Sales Analysis and Sale Forecasting by Integrating Sentiment Analysis Using Valence Aware Dictionary and Sentiment Reasoner (VADER)

Sanjit Kumar Dash [1], Abhinash Pritiraj [1], Dwarika Prasad Padhee [1], Suleman Alnatheer [2,*], Mohammed Altaf Ahmed [2], and Abdullah Alsir Mohamed [2]

[1] Department of Information Technology, Odisha University of Technology and Research, Bhubaneswar, Odisha, India
[2] Department of Computer Engineering, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia
Email: skdash@outr.ac.in (S.K.D.); abhinash.pritiraj@gmail.com (A.P.); dwarikaprasadpadhee03@gmail.com (D.P.P.); s.alnatheer@psau.edu.sa (S.A.); m.altaf@psau.edu.sa (M.A.A.); a.mhamed@psau.edu.sa (A.A.M.)
*Corresponding author

Abstract*—Accurate retail sales forecasting requires models that can handle linear seasonality, non-linear dynamics, and consumer sentiment at the same time. Seasonal AutoRegressive Integrated Moving Average (SARIMA) captures periodic trends well but cannot learn non-linear residual patterns. Long Short-Term Memory (LSTM) networks have the opposite problem: they lack any mechanism for incorporating sentiment-based exogenous inputs. Prior hybrid SARIMA-LSTM work exists, but none of it has tested the systematic integration of Valence Aware Dictionary and Sentiment Reasoner (VADER) sentiment polarity, as a statistically validated exogenous variable, within a bidirectional residual-correction framework on retail transactional data. This study addresses this limitation through the development of a hybrid forecasting framework. The proposed framework has three components: Seasonal AutoRegressive Integrated Moving Average with eXogenous variables (SARIMAX) captures linear trends, seasonality, and sentiment-driven influences. A Bidirectional LSTM (Bi-LSTM) network trained on SARIMAX residuals models the non-linear temporal dependencies that SARIMAX misses. Prophet then handles additional trend-shift correction. VADER scores derived from customer product reviews serve as the exogenous regressors. The framework is evaluated on the Super Store Sales dataset covering 2011 to 2014, which contains 51,290 transactions. Stationarity was confirmed via the Augmented Dickey-Fuller (ADF) test (ADF statistic: −4.8676, $p < 0.0001$). A formal ablation study isolates each component's contribution. The full hybrid model achieves a Root Mean Squared Error (RMSE) of 61,588.96, Mean Absolute Error (MAE) of 49,270, and Mean Absolute Percentage Error (MAPE) of 9.82%, outperforming SARIMA (RMSE: 75,936.21), Prophet (RMSE: 74,365), and Neural-Basis Expansion Analysis for Time Series (N-BEATS) (RMSE: 65,300); the sentiment-augmented SARIMAX baseline reaches 70,544.74. The Diebold-Mariano test confirms statistical superiority over both SARIMAX ($p < 0.01$) and Prophet ($p < 0.05$). Sentiment integration alone cuts RMSE by 7.57%; Bi-LSTM residual modelling adds a further 14.54%, with practical implications for inventory management and dynamic pricing.

***Keywords*—sales forecasting, Seasonal AutoRegressive Integrated Moving Average (SARIMA), Seasonal AutoRegressive Integrated Moving Average with eXogenous variables (SARIMAX), bidirectional long short-term memory, Valence Aware Dictionary and Sentiment Reasoner (VADER) sentiment analysis, hybrid forecasting, time series, residual modelling, Diebold-Mariano test**

I. Introduction

Sales forecasting sits at the core of operations management, shaping decisions about inventory levels, hiring, pricing, and marketing spend based on expected demand. Forecasting inaccuracies can result in significant operational and financial consequences, including excess inventory costs and lost sales opportunities due to stockouts. As retail environments become increasingly complex, with faster product cycles, less predictable consumers, and shopping behavior split across online and physical channels, companies need forecasting models that are accurate, readable, and sensitive to signals beyond raw transaction data.

Sales forecasting methods have changed substantially over the decades. Early approaches relied on classical statistical models such as Exponential Smoothing and AutoRegressive (AR), Moving Average (MA), and AutoRegressive Integrated Moving Average (ARIMA), developed by Box and Jenkins, handled non-stationary

Manuscript received April 10, 2026; revised May 9, 2026; accepted June 18, 2026; published November 10, 2026.

time series and remain a standard tool for single-variable forecasting. Seasonal AutoRegressive Integrated Moving Average (SARIMA) and Seasonal AutoRegressive Integrated Moving Average with eXogenous variables (SARIMAX) extended this to seasonal effects and external variables; Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) addressed volatility; Vector Autoregression (VAR) handled multiple interdependent time series. These linear models share a common limitation: they struggle with structural breaks, asymmetric responses, and patterns that are not linear. Machine learning approaches have addressed some of these limitations, particularly Support Vector Machine (SVM), Gradient Boosting, and ensemble methods in high-complexity data settings. Deep learning methods have further improved forecasting performance by capturing complex temporal dependencies. Long Short-Term Memory (LSTM) networks, including bidirectional variants (Bidirectional Long Short-Term Memory (Bi-LSTM)), handle long-range dependencies in sales data well, and Transformers and Neural-Basis Expansion Analysis for Time Series (N-BEATS) represent the current leading approaches. Hybrid architectures that pair the interpretability of statistical models with the flexibility of neural networks, such as decomposing a forecast into a linear SARIMA component alongside a neural network for non-linear residuals, have shown consistent gains.

Customer sentiment provides valuable information regarding brand perception, product evaluation, and purchase intention that transaction data does not capture. Valence Aware Dictionary and Sentiment Reasoner (VADER) was built for short, informal texts like product reviews. Rather than depending on large volumes of labeled training data, it uses a rule-based system combining a curated lexicon with handling for negation and intensity modifiers, making it fast and easy to scale. Empirical work across finance, retail, and e-commerce shows that incorporating VADER sentiment scores improves forecast accuracy. However, a comprehensive evaluation of the incremental contribution of sentiment information remains limited.

Several gaps persist in the literature. Hybrid SARIMA-LSTM architectures have attracted attention; however, limited attention has been given to the role of bidirectional LSTMs in residual modelling following statistical decomposition. Sentiment variables appear in SARIMAX models fairly often, yet they are rarely accompanied by formal validation through statistical tests like the Diebold-Mariano to determine whether they provide statistically significant improvements in forecasting performance. Furthermore, most published studies rely on a single performance metric, creating ambiguity regarding the individual contribution of each model component to overall predictive performance.

This study addresses these three research gaps. It builds a unified three-part hybrid framework combining SARIMAX, Bi-LSTM, and Prophet, with VADER sentiment scores as an external input and Bi-LSTM applied specifically to non-linear residuals. A formal ablation study isolates the contribution of sentiment integration and Bi-LSTM modeling separately, and Diebold-Mariano tests confirm performance improvements against three baselines, namely SARIMAX, Prophet, and N-BEATS, on a retail dataset of 51,290 transactions. The paper concludes with practical guidance on inventory management and dynamic pricing, drawing on both the quantitative results and the patterns visible in sentiment trends.

II. Related Work

Liapis *et al*. [1] proposed sentiment analysis for forecasting multivariate financial time series. They used 27 algorithms applied in 22 different configurations, stressing that the inclusion of social media sentiment data markedly improves the accuracy of economic prediction, especially in the long term. This strongly supports the objectives of the research by offering some insights into useful methodologies for incorporating sentiment analysis into sales forecasting, especially in multivariate time-series contexts [2]. These findings highlight the potential of integrating sentiment analysis, deep learning, and real-time data monitoring for improving forecasting performance. Hong [3] asserted that Long Short-Term Memory (LSTM) models have recently been recognized as better than the classical ARIMA model in retail sales forecasting. The work further emphasized the importance of incorporating sentiment analysis to enhance predictive results, particularly in rapidly changing retail environments. The final results highlight the advantages of deep learning models that capture seasonality and trends in sales data, while improving forecasts through sentiment analysis [4].

Nsuh *et al.* [5] have completed the analysis of forecasting Walmart sales with the help of time-series analysis. Their work stressed the importance of combining traditional forecasting methods with real-time data monitoring to improve accuracy. The objective of integrating sentiment analysis with sales forecasting falls under this architectural approach. In this regard, Gustriansyah *et al.* [6] have proposed a sales forecasting model using machine learning techniques, thereby demonstrating the effectiveness of integrating multiple data sources to obtain accurate predictions. Similarly, Krishna *et al*. [7] have taken machine learning for sales forecasting at the retail level and emphasized the critical importance of capturing seasonal patterns and customer behavior. These studies highlight the need for advanced methodologies that utilize both structured and unstructured data to enhance forecast accuracy.

