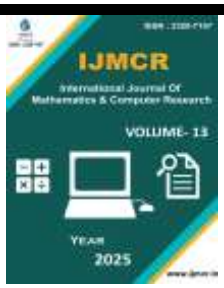
**International Journal of Mathematics and Computer Research**

**ISSN: 2320-7167**

**Volume 13 Issue 07 July 2025, Page no. –5466-5474**

**Index Copernicus ICV: 57.55, Impact Factor: 8.615**

**DOI: 10.47191/ijmcr/v13i7.16**




# Integrating Social Media Sentiments with Extreme Gradient Boosting for Stock Price Prediction: A Case Study on Safaricom PLC

**Joshua Kimani[1], Anthony Karanjah[2], Pius Kihara[3]**

[1,2]Department of Mathematics, Multimedia University of Kenya, Kenya

[3]Department of Financial and Actuarial Mathematics, Technical University of Kenya, Kenya

**ARTICLE INFO**

**Published Online:**
**23 July 2025**

**ABSTRACT**

In an increasingly data-driven financial ecosystem, the fusion of sentiment analysis with machine learning offers new frontiers for stock market prediction. This study explores the predictive power of social media sentiments, extracted from X (formerly Twitter), in modeling stock price movements with an application to Safaricom PLC. Leveraging the VADER lexicon for sentiment scoring and the Extreme Gradient Boosting (XGBoost) algorithm for regression, the analysis examined whether public sentiment could meaningfully enhance forecasting performance. Three cases were evaluated, each varying the input features: from hybrid models combining sentiment and financial data, to pure sentiment-driven predictions. Despite robust modeling and high cross-validation accuracy, the results revealed that sentiment features offered minimal advantage over traditional indicators. The dominance of neutral sentiments and company-specific market dynamics may explain this muted effect. These findings provide a grounded perspective on the practical limitations of sentiment integration and emphasize the need for broader, multi-firm analysis to validate the approach across diverse market contexts.

**Corresponding Author:**
**Joshua Kimani**

**KEYWORDS:** Stock price prediction, sentiment analysis, financial forecasting, Nairobi Securities Exchange (NSE).

## I. INTRODUCTION

This With the rise of the internet and the World Wide Web, internet-based media has become a significant platform for disseminating information. The influence of this vast pool of online content, particularly from social media, on stock prices and financial markets has been the subject of numerous studies [1]. Stock markets are often affected by both domestic and international news events [2]. Traders commonly exhibit behaviors such as selling stocks in response to negative news and purchasing stocks when positive news emerges. Negative events such as poor earnings reports, corporate governance issues, macroeconomic uncertainties, and political instability often lead to panic selling, resulting in a decline in stock prices. Conversely, positive economic indicators, strong earnings reports, new product launches, and corporate acquisitions tend to create buying momentum, driving stock prices higher [3].

Emotions play a crucial role in decision-making, including stock market investment choices [4]. During periods of pessimism or uncertainty, individuals tend to be more cautious, which in turn influences investment decisions. This emotional state can, therefore, be a potential predictor of stock market movements. Recognizing the collective sentiment of the public—capturing the mood of the masses—may offer valuable insights into stock price trends. Researchers [5] explored the relationship between stock volatility and trading volume on internet message boards, emphasizing the impact of public sentiment on market behavior. Utilizing over 20 million posts from the LiveJournal website, a study developed the Anxiety Index to measure the national mood in the US [6]. Findings revealed a correlation between public sentiment, financial news, and stock price movements, showing that shifts in public mood could significantly influence the performance of the S&P 500.

The connection between news events and stock returns has been widely studied by economists. However, there has been no systematic approach to tracking these events and their direct influence on stock market performance [7]. This gap has led to individual initiatives that analyze the sentiment of social media posts and other financial news [8]. The current study adopts content analysis using the Vader lexicon to explore the relationship between positive, neutral, and negative

sentiments expressed in financial news on social media and their potential impact on stock prices [9].

In today's digital era, social media serves as a powerful tool for gauging public opinion and sentiment about current events. Sentiment analysis, which involves understanding the emotions expressed in text, has proven to be a useful method for predicting market trends. Optimistic posts on social media, for instance, could influence investors to purchase stocks, subsequently driving up stock prices. This method has gained significant attention in the finance community, with researchers leveraging linguistic and text analysis tools to evaluate the sentiments of news articles [10] [11]. By incorporating sentiment into existing financial models, these approaches have the potential to enhance market predictions.

