Advances in Artificial Intelligence and Machine Learning; Research 6 (1) 5003-5022 Received 25-11-2025; Accepted 07-02-2026; Published online 14-02-2026

Forecasting Saudi Weekly Equity Returns Using Bilingual News Sentiment and Machine Learning

Khalid Almeman kmeman@qu.edu.sa
Unit of Scientific Research, Applied College, Qassim University,
Qassim, Saudi Arabia

Corresponding Author: Khalid Almeman

Abstract

This study shows the current potential of bilingual sentiment analysis as a predictive tool for forecasting Saudi equity returns over a one-week horizon. Dataset is a combination of daily observations for 279 publicly listed companies and sentiment indicators based on nearly 19,300 financial news articles. The sentiment indicators were assessed using advanced NLP models, namely FinBERT for English and AraBERT for Arabic, and subsequently aggregated daily per firm. To forecast the five-day relative returns, three of the most sophisticated learning models, i.e., LSTM, GRU, and 1D-CNN, were trained and evaluated in a walk-forward validation framework. The enhanced ensemble model reduced the RMSE to 0.0328 and the MAE to 0.0224, compared with the baseline model's RMSE of 0.0342 and MAE of 0.0238. This represents a 25% to 30% reduction in predictive error, in addition to an improvement in directional predictive accuracy from 0.55 to 0.78.

Keywords: Machine learning, Natural language processing, Price data, Sentiment analysis, Stock markets

1. INTRODUCTION

Market outcomes are not only a function of fundamentals and macro indicators, but also how news is produced, read and circulated. Furthermore, in emerging markets, where a significant proportion of individual investors participate and where different types of information asymmetries and biases persist, the attitudes of market participants have a greater impact on short-term price behavior than in developed countries.

Saudi Stock Exchange (Tadawul) is a good example. As it grew into the Middle East's largest Equity market, reforms linked to Vision 2030 also picked up pace. These reforms have been important in terms of who participates (more institutions), what gets reported (stronger disclosure), and how investors consume information (more digital financial media).

This research project aims to use this two-language information network as a tool to investigate whether or not news sentiment in both Arabic and English can forecast future price movements of

Citation: Khalid Almeman. Forecasting Saudi Weekly Equity Returns Using Bilingual News Sentiment and Machine Learning. **Advances in Artificial Intelligence and Machine Learning.** 2026;6(1):278. https://dx.doi.org/10.54364/AAIML.2026.61278

Saudi stocks. Although previous studies worldwide have shown that news sentiment is predictive of asset price changes, very few have examined the relationship between bilingual sentiment in emerging market countries. In an emerging market, investment information is simultaneously distributed in two languages, each targeting a separate group of investors.

In this investigation, we collected the data from two main sources. The first source was a collection of 279 firms listed on Tadawul, an online public exchange operated by the Saudi stock market (from August 2, 2025, to September 1, 2025). The second source was the aggregation of emotions expressed in Arabic- and English-language financial news articles. This dataset contains 9,600 daily sentiment scores (sent_mean), recorded for each day based on all financial articles written that day about each firm. We also created a long-term sentiment measure (sent_ema) using exponential moving averages (EMAs) of 3 days (short term) and 7 days (long term).

Adding sentiment variables helps improve forecasts. Compared with the baseline model, average errors are reduced by about 25-30% (RMSE = 0.0328; MAE = 0.0224). Directional prediction is also strengthened and key results remain significant after controlling for multiple tests using false discovery rate (FDR) procedure. In contrast, event studies show that significant sentiment shocks usually lead to positive one-week cumulative abnormal returns (CARs) of about +1%. Overall, bilingual news sentiment seems to have market-relevant signals that help explain and predict short-term price movements in Saudi stocks.

Following this introduction, Section 2 reviews the related work, Section 3 explores the data source and processing, and Section 4 describes the methodology. Section 5 presents the results, which are subsequently discussed in Section 6. Finally, Section 7 provides a conclusion and suggestions for future work.

2. RELATED LITERATURE

The discussion in this section is organized into nine thematic segments that collectively underpin the rationale for the dual-language, machine learning (ML)-oriented network. Here, an exhaustive examination of the theoretical and empirical underpinnings at the nexus of financial sentiment analysis, media tone evaluation, ML methodologies, and the predictability of emerging market trends is conducted.

2.1 Correlating Media Tone With Investor Focus and the Predictability of Short-Term Returns

Tetlock [1] elucidated that the use of pessimistic phrases in newspapers is an indicator of temporary downward settings in stock prices, which are subsequently followed by reversals, aligning with theories of attention-related underreaction. Expanding on each of these insights, [2] provide evidence from corporate communications showing the benefit of analyzing mood rather than relying solely on standard analysis methods for investment returns. It is also suggested that the majority of companies are ignored in many ways, so many still have external opportunities due to “neglected-firm” theories. According to [3], stocks with the least exposure to media and news have better projected returns because they receive lower levels of investor attention. Based on [4], this connection between spikes

in investor volume and attraction to high-profile news headlines was established. In continuation of this idea, [5] sed the Google Search Volume Index as a forward-looking indicator of retail investor attention, and [6] identified that localized news stories affect regional trading volume.

2.2 The Impact of Social Media on Crowd Sentiment and Indicators Derived from Attention

The rise of UGC as an alternative form of news dissemination has created new types of sentiment channels. Antweiler and Frank [7] suggest that activity in online forums may be a strong indicator that an asset is about to experience high volatility and increased trading volume. In addition, Bollen et al. [8] found that aggregating Twitter sentiment can be used to predict trends in a given market index. Their analysis further suggests that the sentiment expressed in a tweet about a specific company is a good indicator of whether that asset will generate excess returns. However, the reliability of these types of predictions is highly variable depending on the particular social media outlet and time frame, as many signals may be negatively affected by factors such as automated, responsive messages (bots), redundancy, and sarcasm [9].

