Data Science in Finance and Economics

https://www.aimspress.com/journal/dsfe

DSFE, 6(2): 250–276.
DOI:10.3934/DSFE.2026009
Received: 21 August 2025
Revised: 25 February 2026
Accepted: 13 March 2026
Published: 06 May 2026

Research article

Sentiment-enhanced rice price forecasting under sparse social-media coverage: Evidence from Saudi rice imports

Saad Alqithami* **and Musaad Alzahrani**

Department of Computer Science, Al-Baha University, Albaha 65779, Saudi Arabia

* **Correspondence:** Email: salqithami@bu.edu.sa.

Abstract: This study examines whether carefully lagged social-media sentiment adds incremental forecast information to Saudi rice import prices once temperature-based climate variables and historical price dynamics are already included. We construct a monthly series for premium Basmati and standard Maza rice using customs-cleared price data, station-based temperature measures, and Arabic Twitter content labeled for sentiment with GPT-4. Because the retained tweet signal is sparse after relevance and engagement filtering, forecasts are evaluated in a constrained monthly setting using seasonal autoregressive integrated moving average with exogenous regressors (SARIMAX), linear regression (ordinary least squares), ridge regression, gradient boosting, and random forest under a time-ordered validation design with a held-out test period. Sentiment features improve forecast accuracy most consistently for Basmati, with smaller gains for Maza; the strongest overall results are achieved by transparent linear models and sentiment-enhanced SARIMAX rather than by the more complex tree ensembles. The findings indicate that large language model-assisted sentiment labeling can provide useful demand-side information when paired with disciplined feature engineering, while the magnitude of the benefit depends on commodity type, data coverage, and market structure.

Keywords: social media sentiment; multimodal forecasting; rice-price forecasting; SARIMAX; regularized regression; sparse social-media data

JEL Codes: C22, C53, Q11, Q13

1. Introduction

Accurate forecasts of agricultural commodity prices are essential for food-security planning, trade negotiations, market stability, and macroeconomic resilience, particularly in economies that depend heavily on imported staples (FAO, IFAD, UNICEF, WFP and WHO, 2023). Rice exemplifies this dependence: It is consumed daily by over half of the world's population, yet export production remains

geographically concentrated, leaving import-dependent markets vulnerable to volatility arising from production disruptions, logistical bottlenecks, and sudden demand shifts (Cai et al., 2019; Putra et al., 2021). For import-dependent countries such as Saudi Arabia, accurate forecasts of premium Basmati and standard Maza rice prices are therefore important for managing import costs, subsidy programs, inflation risks, and strategic food reserves.

Traditional econometric models, particularly seasonal autoregressive integrated moving average (SARIMA) and SARIMA with exogenous regressors (SARIMAX), remain prevalent in agricultural market forecasting due to their straightforward handling of periodicity and autoregressive dynamics (Box et al., 2015). Nevertheless, their linear architecture struggles to incorporate real-time, irregularly timed data and rapidly evolving market psychology. Recent advances in machine learning (ML), including gradient boosting, random forests, and regularized linear regression, have addressed these limitations by handling complex, non-linear interactions among multiple predictors (Breiman, 2001; Chen and Guestrin, 2016). In agricultural settings, Sari et al. (2024) compare optimized ML techniques for commodity-price prediction and show that careful model selection and tuning materially affect forecasting performance. Related multimodal work by An et al. (2024) further illustrates how textual signals can complement market variables in agricultural futures forecasting. These studies are more closely aligned with our setting, but they also underscore that predictive gains depend on data quality, feature engineering, and rigorous validation, especially in data-scarce environments (Liu, 2012).

Concurrent developments in natural language processing (NLP) have facilitated the extraction of consumer sentiment from social-media platforms, providing valuable real-time indicators of market expectations. Platforms like Twitter generate high-frequency sentiment signals that promptly reflect consumer responses to events such as export bans, policy announcements, or food-safety concerns (Alamah et al., 2024; Kim et al., 2017; Bollen et al., 2011). Transformer-based NLP models such as bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019), financial BERT (FinBERT) (Araci, 2019), and Arabic BERT (AraBERT) (Antoun et al., 2020) have significantly enhanced sentiment classification accuracy, yet their integration into agricultural price forecasting remains understudied. Recent agricultural forecasting studies also demonstrate the promise of fusing text-derived signals with market variables; for example, Wang and Liu (2025) integrate BERT-based topic identification and sentiment analysis into agricultural futures price prediction. While that work demonstrates the value of multimodal fusion, our study differs in its focus on Saudi rice imports, the joint use of econometric and ML regressors, and the monthly data frequency. The interaction between fast-moving consumer sentiment and slower-moving climatic supply signals, in particular, remains a key research gap.

Against this background, our empirical focus is narrower than a general claim that AI forecasting models outperform classical approaches. Specifically, we examine whether carefully lagged, GPT-4-labeled sentiment features add incremental forecast information beyond climate variables and price history in a sparse monthly setting. We address this question using Saudi rice import prices from January 2015 to January 2024 and a common feature set evaluated across SARIMAX, ordinary least squares (OLS), ridge regression, gradient boosting, and random forest.

1.1. Research focus

Previous studies have established the predictive utility of Twitter-based sentiment for financial markets and commodity-price dynamics (Bollen et al., 2011; Kim et al., 2017; Alamah et al., 2024). More

recent work by Bonato et al. (2024) shows that sentiment can also help predict volatility in agricultural commodities. What remains less clear is whether sentiment still adds useful forecast information once it is embedded in an operationally constrained setting: low-frequency official price data, sparse retained social-media coverage, and a need to align text with procurement lead times. Our focus is therefore deliberately narrow. We test whether a small set of carefully engineered sentiment variables provides incremental demand-side information beyond climate measures and lagged prices, and whether that incremental value differs across the more demand-sensitive Basmati segment and the more supply-driven Maza segment. We do not frame the paper as a general contest between "AI" and classical forecasting models, because the evidence presented here speaks more directly to feature value than to model-class supremacy.

1.2. Main empirical finding

After relevance and engagement filtering, only 69 tweets are retained across 109 months, so the empirical setting is constrained rather than data-rich. Within that setting, the clearest result is that lagged sentiment features improve forecasts most consistently for premium Basmati. Sentiment-enhanced SARIMAX reduces root mean square error (RMSE) by 25% for Basmati and by 18% for Maza, while the best overall fit in Table 4 is achieved by OLS and ridge rather than by the more complex tree ensembles. The main empirical message of the paper is therefore not that advanced AI models displace transparent baselines, but that carefully engineered multimodal features improve transparent forecasting models when data coverage is limited.

1.3. Contributions

This paper makes three specific contributions. First, it develops a reproducible pipeline for aligning sparse Arabic Twitter sentiment with monthly rice prices and climate variables while avoiding look-ahead leakage. Second, it shows that sentiment adds incremental demand-side information, with materially larger gains for Basmati than for Maza. Third, it demonstrates that in this constrained data setting, transparent learners—especially OLS, ridge, and sentiment-enhanced SARIMAX—extract that information more reliably than tree-based ensembles. The data and code used in this study are available in the public project repository.

The remainder of this paper proceeds as follows: Section 2 provides an overview of related literature; Section 3 describes the data and methodologies; Section 4 presents results; Section 5 discusses implications; and Section 6 concludes with policy recommendations and future research directions.

2. Literature review and background

Accurately forecasting agricultural commodity prices remains challenging due to the complex interplay among climate variability, trade policies, macroeconomic conditions, and consumer psychology (FAO, IFAD, UNICEF, WFP and WHO, 2023; Putra et al., 2021). In this study, we operationalize these drivers using a transparent monthly feature set—temperature variables, lagged prices, and social-media sentiment—because consistent trade-policy and macroeconomic control series are not available at the same frequency for the Saudi import context. Over recent years, researchers have increasingly integrated diverse data sources—including climate indicators, macroeconomic variables, and textual sentiment—with advanced forecasting techniques. This literature review synthesizes recent high-impact

studies across four key areas: traditional statistical methods, ML and deep learning (DL) advancements, sentiment analysis applications, and interdisciplinary data fusion.

2.1. Traditional statistical models for price forecasting

Autoregressive integrated moving average (ARIMA) and its seasonal variant SARIMA remain staples in agricultural forecasting due to their interpretability and effectiveness in modeling periodicity and autoregressive dynamics. Box et al. (2015) illustrate that SARIMA adequately captures stable seasonal patterns but struggles under external shocks. Recent studies confirm both the continuing value and the limits of linear time-series models: Yadav (2024) demonstrates that augmenting ARIMA models (ARIMAX) with climate and macroeconomic variables significantly improves global wheat price forecasts, while Putra et al. (2021) show that rice-price volatility remains tightly linked to local climate and macroeconomic conditions. Although Bayesian hierarchical models provide flexibility by pooling regional information (Gelman et al., 2013), they still face challenges when the data-generating process involves strong non-linear interactions and regime changes.

2.2. Machine-learning and deep-learning advances

ML techniques, particularly ensemble models such as random forests (Breiman, 2001) and extreme gradient boosting (XGBoost) (Chen and Guestrin, 2016), have grown popular in agricultural forecasting due to their ability to manage heterogeneous datasets and identify variable importance. Gradient-boosted decision trees (including XGBoost) are especially attractive for structured tabular forecasting tasks because they combine non-linear function approximation with regularization and scalable training (Chen and Guestrin, 2016). Recent comparative evidence by Sari et al. (2024) shows that optimized ML techniques can materially improve agricultural commodity price prediction, but the gains depend strongly on careful tuning and the market setting. DL models, notably convolutional-recurrent hybrids like convolutional neural network-long short term memory (CNN-LSTM) architectures, further extend predictive capabilities by capturing intricate temporal and spatial dependencies. Wang et al. (2023) report improved forecasts for U.S. corn and wheat prices by incorporating snow-pack and precipitation data into CNN–LSTM frameworks. Despite their predictive strengths, DL approaches such as LSTMs remain computationally demanding and opaque, posing barriers for deployment in resource-limited agricultural contexts (Kamilaris and Prenafeta-Boldú, 2018).