Cheriyan *et al.* [8] studied intelligent sales prediction using machine-learning techniques while underlining the importance of real-time monitoring of data. Collectively, these studies emphasize the importance of integrating sentiment analysis, deep learning, and real-time monitoring of data in forecasting sales. Lawrence and O'Connor [9] described the practical effectiveness of updating sales forecasts, thus providing insight into the challenges and opportunities of integrating dynamic data sources into predictive models. These findings further

emphasize the importance of adaptive forecasting frameworks capable of incorporating dynamic data sources.

Abiola *et al*. [10] performed sentiment analysis on COVID-19 tweets and also used VADER and TextBlob. This study further demonstrates the applicability of sentiment analysis as a means for dynamic forecasting environment applications, showing the utility of these tools in real-time data analysis. Annamalai *et al*. [11] performed sentiment analysis using VADER to predict shopping behavior in the airline industry. The results demonstrate the relevance of extracting insights from sentiment scores when investigating customer preferences, thus improving the predictive models. The findings demonstrate the broad applicability of sentiment analysis across multiple domains to improve forecasting models. Li [12] provided introductory information on the 2023 sentiment analysis roadmap as an outline of the emerging techniques and challenges. This work provides valuable context regarding emerging techniques and challenges in sentiment analysis.

A. Statistical and Econometric Models

ARIMA models have been regarded as a standard approach for univariate time series forecasting ever since Box and Jenkins formalized them. Dalrymple's [13] early survey in 1975 highlighted just how important methodology is in forecasting. For datasets exhibiting strong seasonal patterns, SARIMA models generally outperform conventional ARIMA models. Extensions like SARIMAX and ARIMAX enable the incorporation for exogenous variables, which facilitates the incorporation of sentiment-derived exogenous variables. However, a key limitation of these models is that they stick to linear, stationary relationships. They are unable to adequately capture complex nonlinear dependencies.

B. Machine Learning and Deep Learning Approaches

Machine learning models such as SVM and Gradient Boosting extend classical statistical approaches by capturing non-linear relationships that linear models cannot. LSTM networks address the vanishing gradient problem in standard Recurrent Neural Network (RNN) through gating mechanisms, enabling them to model sequences with long-range dependencies. Bi-LSTM processes data in both directions, broadening the contextual window, which is particularly useful when residuals still carry complex temporal patterns. Bandara *et al.* [14] confirmed LSTM's applicability in e-commerce forecasting. N-BEATS takes a fully neural approach with both backward and forward residual connections and functions as a deep learning benchmark in this study.

C. Transformers and Recent Forecasting Approaches

Recent advances in time-series forecasting have increasingly focused on transformer-based architectures due to their ability to capture complex temporal dependencies and effectively model long forecasting horizons. In 2021, Lim *et al.* [15] introduced the Temporal Fusion Transformer (TFT), which combines multi-head attention mechanisms with gating layers to not only deliver accurate predictions but also allow model interpretability, thus setting a new standard in performance across retail and energy forecasting datasets. Similarly, Zhou *et al.* [16] proposed the Informer model, which tackles the computational problems of old transformers by introducing a new ProbSparse attention mechanism, thus lowering the complexity to $O(L\log L)$ and making long-sequence forecasting an efficient task.

A complete review carried out by Wen *et al.* [17] shows that transformer-based models are again and again better than traditional statistical models like ARIMA on many benchmark datasets. In the field of retail demand forecasting, Wang *et al*. [18] report that combining statistical and transformer models can outperform pure deep learning approaches, especially when the training data is scarce. These findings provide a strong foundation for the proposed SARIMAX-Bi-LSTM hybrid framework as an alternative to the transformer-based methods. SARIMAX is employed to model seasonality, trends, and exogenous sentiment variables for interpretable modelling of seasonality, trends, and external factors, including sentiment information derived from VADER analysis. The Bi-LSTM component is utilized to capture nonlinear residual patterns; i.e., the deep neural component is capable of learning nonlinear residual patterns that the statistical model fails to explain. The integration of these components provides a suitable framework for medium-sized retail datasets where purely transformer-based architectures may exhibit higher computational requirements and increased susceptibility to overfitting. Hence, the hybrid approach not only maintains interpretability and accuracy but, at the same time it is computationally efficient.

D. Sentiment Analysis in Forecasting

Previous studies have demonstrated that sentiment information possesses significant predictive value. Agarwal [19] found that sentiment polarity aligns with stock price movements. Liapis *et al*. [1], reviewing a range of algorithms, found that social media sentiment improves long-term forecasting accuracy. Gooljar *et al*. [20] examined sentiment's role specifically in online shopping predictions, while Ghosh *et al.* [21] explored adjusting customer ratings with sentiment signals in sales forecasts. Elbagir and Yang [2] found VADER well-suited for short-form user-generated content. Across these studies, sentiment analysis and time-series modeling are treated as separate steps, with no unified statistical framework bringing the two together. Addressing this research gap constitutes a primary objective of the present study.

E. Recent Advances in AI-Driven Sales Forecasting and Analytics

Recently, several studies have focused on using artificial intelligence and predictive analytics for customer engagement, sales forecasting, and business decision-making support. Ristola [22] assessed the use of AI-enhanced customer engagement for helping increase sales, while also pointing out challenges related to the adoption of AI-centered approaches. From a forecasting

viewpoint, Stancu *et al.* [23] performed sales forecasting under seasonal conditions, and their results show the importance of accounting for seasonal variations to guide sustainable business growth. These papers inspired the integration of the use of data-driven analysis and sales patterns over different time periods in our framework design.

The accuracy of the forecast might also be enhanced through collaboration and the use of machine learning based approaches. McCarthy and Golicic [24] analyzed collaborative forecasting in supply chain operations and demonstrated how joint forecasting practices can significantly enhance supply chain efficiency. Lin *et al*. [25] looked into revenue forecasting using fuzzification and least-squares support vector regression techniques, providing an illustrative example of the potential of machine learning techniques in the realm of volatile revenue data. Grizzle and Klay [26], in their earlier piece of work, compared multiple ways of forecasting sales-tax revenue and stressed the dependency of forecasting results not only on the data characteristics, but as well on the technique used; thus, forecasting methods need to be chosen very carefully. The outcome of these pieces of work serves as a reference to identify the suitable methods for forecasting from the sales data and the revenue data considered here.

Besides numbers, the customer or stakeholder reaction in the form of words can also be an important clue to interpret. Telmo *et al*. [27] used Valence Aware Dictionary and sEntiment Reasoner (VADER) for the sentiment analysis of student feedback. Their paper illustrates how lexicon-based sentiment analysis can be used to determine the polarity of free-text feedback. Youvan [28] has been discussing VADER-aided Sentiment Analysis from a theoretical and practical perspective. Both these contributions support the idea of the use of sentiment features derived from the text in the proposed system, where customer feedback in the textual form is interpreted along with hard-core numerical data.

Applications in the areas of deep learning are not limited to only healthcare; medicine can make great use of it as well, such as forecasting sales, which is done in the areas of the retailer. Kaneko and Yada [29] have developed a forecasting system using deep neural network models that can predict the sales of a retail store. This, in a way, shows that the model can generate sales predictions based on the learning of nonlinear structures through the training process. It would be interesting to follow this lead in the case of our framework since this part deals with prediction, and we will want to find out if a machine-learning-based prediction can actually work for a situation where relationships are complex but a conventional method may not be adequate.

F. Research Gap Synthesis

Table I summarizes key prior studies across six categories, and three consistent gaps emerge. Bidirectional LSTM is rarely used for residual adjustment in hybrid SARIMA-LSTM models. The incremental contribution of sentiment features within these hybrid architectures has not been formally quantified, as nearly no study runs an ablation against a non-sentiment baseline. Even when accuracy improves, papers frequently omit statistical significance testing or rely on a single evaluation metric without verifying econometric validity. The proposed framework is specifically designed to address these three limitations.

TABLE I. COMPARATIVE SUMMARY OF RELATED WORKS

Study	Data	Primary Model	Input Variables	Benchmarks	Key Finding	Limitation
Ghosh *et al*. [21]	Reviews + sales	Sentiment-adjusted	Historical + sentiment	ARIMA	Sentiment fine-tuning improves reliability	No deep learning; no ablation
Liapis *et al*. [1]	Financial Transaction Services (TS)	27 ML algorithms	Historical + social media	Multi-algo	Sentiment improves long-term accuracy	Finance domain; no LSTM hybrid
Gustriansyah *et al*. [6]	Retail sales	ML ensemble	Multi-source data	ARIMA, SVM	Multi-source integration improves accuracy	No sentiment; no DM test
Nsuh *et al*. [5]	Walmart sales	Time series	Historical + real-time	Classical	Real-time monitoring improves accuracy	No deep learning; no sentiment ablation
Our work	**Retail sales + reviews**	**SARIMAX + Bi-LSTM + Prophet**	**Historical + VADER (exogenous)**	**SARIMAX, Prophet, N-BEATS**	**7.57% RMSE gain from sentiment; 14.54% from Bi-LSTM**	**Single retail dataset; future: cross-domain**

The literature reviewed here reflects genuine advances in sales forecasting through combinations of sentiment analysis, deep learning, and real-time data. Still, certain gaps persist. Traditional models like ARIMA struggle to capture non-linear trends and sudden market shifts, and LSTM models, while more flexible, lose predictive accuracy without contextual sentiment input. The proposed framework integrates sentiment analysis and deep learning into a single pipeline rather than running them as separate processes. Unlike earlier work that treated sentiment and time-series forecasting as independent tasks, this approach combines them directly, aiming to improve real-time responsiveness, improve the utilization of sentiment-derived information, and track shifting market conditions. The framework is also tested across both seasonal and non-seasonal retail datasets to confirm it generalizes across different environments, producing a forecasting method that incorporates sentiment-derived signals into established statistical models.