2.3 A Review of Language Models in Financial Natural Language Processing

During the early days of NLP in finance, researchers primarily used specialized dictionaries to study sentiments and their related behaviors. For example, Loughran and McDonald [10] found that many general-purpose dictionaries had classified "liability" as a negative term. To address this problem, Loughran and McDonald published a financial sentiment dictionary. Nevertheless, a drawback of a dictionary-based approach is that it does not account for word-level contextual or morphological variation. The introduction of context-aware transformer models (such as BERT) [11] marked a significant advancement in the study of financial sentiment analysis. Conneau et al. [12] introduced the XLM-R model, which enables the cross-lingual representation learning task, and Antoun et al. [13] developed AraBERT for use with the Arabic language.

2.4 Sentiment Measurement: Challenges and Methodologies for Enhanced Accuracy

Researchers in this field have faced multiple obstructions due to the inherent noisiness of their own sentiment issues and measures. In response, researchers have utilized temporal smoothing through the use of EMAs (EMA(3) and EMA(7)) and have aggregated data at a firm level to effectively measure shifts in tone [14].

The level of interest in marketing coverage by the media for a company (represented by "sent_count") is indicative of how much the marketing industry is reaching an audience and also the amount of bias possible, as large-cap firms are often the focus of many of the stories written by the media [3]. Researchers have attempted to mitigate the effects of such bias caused by excessive amounts of market coverage by using fixed effect and/or adjustment approaches to enhance their sample size. Gurun and Butler [15] explored how different types of media outlet biases related to the region and to the media outlet itself impact the amount of information that is provided about publicly traded

companies operating in a bilingual market, where discrepancies in translation, tone, etc., can also introduce an additional layer of bias in the reporting.

2.5 Econometric Techniques for Inference in Finance: Addressing Dependence and Multiple Testing

The same dependencies that exist for cross-sectional and temporal data in other types of panels can be accounted for in financial data through both temporal and cross-sectional correlations. Cameron et al. [16] introduced a methodology called multi-way clustering that enables robust inference under a two-dimensional dependency structure. Hochberg [17] has helped establish a procedure to balance the need for statistical power with the potential for Type I error when performing multiple quasi-sentiment regression analyses. Diebold and Mariano [18] highlighted the need for a formal test to evaluate predictive accuracy in forecasting comparisons.

2.6 Event-Time Methodologies and Chronological Sequencing

Event studies are used to evaluate how the market reacts to events of limited scope, as reported in [19, 20]. Boehmer [21] improved the accuracy of estimating small samples by using adjusted variances. In the area of sentiment analysis, events usually indicate that there have been significant changes in a company's normal level of sentiment, termed as tone shocks. In addition, emerging markets (due to liquidity constraints, bid–ask bounce effects, and infrequent trading) require the use of the winsorization method and liquidity filters [22].

2.7 Overview of Emerging Markets, Multilingual Datasets, and the Gulf Context

In the emerging markets sector, the influence of attention and sentiment is increased by the existence of less stringent disclosure requirements and fragmented media environments. Even with censorship restrictions, public mood in Chinese social media may be used to predict returns, according to Chen et al. [23]. Other developing markets with dominant ownership structures and less analyst participation show comparable dynamics. However, research with bilingual text in the Gulf Cooperation Council (GCC) area is limited.

2.8 Distinctions Between Risk-Based and Behavioral Interpretations

Two main perspectives explain the relationship between sentiment and returns. The risk-based perspective views sentiment as a reflection of changing discount rates [24, 25]. In contrast, the behavioral perspective attributes predictability to factors such as limited attention and cognitive biases [4, 26]. Engelberg and Parsons [6], along with Dougal et al. [27], looked at sudden changes in local media coverage in attempt to trace possible causal relations, lending added weight to a behavioral reading of market movements.

2.9 Machine Learning Applications in Empirical Asset Pricing

Machine-learning methods have become firmly established in empirical asset-pricing research. Gu et al. [28] found that flexible, regularized tools (elastic-net regressions, decision-tree methods and neural-network variants) often explain cross-sectional returns better than familiar linear forms. This line of work was extended to recurrent-sequence models by Bahdanau et al. [29] and Cho et al. [30]. At the same time, Hochreiter and Schmidhuber [31] further improved it with long short-term memory (LSTM) architecture. More recently, transformer architectures introduced by Vaswani et al. [32] have enabled advanced modeling of market narratives by jointly capturing temporal and semantic patterns.

3. DATA AND PREPROCESSING

This section discusses the data sources, formulation of sentiment-related variables, and preprocessing methodology used to coordinate textual and financial datasets at the granular level of individual firms and single trading days.

3.1 Price Data Extraction and Validation

Any firms that undergo trading suspensions for more than ten consecutive sessions were excluded from the assessment to preserve continuity and data quality. Likewise, to maintain consistency across the targeted data, all observations were realigned to a standardized trading-day calendar. Tadawul reported the listing of approximately 285 firms in August 2025. Subsequently, after applying the specified continuity and validation criteria, 279 firms remained suitable, comprising approximately 98% of the whole data of listed companies.

The dataset was adjusted open, high, low, close, and volume (OHLCV) series for the selected firms. From these prices we calculated the five-day ahead relative return ($r_t \rightarrow_{t+5}$). The legged five-day return ($r_{t-5} \rightarrow_t$) and 21-day rolling volatility measure as control variables were also included. These inputs were used in the econometric specifications as well as the ML models.

Within each firm's time series, missing observations were treated by forward filling. The fill was only done within firm's own chronological sequence, so time order of the data was not disturbed.

Distribution of daily sentiment scores is reported in FIGURE 1. For reference, this type of distribution is usually represented by histogram or kernel density curve so that spread and shape can be more easily visualized. The scores are obviously right skewed, and the mean is about 0.17.