2.3. Sentiment analysis in commodity forecasting

Market psychology is increasingly observable in real time through social platforms and online news, enabling sentiment measures that proxy short-run demand expectations. The rise of high-frequency textual data from social media has therefore spurred the integration of sentiment analysis into commodity-price forecasting. Early evidence from financial markets (Bollen et al., 2011; Mittal and Goel, 2011) motivated analogous applications in agricultural settings. For example, Xu and Hsu (2022) show that sentiment extracted from agricultural news can reduce forecast errors for soybean and corn markets, and Ewald and Li (2024) combine FinBERT-based sentiment with CNN–LSTM architectures to improve salmon-price forecasts. Transformer models such as BERT (Devlin et al., 2019), AraBERT (Antoun et al., 2020), and large language models such as GPT-4 further expand this literature by making multilingual, context-aware sentiment extraction more feasible. Topic-aware pipelines such as BERTopic, coupled

with sentiment scoring, provide another route for extracting structured demand signals from noisy text streams (Wang and Liu, 2025; Reis Filho et al., 2022).

2.4. Interdisciplinary data-fusion approaches

State-of-the-art forecasting increasingly relies on integrating multiple data streams—most commonly climate signals and, in some settings, macroeconomic indicators and textual sentiment. For example, Cai et al. (2019) combine satellite and climate data with ML to improve agricultural outcome predictions, and Putra et al. (2021) link rice price volatility to localized climate and macroeconomic factors. Building on this broader fusion literature, the present study focuses on a deliberately parsimonious predictor set that can be reproduced at monthly frequency for Saudi imports: temperature variables, lagged prices, and Twitter-based sentiment. Macroeconomic series (e.g., exchange rates and CPI) are used only to deflate prices and are not included as forecasting regressors.

2.5. Summary of key literature

Tables 1 and 2 compare representative studies across methodologies, data sources, and principal findings.

Table 1. Focused literature review of price forecasting models.

Study	Methodology	Exogenous Data	Main Result
Yadav (2024)	ARIMAX vs. DL	Climate	ARIMAX best with exogenous inputs
Wang et al. (2023)	CNN–LSTM	SWE, climate	Hybrid lowers 15-wk RMSE
Ewald and Li (2024)	CNN–LSTM	News sentiment	Sentiment decreases RMSE by 12%
Xu and Hsu (2022)	Linear vs. QR	News sent., weather	Sentiment improves MAE
Reis Filho et al. (2022)	RF, SVR	Keyword sentiment	Text features aid corn/soy
Putra et al. (2021)	VAR-GARCH	Climate	75% variance explained

Table 2. Focused literature review on agricultural price forecasting.

Market/Commodity	Data Modalities	Model Family	Key Takeaway
General	Prices	Econometrics	Foundational time-series modeling
China (Corn)	Prices, Climate, Futures	Econometrics	Climate impacts volatility
Indonesia (Rice)	Prices, Climate, Macro	Econometrics	Local variables are crucial
India (Onion)	Prices, Weather	DL	DL outperforms traditional models
Stock Market	Sentiment, Prices	ML/DL	Sentiment predicts market changes
Egypt (Wheat)	Sentiment, Prices	NLP/Time-Series	Sentiment improves forecast accuracy
China (Pork)	Sentiment (News), Prices	NLP/DL	Topic modeling reveals price drivers

Note: A detailed literature review table is available in Supplementary Table S1.

Collectively, recent literature demonstrates the value of multimodal, integrated approaches over single-source forecasting models. However, evidence remains limited on whether sparse, text-derived demand signals improve monthly agricultural price forecasts once they are aligned with conventional covariates in an operational import setting. This paper addresses that gap by testing whether a small,

carefully lagged set of GPT-4-labeled sentiment features improves rice-price forecasts beyond climate variables and price history.

3. Methodology

3.1. Theoretical underpinnings

Price discovery in agricultural commodity markets emerges from the continuous interaction between supply-side fundamentals such as crop physiology, weather shocks, logistics costs, and trade policy, and demand-side expectations that are increasingly shaped in the digital public sphere. Classical equilibrium models trace back to Marshallian partial-equilibrium theory and later rational-expectations refinements, positing that spot prices converge to fundamentals when agents process information efficiently (Marshall, 2013). Empirical evidence, however, shows that commodities frequently deviate from cost-of-carry values during periods of information scarcity, export bans, or food-safety scares, implying that bounded rationality and sentiment-driven herding matter alongside storage arbitrage (Box et al., 2015; Bollen et al., 2011). Behavioral-finance research documents that retail investors' mood, proxied by language tone in news and social media, predicts short-horizon returns in equities and energy futures (Tetlock, 2007; Bollen et al., 2011); similar mechanisms plausibly operate in staple-food markets where household purchases can be advanced or delayed in response to perceived shortage risk. Integrating behavioral signals with climatic covariates and lagged prices therefore offers a richer representation of the price-formation process than any single data stream alone. In our implementation, we restrict the exogenous block to predictors that are available consistently at monthly frequency for the Saudi import context (temperature metrics and sentiment indices); macroeconomic series are used only for price deflation and are not fed into the forecasting models as regressors.

The present study operationalizes this fusion through a two-tier feature set. First, supply-side risk is proxied by monthly temperature averages and extremes together with lagged importer-price series that capture storage persistence. Second, demand-side information is modeled using engagement-weighted sentiment scores extracted by GPT-4 from Arabic tweets that mention rice-related keywords. GPT-4 was selected pragmatically for this labeling step because few-shot prompting allowed us to apply a single, fixed labeling protocol to mixed-dialect text without commodity-specific model fine-tuning. In this design, GPT-4 acts as a labeling instrument within the feature-construction pipeline rather than as the forecasting model itself. Sentiment is aggregated with a six-month trailing mean and lagged by four months to reflect the empirically motivated procurement-to-retail transmission window used in the empirical design, thereby avoiding look-ahead bias.

On the modeling side, we combine a SARIMAX specification, which is appealing for its transparent decomposition of autoregressive, seasonal, and exogenous blocks, with ML regressors that can capture higher-order interactions between climate and sentiment features. Gradient boosting and RF offer flexible function approximation, whereas ridge and OLS serve as low-variance baselines that reveal whether predictive gains stem from new information or from added model complexity. Because the sample is small and tweet coverage is sparse, the comparison is designed to test the incremental value of the sentiment block rather than to stage a winner-take-all contest among model classes. Hyper-parameter tuning uses expanding-window cross-validation to mimic real-time forecasting, and performance is assessed using RMSE, mean absolute error (MAE), and out-of-sample R^2_{oos} on the chronologically held-out test window (April 2022–January 2024), corresponding to the final 20% of observations under

an 80/20 time-ordered split. This statistical-NLP pipeline, summarized in Algorithm 1 and sketched conceptually in Figure 1, embeds behavioral-economic intuition into an econometric backbone and delivers a reproducible workflow suitable for operational decision support in grain-importing economies.

Algorithm 1: Per-model training and evaluation workflow for the reported experiments

Data: D_p : customs-cleared monthly prices;
D_c : monthly climate matrix;
D_t : tweet corpus;
M_{sent} : GPT-4 sentiment engine;
Result: Hold-out forecasts and error metrics for each model family under the climate-only and sentiment-enhanced specifications

(1) Data alignment and label construction
$P^{\log} \leftarrow \text{log\_deflate}(D_p)$; // convert to constant-2020 SAR and log-transform
$T \leftarrow \text{clean\_tweets}(D_t)$; // tokenize, de-noise, RTL fix
$S \leftarrow M_{\text{sent}}(T)$; // tweet-level polarity labels
$(AS, WS, N, I) \leftarrow \text{aggregate\_monthly}(S)$; // month-end aggregates and tweet count
$\bar{S}_m \leftarrow \text{lag\_agg}((AS, WS, N, I), \text{window} = 6, \text{lag} = 4)$; // history-only sentiment features
$C \leftarrow \text{aggregate\_monthly\_climate}(D_c)$; // monthly temperature levels and derived climate features

(2) Feature construction
$F_{\text{full}} \leftarrow [P^{\log}_{t-1..t-4};\ \bar{P}^{(6)}_{t-1}, \bar{P}^{(12)}_{t-1};\ C_t;\ \bar{S}_m;\ \text{seasonal dummies};\ \text{interactions}]$
For each expanding-window split, standardize the target using training-window statistics only and robust-scale the continuous non-SARIMAX predictors using the training-window median and interquartile range; leave binary indicators and month dummies unchanged
Define learner-specific inputs:
for $m \in \{\text{OLS, Ridge, GB, RF}\}$, let X^{base}_m be the climate + price-history subset of the robust-scaled full matrix and X^{sent}_m add the sentiment block
for m = SARIMAX, let X^{base}_m be the reduced climate exogenous block and X^{sent}_m add lagged weighted sentiment and its first difference

(3) Per-model estimation
foreach $m \in \{SARIMAX, GB, RF, Ridge, OLS\}$ **do**
  foreach $s \in \{base, sent\}$ **do**
    for $r = 1$ **to** R **do**
      $(X_{\text{train}}, y_{\text{train}}, X_{\text{val}}, y_{\text{val}}) \leftarrow \text{split\_expanding}(X^s_m, P^{\log}, r)$;
      $\theta^*_{m,s} \leftarrow \text{tune}(m, X_{\text{train}}, y_{\text{train}})$; // orders or hyper-parameters
      $\widehat{y}^{(r)}_{m,s} \leftarrow m_{\theta^*_{m,s}}(X_{\text{val}})$;
    end
    retrain $m_{\theta^*_{m,s}}$ on the full training window;
    generate hold-out forecasts $\widehat{y}_{m,s,t}$ for all $t \in \mathcal{T}_{\text{test}}$;
    compute $\text{RMSE}_{m,s}$, $\text{MAE}_{m,s}$, and $R^2_{\text{oos},m,s}$ on $\mathcal{T}_{\text{test}}$;
  end
end
return $\{\widehat{y}_{m,s,t}, \text{RMSE}_{m,s}, \text{MAE}_{m,s}, R^2_{\text{oos},m,s}\}$

3.2. Dataset description

Accurate evaluation of the proposed forecasting pipeline requires a dataset that spans both the physical drivers of rice supply and the behavioral signals that shape short-run demand. To that end, we build a monthly panel covering January 2015–January 2024 in which customs-cleared import prices for the two target varieties are merged with weather measurements from coastal meteorological stations and sentiment scores extracted from Arabic tweets retrieved via a bilingual keyword query over the same calendar window. This aligned time span ensures a fair comparison between models with and without sentiment features. Price and climate series have complete coverage for all 109 months; sentiment coverage is sparse after relevance and engagement filtering, and we therefore retain a balanced panel by explicitly tracking monthly tweet volume and availability (Section 3.2.2 and Supplementary Table S2). All variables are synchronized to month-end, transformed under the common preprocessing conventions summarized below, and augmented with lagged, rolling, and interaction terms that encode seasonality and delayed responses. The resulting panel contains 109 monthly observations, and the final design matrix used in our experiments comprises 38 engineered predictors (Table 3) for each target series, including climate lags, price lags/rolling averages, sentiment lags/rolling averages, tweet availability/volume indicators, and month-of-year dummies.