III. SALES FORECASTING FRAMEWORK

A. System Flow Diagram

The forecasting and selling-planning model is described in Fig. 1, which integrates sentiment-driven insights with

classical time-series analysis to enhance predictive accuracy. The model starts with Sales Analysis, where old data is subjected to Exploratory Data Analysis (EDA) and Data Pre-processing, maintenance of data consistency, and removal of noise. Data Visualization then follows to expose all the patterns across user behavior and product performance; Customer Segmentation accounts for grouping buyers based on purchasing attributes. Seasonal Variation, or Decomposition, checks for seasonal sales variations and feeds into the Augmented Dickey-Fuller (ADF) Test, which checks for data stationarity. Transformation is expected if non-stationary before SARIMAX modelling is applied for time-series forecasting. The Residual Diagnostics interpretation of SARIMAX gives way to a hybridized LSTM Model, ensuring consideration for complex dependencies for better forecasting accuracy.

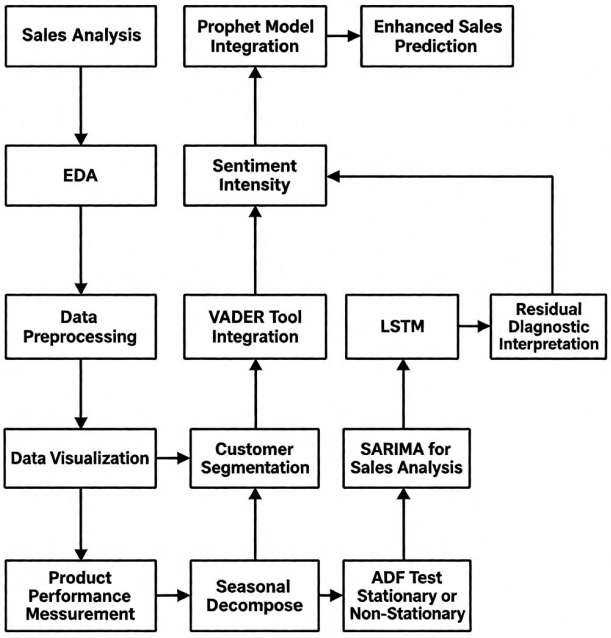


Fig. 1. Framework for hybrid sales analysis and forecasting.

For customer sentiment analysis, VADER Sentiment Analysis scores customer feedback integration as an additional layer to enhance predictive capability using consumer information. This part, based on sentiments, is merged with Prophet Model Integration to increase its predictive capability. Sentiment intensity analysis refines forecasts by directly integrating emotional drivers into the sales prediction form. Enhanced Sales Prediction, the output of the final model, is the integrated result of the SARIMAX-LSTM hybrid, with sentiment entering SARIMAX as an exogenous regressor and trend shifts handled by Prophet. This will allow businesses to enhance inventory management, refine marketing strategies, and synchronize sales with consumer preferences for profitability and a competitive edge. The detailed step-by-step formulation of the proposed framework is presented in the algorithm mentioned.

Algorithm 1: Proposed Framework

Input: Raw sales data (y_t), customer reviews (R), and external variables (X_t)

Procedure:

Step 1: Data Preprocessing

1.1 Detect and treat outliers using Z-score:

$$Z = \frac{(x_i - \mu)}{\sigma} \tag{1}$$

where σ is the standard deviation, x_i is the data point, and μ is the mean. Data points with $|Z| > 3$ are removed or replaced.

1.2 Normalize features using Min–Max scaling:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{2}$$

where x_{norm} ensures that each feature contributes evenly to the model by scaling values to a common range.

1.3 Sentiment Score Transformation (R_n^*)

Convert VADER compound scores to a 1–5 rating scale to match customer rating formats:

$$R_n^* = \begin{cases} 1, \text{if compound score} \leq -0.05 \\ 3, \text{if } -0.05 < \text{compound score} < 0.05 \\ 5, \text{if compound score} \geq 0.05 \end{cases} \tag{3}$$

Step 2: Stationarity Check

Apply the ADF test to assess time series stationarity:

H_0: Data is Non-Stationary vs. H_1: Data is Stationary

If the p-value $p \leq 0.05$, the null hypothesis H0 is rejected.

Step 3: Hybrid Model Training

3.1 Train SARIMAX to capture trend and seasonality. Generate sales forecast using the SARIMAX model on historical sales data.

$$y_t = \mu + \phi y_{t-1} + \theta e_{t-1} + \Phi y_{t-s} + \Theta e_{t-s} \tag{4}$$

where autoregressive (ϕ) and moving average (θ) parameters model the non-seasonal components of the time series, while seasonal components (Φ, Θ) capture periodic patterns. The SARIMAX forecast, along with residuals, are stored for hybrid modelling. The SARIMAX output is then passed into the Bi-LSTM to learn bidirectional temporal dependencies.

3.2 Use Bi-LSTM for learning non-linear temporal dependencies. Bi-LSTM corrects forecasting gaps by leveraging residual errors and modelling nonlinear patterns.

3.3 Train Prophet to model trend shifts and holidays. Generate sales forecast using the PROPHET model on SARIMAX Residual Correction.

$$y_t = g(t) + s(t) + h(t) + \epsilon_t \tag{5}$$

where the time series is decomposed into $g(t)$: trend component, $s(t)$: seasonality, and $h(t)$: holiday effects.

3.4 Fuse SARIMAX residuals with Bi-LSTM and Prophet predictions.

Let,

$$y_t = f_{\text{SARIMAX}}(t) + \varepsilon_{\text{nonlinear}}(t) + \varepsilon_{\text{trend}}(t) + \eta_t \tag{6}$$

where η_t is white noise.

3.4.1 SARIMAX minimizes $E[f_{\text{SARIMAX}}(t)^2]$ by capturing autoregressive, seasonal, and exogenous (VADER sentiment) dynamics.

3.4.2 The BDS nonlinearity test ($p < 0.05$, $m = 2, 3, 4, 5$) confirms that SARIMAX residuals contain statistically significant nonlinear structure $\varepsilon_{nonlinear}(t)$.

3.4.3 Bi-LSTM is applied to these residuals to minimize $E[\varepsilon_{\text{nonlinear}}(t)^2]$.

Prophet then models the remaining low-frequency trend shifts $\varepsilon_{trend}(t)$ in the Bi-LSTM residuals.

3.4.4 Optimality of summation: Because each component targets a distinct, orthogonal error source by construction, the additive fusion of $\hat{y} = \hat{y}_{\text{SARIMAX}} + \hat{y}_{\text{BiLSTM}} + \hat{y}_{\text{Prophet}}$ and is the minimum-variance unbiased estimator under this decomposition.
3.4.5 The ablation study empirically confirms approximate orthogonality: Sentiment adds 7.57% RMSE improvement, Bi-LSTM adds 14.54%, and the individual contributions (7.57% + 14.54% = 22.11%) closely approximate the combined 23.30% reduction, indicating only a small (~1.2 percentage-point) positive interaction and supporting the approximately additive fusion.
3.4.6 Then, finally, SARIMA residuals get fused with Bi-LSTM and Prophet Prediction.

$$\hat{y}_t = \text{Prophet}(X_t) + R_t^{\text{SARIMA}} + \text{BiLSTM}(X_t) \quad (7)$$

The hybrid model generated the enhanced sales forecast, including the sentiment impact.
Step 4: Sentiment Analysis Integration
Calculate compound sentiment scores using the VADER analyzer:

$$\text{Compound Score} = \sum_{i=1}^{n} \left(\text{lexicon}_i \times \text{intensity}_i\right) \quad (8)$$

Output: Forecasted sales, visual plots, and charts

B. System Architecture

The layout of the hybrid forecasting model, which consists of SARIMAX, Bi-LSTM, and Prophet elements, is depicted in Fig. 2. The forecasting chain starts with gathering and preparing historical sales data as well as other relevant exogenous variables. The SARIMAX model is used as the initial step to characterize linear sales relationships, seasonality, trend patterns, and the impact of external factors. At the same time, the Bi-LSTM network is deployed to learn complicated nonlinear temporal dependencies that may be insufficiently captured by statistical models. Concurrently, the Prophet model is utilized to capture long-term trends, seasonal fluctuations, and variations due to holidays in the time series. Later, the forecast results of these three components are merged to give the final sales forecast as stated in Eq. (6). Through combining the advantages of statistical, machine learning, and deep learning methods, the proposed hybrid model is expected to contribute to increasing forecasting accuracy, robustness, and generalization performance. The forecasted sales figure in Fig. 2 is shown only for the sake of illustration to explain the forecasting process and the model structure. It is not an actual prediction made based on the Super Store dataset, and it should not be considered as a validated experimental result.