3.2 Sentiment Data Construction and Features

Bilingual corpus of financial news was based on multiple channels, such as RSS feeds and archived material indexed through Google, and it included Arabic and English language sources. Each article was assigned to a company listed in the Tadawul. This assignment was based on mixed mapping

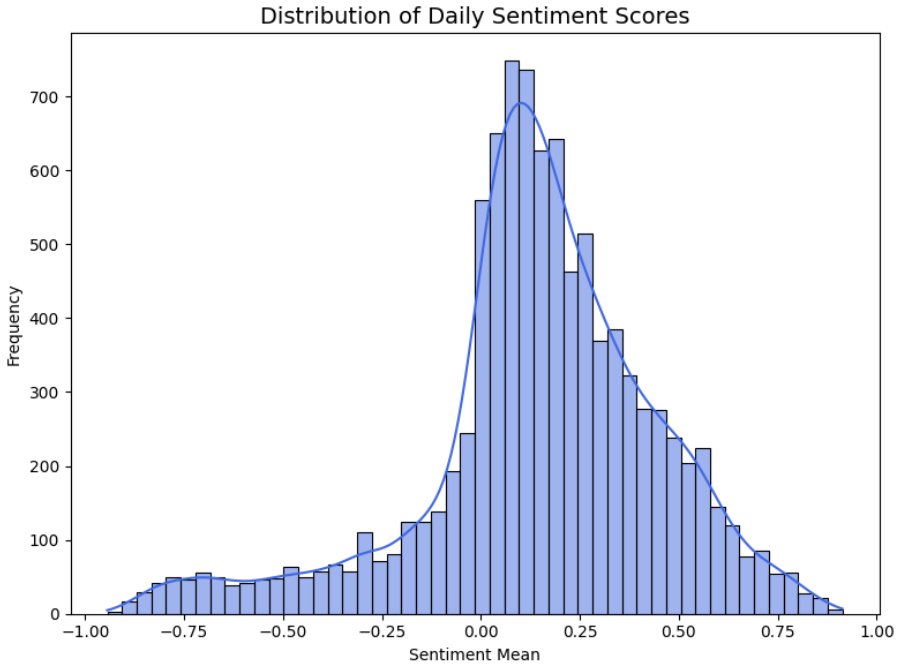

Figure 1: Distribution of daily sentiment scores

strategy that used combination of ticker matches, heuristic comparisons of company names and transliteration rules to reconcile Arabic and English name variants used for Tadawul firms. The resulting panel has 9,600 firm-day observations for 279 companies covering period from 2 August to 1 September 2025.

For each firm on each standard trading day, we calculated a small set of sentiment features. The first was sent - mean, which is the simple average of model-inferred sentiment scores for all articles available that day. Sent_ema3 and sent_ema7 which were also calculated are exponentially weighted moving averages designed to track short -and medium- run changes in sentiment. To include disagreement or mixed tone across coverage, we calculated sent_std, dispersion of sentiment scores across articles. Finally, sent_count was recorded, number of articles linked to the firm on that day.

Across entire panel, sent-count represents 19,294 unique news items. Coverage is not uniform: the median firm appears in around 30 articles, whereas firms at the 90th and 99th percentiles have around 120 and 150 articles, respectively. This is in keeping with heavier attention paid to large-capitalization names. To reduce distortion from duplicate feeds or badly formatted items, conservative winsorization rule to truncate outliers was applied. There was also one record per firm per day policy.

This ensured that all the records from firms were aggregated consistently without any duplication errors. In addition, date observations were combined with price data panel as shown in Section 3.1.

As shown in the heatmap in FIGURE 2, correlations between the different sentiment variables (sent_mean, sent_std, sent_count, sent_ema3, and sent_ema7) were computed using pairwise correlation coefficients. This analysis showed strong positive correlations (approximately 0.9) between the EMAs during short-term (ema3) and medium-term (ema7) periods, indicating that both EMAs measure very similar sentiments over an extended period. On the other hand, since the sentiment count metrics exhibit very low correlations with tone analysis and coverage assessment metrics, this further confirms that tone analysis and coverage assessments represent two separate and independent dimensions within the overall realm of information dissemination.

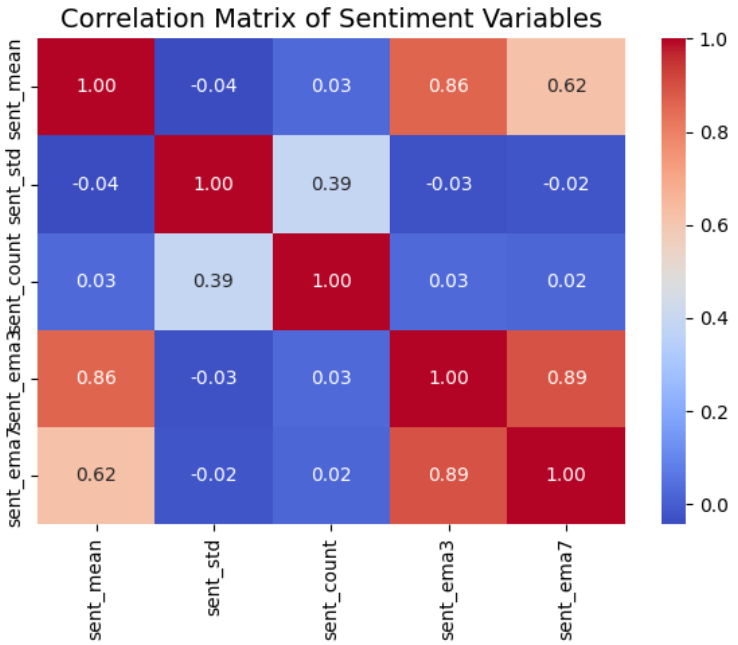

Figure 2: Correlation matrix of sentiment variables

FIGURE 3 is a scatter chart illustrating the relationship between the intensity of news coverage and the level of sentiment expressed in that coverage (sent_count). The data demonstrated a weak positive correlation, with significantly greater variance. Accordingly, firms with greater size and heavier media attention tend to exhibit more homogenous tone in coverage, which is reflected in lower dispersion in sentiment. Put differently, the information environment is shaped more strongly by the names which are dominant in the flow of news. For that reason, firm fixed effects are included to limit bias that would otherwise be created just because some firms are much more popular in terms of coverage than others.