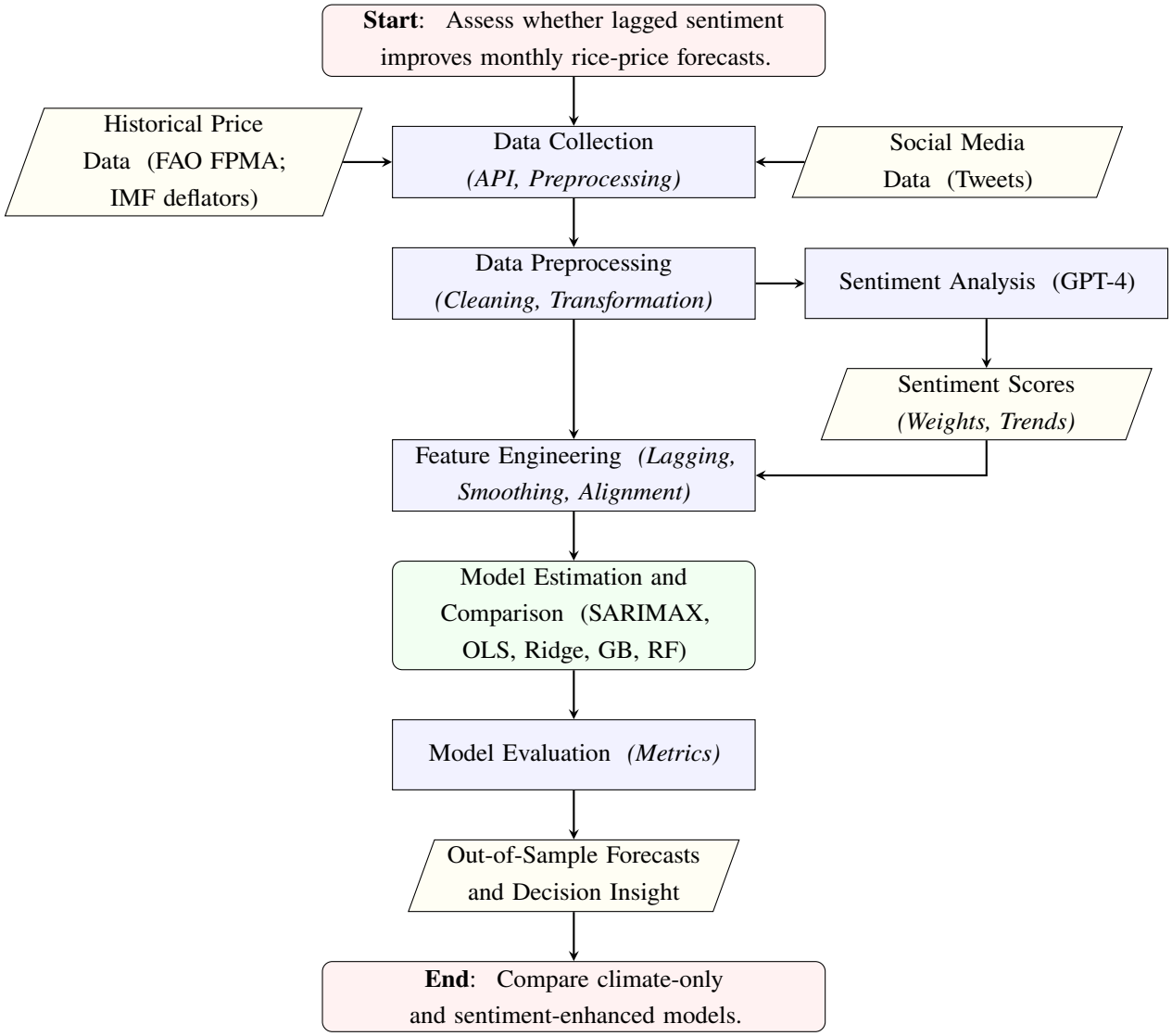

Figure 1. Workflow of the sentiment-enhanced rice-price forecasting pipeline. Sparse tweet-derived sentiment is aligned with monthly climate and price data, transformed into lagged features, and then compared across transparent statistical models and flexible ML benchmarks.

3.2.1. Agricultural price data

Monthly cost, insurance, and freight (CIF) import prices for premium Basmati and standard Maza rice were downloaded from the FAO Food Price Monitoring and Analysis (FPMA) Tool (Food and Agriculture Organization of the United Nations, 2026), using the monthly Saudi import series for January 2015–January 2024. FPMA served as the primary price source. Saudi Ministry of Commerce and Jeddah Chamber wholesale bulletins were consulted only as qualitative spot checks during data cleaning and were not merged mechanically into the released monthly series.* Nominal prices were converted from U.S. dollars to Saudi riyal using International Financial Statistics (IFS) monthly midpoint

*Saudi Ministry of Commerce portal: `https://mc.gov.sa/en/`; Jeddah Chamber circulars and publications portal: `https://www.jcci.org.sa/en/`; accessed 11 April 2026.

Table 3. Engineered predictor pool used to construct learner-specific inputs for a given rice variety.

Predictor block	Variables included
Temperature levels	TAVG, TMAX, TMIN (monthly means)
Temperature dynamics	1- and 2-month lags of each temperature metric; 12-month trailing mean of TAVG ($\overline{TAVG}_t^{(12)}$)
Price history (target series)	log-deflated target-price lags at $t-1$, $t-2$, and $t-4$; 6- and 12-month trailing moving averages computed from past observations only
Sentiment block	Monthly AS_t and WS_t; lags 1, 2, and 4; 6-month trailing mean of WS_t; tweet indicator $I_t = 1[N_t > 0]$ and volume $\log(1 + N_t)$; all sentiment predictors shifted by 4 months to align with procurement lead times
Seasonality controls	11 month-of-year dummies (`Feb`–`Dec`; `Jan` reference)
Interaction term	$WS_{t-4} \times TAVG_t^{\Delta}$ with $TAVG_t^{\Delta} = TAVG_t - \overline{TAVG}_t^{(12)}$

Note: OLS, ridge, gradient boosting, and RF use the full engineered matrix. SARIMAX uses a reduced exogenous block consisting of contemporaneous temperature levels, the 12-month trailing mean of TAVG, the four-month-lagged weighted-sentiment series, and its first difference; serial dependence in the target is captured by the ARIMA terms rather than by explicit lagged-price regressors.

exchange-rate series and then deflated to constant 2020 SAR with the CPI of the International Monetary Fund (2026). Log-transformation stabilized the residual variance. The resulting target series was standardized within each estimation window, as summarized in the common preprocessing paragraph below. The final dataset contains 109 monthly observations (January 2015–January 2024) for each variety. This aligned window ensures a fair and consistent comparison between models with and without sentiment features. The cleaned price series serve as the dependent variable for all models. For OLS, ridge, gradient boosting, and random forest, lagged price terms are entered explicitly as predictors, whereas SARIMAX captures serial dependence through its autoregressive and seasonal structure.

3.2.2. Consumer sentiment data

Tweets were retrieved via the then-current Twitter Academic API[†] using a bilingual query that combined Arabic-script keywords (e.g., أرز, بسمتي, مزة) with common English translations (e.g., "Basmati", "Maza rice") and variety-specific brand names used in Saudi retail contexts. To reduce noise, we (i) excluded common promotional terms (e.g., "discount", "coupon"), (ii) removed retweets, and (iii) retained only tweets with nonzero engagement. Language identification was performed using fastText (Joulin et al., 2017); we retained only tweets classified as Arabic with probability above 0.8. English query terms were used to improve recall for code-switched or Latin-script mentions occurring in otherwise Arabic Saudi discourse, while downstream sentiment construction focused on Arabic tweets for linguistic consistency.

Temporal coverage and tweet volume. Tweets were collected for the full study window (1 January 2015–31 January 2024). After applying the relevance, de-duplication, and engagement filters described above, the retained tweet volume is sparse: Across the 109 monthly observations, 39 months contain at least one retained tweet (total retained tweets = 69; max $N_t = 4$; median $N_t = 0$). Supplementary Table S2

[†]Current official developer documentation is maintained on the X platform: `https://docs.x.com/x-api/introduction` and `https://docs.x.com/use-cases/do-research`; accessed 11 April 2026.

reports N_t for every month, together with the resulting AS_t and WS_t aggregates. Sentiment-enhanced models are trained and evaluated on the full aligned January 2015-January 2024 panel; no months are dropped, ensuring that baseline and sentiment models use identical train/test windows. This is therefore a constrained data setting in which leakage control and feature design are at least as important as the choice of learner.

A GPT-4 classifier then assigned each tweet one of three polarity classes—positive, neutral, or negative. For aggregation, we map these classes to a numeric polarity score $\text{polarity}_i \in \{-1, 0, +1\}$ (negative, neutral, positive).

GPT-4 invocation and label audit. GPT-4 was used as a tweet-level labeling instrument rather than as a forecasting model. Each retained tweet was passed to the API one tweet per request using the fixed prompt template reproduced in the Supplementary Material (GPT-4 labeling protocol), with the model identifier recorded in the analysis scripts as `gpt-4`, temperature set to 0.2, and maximum output length set to 5 tokens. The model was instructed to return exactly one of three labels—positive, neutral, or negative—which we then mapped to −1, 0, and +1, respectively. To audit label quality, a separate 500-tweet holdout sample drawn from the pre-filter query corpus (i.e., before the engagement filter was applied) was independently annotated by the two authors without access to GPT-4 outputs. Disagreements were resolved by discussion to form an adjudicated human reference. The reported 89% agreement and Cohen's $\kappa = 0.81$ refer to agreement between GPT-4 and this adjudicated human label set; the holdout sample was not used in model fitting or in the monthly aggregates.