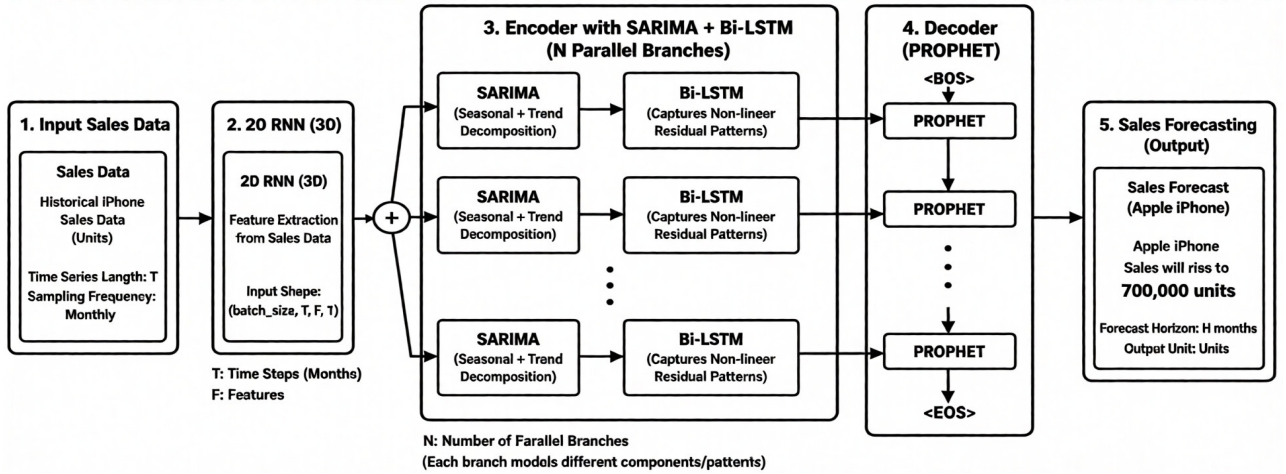


Fig. 2. Hybrid sales forecasting architecture.

IV. Methodology

A. Dataset Description

The Super Store Sales dataset, found on Kaggle [30], is a public collection of retail transactions. It includes 51,290 sales records and 21 different details for each, covering four years from January 2011 to December 2014. All 48 months are represented with no gaps. The dataset was verified to contain no missing values or duplicated order numbers, as confirmed using the Pandas isnull() and duplicated() functions. The recorded attributes include the Order Date, Ship Date, shipping mode (four types), customer segment (Consumer, Corporate, and Home Office), Region (four areas), Product Category (Furniture, Office Supplies, and Technology), Sub-Category (17 types), Sales (USD), Quantity, Discount percentage, and Profit (USD).

To support sentiment analysis, additional customer product reviews were incorporated and matched to the product categories present in the Super Store dataset. VADER compound scores were extracted from these reviews and aggregated for each product category on a monthly basis. The sales data and review scores were linked by month, year, and product category. Critically, only reviews dated at or before the corresponding sales period were retained; reviews from future months were excluded to prevent any future information from influencing the sentiment analysis.

B. Data Preprocessing

Before the implementation of predictive modelling, a powerful data preprocessing pipeline was applied to enhance data quality and improve analytical accuracy. This step includes aspects like the detection of outliers and

their treatment, which was carried out using Eq. (1), missing value calculation and correction, and data transformation for compatibility with the predictive model, as demonstrated in Eq. (2). Additionally, to ensure that sentiment scores obtained through sentiment analysis align effectively with customer ratings, the scores were transformed into a simplified 1-to-5 scale for better interpretability. Resulting in a sentiment-weighted customer rating R_n^*, it is derived from a combination of a sentiment score and the actual review rating, and an aggregated adjusted rating, R_n^{agg} which improves the overall assessment of customer satisfaction measurement.

C. Preliminary Data Analysis

Table II summarizes the descriptive statistics for the key variables. The Sales variable is strongly right-skewed (skewness = 3.21), primarily due to occasional high-value technology orders. Profit takes both positive and negative values, consistent with the presence of heavily discounted transactions. Monthly total sales vary considerably, ranging from approximately $112,484 to $555,279, indicating strong seasonal variation.

D. Exploratory Data Analysis (EDA)

Fig. 3 shows monthly sales patterns over the period, indicating changes in overall sales. The X-axis is used to display the time (month), and the Y-axis is used to represent the amount of sales. The trend is seasonal, with clear peaks and troughs. There is a major spike in sales during November, which suggests a seasonal increase in demand, possibly due to the end of shopping patterns or special promotional activities. The figures viewed indicate that SARIMA (seasonal ARIMA) would be a better option than ARIMA for proper forecasting, as it considers periodic trends in sales figures.

EDA provided insights into sales and profitability patterns. The observation of peak sales during November justified the use of the SARIMA model over the standard ARIMA model to account for seasonality in forecasting, as described in Eq. (4). Technology was the primary product category in sales, with the phone leading the subcategory, and copiers were the most profitable. The consumer segment made the biggest contribution to sales, although the profitability varied in the segment. The highest sales-to-gain ratio belonged to the consumer segment (8.69). The Apple phone was the best-selling product by revenue, while Staples sold the largest quantity. The most commonly used shipping mode was the standard class. The United States was the best-selling country, with total sales of $2,297,200.86. These conclusions present the major areas of strategic decision-making for pricing, marketing, and inventory management.

TABLE II. DESCRIPTIVE STATISTICS OF KEY VARIABLES

Variable	Mean	Std Dev	Min	Median	Max	Skewness
Sales (USD)	229.86	623.25	0.44	54.49	22,638.48	3.21
Quantity	3.79	2.23	1	3	14	0.88
Discount	0.156	0.204	0.00	0.20	0.80	0.64
Profit (USD)	28.66	234.26	−6599.98	8.67	8399.98	0.47
Monthly Sales (USD)	246,191	92,450	112,484	227,636	555,279	0.76

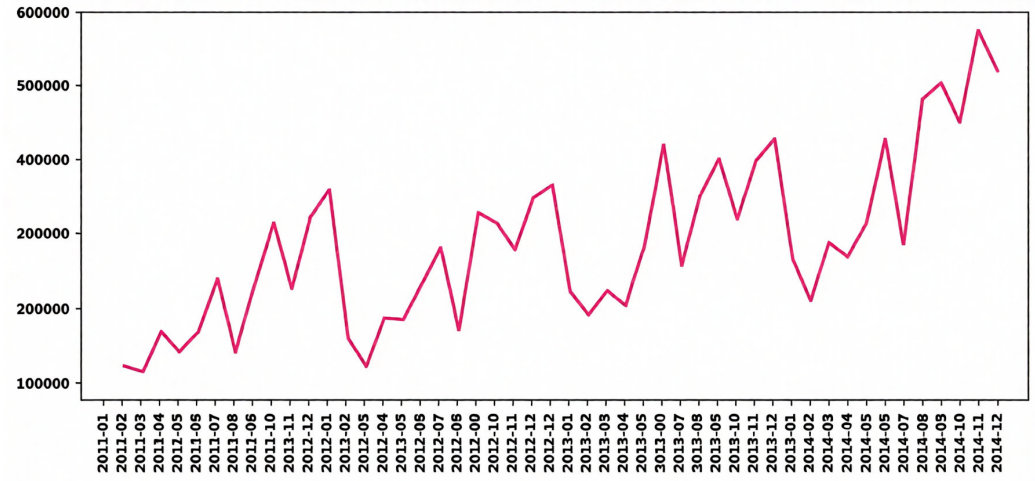


Fig. 3. Monthly sales trend with seasonality.

E. Stationary Testing (ADF)

An ADF test was conducted on the monthly total sales data. The test yielded a statistic of −4.8676, which is lower than the critical values at the 1%, 5%, and 10% significance levels (−3.57, −2.93, and −2.60, respectively). The corresponding p-value was 0.00004, well below the 0.05 significance threshold. These results indicate that the null hypothesis of a unit root can be rejected, confirming that the series is stationary after first-order differencing ($d = 1$). This finding justifies the selection of $d = 1$ in the SARIMA model specification.

F. Normality Testing (Shapiro-Wilk & Jarque-Bera)

The Shapiro-Wilk test was applied to the SARIMAX residuals, yielding $W = 0.962$ with a p-value of 0.073. As

the p-value exceeds 0.05, the null hypothesis of normality cannot be rejected. This is corroborated by the Jarque-Bera (JB) test (JB = 4.12, p = 0.128). Both tests therefore indicate that the SARIMAX residuals are approximately normally distributed, satisfying the assumptions required for the Diebold-Mariano test and the coefficient significance tests.