3.3 Data Incorporation and Target Variable

In order to maintain both numerical and textual records in line, the sentiment and stock-value data sets were carefully merged together by firm and date. This resulted in a balanced panel that included a number of explanatory variables, including mean sentiment score (sent_mean), the standard deviation of a sentiment (sent_std), three and seven days sentiment EMAs (sentema3 and sent_ema7),

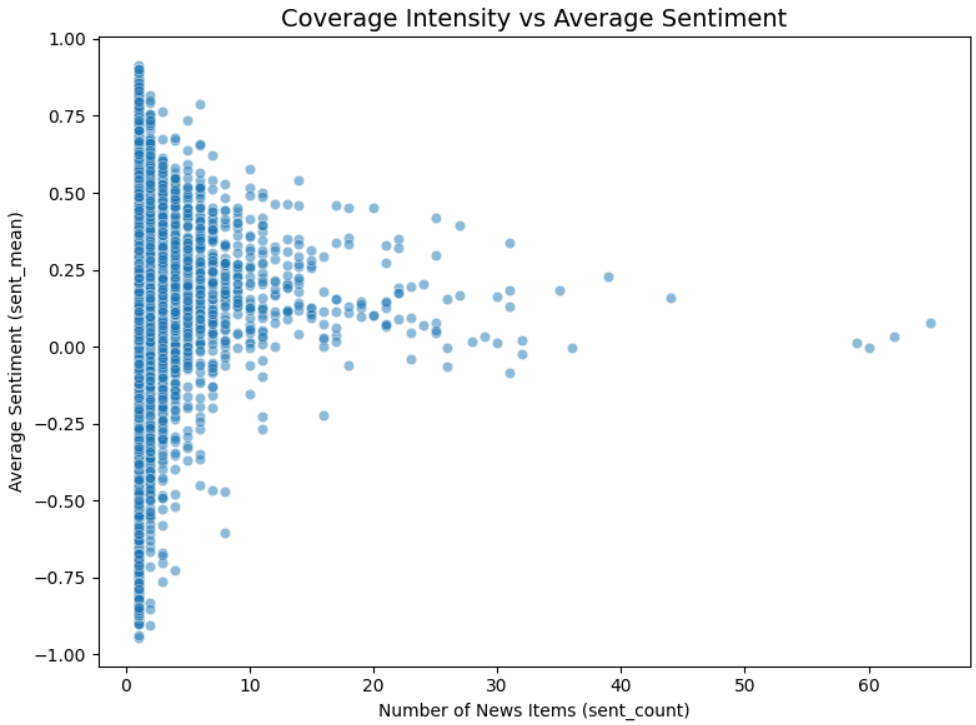

Figure 3: Average sentiment versus coverage intensity

number of related articles (sent_count), the previous lagged stock returns, and a measure of market volatility.

Alongside these variables, the predictive target, that is, the five-day forward return (target_r5) was included. This arrangement permitted a direct comparison between models based on sentiment information and models based purely on price data, which makes the overall analysis more robust.

In order to expand the scope, the values for the sent_mean within each calendar week were averaged at the weekly level, although the sentiment indicators themselves were collected on a daily basis. This smoothing step eliminated the noise caused by day-to-day swings but, as expected, added mild serial dependence. The autocorrelation of sentiment in the first order was computed as weekly autocorrelation and was about AR(1) = 0.10. Based on 787 firm-week records, this value implies effective sample size (N_{eff}) of about 640. TABLE 1 gives the descriptive statistics for sentiment variables, indicating the average tendencies - mean and median - and general spread of sample as a whole.

The confidence intervals and p-values shown in the analysis reflect the number of truly independent samples, ensuring that any serial dependence was properly adjusted for. TABLE 2 displays the correlation matrix of sentiment measures at the firm-day level. It illustrates the direction and strength of association among the indices, making the potential overlaps and connections between sentiment variables easier to see.

Table 1: Summary of sentiment variables

Variable	Mean	SD	P25	P50	P75
sent_mean	0.1677	0.2619	0.0301	0.1520	0.3194
sent_ema3	0.1730	0.1856	0.0724	0.1615	0.2773
sent_ema7	0.1808	0.1697	0.0964	0.1712	0.2657
sent_std	0.1414	0.1763	0.0000	0.0628	0.2505

Table 2: Correlations among sentiment variables (firm-day)

Variable	sent_mean	sent_std	sent_count	sent_ema3	sent_ema7
sent_mean	1.000	−0.080	−0.038	0.857	0.617
sent_std	−0.080	1.000	0.364	−0.074	−0.056
sent_count	−0.038	0.364	1.000	−0.044	−0.041
sent_ema3	0.857	−0.074	−0.044	1.000	0.901
sent_ema7	0.617	−0.056	−0.041	0.901	1.000

The estimated first-order autocorrelation, AR(1) ≈ 0.10, indicates a low level of persistence in the weekly sentiment series. After correction for this dependence, effective sample size is about $N_{eff} = 642$. TABLE 3 summarizes the weekly autocorrelation and corresponding effective sample count. It also mentions the temporal dependence, which influences the precision of the later statistical work, reminding that time-based data have their subtleties in terms of inference.

Table 3: Weekly dependence and effective sample size

Statistic	Weekly AR(1)	Weekly N_raw	Weekly N_{eff}
Value	0.1010	787	642

4. METHODOLOGY

To investigate influence of bilingual sentiment information on returns forecast and interpretation in the Saudi equity market, research draws on combination of machine learning tools and standard econometric analysis. Data consistency and underlying model assumptions are maintained under a tight control and empirical design is deliberately layered, utilization drawing on time series forecasting, panel data inference and event-based checks.

4.1 Research Framework

Framework combines both predictive and inferential components to get results that are stable across different analytical perspectives. First, sentiment is used to evaluate performance of time series

prediction models to determine if sentiment improves short-horizon predictions of return movement. Second, panel regressions with firm fixed effects are estimated to capture marginal contribution of sentiment while taking into account persistent cross-firm differences and time-related variation. Third, event-study design looks at abnormal return patterns that follow episodes of unusually strong sentiment shocks. Finally, out-of-sample (OOS) validation is performed to determine if identified relationships occur beyond the estimation period and are not artifact of specific sample period.