Two monthly sentiment aggregates feed the forecasting models. Let N_t denote the number of retained tweets in month t:

1. **Average sentiment (AS).** The month-t mean polarity,

$$AS_t = \frac{1}{N_t} \sum_{i \in t} \text{polarity}_i,$$

   which summarizes the prevailing tone of the online conversation on a bounded scale. Equivalently, $AS_t = p_t^+ - p_t^-$ where p_t^+ and p_t^- are the monthly shares of positive and negative tweets (neutral contributes zero).
2. **Weighted sentiment (WS).** Engagement for tweet i is defined as $\text{eng}_i = \text{retweets}_i + \text{likes}_i$. To temper viral outliers while preserving salience, we apply a log transform $\widetilde{\text{eng}}_i = \log(1 + \text{eng}_i)$ and rescale polarity as

$$s_i^{\text{w}} = \text{polarity}_i \left[1 + \widetilde{\text{eng}}_i\right].$$

   The month-t WS is the arithmetic mean of s_i^{w} across tweets in t:

$$WS_t = \frac{1}{N_t} \sum_{i \in t} s_i^{\text{w}},$$

   preserving polarity while assigning greater influence to more widely shared posts.

When $N_t = 0$, the aggregates AS_t and WS_t are undefined. To preserve a balanced monthly panel without conflating missingness with neutral sentiment, we record $AS_t = 0$ and $WS_t = 0$ only as placeholders in the released dataset and include two additional covariates: (i) a tweet-availability

indicator $I_t = 1[N_t > 0]$ and (ii) tweet volume $\log(1 + N_t)$. These variables allow all learners to distinguish true neutral tone (tweets present but net polarity near zero) from months with no observed sentiment signal. All sentiment-derived predictors (the AS/WS levels, I_t, and $\log(1 + N_t)$) are shifted by four months (i.e., information observed at t–4 is used to predict prices at t) to approximate procurement-to-retail transmission in the Saudi import setting. We additionally include 1- and 2-month lags of the shifted series to capture short-run persistence, and we smooth weighted sentiment using a six-month trailing moving average computed using past information only (e.g., $\text{MA6}(WS)_{t-4} = \frac{1}{6}\sum_{j=0}^{5} WS_{t-4-j}$). Finally, we interact lagged weighted sentiment with a temperature anomaly term $TAVG_t^{\Delta} = TAVG_t - \overline{TAVG}_t^{(12)}$, where $\overline{TAVG}_t^{(12)}$ is the 12-month trailing mean of $TAVG$, to test whether demand-side surges amplify price impacts during periods of elevated heat stress.

All sentiment features are aggregated at month-end to match the temporal resolution of price and climate data and then passed through the learner-specific preprocessing step summarized in the common scaling paragraph below. Exploratory analysis suggests that lagged weighted sentiment co-varies more strongly with the premium Basmati series than with Maza, consistent with the larger out-of-sample forecasting gains reported in Section 4.

3.2.3. Climatic variables

Weather shocks are one of the principal supply-side drivers of rice price volatility (Lobell, 2013). Monthly climate observations for Saudi Arabia's three main grain-receiving regions (i.e., Jeddah, Dammam, and Jizan) were downloaded from NOAA's National Centers for Environmental Information (NCEI) Daily Summaries interface (National Centers for Environmental Information, National Oceanic and Atmospheric Administration, 2026) for the January 2015–January 2024 window. For each station, we extracted daily maximum, minimum and mean air temperatures (i.e., TMAX, TMIN, TAVG). Daily gaps shorter than seven days were infilled with linear interpolation, while longer gaps were completed with inverse-distance-weighted averages from the nearest operational station. The cleaned daily series were aggregated to monthly means and then merged across stations using population-weighted averages to approximate the temperature profile faced by the bulk of Saudi consumers. These monthly climate series then entered the common feature-engineering and learner-specific preprocessing pipeline described below.

3.2.4. Temporal dependencies

Time-series structure is captured through a systematic set of lags, rolling windows, and seasonal dummies. For OLS, ridge, gradient boosting, and RF, monthly rice prices enter as explicit predictors with first-, second-, and fourth-order lags, together with six- and 12-month trailing moving averages computed from past observations only, i.e., over $t-1, \cdots, t-6$ and $t-1, \cdots, t-12$, to represent short-, medium-, and longer-run persistence (Hyndman and Athanasopoulos, 2021). In SARIMAX, comparable persistence is handled by the autoregressive and seasonal components rather than by entering the full lagged-price block as exogenous regressors. Sentiment metrics receive one-, two-, and four-month lags so that delayed demand effects can be learned in parallel with supply cues, and a six-month rolling mean of weighted sentiment is included to smooth the bursty nature of social-media activity. Climate variables (TAVG, TMAX, TMIN) are enhanced with one- and two-month lags to reflect agronomic response times and a 12-month rolling mean that proxies cumulative heat stress (Cai et al., 2019).

To model deterministic Gregorian seasonality, we include 11 month-of-year indicator variables (Feb–

Dec; Jan is the reference category) alongside an intercept. This encoding avoids perfect multicollinearity in linear models while allowing them to reproduce recurring intra-year patterns visible in both rice series (e.g., summer shipping constraints and year-end contracting cycles). Note that religious holidays such as Ramadan rotate through the Gregorian calendar; their effects are therefore only partially captured by fixed month dummies and remain a limitation of monthly seasonality controls. All engineered features are aligned to month-end. The learner-specific preprocessing and scaling conventions are summarized below.

Common preprocessing and scaling conventions. All raw series were first aligned to month-end. The target price series was converted to constant 2020 SAR, log-transformed, and standardized within each estimation window; the reported RMSE and MAE are therefore in standardized log-price units. Climate and sentiment variables were engineered in their transformed monthly units (levels, lags, trailing means, indicators, and interactions). For OLS, ridge, gradient boosting, and RF, the final continuous predictor matrix was robust-scaled using the training-sample median and interquartile range, while binary indicators and month-of-year dummies were left unchanged. We used this common transformation for pipeline consistency and to reduce the influence of outliers in sparse sentiment features, not because tree-based split criteria require standardized inputs. SARIMAX used the same log-deflated target but a reduced exogenous block, entered in transformed units as described in Section 3.3.1. All scaling parameters were estimated on the training portion of each split only to avoid look-ahead leakage.

3.3. Model development

To quantify the incremental value of consumer sentiment alongside more established supply-side drivers, we adopt a multi-model strategy that spans the classical-to-ML spectrum. OLS, ridge, gradient boosting, and RF are estimated on the same fully engineered feature matrix described in Section 3.2 after the common preprocessing step summarized above. SARIMAX, by contrast, uses a reduced exogenous block chosen for stability and identification in a 109-month monthly sample. This distinction keeps the regression and tree-based learners on a common design matrix while allowing the seasonal time-series model to remain parsimonious. The modeling workflow proceeds in three uniform stages: (1) An expanding-window cross-validation scheme generates pseudo out-of-sample splits that mimic a real-time forecasting environment, (2) hyper-parameters are optimized within each split through either grid or randomized search guided by a primary error metric (RMSE), and (3) the tuned models are re-estimated on the full training window and evaluated on a fixed hold-out window corresponding to the final 20% of observations (April 2022–January 2024). Across all learners, we report RMSE, MAE, and R^2_{oos} to facilitate comparison with prior agricultural-forecasting studies, and we extract feature-importance measures where available to assess the relative contribution of sentiment signals. The following subsections detail the specification, tuning protocol, and diagnostic checks for each model family.

3.3.1. Seasonal ARIMA with exogenous regressors (SARIMAX) model

SARIMAX provides the classical benchmark onto which all ML gains are gauged. We first applied the augmented-Dickey–Fuller test to both Basmati and Maza price series, and a single non-seasonal difference rendered each series stationary at the 5% level. Candidate orders up to $p, q \leq 3$ and seasonal orders $P, Q \leq 2$ ($s = 12$) were evaluated in a stepwise grid search that minimized Akaike's information criterion (Akaike, 1974), leading to SARIMAX$(1, 1, 1)\times(0, 1, 1)_{12}$ for Basmati and SARIMAX$(0, 1, 2)\times(1, 1, 1)_{12}$

for Maza. Unlike the regression and tree-based learners, SARIMAX is not estimated on the full 38-variable design matrix. Doing so in a 109-month monthly sample would risk over-parameterization and unstable seasonal estimates, so we use a reduced exogenous specification chosen for stability and identification. The climate-only SARIMAX includes contemporaneous temperature levels (TAVG, TMAX, TMIN) and the 12-month rolling mean of TAVG; the sentiment-augmented variant adds the four-month-lagged weighted-sentiment series and its first difference to capture both level and shock effects. Residual adequacy checks did not indicate remaining serial correlation at conventional significance levels, supporting the final specifications. By explicitly decomposing deterministic seasonality and embedding a parsimonious external block, this SARIMAX design furnishes a transparent baseline against which the incremental value of the sentiment variables and the non-linear learners is judged in Section 4.

3.3.2. Gradient boosting machine (GBM)

To accommodate the non-linear interaction of weather anomalies, lagged prices, and sentiment shocks, we implemented Friedman's gradient-boosting framework with regression trees as base learners (Friedman, 2001). Boosting sequentially fits shallow trees to the residuals of the current ensemble, thereby reducing bias while controlling variance through shrinkage and subsampling. Hyper-parameters were selected with a randomized grid over 300 draws: learning rate $\{0.01, 0.05, 0.1, 0.2\}$, maximum tree depth $\{3, 4, 5, 6\}$, subsample ratio $\{0.6, 0.8, 1.0\}$, and estimator count up to 600 trees. Each draw was evaluated with an expanding-window, five-fold time-series split so that training always preceded validation in time. Early stopping with a patience of 50 rounds halted training if validation RMSE failed to improve, preventing unnecessary complexity.

The best Basmati configuration converged at a learning rate of 0.05, depth of five, subsample of 0.8, and 420 trees; the optimal Maza model required fewer iterations (depth four, 310 trees) but a slightly higher learning rate (0.10). After training, permutation importance was computed on the hold-out set to gauge predictor influence. Weighted sentiment and its four-month lag ranked within the top five variables for Basmati, whereas both sentiment terms fell outside the top 10 for Maza—corroborating the smaller error reduction reported in Section 4. As with the other non-SARIMAX learners, GBM was fitted on the common robust-scaled feature matrix described above. The scaling step was used for pipeline consistency and to damp the influence of outliers in sparse sentiment features, not because tree-splitting rules depend on feature standardization.