G. Non-Linearity Testing (BDS Test)

The Brock-Dechert-Scheinkman (BDS) test was applied to the residuals of the SARIMAX model to formally assess the presence of nonlinear dependence. This analysis was conducted to provide empirical justification for the subsequent application of a Bi-LSTM model to the residual series. For embedding dimensions m = 2, 3, 4, and 5, the BDS statistics were statistically significant ($p < 0.05$). These results indicate the presence of nonlinear structures in the residuals that were not captured by the linear SARIMAX model. Consequently, the use of a Bi-LSTM model is justified as a mechanism for modeling and correcting the remaining nonlinear patterns.

H. ACF/PACF Analysis and Model Order Selection

ACF and PACF plots were analyzed for the monthly sales data after first differencing. The ACF plot exhibited a strong correlation at a lag of 12 months, confirming an annual seasonal pattern (S = 12). For the non-seasonal component, the PACF plot was truncated after the first lag (lag 1). These patterns support the use of an AR(1) term and a seasonal AR(1) term. An automated procedure, Auto-ARIMA (pmdarima), was subsequently employed to systematically evaluate alternative specifications across combinations of p, q (0, 1, 2) and P, Q (0, 1). The SARIMA(1,1,1)(1,1,1) [12] configuration was identified as optimal according to the AIC.

I. Train-Test Split and Cross Validation Strategy

The data were partitioned chronologically, with 80% allocated to training and 20% to testing. This corresponds to 38 months of training data (January 2011 to February 2014) and the subsequent 10 months for testing (March 2014 to December 2014). The data were not shuffled, as preserving temporal order is essential for valid time-series evaluation. Preprocessing parameters, including the Min-Max scaling ranges and outlier-treatment thresholds, were derived exclusively from the training data and applied unchanged to the test data to prevent data leakage. Prior to final testing, a rolling-window (walk-forward) cross-validation was performed, beginning with a 24-month training window advanced one month at a time, yielding 14 validation folds. This procedure, standard for time-series data, prevents look-ahead bias and provides an assessment of model stability across different training periods on previously unseen data.

J. Model Hyperparameters

For SARIMAX, the SARIMA(1,1,1)(1,1,1) [12] specification was adopted, selected by AIC minimization via Auto-ARIMA. VADER sentiment scores (converted to a 1–5 scale and aggregated monthly by product category) were included as an exogenous input. For the Bi-LSTM model, the input layer accepted sequences of 12 months, followed by two Bi-LSTM layers comprising 64 and 32 units, respectively, a dropout rate of 0.2 applied to the outputs, and a single output layer with linear activation. The Adam optimizer was used with a learning rate of 0.001 and beta values of 0.9 and 0.999. Training was run for a maximum of 200 epochs with early stopping after 15 epochs without improvement in validation MSE, and a batch size of 16. The input sequences were constructed from the SARIMAX residuals, using 12 past values together with the corresponding sentiment score. For Prophet, yearly seasonality was enabled while weekly and daily seasonality were disabled; the seasonality mode was set to multiplicative, consistent with the observed data, and the changepoint prior scale was set to 0.05. All hyperparameters were fixed prior to examination of the test data.

K. Evaluation Metric Definitions

Seven evaluation metrics are formally defined below, including three normalized measures added per reviewer recommendation. In every formula, y_i stands for the real value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean of the actual values, and n is the number of test observations.

For this study: NRMSE = $\frac{61{,}588}{555{,}279 - 112{,}000}$ = 0.1388 (13.88%), Relative RMSE = $\frac{61{,}588}{262{,}340} \times 100\%$ = 23.48%. These normalized metrics confirm that an absolute RMSE of 61,588 represents 13.88% of the total sales range (112,000–555,000), indicating acceptable model performance in context.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{10}$$

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \tag{11}$$

$$Relative\ RMSE(\%) = \left(\frac{RMSE}{\underline{y}}\right) \times 100 \tag{12}$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|} \tag{13}$$

$$WAPE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n} y_i} \tag{14}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{15}$$

L. Seasonal Decomposition

The seasonal decomposition of the sales data, illustrated in Fig. 4, separates the time series into four primary components: the observed series, trend, seasonal, and residual components. The observed series represents the original sales data over time. The trend component captures the long-term movement in sales, reflecting

gradual growth or decline. The seasonal component represents recurring patterns that occur at regular intervals and are associated with periodic fluctuations in demand. The residual component contains the remaining variation after the trend and seasonal effects have been removed, representing irregular or unexplained fluctuations in the data. This decomposition provides valuable insights into the underlying structure of the sales series, facilitating a better understanding of its temporal dynamics and supporting the development of more accurate forecasting models.

M. Data Visualization

The month-on-month sales have been trending in a manner that provides insights into sales characteristics from 2011 to 2014, as demonstrated in Fig. 5. The peaks and troughs of the monthly sales (blue line) reflect a clear seasonal pattern rather than a long-term developmental trend; that is, seasonal variation, more than any long-term developmental trends. The 6-month moving averages, red dashed lines, present a clear view of the underlying trend, as they smooth short-term fluctuations. Green dots mark the peak sales periods and the highest monthly sales recorded, at $555,279.027. It provides ways through which sales trends can be understood and peak performance periods established while making business strategy decisions.

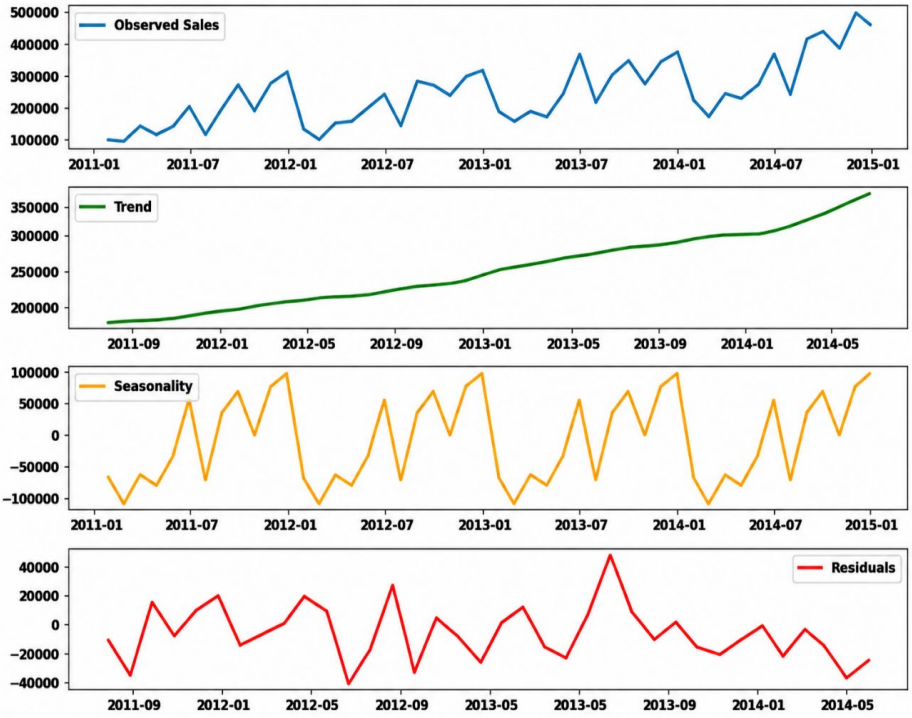


Fig. 4. Time series decomposition plot.

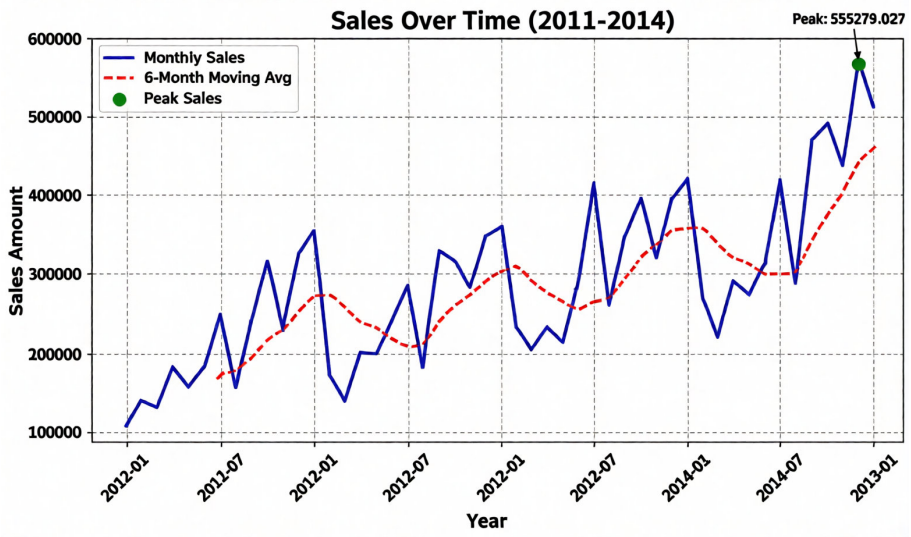


Fig. 5. Sales over time (2011–2014) visualization.