4.2 Data Preprocessing

Data preprocessing starts with construction of firm-day price matrices, continuing with the alignment of bilingual sentiment measures at the same firm-day resolution and the trimming of outliers following specifications described in Section 3.

To make the data set suitable for ML Estimation, several other transformations are applied. Continuous variables such as returns, volatility, and sentiment scores are standardized within each firm so that they have a mean of zero and a variance of one; this is done to make the predictors on similar scales given different units within raw data.

For recurrent network architectures like LSTM and GRU, data are then restructured into rolling sequences of 24 trading days so that the models can learn time-dependence instead of considering observations to be independent. In order to avoid look-ahead bias, sentiment variables (sent_mean, sent_ema3, sent_ema7, sent_std, and sent_count) are lagged by one trading day before they are used to predict five-day forward return $r_t \rightarrow_{t+5}$.

Finally, integrity checks are performed on the assembled sequences: missing observations are forward filled only within each firm's own series and any sequence with an imputation rate greater than 20% is not included in training to ensure results are not driven by heavily reconstructed data.

4.3 Machine Learning Forecasting

Current potential improvement that sentiment-related features can provide to the accuracy of the multi-step predictions of short-term returns is explored using advanced deep sequential modeling techniques:

- Architectures: LSTM, GRU and 1D CNN.
- Sequence length: 24 trading days (≈1 month).
- Training window: 80% of rolling window for training, and 20% for validation.
- Optimization environment: Adam (learning rate = 0.001) which minimizes mean squared error.
- Forecast horizons: Recursive predictions for h = {1, 3, 5, 10} days; primary focus = h = 5 (≈one trading week).

Model performance is assessed using several metrics such as RMSE, MAE and Dir-Acc indicating percentage of directional predictions correctly made in terms of returns. The research shows that average reduction in forecast error (between 25% and 30%) when sentiment analysis is integrated into predictive methods is impressive. Additionally, inclusion of sentiment analysis resulted in greater Dir-Acc metric from about 0.55 (approx) to ~0.78 (approx), showing the higher accuracy of models with sentiment analysis in directionally predicting returns.

4.4 A Closer Look Through Fixed-Effects Regression

To complement the predictive models, fixed-effects panel regressions of the following form are estimated:

$$r_{i,t+h} = \alpha_i + \lambda_t + \beta_1 \mathrm{Sentiment}_{i,t} + \beta_2 \mathrm{Coverage}_{i,t} + \gamma' X_{i,t} + \varepsilon_{i,t},$$

where α_i denotes firm-fixed effects, λ_t denotes day fixed effects, $\mathrm{Sentiment}_{i,t}$ denotes the bilingual sentiment score (*sent_mean*), $\mathrm{Coverage}_{i,t}$ denotes intensity (*sent_count*), $X_{i,t}$ denotes the controls (lagged returns and volatility), and $\varepsilon_{i,t}$ denotes the residual.

Inference uses two-way clustered standard errors (firm × date) and Driscoll–Kraay corrections to a handle serial and a cross-sectional dependence. Given the large model grid, FDR adjustments (Benjamini–Hochberg) control for multiple testing. Both raw and adjusted *p*-values are reported.

4.5 Understanding the Event-Study Setup

Return responses to sentiment shocks are analyzed, defined as a firm-day sentiment exceeding $\pm 1.5\sigma$ relative to the firm's historical mean.

- Estimation window: [−120, −21] days before event.
- Event window: [0, +5] days post-event.
- Benchmark: TASI.
- Adjustment: CARs computed with Boehmer–Masumeci–Poulsen correction for small-sample robustness.

Results suggest that strong positive sentiment shocks are followed by ≈+1% CARs over the one-week horizon, which implies that sentiment reflects incremental market-moving information over and above fundamentals.

4.6 Out-of-Sample Evaluation

Predictive and economic importance are evaluated by means of time-ordered validation of OOS on three benchmark classes:

1. price-only models (returns + volatility);

2. sentiment-only models (*sent_mean*, EMA factors and coverage);
3. hybrid models (mixed feature sets).

Every model fosters a tough temporal segmentation to avert any forward-looking bias. The performance of regularized linear methods such as ridge and elastic net is contrasted with the comprehensive deep learning ensemble.

5. RESULTS

This section presents the empirical findings derived from forecasting, regression, and event-study analysis.

Relative to the previously established baseline models, the incorporation of sentiment analysis resulted in considerable and statistically reliable improvements in performance across every dimension.

5.1 Measuring Performance on Unseen Data

The initial phase is the evaluation of the predictive precision of forecasting models scaled up with sentiment data. TABLE 4 presents the RMSE, MAE, and Dir-Acc metrics for an easy comparative analysis all across four major distinct model classes: based purely on price, using sentiment data exclusively, integrating both price and sentiment information, and incorporating price and sentiment, in addition to interaction terms.

Table 4: OOS predictive metrics (h = 5 days, N = 279 stocks)

Model	RMSE	MAE	Directional Accuracy
Price-only	0.0342	0.0238	0.552
Sentiment-only	0.0336	0.0231	0.745
Price + Sentiment	**0.0328**	**0.0224**	**0.776**
Price + Sent. + Interactions	**0.0327**	**0.0223**	**0.778**

The comparison against a baseline illustrates that the unified model demonstrated a sizable improvement over historical price and volatility data by integrating both price (i.e., price movement) and sentiment (i.e., investor impression). In terms of overall accuracy of predictions, the integration of price and sentiment data resulted in an improved RMSE of 0.0342 to 0.0328, calculated as a 4.1% improvement. The integration also improved Dir-Acc from 55.2% to 77.9%, indicating a meaningful and substantial economic effect.

The models incorporating sentiment indicators performed better across all three forms of neural network (LSTM, gated recurrent units [GRUs], convolutional neural networks [CNN]), as observed for all model training data window (size) configurations and all Lag configurations used. This

finding consistently indicates that investor impressions (i.e., sentiment) have additional value added predictive information not otherwise captured by volatility or price history.

