture nonlinear dependencies and sequential patterns. The output is a predicted trend label $L\hat{}ij$, representing the future market direction. The model performance is improved through hyperparameter tuning using either grid search or random search strategies. Parameters such as learning rate, batch size, and number of LSTM units are optimization on validation sets. The final predicted stock index value $X\hat{}i$ is obtained by adding the ARIMA forecast and the LSTM-predicted residual:

$$\hat{X}_i = \hat{X}_i^{ARIMA} + \hat{R}^{LSTM}$$

This kind of mixed modeling catches efficiently both the linear and nonlinear trends observed in the financial time-series data. Such classification measures are used to measure the performance of the model: Hit Rate (HR), Positive Accuracy, and Negative Accuracy. Moreover, the assessment of the financial performance is analyzed based on such key indicators as Annualized Return (AR), Average Return, Maximum Drawdown (MDD) as well as Sharpe Ratio. Such measures not only make the model useful in predicting the accuracy of the model but also test the model usefulness in the real world of the investment decision making process.

$$\text{F1} - \text{Score} = \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

The hybrid model based on combining the ARIMA and LSTM models combines the statistical and deep learning methods in order to make robust predictions of the time series especially in the stock price prediction. Valuation of the sentiment indicators and capital flow information increases the contextual intelligence about market behavior and, therefore, the predictions made using the model are not only accurate but executable as well.

$$\phi(B)(1-B)^d X_t = \theta(B)\varepsilon_t, with\ \phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p, \theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$$

Figure 3 shows the correlation between the real values of the stock-index and the forecasts made by ARIMA: LSTM model. The high matching close visual effect of the two curves shows that the model has been effective in its ability to track the movements that are happening in the market, providing a continuous linear change and a nonlinear change. Such an outcome is in support of the hypothesis that the combination of ARIMA and LSTM, in the trend estimation and residual modeling, respectively, would improve the overall predictive power in volatile market condition like the Indian stock indices.

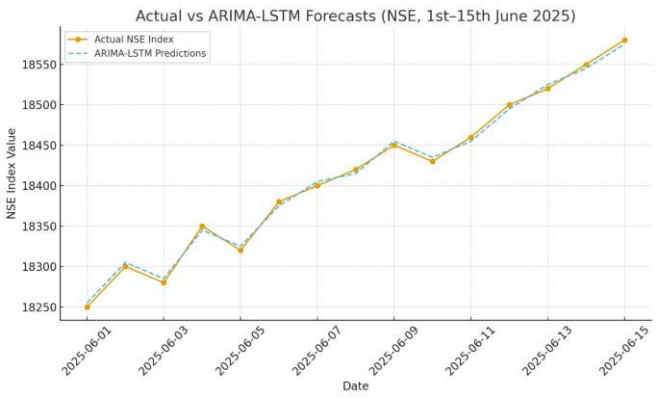

Figure 3. Actual Index Values vs. ARIMA-LSTM Combined Predictions

The validation process for the proposed framework performance is assessed through widely accepted metrics, including confusion Matrix, sensitivity (detecting positives correctly), specificity (detecting negatives correctly), positive predictive value and F1-score (balance precision & recall). These results clearly show that the hybrid forecasting model is more accurate than its constitutive counterparts in predicting. The combination of the explanatory technique of the autoregressive integrated moving-average (ARIMA) to capture the linear elements and the ability to model nonlinear tendencies of residuals in the hybrid framework results in a smaller magnitude of forecasting errors, thus confirming its effectiveness in consideration of the stock index dynamics. The main asset of the ARIMA and LSTM hybrid model is that it has a high predictive accuracy. The majority of the errors in predicting the index value are within the acceptable threshold range of 2 to 30 points which is relatively good in forecasting the stock index values. This proves the high capability of the model to reflect the overall trends in the market and the scatter in the market.

Another highlight of such hybrid scheme was the inclusion of LSTM network to improve forecast performance by learning and adjusting the nonlinearity residuals that could not be completely fitted by the ARIMA component. This especially comes in handy on days when the markets are prone to sudden or unexpected movements, on which the conventional statistical models tend to perform inadequately. It is the memory-based structure of the LSTM that enables it to react to sudden shifts and complicated time relevancies in prices making the general structure more robust and adaptable against the abrupt changes in the market.