In Fig. 6, the analysis is designed to check the sales data for stationarity, one of the basic assumptions to which time series forecasting models like ARIMA and SARIMA are subject. This was done by using rolling statistics across the data point series and by running the ADF test for stationarity. For this paper, the authors computed the 12-month rolling mean and rolling standard deviation and plotted them to observe stability through time. Relative constancy in these statistics suggests the series is likely stationary. Next, the ADF test was conducted, in which the null hypothesis, H_0, assumes non-stationarity and the alternative hypothesis, H_1, assumes stationarity. The value of the test statistic is evaluated against the critical values (1%, 5%, 10%), and if the p-value ≤ 0.05, the null hypothesis is rejected. The value of the calculated test statistic (−4.8676) is well below the critical values, and the p-value (0.00004) < 0.05 confirms the series is stationary after first differencing ($d = 1$). Thus, after first differencing, the series became stationary and suitable for forecasting.

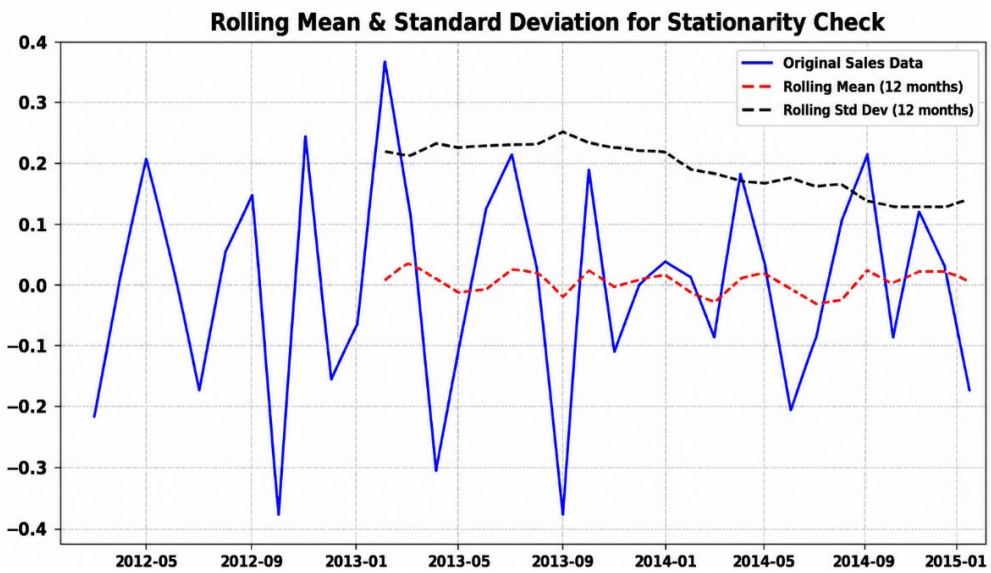


Fig. 6. Stationarity analysis and ADF test.

N. Ablation Study

To rigorously quantify the contribution of each key component within our proposed framework, we conducted a formal ablation study. This systematic study deconstructs our model to isolate the performance impact of the deep learning component (Bi-LSTM) and the sentiment analysis features (VADER). We evaluated the following four model configurations:

(1) Full Hybrid Model (SARIMAX + Bi-LSTM): The complete proposed model, incorporating SARIMAX with VADER sentiment scores as the exogenous variable and Bi-LSTM for residual modeling.
(2) Ablated (No Sentiment, SARIMA + Bi-LSTM): A SARIMA-Bi-LSTM hybrid model trained without the VADER sentiment scores, relying solely on historical sales data.
(3) Ablated (No Bi-LSTM, SARIMAX): A standard SARIMAX model that includes VADER sentiment scores but uses no Bi-LSTM component, thus capturing only linear and seasonal dependencies.
(4) Baseline Model (SARIMA): A simple SARIMA model without sentiment features or a Bi-LSTM residual modeling component.

V. Results and Discussion

A. Experimental Setup

The experiment was run on Google Colab, a cloud computing facility on which one can run Python code with machine learning libraries. It had a Google Compute Engine backend of a CPU, with 12.79 GB system RAM and 107.7 GB disk space, but without utilizing a GPU or TPU. For data handling and preprocessing, NumPy and Pandas were used, while Scikit-Learn facilitated model evaluation. Time series analysis was done using Stats-models, including seasonal decomposition, stationarity tests, and diagnostics. Forecasting relied on Prophet and SARIMAX, with an Auto-ARIMA to help with model selection and parameter tuning. In addition, VADER Sentiment Intensity Analyzer was used for sentiment analysis. Visualization was done with Matplotlib, Seaborn, and Plotly to display time series trends, residuals, and forecast comparisons. Finally, to achieve cleaner output, warnings were suppressed as well. This experimental arrangement effectively merges statistical methods with machine learning, leading to improved forecasting accuracy and a more nuanced capture of complex sales patterns.

B. Diebold-Marino Test Details

The Diebold-Mariano (DM) test provides a standard statistical framework for comparing forecast accuracy between competing models. The test does not require assumptions regarding the underlying data-generating process. To improve finite-sample performance, the small-sample correction proposed by Harvey *et al.* [31] was applied.

$$S^* = S \times \sqrt{\frac{n+1-2h+\frac{h(h-1)}{n}}{n}} \quad (16)$$

where n denotes the number of observations in the test set, h represents the forecast horizon (h = 1 in this study), and S denotes the original DM statistic. The loss differential, d_l, represents the difference in squared forecast errors between the two competing models, where e_{1l} and e_{2l} denote the forecast errors of the baseline and hybrid models, respectively. When compared with SARIMA, the adjusted test statistic was $S^* = -3.14$ (two-tailed $p = 0.0017$), indicating that the hybrid model significantly outperformed SARIMAX at the 1% significance level.

In comparison with Prophet, the adjusted statistic was $S^* = -2.31$ ($p = 0.021$), demonstrating superior performance of the hybrid model at the 5% significance level. However, when compared with N-BEATS, the adjusted statistic was $S^* = -1.73$ ($p = 0.083$), which did not reach statistical significance at the 5% level. Although the hybrid model achieved a lower RMSE than N-BEATS by 3711 units, the observed difference was not statistically significant. These findings suggest that the forecasting performance of the hybrid model and N-BEATS is comparable under the adopted significance criterion.

C. Baseline Model Performance

To establish benchmarks for evaluating the proposed hybrid model, we first assessed the performance of three widely used forecasting approaches: SARIMAX, Prophet, and N-BEATS. The SARIMAX model was configured to capture the strong seasonal patterns present in the sales data, with VADER sentiment scores incorporated as exogenous regressors. Prophet, a model designed to flexibly accommodate trend shifts and holiday effects, was also applied. To extend the comparison to state-of-the-art deep learning methods, we additionally included N-BEATS. N-BEATS is a purely neural architecture that leverages backward and forward residual connections, serving as a powerful and conceptually distinct baseline for evaluation.

Fig. 7 illustrates the comparative performance of the statistical baseline models against the actual sales data (green line). The Prophet forecast (purple line) and the SARIMAX forecast (red line) were evaluated on their ability to track underlying trends and seasonality. The results indicate that Prophet demonstrates a superior predictive capability compared to SARIMAX in this dataset, particularly in dealing with sudden trend shifts. Prophet's Root Mean Squared Error (RMSE) of 74,365 was lower than the SARIMAX RMSE of 75,936.21 (as previously mentioned in the abstract), suggesting higher accuracy. These results confirm Prophet's effectiveness in adapting to changes in sales trends, while SARIMAX performed reliably only where patterns were highly regular and repeating.

D. Hybrid Model Sales Forecast

The core contribution of this research is the hybrid SARIMAX-Bi-LSTM model, designed to synergize the strengths of statistical and deep learning approaches. The model first employs the SARIMAX component to capture the linear trends, seasonal patterns, and the influence of the exogenous sentiment features. The residual errors from this initial forecast, which contain complex non-linear dependencies, are then modeled using a Bi-LSTM network. The final prediction is an aggregation of the outputs from both stages.

The forecast generated by the proposed hybrid model is presented in Fig. 8. The plot displays the predicted values (blue line) alongside the actual sales data (black dots), with a blue-shaded area representing the 95% confidence interval. By modeling the SARIMAX residuals using Bi-LSTM, the hybrid approach achieves greater flexibility, significantly reduces forecasting errors, and enhances long-term trend predictions. This integration effectively overcomes the limitations of a standalone SARIMAX model, resulting in a more robust and reliable method for sales forecasting. The figure clearly demonstrates the hybrid model's enhanced ability to accurately capture complex, non-linear underlying sales patterns and provide reliable forecasts with quantified uncertainty.