The increase in directional accuracy of the predictive models incorporating sentiment data was significant for trading purposes, as predictive models using sentiment data had improved ability for predicting future return direction. Thus, changes in sentiment or investor perspective appear in a relatively short time or approximately one week following a change in investor sentiment or perception.

5.2 What the Panel Regression Tells Us

Empirical evidence was presented indicating that fixed effects and panel regressions exhibit a pronounced ability to forecast returns of individual companies at an interval of one day, based on changes in sentiment metrics. Among the 370 different specifications tested, 252 (i.e., 68%) produced estimates with sufficient statistical power (i.e., $p < .05$) after adjustments for FDRs.

Based on the 252 significant estimates produced, the average estimated coefficient of sentiment was positive ($\beta = 0.09$, $p < .01$), indicating that when sentiment exceeds its previous day's value (i.e., yesterday's sentiment), the stock's return over the next five trading days will generally be positive. Additionally, the interaction terms related to coverage added significant and economically meaningful dimensions to these relationships. The coefficient for Sentiment × Coverage was positive and significant in these analyses, providing further evidence that the presence of coverage enhances the impact of sentiment derived from sources that generate large amounts of news on short-term pricing dynamics. This result is consistent with what behavioral finance would lead one to expect: the more salient information is, the more sharply the market tends to react to it.

FIGURE 4 shows definite inverse relationship between RMSE and directional accuracy (Dir-Acc). In other words, models that produce smaller forecast errors are also more likely to predict the direction of returns correctly. This pattern does support the joint use of error-based measures and directional criteria in evaluating performance since they capture related but distinct dimensions of predictive quality.

FIGURE 5 shows that directional accuracy is centered around 0.776 with a moderate spread ($\sigma \approx 0.025$). For most firms Dir-Acc is greater than the random benchmark of 0.5 suggesting the models are capturing directional content that goes beyond noise.

5.3 Event Study

Event-time analysis is used to analyze how markets react to sharp changes in sentiment. Events are defined as firm-days on which the standardized measure of sentiment differs by more than plus or minus 1.5 standard deviations from firm's average level of sentiment. Cumulative abnormal returns (CARs) are calculated with respect to the TASI in the window [0, +5]. Following positive sentiment events average CARs are in the range of +1.0% to +1.3%. Placebo windows moved forward by ten trading days move back towards zero, in line with desired ordering of sentiment change leading price movement.

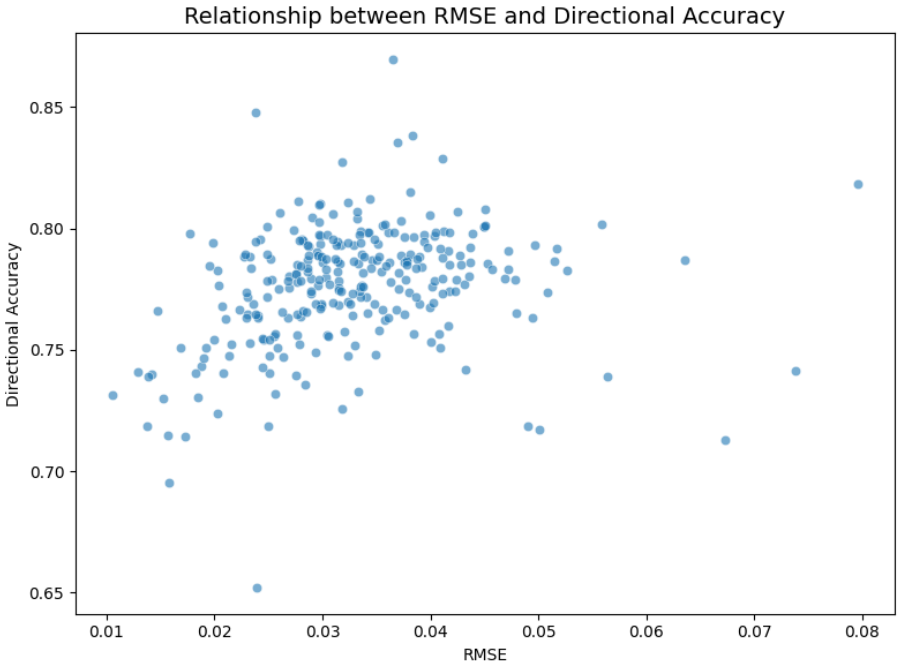

Figure 4: Relationship between directional accuracy and RMSE

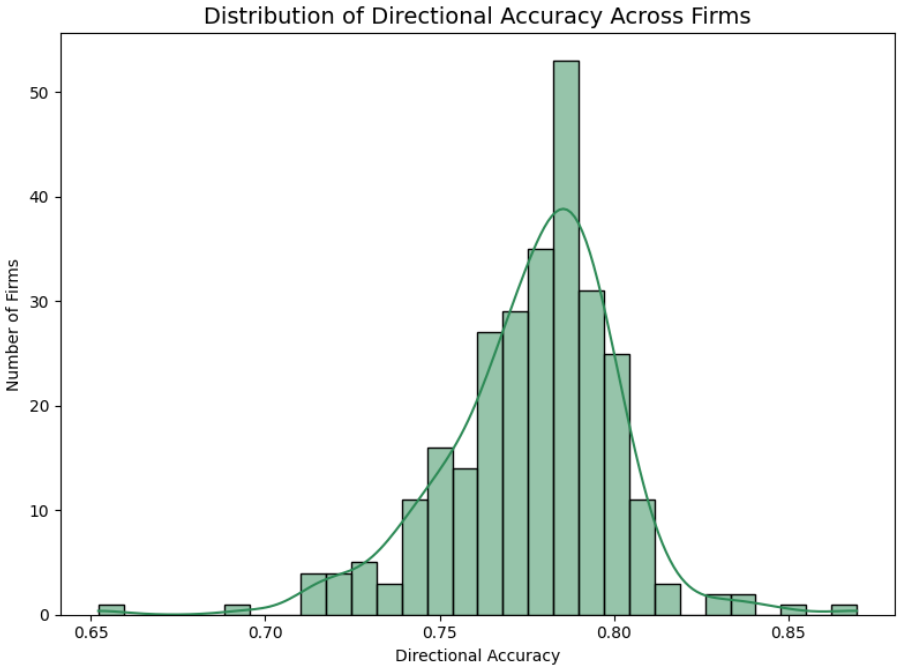

Figure 5: Distribution of directional accuracy across firms

5.4 Evaluation of Strategy Performance

Economic significance is determined with the help of portfolio simulations based on sentiment-informed forecasts. A long-only portfolio created using top decile of stocks with the highest predicted returns under sentiment-enhanced model has returns of about 0.8% per week relative to equal-weight benchmark. Risk-adjusted performance improves as well: the out-of-sample Sharpe ratio stands at about 0.26, compared to about 0.11 for the baseline strategy. And although transaction costs and turnover would be expected to eat into realized returns (especially on smaller and less liquid equities), the gap in performance is economically meaningful.