This makes it good to use in short term trading activity, volatility tracking, as well as risk management decision making where the reliability over a very short time frame would be critical. Figure 3 will show the close prices of Nifty 50 index between 1 st to 15 th June 2025. It represents the work of the hybrid model ARIMA and LSTM applied to predicting the Nifty 50 indexes value during the same period. The X-axis corresponds to daily dates and the Y-axis corresponds to the Nifty 50 index values which are represented as having a range of 22,500 to 23,500 (simulated).

Two time-series lines are plotted for comparison:

- The actual Nifty 50 closing prices are shown as a solid blue line with circular markers;
- The predicted values generated by the ARIMA and LSTM hybrid model are represented by an orange dashed line with square markers.

The plot comparison given shows that the financial forecasted value almost matches to the realistic trend with a slight variation. This correspondence denotes that the model can capture nonlinear and linear patterns of the index movement. The prediction is further enhanced by the residual learning capacity of the LSTM element through the ability to adapt to short term changes too infrequently modelled by the ARIMA element alone. The visual similarity of the two lines is an indication and confirmation of the strength and the predictive power of the suggested hybrid approach over the period in question.

5. Conclusions and future work

The research paper is an investigation on the relevance of combining the quantitative market data with the qualitative sentiment analysis in easing stock index forecasting in the NSE context of a financial scenario. The data is analyzed on the basis of NIFTY 50 indices, and a non-linear model incorporating the advantages of deeper learning with the statistical paradigm; the ARIMA -LSTM model is used to capture nonlinear residual patterns. The sentiment of advisor with respect to a domain-based sentiment dictionary is generated by calibrating dictionary to discourse in the Indian financial sector and is deemed to be a necessary part in improving the predictive performance. The empirical evidence clearly indicates that a significant decrease in the error of forecasting can be achieved using sentiment factors, including regional news articles, corporate disclosures, and financial posts on the social media. In general, the findings demonstrate the pivotal nature of the investor sentiment when it comes to the determination of the market dynamics. They also reveal the fact that the models using a mixture of traditional sequence analysis methods and sentiment-conscious deep learning turn out to be better especially in the case of the new market circumstances

as the National Stock Exchange (NSE) index portrays. This knowledge of complex market action is more precise and helps better understanding since it was not based solely on the known law of supply.

REFERENCES

Bao, S., Yue, J. & Rao, Y. (2017) A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PLoS One*. 12(7), 1-24. e0180944. doi:10.1371/journal.pone.0180944.

Brown, G.W. & Cliff, M.T. (2004) Investor sentiment and the near-term stock market. *Journal of Empirical Finance.* 11(1), 1–27. doi:10.1016/j.jempfin.2002.12.001.

Chen, M., Guo, Z., Abbass, K. & Huang, W. (2022) Analysis of the impact of investor sentiment on stock price using the latent dirichlet allocation topic model. *Frontiers in Environmental Science*. 10:1068398, 1-16. doi:10.3389/fenvs.2022.1068398.

Chen, T. & Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), 13–17 August 2016, San Francisco, CA, USA*. ACM. pp. 785–794. doi:10.1145/2939672.2939785.

Chen, W., Zhang, H., Mehlawat, M. K. & Jia, L. (2021) Mean–Variance Portfolio Optimization Using Machine Learning-Based Stock Price Prediction. *Applied Soft Computing*. 100, 106943. doi:10.1016/j.asoc.2020.106943.

D'Angelo, G. & Palmieri, F. (2021) GGA: A Modified Genetic Algorithm with Gradient-Based Local Search for Solving Constrained Optimization Problems. *Information Sciences*. 547, 136–162. doi:10.1016/j.ins.2020.08.040.