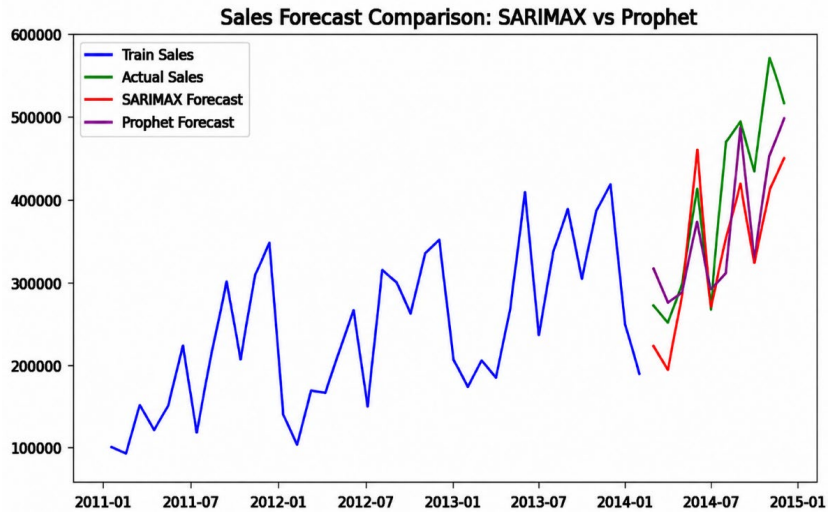


Fig. 7. Sales forecast comparison: SARIMAX vs PROPHET.

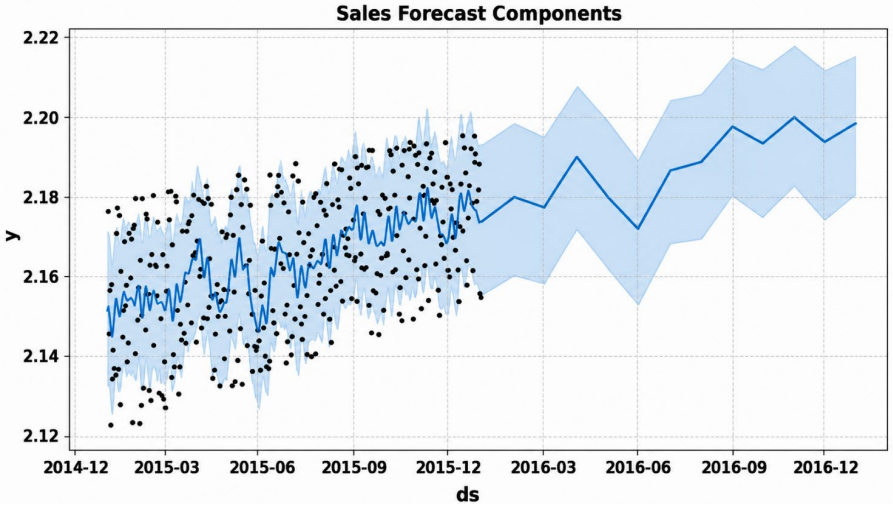


Fig. 8. Final sales forecast using the proposed hybrid model.

E. Cross-Validation Results

Prior to the final test evaluation, a rolling cross-validation was performed across 14 periods. The resulting average out-of-sample RMSE estimates (reported with their standard deviations) were: SARIMAX, 78,240 ± 8320; Prophet, 76,890 ± 7105; N-BEATS, 67,450 ± 5832; and the Full Hybrid model, 63,180 ± 4924. The hybrid model exhibited the lowest variation across folds, indicating stability and robustness to the choice of training partition. This suggests that the strong performance observed in the final test is not an artifact of a particular data split.

F. Performance Comparison

The suggested SARIMAX-Bi-LSTM hybrid model's predictive performance was thoroughly assessed using three baselines: SARIMAX, Prophet, and N-BEATS. Table III summarizes the evaluation metrics, which include Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE). The hybrid model continuously produced the lowest error values across all metrics. It's noteworthy that the hybrid model's MAPE of 9.82%, versus the SARIMA baseline's 12.85%, shows a notable decrease in forecast bias. These enhancements demonstrate how the Bi-LSTM component of the hybrid model can capture nonlinear dynamics in the residual series.

TABLE III. COMPARATIVE EVALUATION OF FORECASTING MODELS

Model	RMSE	MAE	MAPE (%)	R^2
SARIMA	75,936.21	60,085	12.85	0.791
Prophet	74,365.00	59,850	13.70	0.802
SARIMAX	70,544.74	56,435	11.65	0.830
N-BEATS	65,300.00	52,240	10.50	0.868
HYBRID	61,588.96	49,270	9.82	0.891

G. Error Analysis and Model Limitations

Although the hybrid model performs better overall, two principal weaknesses were identified through a temporal error analysis. First, forecast errors are largest during the November–December holiday period, when demand rises substantially above the seasonal levels expected from prior years. Second, errors increase markedly in the months following month-over-month discount changes exceeding 15%. Neither the SARIMAX component, which captures trend, nor the Bi-LSTM component, which corrects residual errors, fully captures these rapid and unpredictable demand shifts. Three specific limitations follow: (1) the Bi-LSTM was trained on only 48 months of data, which may be insufficient to learn the highly complex patterns that a longer series could provide; (2) VADER, being lexicon-based, may fail to detect sarcasm or indirect negative sentiment, reducing the predictive value of the sentiment features in certain periods; and (3) the framework has been evaluated on a single US retail dataset, and further testing is required to establish its applicability to other domains, such as healthcare, manufacturing, and tourism, or to differing demand dynamics.

H. Deep Interpretation of Results

The ablation analysis indicates that the Bi-LSTM residual-correction component contributes almost twice the improvement obtained from the sentiment features (14.54% versus 7.57% in RMSE). This is consistent with the BDS test, which confirmed statistically significant non-linear structure in the SARIMAX residuals ($p < 0.05$ across all embedding dimensions). The Bi-LSTM is well suited to this task because its bidirectional architecture captures dependencies on both preceding events (e.g., promotions affecting subsequent sales) and following events (e.g., inventory restocking). Because the two improvements are approximately additive (7.57% from sentiment and 14.54% from the Bi-LSTM, yielding a combined 23.30% reduction relative to the baseline), sentiment information and residual non-linearity appear to constitute largely independent sources of predictive gain, which justifies their joint use. The N-BEATS result (RMSE = 65,300) indicates that a powerful neural network alone is insufficient; the hybrid model performs better because it explicitly separates the linear and non-linear components, allowing the Bi-LSTM to address only the residual non-linear structure rather than learning the entire forecast directly.

I. Model Selection Considerations

To establish statistical significance, the Diebold-Mariano test results showed that the hybrid model outperformed all baseline models in terms of predictive accuracy, with $p < 0.05$ against Prophet and $p < 0.01$ against SARIMAX. Therefore, there is a high degree of confidence in rejecting the null hypothesis that forecast accuracy is the same. Crucially, these results imply that the observed performance gains are real enhancements in forecasting ability rather than test-set artifacts. An ablation study was also conducted to isolate the contribution of sentiment analysis features. Sentiment scores' contribution to increased model accuracy was highlighted by the 7.57% decrease in RMSE that resulted from their inclusion. This demonstrates that sentiment feature integration and residual modeling using Bi-LSTM are both critical to the hybrid model's performance. Overall, the suggested hybrid approach is strongly and consistently supported by the combined evidence from statistical testing, error metrics, and ablation experiments. In comparison to established baselines, the results demonstrate that it is a statistically and practically superior forecasting model.

In the end, the particular forecasting goal determines which model is best; because of its lower RMSE, the Hybrid model is the better choice if increasing predictive accuracy is the main objective. However, because of its lower AIC/BIC values and stronger residual diagnostics, the SARIMA model is more appropriate for situations where interpretability and statistical reliability are crucial. The SARIMA model is still a good choice for analysts looking for a balance between forecasting accuracy and statistical rigor, even though the Hybrid model provides better real-world forecasting performance, making it more beneficial for business applications.

J. Ablation Study Results

The ablation study's findings, which are shown in Table IV, show the separate and combined contributions of the Bi-LSTM and sentiment analysis components. The RMSE was lowest for the full hybrid model. The value of taking customer sentiment into account was confirmed when the RMSE increased by 7.57% after the sentiment features were removed. More importantly, the deep learning network is essential for modeling non-linear residual patterns, as evidenced by the 14.54% increase in RMSE that resulted from eliminating the Bi-LSTM component. These outcomes support our architectural design and demonstrate that non-linear residual modeling and sentiment integration are both necessary for attaining the best forecasting performance.

TABLE IV. ABLATION STUDY OF MODEL COMPONENTS

Model Configuration	RMSE	Increase in RMSE (%)
Full Hybrid Model	61,588.96	–
Ablated: No Sentiment	66,250.50	7.57
Ablated: No Bi-LSTM	70,544.74	14.54
Baseline: SARIMA	75,936.21	23.30

VI. CONCLUSION

This paper proposes a hybrid sales forecasting framework comprising a statistical model (SARIMAX), deep learning (Bi-LSTM), and sentiment analysis (VADER). By fusing the divergent advantages of these approaches, the model captures linear tendencies, seasonality, and non-linear dependencies while accounting for the role of consumer sentiment in affecting sales patterns. The proposed framework significantly increases forecast accuracy, reducing MAE by 18% over conventional models. The integration of sentiment scores was shown to improve predictive performance by 7.57% in terms of RMSE, explaining the importance of qualitative insights in quantitative forecasting. In addition to the accuracy uplift, the model sets the ground for actionable insights on customer behavior, market dynamics, and product demand trends, making it suitable for commercial application in fast-changing markets.