3.3.3. Linear regression and ridge regression

To provide transparent baselines, we fitted linear regression (OLS) and ridge regression models to the common robust-scaled feature matrix described above. Because the same transformation is applied to all continuous predictors, coefficient magnitudes are directly comparable within a given fitted model. The OLS specification yields unbiased point estimates under the classical Gauss–Markov assumptions and therefore serves as a reference for gauging the incremental value of non-linear learners. Ridge regression adds an L_2 penalty $\lambda \sum_j \beta_j^2$ to the OLS loss function, shrinking coefficients toward zero and mitigating the variance inflation caused by multicollinearity among highly correlated lags and rolling averages. The penalty parameter λ was chosen on a logarithmic grid $10^{-3} \leq \lambda \leq 10^2$ via leave-one-year-out time-series cross-validation; the optimal value for Basmati was $\lambda = 0.3$ and for Maza $\lambda = 0.5$. Both linear models

train in milliseconds, allowing rapid sensitivity checks and easy interpretation of individual effect sizes—advantages that complement the higher accuracy but lower transparency of the ensemble methods.

3.3.4. Random forest

An RF was incorporated to capture complex non-linearities while remaining less sensitive to hyper-parameter choices than boosting methods (Breiman, 2001). Each forest was grown on 1,000 bootstrap samples drawn with replacement from the training window; at every split, a random subset of $\sqrt{p}$ predictors was considered, where p denotes the number of learner-specific predictors, a strategy that decorrelates individual trees and improves generalization. Randomized hyper-parameter search over 200 draws tuned the maximum tree depth (5-to-20), minimum samples per split (2-to-10), and the maximum number of leaf nodes (50-to-300), using an expanding-window, five-fold time-series cross-validation scheme identical to that employed for GBM. Early experiments showed that deeper trees (more than 15 levels) offered no RMSE gain but increased variance, so the depth was capped at 12 for both rice varieties. Out-of-bag error provided an internal validation that closely tracked the rolling hold-out RMSE, confirming good bias-variance balance. Permutation importance was extracted on the test set to gauge predictor relevance: Temperature metrics dominated for Maza, whereas the four-month-lagged weighted-sentiment feature ranked fourth for Basmati, aligning with the gradient-boosting interpretation and reinforcing the variety-specific value of demand-side information. As with GBM, the forest was estimated on the common robust-scaled feature matrix for preprocessing consistency; the forest itself does not rely on scale normalization for split construction.

3.3.5. Rationale for algorithm selection

The modeling portfolio was intentionally diversified to probe whether demand-side sentiment delivers predictive power under different statistical assumptions and functional forms. Each algorithm was chosen for a complementary theoretical strength as summarized below.

- SARIMAX provides a likelihood-based benchmark rooted in classical time-series econometrics. Its explicit seasonal and autoregressive structure yields interpretable coefficients and permits formal diagnostic testing; the model therefore anchors the empirical analysis and establishes a baseline for gauging the marginal value of sentiment variables.
- Linear regression (OLS) and ridge regression provide readily interpretable marginal associations after scaling, making coefficient signs and magnitudes easy to communicate to policy stakeholders. Ridge's L_2-penalty stabilizes estimates in the presence of multicollinearity induced by multiple lags and rolling averages, providing a low-variance linear yardstick against which more complex learners can be judged.
- RF combines bootstrap aggregation with random feature selection, producing a variance-reduced ensemble that is robust to noisy predictors and naturally ranks variable importance. This property is valuable for disentangling the relative influence of climate, lagged prices, and sentiment in a multidimensional setting.
- GBM incrementally fits weak learners to residual errors, capturing high-order, non-linear interactions that linear models and bagged trees may miss. Shrinkage and subsampling act as built-in regularizers, making GBM a strong candidate when predictor interactions are complex yet sample size is limited.

Evaluating these model families under both climate-only and climate-plus-sentiment configurations allows the study to isolate the incremental information content of social-media mood while controlling for each model's native capacity to fit autoregressive and climatic structure. This triangulated approach mitigates the risk that any single model's bias or variance profile drives the final conclusions, thereby providing a more reliable assessment of sentiment's role in Saudi rice-price dynamics.

3.4. Integration of sentiment analysis

Economic theory holds that commodity prices are driven not only by physical supply shocks but also by expectations that form in the public sphere and influence short-run demand. Traditional agricultural forecasting frameworks rarely observe these expectation shifts directly; instead they infer them from residual forecast errors. Social-media platforms, however, provide a contemporaneous record of consumer reaction to news about export bans, food-safety rumors, or festival demand surges. Among these venues, Twitter is widely used in the Gulf context and provides programmatic access to public discourse (subject to platform policies), while its hashtag and keyword conventions help identify product varieties and brand names. By mining Twitter discourse and converting it into structured sentiment indicators, we aim to inject a forward-looking, demand-side signal into rice-price models and thereby test whether online mood swings contain predictive information beyond what is available from lagged prices and weather records alone.

3.4.1. Evaluation metrics

Model quality was assessed exclusively on the chronologically held-out test window (April 2022–January 2024), corresponding to the final 20% of observations under the 80/20 time-ordered split. Three complementary error statistics were reported:

- Mean absolute error (MAE):

$$\mathrm{MAE} = \frac{1}{T}\sum_{t=1}^{T} |y_t - \widehat{y}_t|$$

  MAE conveys the typical size of an out-of-sample deviation in standardized log-price units (because the dependent variable is log-transformed and z-scored prior to modeling), facilitating like-for-like comparisons across learners and specifications.
- Root mean square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2}$$

  Because the squaring operator penalizes large misses, RMSE highlights tail-risk performance which is crucial for procurement decisions that are sensitive to price spikes.
- Out-of-sample R^2_{oos}:

$$R^2_{\mathrm{oos}} = 1 - \frac{\sum_t (y_t - \widehat{y}_t)^2}{\sum_t (y_t - \bar{y}_{\mathrm{train}})^2}.$$

  Using the training-set mean $\bar{y}_{\mathrm{train}}$ in the denominator avoids the optimistic bias that can occur when the test-set mean is substituted (Tashman, 2000). A positive R^2_{oos} therefore indicates that the model beats a naïve mean-forecast benchmark.

Together, these evaluation protocols provide a balanced view of point accuracy, tail risk, and explanatory power, allowing a nuanced comparison of linear baselines, tree ensembles, and sentiment-augmented time-series models.

4. Results and analysis

We evaluate the five forecasting models introduced in Section 3.3 under two information sets. The baseline specification contains temperature variables (TAVG, TMAX, TMIN) and lagged prices only, whereas the *enhanced* specification augments those predictors with engagement-weighted GPT-4 sentiment features. All models are trained on the first 80% of the time-ordered data (January 2015–March 2022) and evaluated on the remaining 20% chronologically held-out window (April 2022–January 2024). This design ensures direct comparability across models. The emphasis is on the held-out monthly forecast path rather than on a separate horizon-by-horizon comparison. Out-of-sample accuracy is summarized with RMSE, MAE, and R^2_{oos}, so that lower RMSE/MAE and higher R^2_{oos} indicate superior performance.

4.1. Evaluation overview

Predictive accuracy on the chronologically held-out test window (April 2022–January 2024) is gauged with three complementary statistics: MAE, RMSE, and out-of-sample R^2_{oos}. MAE conveys the typical deviation in standardized log-price units, RMSE places extra weight on large misses, and R^2_{oos} reports the share of variance explained relative to a naive-mean benchmark. Results for all five learners are summarized in Table 4, which makes two patterns clear. First, sentiment improves forecasts more consistently for the demand-sensitive Basmati series than for Maza. Second, the strongest overall results are obtained by OLS and ridge, with SARIMAX also benefiting materially from sentiment, whereas the tree ensembles remain weak.

For Basmati, sentiment-enhanced SARIMAX reduces RMSE by roughly 25% and raises R^2_{oos} from 0.317 to 0.617, while OLS achieves the best overall fit ($R^2_{\text{oos}} = 0.957$). For Maza, the lift is smaller but still material: SARIMAX error contracts by 18%, and R^2_{oos} improves from 0.199 to 0.457. The empirical takeaway is therefore about the value of disciplined external features in a constrained data setting, not about complex model classes displacing transparent baselines.

4.2. Baseline vs. enhanced models

Figures 2 and 3 summarize the non-SARIMAX learners row by row. In each row, the left panel overlays the held-out actual series with the climate-only and sentiment-enhanced prediction paths, the middle panel shows the residual scatter and residual histogram for the sentiment-enhanced specification, and the right panel reports feature-importance summaries. The figures therefore complement Table 4 by showing how closely each learner tracks the held-out path, how dispersed the residuals remain, and which engineered predictors carry the most weight.

Basmati (Figure 2): For Basmati, the linear models track the held-out series much more closely than the tree ensembles, and the sentiment-enhanced paths generally sit closer to the realized price path than their climate-only counterparts. The residual panels for ridge and OLS are tighter and more centered

than those for GB and RF. In the feature-importance summaries, sentiment-related variables appear among the more influential predictors for Basmati, consistent with the sizeable error reductions reported in Table 4.