Deng, S., Xiao, C., Zhu, Y., Peng, J., Li, J., Liu, Z. (2023) High-Frequency Direction Forecasting and Simulation Trading of the Crude Oil Futures Using Ichimoku Kinko-Hyo and Fuzzy Rough Set. *Expert Systems with Applications*. 219, 119326. doi:10.1016/j.eswa.2022.119326.

Fischer, T. & Krauss, C. (2018) Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*. 270(2), 654–669. doi:10.1016/j.ejor.2017.11.054.

Gite, S. et al. (2021) Explainable stock prices prediction from financial news articles using sentiment analysis. *PeerJ Computer Science*. 7, e613. https://doi.org/10.7717/peerj-cs.340.

Halder, S. (2022) FinBERT-LSTM: Deep Learning based stock price prediction using News Sentiment Analysis. To be published in *Statistical Finance*. [Preprint] https://doi.org/10.48550/arXiv.2211.07392.

Hu, Y., Shao, L., La, L. & Hua, H. (2021) Using Investor and News Sentiment in Tourism Stock Price Prediction Based on Xgboost Model. *In Proceedings of the 2021 IEEE/ACIS 6th International Conference on Big Data, Cloud Computing and Data Science (BCD), 13-15 September 2021, Zhuhai, China*. IEEE. pp. 20–24. doi:10.1109/bcd51206.2021.9581619.

Jiang, C. (2024) Comparative Analysis of ARIMA and Deep Learning Models for Time Series Prediction. *In Proceedings of the 2nd International Conference on Data Analysis and Machine Learning DAML, Kuala Lumpur, Malaysia.* SciTePress. 1, 306-310. doi:10.5220/0013516200004619.

Kasture, P. & Shirsath, K. (2024) Enhancing Stock Market Prediction: A Hybrid RNN-LSTM Framework with Sentiment Analysis. *Indian Journal of Science and Technology*. 17(18), 1880-1888. doi:10.17485/IJST/v17i18.466.

Lim, S., Park, J., Kim, S., Wi, H., Lim, H., Jeon, J., Choi, J. & Park, N. (2023) Long-term Time Series Forecasting based on Decomposition and Neural Ordinary Differential Equations. To be published in *Machine Learning*. [Preprint] https://doi.org/10.48550/arXiv.2311.04522.

Livieris, D., Pintelas, E. & Pintelas, P. (2020) A CNN–LSTM model for gold price time-series forecasting. *Neural Computing and Applications*. 32, 17351–17360. doi:10.1007/s00521-020-04867.

Nelson, J., Pereira, A. & de Oliveira, M. (2016) Stock Market's Price Movement Prediction with LSTM Neural Networks and Technical Indicators. *In Proceedings of the 31st Brazilian Symposium on Databases (SBBD), 4–7 October 2016, Salvador, Brazil*. SBC. pp. 1–6.

Siami-Namini, M., Tavakoli, N. & Siami Namin, A. (2018) A Comparison of ARIMA and LSTM in Forecasting Time Series. *In Proceedings of the 17th IEEE International Conference on Machine Learn- ing and Applications (ICMLA), 17-20 December 2018, Orlando, FL, USA*. IEEE. pp. 1394–1401. doi:10.1109/ICMLA.2018.00227.

Xiao, R., Feng, Y., Yan, L. & Ma, Y. (2022) Predict stock prices with ARIMA and LSTM. To be published in *Statistical Finance*. [Preprint] https://doi.org/10.48550/arXiv.2209.02407.

Zhang, G.P. (2003) Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*. 50, 159–175. doi:10.1016/S0925-2312(01)00702-0.

Mani PADMANABHAN is an Associate Professor Senior, Faculty of Computer Applications, School of Social Sciences and Languages (SSL), Vellore Institute of Technology (VIT), Vellore, India (Profile: [https://orcid.org/0000-0002-2402-7684]). He has over 14 years of teaching and research experience. His academic qualifications include a Ph.D. in Software Testing, an M.Tech in Computer Science and Engineering, and an M.S. in Software Engineering (5-year Integrated Program). His current research areas focus on Applications of AI in Finance, Machine Learning in Finance, Design Thinking in Media and Entertainment,Automatic Software Testing, and Software Project Management.