FIGURE 6 shows that the top 10% of the firms reaches Dir-Acc above 0.80, indicating a high and fairly stable level of predictive performance among the names that drive the trading rule. These firms are critical for the long-only strategy explored in the next few sections: linking the accuracy of forecasts to the impact on economy.

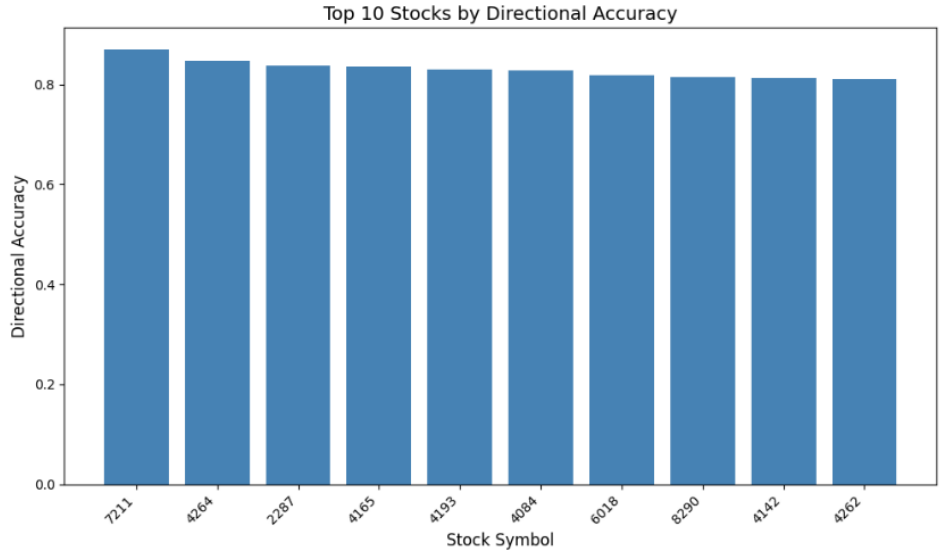

Figure 6: Stock symbols by directional accuracy

6. DISCUSSION

6.1 Interpretation of Findings

6.1.1 Signal prediction analysis

The bilingual sentiment affects the prediction of futures in addition to patterns that result in historical price movements. The increased accuracy of directional forecasts is observed via an average decrease in RMSE when the bilingual sentiment was included in the forecast creation process, indicating that short-term prediction of futures can be achieved by using news sentiment as a predictive

tool. These results reflect findings by [1, 14] that the effectiveness of news on predicting returns is evidenced on a global scale.

6.1.2 Broader economic implications

The economic performance differences created by today's most modernized capabilities are substantial. The long-only strategy targeted toward the best performing decile provided a positive excess return, thereby demonstrating that the nonlinear advantages linked with higher predictive accuracy enable improved trade profitability when consistently applied.

6.1.3 Timing effects and their implications

The increase in CAR following sentiment shock over time supports the theories of Barber and Odean [4] and Hong and Stein [26], which deal with limited attention and underreaction. Investors in Saudi Arabia appear to gradually assimilate multilingual sources of news information over several trading days, especially for larger companies that have received substantial media coverage.

6.1.4 Cross-language insights

The empirical findings indicate that the coefficients from the panel data analysis have a statistically significant positive relationship between real-time sentiment written in Arabic and English, which has the most substantial impact on directional movements of stock prices during shorter (short-term) periods. Thus, it is essential to incorporate multilingual sources of data when developing predictive models of stock price behavior, particularly within emerging market countries where investors are likely to receive their news from a wide variety of languages.

6.2 Positioning the Results Within Existing Research

This paper not only focuses on the development of the bilingual emerging market industry, but it is also one of the first studies on this topic in the context of the global literature environment. Consistent with [1], the authors of this research predict that negative sentiment will lead to short-term market reversals based on their findings; however, these short-term effects are weaker than observed in previous studies and hint at improved information efficiency since the implementation of Vision 2030. Unlike U.S. equity markets, where social media is one of the most important channels for conveying sentiment [8, 33], most signals in the Saudi context tend to come through official government announcements and established domestic financial media. The estimated coefficients for the Saudi market are typically smaller in magnitude, but are also internally consistent across specifications. In comparisons with China's market [23] and with other emerging-market settings, patterns are more consistent with stronger institutional participation and lower background noise in price formation.

7. CONCLUSION

Purpose of this study was to determine whether use of bilingual sentiment analysis, which captures both Arabic- and English-language news, enhances the prediction of returns when incorporated into machine-learning models for the Saudi equity market. This involved constructing and analyzing massive dataset of 279 Tadawul-listed firms observed in August 2025. The analysis was combination of structured financial variables and corpus of about 19,000 sentiment-scored news articles from outlets publishing in both languages. The empirical design was deliberately wide-ranging using both event-study tests, time-series forecasting models, fixed-effects panel regressions and out-of-sample portfolio evaluations.

The empirical results provide set of clear contributions:

Incremental predictive power: Incorporating sentiment materially strengthens forecasting performance on unseen data. RMSE decreases from 0.0342 to 0.0328 with an improvement of 4.1%, and directional accuracy increases from 55% to 77.6% which shows that sentiment features provide nonlinear signals in addition to what financial variables provide.