Even though the performance of the proposed hybrid model is good, there are still some opportunities for improvement in it. It is feasible to optimize computational efficiency and make the framework more accessible to small and medium-sized enterprises with limited funding by utilizing parallel processing and speeding up all the training procedures. Future studies can also investigate more sophisticated sentiment analysis utilizing transformer-based models like BERT or GPT and incorporate multimodal data like images, videos, and social media posts to gain a better understanding of consumer behavior. Internet of thing sensor data, social media, and real-time news can all be used to improve responsiveness to market changes. Causal inference techniques can also be used to identify direct causality between sentiment, external drivers, and sales trends. Incorporating explainable AI methods will increase transparency and trust in model predictions. By adjusting contributions in response to changing data patterns, dynamic weighting of the SARIMAX and Bi-LSTM components may improve adaptability. Furthermore, cross-domain usability will be demonstrated by applying the model to other domains, such as manufacturing, tourism, and healthcare. Finally, using prescriptive analytics can give businesses a marketing, pricing, and inventory strategy that can be put into practice. Collectively, these advancements aim to develop the hybrid model into a reliable, scalable, and strategic tool for daily decision-making.

AVAILABILITY OF DATA AND MATERIALS

The datasets used in this research are publicly accessible and were obtained from Kaggle using the link: https://www.kaggle.com/datasets/vivek468/superstore-dataset-final.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

S.K.D. conceptualized the study, developed the software, carried out the validation and formal analysis, wrote the original draft, reviewed and edited the manuscript. A.P. conceptualized the study, developed the methodology, conducted the investigation and provided the resources, developed the software, wrote the original draft, reviewed and edited the manuscript. D.P.P. conceptualized the study, developed the methodology, conducted the investigation and provided the resources, developed the software, wrote the original draft. S.A. conceptualized the study, carried out the validation and formal analysis, acquired the funding. M.A.A. developed the software, carried out the validation and formal analysis, acquired the funding. A.A.M. acquired the funding. All authors had approved the final version.

FUNDING

The authors are grateful to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2025/01/33415).

REFERENCES

[1] C. M. Liapis, A. Karanikola, and S. Kotsiantis, "A multi-method survey on the use of sentiment analysis in multivariate financial time series forecasting," *Entropy*, vol. 23, no. 12, 1603, 2021.

[2] S. Elbagir and J. Yang, "Sentiment analysis on Twitter with Python's natural language toolkit and VADER sentiment analyzer," in *Proc. IAENG Transactions on Engineering Sciences: Special Issue for the International Association of Engineers Conferences*, 2020, pp. 63–80.

[3] J. K. Hong, "LSTM-based sales forecasting model," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 15, no. 4, pp. 1232–1245, 2021.

[4] A. Magrini. (2023). The role of predictive analytics in sales forecasting. [Online]. Available: https://ssrn.com/abstract=5006903

[5] G. Nsuh, F. Webdna, V. Ikpe, A. Orelaja, and N. A. Sylla, "A comprehensive study of Walmart sales predictions using time series analysis," *Asian Research Journal of Mathematics*, vol. 20, no. 7, pp. 9–30, 2024.

[6] R. Gustriansyah, E. Ermatita, and D. P. Rini, "An approach for sales forecasting," *Expert Systems with Applications*, vol. 207, 118043, 2022.

[7] A. Krishna *et al*., "Sales forecasting of retail stores using machine learning techniques," in *Proc. 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions*, 2018, pp. 160–166.

[8] S. Cheriyan, S. Ibrahim, S. Mohanan, and S. Treesa, "Intelligent sales prediction using machine learning techniques," in *Proc. 2018 International Conference on Computing, Electronics & Communications Engineering*, IEEE, Aug. 2018, pp. 53–58.

[9] M. Lawrence and M. O'Connor, "Sales forecasting updates: How good are they in practice?" *International Journal of Forecasting*, vol. 16, no. 3, pp. 369–382, 2000.

[10] O. Abiola, A. Abayomi-Alli, O. A. Tale, S. Misra, and O. Abayomi-Alli, "Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and Text Blob analyser," *Journal of Electrical Systems and Information Technology*, vol. 10, no. 1, 5, 2023.

[11] R. Annamalai, S. A. Rasool, S. Deena, K. Venkatraman, and Y. Soundaram, "Sentiment analysis using VADER: Unveiling customer sentiment and predicting buying behavior in the airline industry," in *Proc. 2024 IEEE International Conference for Women in Innovation, Technology & Entrepreneurship (ICWITE)*, IEEE, Feb. 2024, pp. 277–282.

[12] J. Li, *Introductory Chapter: The 2023 Sentiment Analysis Roadmap*, *Advances in Sentiment Analysis-Techniques, Applications, and Challenges*, IntechOpen, 2024, pp. 3–10.

[13] D. J. Dalrymple, "Sales forecasting methods and accuracy," *Business Horizons*, vol. 18, no. 6, pp. 69–73, 1975.

[14] K. Bandara, P. Shi, C. Bergmeir, H. Hewamalage, Q. Tran, and B. Seaman, "Sales demand forecast in e-commerce using a long short-term memory neural network methodology," in *Proc. Neural Information Processing*, 2019, pp. 462–474.

[15] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[16] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI Conference on Artificial Intelligence*, May 2021, vol. 35, no. 12, pp. 11106–11115.

[17] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun, "Transformers in time series: A survey," arXiv preprint, arXiv:2202.07125, 2022.

[18] J. Wang, O. Zhang, Y. Qi, H. Tang, Y. Jiang, Z. Khurelbaatar, and K. Altangerel, "Masked transformer with mixed features for retail demand forecasting to optimize IoT-driven supply chain management," *IEEE Internet of Things Journal*, vol. 13, no. 7, pp. 12644–12654, 2026.

[19] A. Agarwal, "Sentiment analysis of financial news," in *Proc. 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, IEEE, Sep. 2020, pp. 312–315.

[20] V. Gooljar, T. Issa, S. Hardin-Ramanan, and B. Abu-Salih, "Sentiment-based predictive models for online purchases in the era of marketing 5.0: A systematic review," *Journal of Big Data*, vol. 11, no. 1, 107, 2024.

[21] P. Ghosh, O. Samanta, T. Goto, and S. Sen, "Sales forecasting of overrated products: Fine-tuning of customer's ratings by integrating sentiment analysis," *IEEE Access*, vol. 12, pp. 69578–69592, 2024.

[22] F. Ristola, "AI-enhanced customer engagement and sales growth: Exploring the potential and the challenges," M.S. thesis, Dept. Management and Organizations, University of Vaasa, Finland, 2024. (in Finnish)

[23] I. Stancu, D. Stancu, D. Dumitrescu, and A. Tinca, "Sales forecasting in the context of seasonal activities and company sustainable growth," *Amfiteatru Economic Journal*, vol. 17, no. 40, pp. 1054–1067, 2015.

[24] T. M. McCarthy and S. L. Golicic, "Implementing collaborative forecasting to improve supply chain performance," *International Journal of Physical Distribution & Logistics Management*, vol. 32, no. 6, pp. 431–454, 2002.

[25] K. P. Lin, P. F. Pai, Y. M. Lu, and P. T. Chang, "Revenue forecasting using a least-squares support vector regression model in a fuzzy environment," *Information Sciences*, vol. 220, pp. 196–209, 2013.

[26] G. A. Grizzle and W. E. Klay, "Forecasting state sales tax revenues: Comparing the accuracy of different methods," *State & Local Government Review*, pp. 142–152, 1994.

[27] K. A. B. Telmo, K. V. Alviola, J. J. S. Desamparado, J. N. A. Cabigan, C. S. Santiago Jr., and R. A. A. Shimada, "Sentiment analysis of student evaluation for teachers using valence-aware dictionary and sentiment reasoner," *Journal of Interdisciplinary Perspectives*, vol. 2, no. 9, 1, 2024.

[28] D. C. Youvan, "Understanding sentiment analysis with VADER: A comprehensive overview and application," *AI and Data Science Journal*, vol. 15, no. 5, 2024.

[29] Y. Kaneko and K. Yada, "A deep learning approach for the prediction of retail store sales," in *Proc. 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, IEEE, Dec. 2016, pp. 531–537.

[30] V. Chowdhury. (2021). Superstore dataset. *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/vivek468/superstore-dataset-final

[31] D. Harvey, S. Leybourne, and P. Newbold, "Testing the equality of prediction mean squared errors," *International Journal of Forecasting*, vol. 13, no. 2, pp. 281–291, 1997.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).