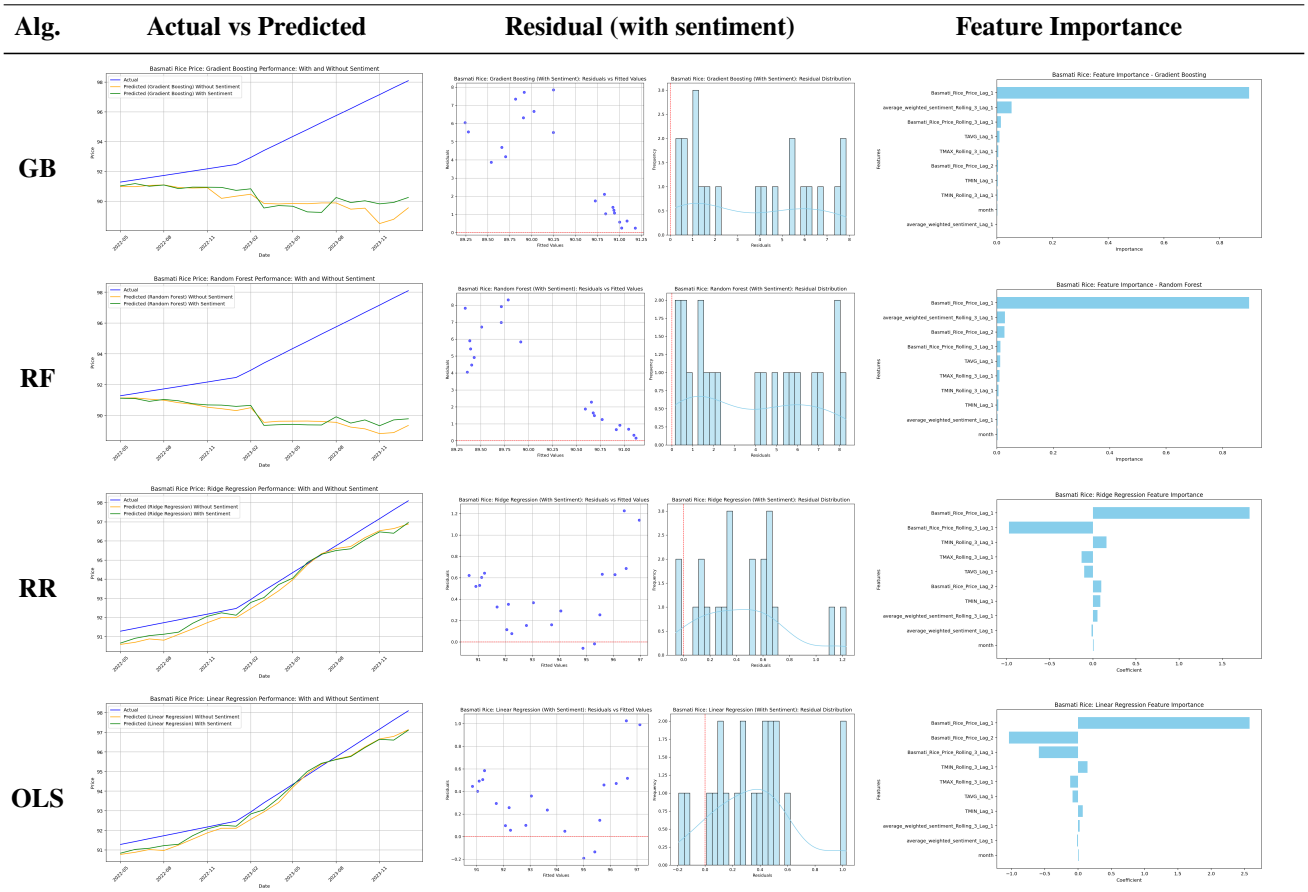

Figure 2. Basmati results for gradient boosting (GB), random forest (RF), ridge regression (RR), and linear regression (OLS): held-out actual and predicted paths under the climate-only and sentiment-enhanced specifications (left), residual scatter and histogram for the sentiment-enhanced specification (middle), and feature-importance summaries (right).

Maza (Figure 3): For Maza, the sentiment-enhanced paths show only modest improvement over the climate-only versions, and the gap between the linear models and the tree ensembles remains evident. Residual spreads are generally wider than for Basmati, and the importance rankings are led primarily by temperature and lagged-price variables, with sentiment appearing lower in the ordering. This visual pattern matches the smaller quantitative gains for Maza in Table 4.

Cross-model patterns: Across both varieties, OLS and ridge offer the closest held-out tracking, while GB and RF remain comparatively unstable in this small-sample setting. Figures 2 and 3 therefore reinforce the main numerical message of Table 4: Sentiment is more useful for the demand-sensitive Basmati series, and the clearest gains are realized in the more transparent model classes.

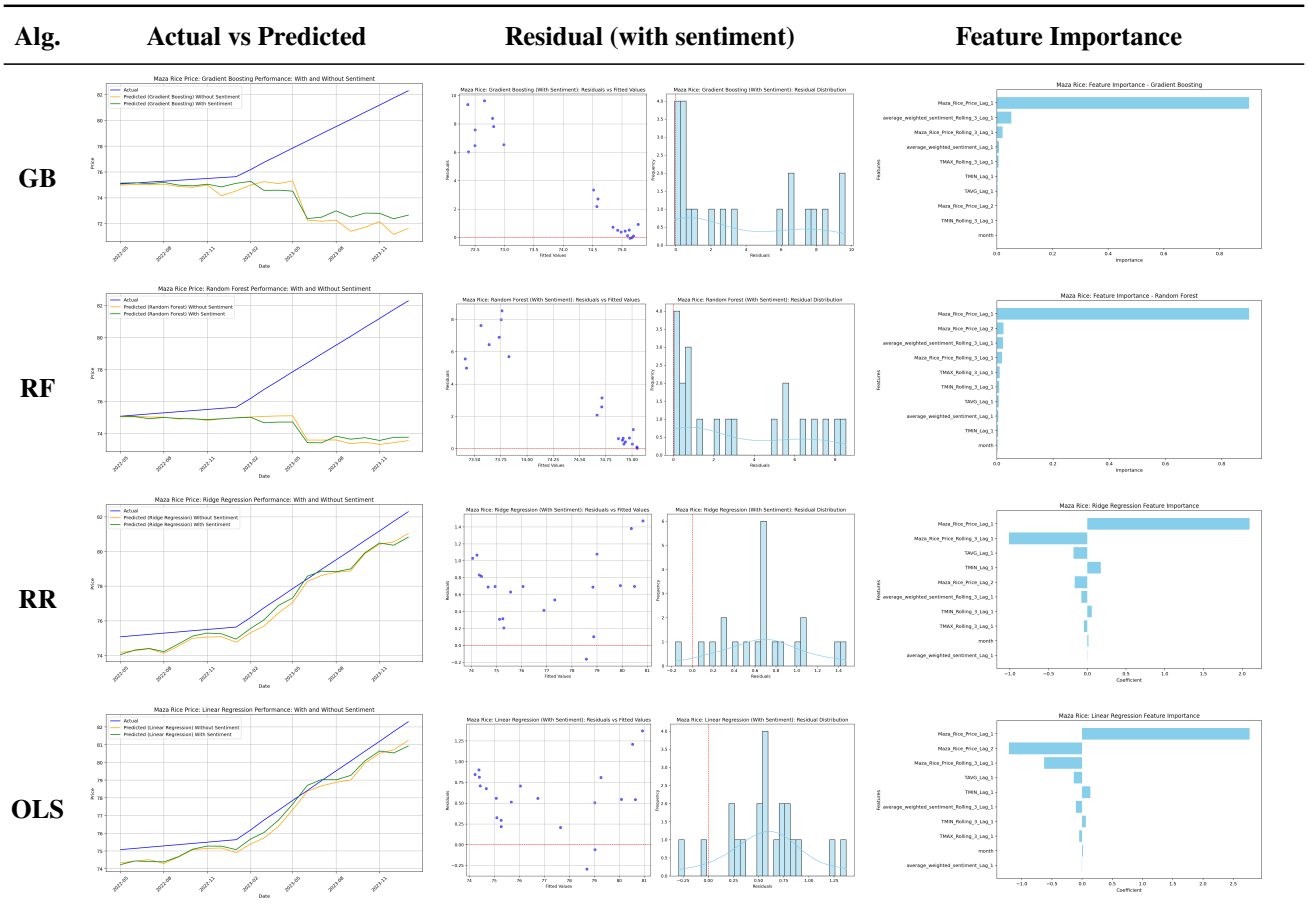

Figure 3. Maza results for gradient boosting (GB), random forest (RF), ridge regression (RR), and linear regression (OLS): held-out actual and predicted paths under the climate-only and sentiment-enhanced specifications (left), residual scatter and histogram for the sentiment-enhanced specification (middle), and feature-importance summaries (right).

4.3. Impact of sentiment integration

Figure 4 isolates the SARIMAX comparison and is the only results figure that includes that model. The left panels compare the held-out actual series with the climate-only and sentiment-enhanced SARIMAX paths, the middle panels show residual scatter and residual histograms for the sentiment-enhanced specification, and the right panels summarize the relative importance of the exogenous terms.

For Basmati, the sentiment-augmented SARIMAX path follows the held-out series more closely than the climate-only path, consistent with the reduction in RMSE from 1.81 to 1.36 and the rise in R^2_{oos} from 0.317 to 0.617 in Table 4. The corresponding residual distribution is also tighter and more centered. For Maza, the improvement is smaller: The sentiment-augmented path still outperforms the climate-only benchmark, but the gap is narrower, matching the smaller gain from 2.16 to 1.78 in RMSE.

Table 4. Out-of-sample forecasting performance with and without sentiment features (reported on the held-out test window; lower RMSE/MAE and higher R^2_{oos} indicate better performance).

Rice Type	Model	Without Sentiment			With Sentiment		
		RMSE	MAE	R^2_{oos}	RMSE	MAE	R^2_{oos}
Basmati	SARIMAX	1.812	1.425	0.317	1.357	1.080	0.617
	Gradient Boosting	4.790	3.830	−3.768	4.501	3.621	−3.210
	Random Forest	4.884	3.920	−3.958	4.694	3.793	−3.580
	Ridge Regression	0.622	0.555	0.920	0.551	0.448	0.937
	Linear Regression (OLS)	0.497	0.441	0.949	0.456	0.372	0.957
Maza	SARIMAX	2.158	1.624	0.199	1.777	1.347	0.457
	Gradient Boosting	5.424	3.822	−4.061	4.958	3.516	-3.229
	Random Forest	4.423	3.158	−2.365	4.346	3.160	−2.249
	Ridge Regression	0.848	0.798	0.876	0.782	0.692	0.895
	Linear Regression (OLS)	0.738	0.686	0.906	0.682	0.603	0.920

Note: Negative R^2_{oos} indicates performance below the naive mean benchmark on the test window.

The rightmost panels in Figure 4 also show a sharper role for sentiment in Basmati than in Maza: Weighted sentiment ranks more prominently among the exogenous terms for Basmati, whereas temperature variables remain more dominant for Maza. Taken together with the OLS and ridge results in Table 4, the SARIMAX visuals support the same controlled conclusion as the broader model comparison: A compact set of lagged sentiment measures adds useful forecast information, but most clearly for the demand-sensitive Basmati series.