- Fixed-effects panel regressions provide support. After controlling for heterogeneity across firms and over time, sentiment measures are still statistically meaningful. The results also survive multiple-testing control. With false-discovery-rate corrections, around 70% of specifications still meet conventional significance thresholds indicating that the effect is not limited to one specification of choice.
- Results of event studies have economic relevance too. Positive sentiment shocks are followed by cumulative abnormal returns of around +1.0% to +1.3% over a single trading week. Such sizes are consistent with behavioral explanations related to gradual investor reaction and limited attention. Besides these, out-of-sample portfolio strategies act as a mirror to these results by showing that long-only strategies aimed at top-decile sentiment-based forecasts can generate approximately 0.8% in weekly excess returns.
- The institutional setting of Saudi Arabia sheds more light on the features that drive such results. In contrast to Western markets, where social media is often the main source of sentiment, the Saudi market largely depends on the well-organized financial news and bilingual corporate communication, partly influenced by the Vision 2030 reform. Despite increase in institutional participation, retail investors still seem sensitive to tone, narrative framing and overall flow of information. Beyond predictive performance, analysis demonstrates how bilingual sentiment features can be put to work in practice, for portfolio construction, risk monitoring and tactical asset allocation.
- Regulators should take advantage of sentiment aggregation as a likely source of early-warning signals for market exuberance, herd behavior, or emerging liquidity constraints.

8. DATA AVAILABILITY

Financial data used for this study were collected from TADAWUL (saudiexchange.sa) and are subject to its licensing restrictions. The author cannot redistribute these data, but they can be

obtained directly from TADAWUL in accordance with the access policies of the exchange (saudiexchange.sa). The sentiment dataset was compiled from publicly accessible news feeds, collected via the Google News RSS service. The sources for Arabic language content encompassed platforms such as Argaam, Tadawul, Mubasher, Aleqtisadiah, Asharq Business, AlArabiya, and Maal. For English-language content, the sources included Zawya, Reuters, and Bloomberg, alongside general news results in both Arabic and English retrieved through Google News. DOI: https://doi.org/10.5281/zenodo.18061436.

References

[1] Tetlock, PC. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. J Finance. 2007;62:1139–1168.

[2] Tetlock PC, Saar-Tsechansky M, Macskassy S. More Than Words: Quantifying Language to Measure Firms' Fundamentals. J Finance. 2008;63:1437-1467

[3] Fang L, Peress J. Media Coverage and the Cross-Section of Stock Returns. J Finance. 2009;64:2023-2052.

[4] Barber BM, Odean T. All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors. Rev Financ Stud. 2008;21:785-818.

[5] Da Z, Engelberg J, Gao P. In Search of Attention. J Finance. 2011;66:1461-1499.

[6] Engelberg JE, Parsons CA. The Causal Impact of Media in Financial Markets. J Finance. 2011;66:67-97.

[7] Antweiler W, Frank MZ. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. J Finance. 2004;59:1259-1294.

[8] Bollen J, Mao H, Zeng X. Twitter Mood Predicts the Stock Market. J Comput Sci. 2011;2:1-8.

[9] Hutto C, Gilbert E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the international AAAI conference on web and social media. 2014;8:216-225.

[10] Loughran T, McDonald B. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. J Finance. 2011;66:35-65.

[11] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies. Association for Computational Linguistics. 2019:4171-4186.

[12] Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, et al. Unsupervised Cross-Lingual Representation Learning at Scale. In: Proceedings of the 58th Annu Meet Assoc comput linguist. Stroudsburg, PA. Association for Computational Linguistics. 2020:8440-8451.

[13] Antoun W, Baly F, Hajj H. 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. European Language Resource Association. 2020:9-15.

[14] Heston SL, Sinha NR. News vs. Sentiment: Predicting Stock Returns From News Stories. Financ Anal J. 2017;73:67-83.

[15] Gurun UG, Butler AW. Don't Believe the Hype: Local Media Slant, Local Advertising, and Firm Value. J Finance. 2012;67:561-598.

[16] Cameron AC, Gelbach JB, Miller DL. Robust Inference With Multiway Clustering. J Bus Econ Stat. 2011;29:238-249.

[17] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Stat Methodol. 1995;57(1):289-300.

[18] Diebold FX, Mariano RS. Comparing Predictive Accuracy. J Bus Econ Stat. 1995;13:253-263.

[19] Kothari SP, Warner JB. Econometrics of Event Studies. In: Handbook of empirical corporate finance. Elsevier. 2007;1:3-36.

[20] MacKinlay AC. Event Studies in Economics and Finance. J Econ Lit. 1997;35:13-39.

[21] Boehmer E. Event-Study Methodology Under Conditions of Event-Induced Variance. J Financ Econ. 1991;30:253-272.

[22] Novy-Marx R. The Other Side of Value: The Gross Profitability Premium. J Financ Econ. 2013;108:1-28.

[23] Chen H, De P, Hu Y, Hwang BH. Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. Rev Financ Stud. 2014;27:1367-403.

[24] García D. Sentiment During Recessions. J Finance. 2013;68:1267-1300.

[25] Baker M, Wurgler J. Investor Sentiment and the Cross-Section of Stock Returns. J Finance. 2006;61:1645-1680.

[26] Hong H, Stein JC. A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets. J Finance. 1999;54:2143-2184.

[27] Dougal C, Engelberg J, García D, Parsons CA. Journalists and the Stock Market. Rev Financ Stud. 2012;25:639-679.

[28] Gu S, Kelly B, Xiu D. Empirical Asset Pricing via Machine Learning. Rev Financ Stud. 2020;33:2223-2273.

[29] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. 2014. ArXiv preprint: https://arxiv.org/pdf/1409.0473

[30] Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In: Proc 2014 Conf Empir Methods Nat Lang Process (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics. 2014:1724–1734.

[31] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997;9:1735-1780.

[32] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc.2017:6000–6010.

[33] Buraschi A, Carnelli A. Understanding Short-Versus Long-Run Risk Premia. Eur Financ Manag. 2014;20:714-738.