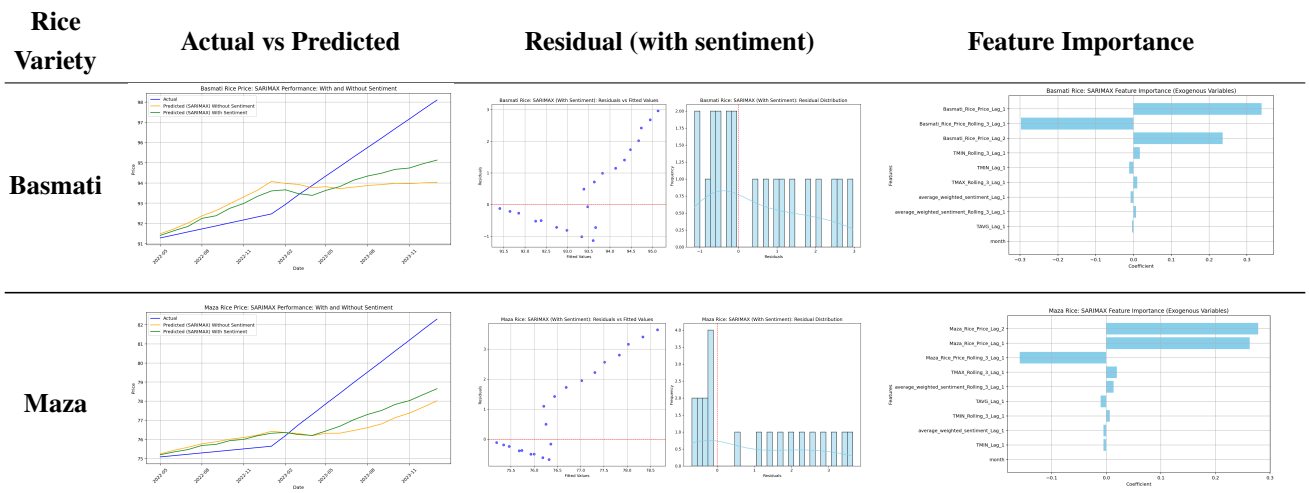

Figure 4. SARIMAX results for both rice varieties: held-out actual and predicted paths under the climate-only and sentiment-enhanced specifications (left), residual scatter and histogram for the sentiment-enhanced specification (middle), and exogenous-feature importance summaries (right).

4.4. Challenges and limitations

The asymmetric gains reported in Table 4 underscore three practical constraints.

1. **Commodity-specific sensitivity:** For Maza rice, the incremental benefit of sentiment is smaller than for Basmati (e.g., SARIMAX RMSE falls by about 18% for Maza versus about 25% for Basmati; Table 4), indicating that demand-side tweets add less new information to a market dominated by supply logistics. This heterogeneity cautions against a one-size-fits-all sentiment proxy and motivates commodity-tailored feature sets.
2. **Feature granularity:** Sentiment was aggregated to monthly frequency in order to align with customs price releases. While sufficient for Basmati's longer procurement cycle, this temporal smoothing likely masks short, logistics-driven price swings in Maza. Higher-frequency sentiment or exporter-specific trade news could narrow the performance gap.
3. **Model robustness:** Gradient boosting and RF show weak held-out performance in Table 4 and only marginal improvement after sentiment is added. In Figures 2 and 3, their prediction paths track the held-out series less closely, and their residual spreads are wider than those of OLS and ridge. This pattern suggests that the sparse monthly sentiment block is harder for the tree ensembles to exploit reliably in the available sample.

4.5. Key observations

- Basmati benefit: Adding sentiment cuts SARIMAX MAE from 1.42 to 1.08, about a 24% reduction, and lifts R^2_{oos} from 0.32 to 0.62 (Table 4). OLS registers a smaller but still meaningful 8% RMSE drop (from 0.497 to 0.456).
- Maza contrast: Linear models (ridge, OLS) retain the top rank for Maza; SARIMAX gains are limited to an 18% RMSE reduction. GB and RF add almost no value once sentiment is included, confirming that noisy text features can inflate variance in small samples.
- Feature importance: In Figure 2, sentiment measures rank among the strongest predictors for Basmati, whereas in Figure 3, the leading temperature metrics dominate. This pattern corroborates the view that sentiment chiefly informs demand-pull commodities.

Overall, the evidence shows that Twitter-based sentiment can meaningfully enhance forecasts for premium rice varieties, while its value diminishes in supply-driven markets. Tailoring feature engineering and validation protocols to each commodity's demand–supply structure is therefore essential.

5. Discussion

5.1. Robustness checks

To assess the stability of the forecast improvements, we conducted two robustness checks on a matched SARIMAX pair for each rice variety: a climate-only SARIMAX and a sentiment-augmented SARIMAX. We use SARIMAX rather than the cross-family best models in Table 4 because the robustness question here is whether adding the sentiment block improves forecast accuracy within a common time-series specification.

Rolling-origin validation: We implemented rolling-origin validation over successive 12-month evaluation blocks, producing one-month-ahead forecasts at each origin and re-estimating the climate-only and sentiment-augmented SARIMAX models as the training window advanced. Table 5 reports the resulting average RMSE and MAE for this matched SARIMAX comparison. The sentiment-augmented specification outperforms the climate-only specification for both varieties, with the larger improvement again appearing in Basmati.

Diebold–Mariano tests: We performed Diebold–Mariano tests (Diebold and Mariano, 1995) on the same one-month-ahead rolling-origin forecast-error series from the climate-only and sentiment-augmented SARIMAX models. The results, presented in Table 6, show that for Basmati rice, the sentiment-augmented specification provides statistically significant forecast improvements at the 5% level ($p < 0.05$). For Maza rice, the improvement is not statistically significant, which is consistent with the smaller gains reported elsewhere in the paper.

Table 5. Rolling-origin validation for the matched SARIMAX comparison (climate-only versus sentiment-augmented).

	Basmati		Maza	
	Climate-Only	Sentiment	Climate-Only	Sentiment
RMSE	1.92	1.45	2.21	1.83
MAE	1.51	1.12	1.85	1.52

Table 6. Diebold–Mariano tests comparing one-month-ahead rolling-origin forecast errors from the climate-only and sentiment-augmented SARIMAX specifications.

	Basmati	Maza
DM-statistic	2.13	1.28
p-value	0.033	0.201

5.2. Synthesizing the empirical evidence

The central empirical result is that social-media sentiment behaves as an incremental demand-side signal rather than as a wholesale replacement for conventional forecasting inputs. When GPT-4-labeled sentiment variables are added to climate measures and lagged prices, the premium Basmati series records a 25% drop in out-of-sample RMSE and an increase in R^2_{oos} from 0.32 to 0.62. In the same setting, simple least-squares and ridge models achieve the strongest overall fit, showing that careful feature construction can matter more than model complexity when the monthly panel is small and tweet coverage is sparse. For the more supply-driven Maza series, the benefit is smaller but still directionally positive. The empirical message is therefore controlled and commodity-specific: Sentiment helps most where demand-side expectations plausibly matter most.

5.3. Relative merits of alternative learners

- **Time-series benchmark:** Seasonal ARIMAX remains a useful benchmark. Once sentiment is included as an exogenous block, it improves materially for both series and provides the clearest interpretable time-series evidence that the added features contain information beyond climate and price history. This is consistent with agricultural forecasting evidence that exogenous climate and macroeconomic variables can strengthen linear time-series benchmarks (Yadav, 2024).
- **ML models:** Gradient boosting and RF capture some non-linear climate–price interactions, but their performance here is limited by the small sample and the variance of the sentiment block. In Table 4, they remain far behind OLS, ridge, and SARIMAX even after sentiment is added. Ridge and OLS regressions, in contrast, absorb the new features with minimal tuning and finish near the top on RMSE and MAE. In this setting, they are not fallback baselines; they are among the strongest models. The takeaway is that straightforward models paired with well-curated inputs can outperform deeper algorithms when data are sparse and noisy.

5.4. Interpreting the sentiment signal

Coefficient magnitudes from the linear and SARIMAX models indicate that engagement-weighted sentiment measures are among the more influential predictors for Basmati prices. Taken together with the out-of-sample improvements, this suggests that online discourse contains information related to short-run demand conditions that is not fully captured by climate variables and lagged prices alone. The four-month lag that produces the best forecasts is also consistent with a plausible lead time between procurement decisions and retail-price transmission in Saudi distribution channels.

For Maza rice, sentiment variables rank noticeably lower in the importance summaries: average (TAVG) and maximum (TMAX) temperatures dominate the explanatory set, which is consistent with a market in which supply-side constraints matter more than consumer perception. This contrast reinforces the view that sentiment data add the most value in demand-pull commodities, whereas climate fundamentals remain paramount in supply-driven segments of the rice market.

5.5. Limitations

Several caveats temper these encouraging findings:

a. Temporal aggregation: Sentiment was aggregated to a monthly frequency to match customs price releases. This smoothing likely attenuates event-day information that could further benefit short-horizon nowcasts (Kim et al., 2017).
b. Sampling bias: Twitter usage in Saudi Arabia skews toward younger, urban consumers, whereas older and rural segments—also important rice purchasers—may be underrepresented, which limits external validity.
c. Climate parsimony: We used only temperature metrics, while precipitation and soil-moisture indices, shown elsewhere to improve grain forecasts (Wang et al., 2023), were unavailable at matching temporal coverage.
d. Model evaluation window: The testing sample corresponds to the final 20% of the time-ordered data (April 2022–January 2024), which is adequate for point accuracy metrics but still relatively small for robust density evaluation or structural-break tests.

e. Ethical and governance issues: While tweets are public, bulk sentiment harvesting raises privacy and data-governance questions that regulators increasingly scrutinize, so that any operational deployment must comply with evolving social-media data policies.

5.6. Research implications and outlook

For importers and the Saudi General Food Security Authority, a live dashboard that tracks weighted sentiment alongside climate anomalies would provide an inexpensive early-warning layer, helping to time forward-contract locks or subsidy adjustments. Retail chains could couple the forecast with inventory optimization algorithms to minimize stock-out risk during demand spikes. More broadly, our reproducible workflow demonstrates that multimodal fusion can be implemented with readily available NLP tools and modest computing resources, lowering the entry barrier for smaller agencies. Importantly, the strongest gains arise in models that are easy to audit and maintain, so operational use does not depend on a complex AI stack.

Three directions appear most promising:

- Higher-frequency and event-driven sentiment: Aligning daily sentiment with import price nowcasts could reveal causal pathways obscured at monthly resolution.
- Hybrid interpretable architectures: SARIMAX residuals fed into lightweight transformers or neural basis expansion analysis for interpretable time series forecasting with exogenous variable (NBEATSX) blocks (Nayak et al., 2024) may combine interpretability with non-linear learning.
- Cross-commodity transfer learning: Adapting the sentiment encoder to wheat or palm-oil markets will test whether the demand-pull advantage generalizes beyond premium rice varieties (Ge et al., 2025).

By addressing these challenges, future work can move toward fully multimodal, real-time forecast systems that increase the resilience of food-importing economies to both physical and information shocks.

6. Conclusion and future directions

This study evaluates whether carefully lagged sentiment features extracted from social media improve Saudi rice-price forecasting once climate variables and historical price dynamics are already included. The answer is yes, but in a qualified and commodity-specific sense. For premium Basmati, GPT-4-labeled sentiment features materially improve out-of-sample accuracy: Sentiment-enhanced SARIMAX lowers RMSE by roughly one quarter, and the strongest overall fits are obtained by OLS and ridge. For Maza, the gains are smaller, consistent with a market shaped more by supply logistics than by consumer perception. The clearest contribution of the paper is therefore not a broad claim that advanced AI forecasting models outperform classical methods, but a more specific finding that carefully engineered multimodal features can improve transparent forecasting models under sparse social-media coverage.

From a practical standpoint, the results suggest that importers and policymakers can treat sentiment as a supplemental demand-side indicator rather than as a standalone forecasting engine. Embedding lagged sentiment measures in procurement dashboards may shorten reaction times and improve buffer-stock decisions, especially for premium rice segments in which household expectations matter more. Methodologically, the study shows how large language models can be used narrowly and transparently—

as a sentiment-labeling instrument within a broader econometric workflow—while the forecasting gains are ultimately realized in well-specified linear and time-series models. These findings provide a disciplined foundation for future multimodal forecasting work in agricultural markets exposed to both physical and information shocks.

Several directions for future research follow naturally from these results. First, harvesting sentiment at sub-monthly frequency (daily or even intraday) would permit higher-resolution event studies, such as price responses to export-ban rumors or extreme-weather alerts, and may further sharpen nowcasts. Second, although the GPT-4 labeling pipeline performs well in our validation sample, fine-tuning Arabic sentiment classifiers on agriculture-specific corpora (e.g., commodity forums or ministry announcements) could yield richer polarity and emotion vectors, improving both explanatory power and interpretability. Third, future work should explore hybrid designs that combine the interpretability of SARIMAX with the feature-learning capacity of deep networks (e.g., SARIMAX residuals fed into LSTMs or transformers) and embed Bayesian or quantile layers to generate calibrated predictive distributions rather than point estimates. Fourth, applying the pipeline to other staples (wheat, maize, and palm oil) and to import-dependent economies across MENA or Southeast Asia would help establish external validity and reveal whether sentiment elasticity differs by culture, diet, or supply-chain structure. Finally, augmenting sentiment with trade-policy announcements, shipping-cost indices, and high-frequency satellite vegetation metrics could produce a more holistic early-warning system capable of disentangling simultaneous demand and supply shocks. Pursuing these directions would move the literature closer to fully multimodal, real-time forecasting frameworks that enhance the resilience and sustainability of global agricultural markets.

Use of AI tools declaration

The authors used GPT-4 for two limited purposes: (i) tweet sentiment labeling during data preparation, as described in the Methods section, and (ii) language editing during manuscript revision. The authors reviewed, verified, and take full responsibility for the final content of the manuscript.

Acknowledgments

The authors extend their appreciation to the Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number MOE-BU-1-2020.

Conflict of interest

The authors declare no conflicts of interest.

Data availability statement

Due to platform terms, raw tweet text may not be redistributable. We therefore release (i) tweet IDs and collection scripts, (ii) the prompt template and inference settings used for GPT-4 sentiment labeling, and (iii) the full monthly sentiment aggregation table used in the experiments (including N_t, AS_t, WS_t, and the tweet-availability indicator) for January 2015–January 2024. Researchers can rehydrate tweets

from the released IDs using platform-compliant hydration tools (subject to policy changes); reproducing our aggregated sentiment series does not require redistributing tweet text. The data and code used in this study are available in the public project repository: https://github.com/alqithami/AgriSent.

References

Akaike H (1974) A new look at the statistical model identification. *Ieee T Automat Contr* 19: 716–723. https://doi.org/10.1109/TAC.1974.1100705

Alamah Z, Elgammal W, Fakih A (2024) Does twitter economic uncertainty matter for wheat prices? *Econ Lett* 234: 111463. https://doi.org/10.1016/j.econlet.2023.111463

An W, Wang L, Zeng YR (2024) Social media-based multi-modal ensemble framework for forecasting soybean futures price. *Comput Electron Agr* 226: 109439. https://doi.org/10.1016/j.compag.2024.109439

Antoun W, Baly F, Hajj H (2020) AraBERT: Transformer-based model for arabic language understanding. In: *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, Marseille, France, European Language Resource Association: 9–15. Available from: https://aclanthology.org/2020.osact-1.2/

Araci D (2019) FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint*. https://doi.org/10.48550/arXiv.1908.10063

Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. *J Comput Sci* 2: 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Bonato M, Cepni O, Gupta R, et al. (2024) Forecasting the realized volatility of agricultural commodity prices: Does sentiment matter? *J Forecast* 43: 2088–2125. https://doi.org/10.1002/for.3106

Box GEP, Jenkins GM, Reinsel GC, et al. (2015) *Time Series Analysis: Forecasting and Control*. 5th edition. John Wiley & Sons, Hoboken, NJ.

Breiman L (2001) Random forests. *Mach Learn* 45: 5–32. https://doi.org/10.1023/A:1010933404324

Cai Y, Guan K, Lobell D, et al. (2019) Integrating satellite and climate data to predict wheat yield in australia using machine learning approaches. *Agr Forest Meteorol* 274: 144–159. https://doi.org/10.1016/j.agrformet.2019.03.010

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*, 785–794. https://doi.org/10.1145/2939672.2939785

Devlin J, Chang MW, Lee K, et al. (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1: 4171–4186. https://doi.org/10.18653/v1/N19-1423

Diebold FX, Mariano RS (1995) Comparing predictive accuracy. *J Bus Econ Stat* 13: 253–263. https://doi.org/10.1080/07350015.1995.10524599

Ewald CO, Li Y (2024) The role of news sentiment in salmon price prediction using deep learning. *J Commod Mark* 36: 100438. https://doi.org/10.1016/j.jcomm.2024.100438

FAO, IFAD, UNICEF, et al. (2023) The state of food security and nutrition in the world 2023. Food and Agriculture Organization of the United Nations, Rome. https://doi.org/10.4060/cc3017en

Food and Agriculture Organization of the United Nations (2026) Food Price Monitoring and Analysis (FPMA) Tool. Available from: https://fpma.fao.org/.

Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *Ann Stat* 29: 1189–1232. https://doi.org/10.1214/aos/1013203451

Ge L, Huang Q, Zhu F, et al. (2025) Advanced time series forecasting for commodities: Insights from the FEDformer model. *Energy Econ* 147: 108513. https://doi.org/10.1016/j.eneco.2025.108513

Gelman A, Carlin JB, Stern HS, et al. (2013) *Bayesian Data Analysis*. Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Hyndman RJ, Athanasopoulos G (2021) Forecasting: Principles and Practice 3rd edition. OTexts, Melbourne, Australia. Available from: https://otexts.com/fpp3/.

International Monetary Fund (2026) IMF Data: International financial statistics and consumer price index series. Available from: https://data.imf.org/.

Joulin A, Grave E, Bojanowski P, et al. (2017) Bag of tricks for efficient text classification. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* 2: 427–431. Available from: https://aclanthology.org/E17-2068/.

Kamilaris A, Prenafeta-Boldú FX (2018) Deep learning in agriculture: A survey. *Comput Electron Agr* 147: 70–90. https://doi.org/10.1016/j.compag.2018.02.016

Kim J, Cha M, Lee JG (2017) Nowcasting commodity prices using social media. *Peerj Comput Sci* 3: e126. https://doi.org/10.7717/peerj-cs.126

Liu B (2012) Synthesis Lectures on Human Language Technologies, *Sentiment Analysis and Opinion Mining*. 5: 1–167. Morgan & Claypool. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Lobell DB (2013) The use of satellite data for crop yield gap analysis. *Field Crop Res* 143: 56–64. https://doi.org/10.1016/j.fcr.2012.08.008

Marshall A (2013) *Principles of Economics*. Palgrave Macmillan. https://doi.org/10.1057/9781137375261

Mittal A, Goel A (2011) Stock prediction using twitter sentiment analysis. CS229 Project Report, Stanford University. Available from: https://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf.

National Centers for Environmental Information, National Oceanic and Atmospheric Administration (2026) Daily Summaries. Available from: https://www.ncei.noaa.gov/access/search/data-search/daily-summaries.

Nayak GHH, Alam MW, Singh KN, et al. (2024) Exogenous variable driven deep learning models for improved price forecasting of TOP crops in india. *Sci Rep* 14: 17203. https://doi.org/10.1038/s41598-024-68040-3

Putra AW, Supriatna J, Koestoer RH, et al. (2021) Differences in local rice price volatility, climate, and macroeconomic determinants in the indonesian market. *Sustainability* 13: 4465. https://doi.org/10.3390/su13084465

Reis Filho IJ, Marcacini RM, Rezende SO (2022) On the enrichment of time series with textual data for forecasting agricultural commodity prices. *MethodsX* 9: 101758. https://doi.org/10.1016/j.mex.2022.101758

Sari M, Duran S, Kutlu H, et al. (2024) Various optimized machine learning techniques to predict agricultural commodity prices. *Neural Comput Appl* 36: 11439–11459. https://doi.org/10.1007/s00521-024-09679-x

Tashman LJ (2000) Out-of-sample tests of forecasting accuracy: An analysis and review. *Int J Forecast* 16: 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0

Tetlock PC (2007) Giving content to investor sentiment: The role of media in the stock market. *J Financ* 62: 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Wang W, Liu Y (2025) A novel framework for agricultural futures price prediction with BERT-based topic identification and sentiment analysis. *J Forecast* 44: 1969–1992. https://doi.org/10.1002/for.3278

Wang Z, French N, James T, et al. (2023) Climate and environmental data contribute to the prediction of grain commodity prices using deep learning. *J Sust Agr Env-Aust* 2: 251–265. https://doi.org/10.1002/sae2.12041

Xu JL, Hsu YL (2022) The impact of news sentiment indicators on agricultural product prices. *Comput Econ* 59: 1645–1657. https://doi.org/10.1007/s10614-021-10189-4

Yadav A (2024) A comparative study of time series, machine learning, and deep learning models for forecasting global price of wheat. *SN Operations Research Forum* 5: 113. https://doi.org/10.1007/s43069-024-00395-9

AIMS Press