Extending prior research, analyzed social media posts about Safaricom PLC, specifically focusing on posts from X (formerly Twitter), were classified using the VADER lexicon [12]. While the study suggested a potential link between online sentiment and stock price movements, it stopped short of integrating sentiment data into a predictive model. The current study builds on this foundation by combining sentiment analysis with Extreme Gradient Boosting (XGBoost), a powerful machine learning model, to predict stock price movements for Safaricom PLC. This integration seeks to empirically validate the impact of social media sentiment on financial forecasting and improve the accuracy of stock price prediction models within the
Kenyan market context.

## II. RELATED WORKS

Stock price prediction has gained significant attention in recent years due to its implications for financial decision-making and investment strategy. Accurately forecasting stock movements remains challenging given the nonlinear, volatile, and complex behavior of financial markets. In response, a wide range of machine learning techniques, ranging from traditional statistical models to advanced deep learning architectures, have been explored to address this challenge.

Deep learning models have emerged as powerful tools in this domain, capable of capturing intricate temporal and semantic patterns in financial and textual data. For instance, studies have shown that deep learning models outperform Support Vector Machines (SVMs) in stock market forecasting tasks, particularly when enhanced with neural tensor networks that incorporate word and structured event embeddings [13]. Further improvements were realized by combining convolutional and recurrent neural layers with pre-trained word vectors to predict major indices like the S&P 500, resulting in enhanced forecasting accuracy [14]. In parallel, Bidirectional Encoder Representations from Transformers (BERT) were fine-tuned on financial news data to extract sentiment features, underlining the utility of natural language processing in market prediction [15]. Additionally, hybrid models combining CNNs and LSTMs have been developed to better handle short-term volatility, while attention mechanisms have been incorporated to dynamically weigh the relevance of input features, improving both accuracy and interpretability [16]. Nevertheless, challenges such as data scarcity, noise, and market instability continue to limit the effectiveness of these models, particularly in less liquid markets [17].

In the context of emerging markets like Kenya, deep learning models have shown promising results on the Nairobi Securities Exchange (NSE). Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models have been successfully applied to forecast stock prices on the NSE, with reported accuracies as high as 86.7% and 89.2% respectively [18, 19, 20]. These findings affirm the feasibility and relevance of applying sophisticated machine learning models to local financial markets.

Traditional machine learning algorithms such as SVMs have also demonstrated strong performance in both global and Kenyan contexts. For instance, the use of SVMs with Radial Basis Function (RBF) kernels outperformed other techniques like Random Forest and Gradient Boosting in predicting stock movements in the Saudi stock market [21]. Locally, SVMs have been used to model the stock prices of NSE-listed firms by integrating technical indicators and macroeconomic variables such as inflation and exchange rates, with results showing high levels of predictive accuracy and model robustness [22, 23]. Similarly, logistic regression has been utilized for stock price prediction using a mix of financial ratios and technical indicators. Its performance has been found to be competitive, particularly when models are extended to incorporate macroeconomic factors and historical data, including in the Kenyan context [24, 25].

Other algorithms such as the K-Nearest Neighbors (KNN) have also been adapted for financial prediction. A dynamic KNN approach, where the number of neighbors varies based on market volatility, demonstrated superior accuracy compared to static versions [26]. On the NSE, KNN models have achieved up to 60% accuracy, effectively capturing nonlinear dependencies in local stock trends [27]. Ensemble methods like Random Forests have also gained popularity due to their robustness and generalizability. When combined with technical indicators, Random Forests have consistently outperformed single-model approaches in both developed and developing markets [28, 29]. In the Kenyan market, Random Forest models were used to predict the volatility of the NSE-20 index, yielding promising results with a mean absolute error of 0.28% [30].

Building on these foundations, this study incorporates previously classified Twitter sentiments related to Safaricom PLC, where tweets were categorized based on their relevance to market behavior [12]. These sentiment classifications are integrated into an Extreme Gradient Boosting (XGBoost) model to predict the performance of Safaricom's stock. By combining social media sentiment with ensemble learning, the study offers a novel approach to stock price prediction within the Kenyan financial landscape, focusing on

Safaricom PLC, one of the most actively traded equities on the Nairobi Securities Exchange.

## III. METHODOLOGY

This section presents the methodological framework employed to predict stock performance. The approach integrates sentiment analysis of social media posts with a machine learning based predictive model. The methodology comprises three core components: the research design, the application of the Extreme Gradient Boosting (XGBoost) algorithm, and the subsequent testing and validation of the predictive model. All analyses were implemented using Python and its relevant packages.

### *A. Research Design*

This study employs a machine learning approach to model and predict the performance of Safaricom PLC stock by integrating social media sentiment data with historical stock market data. The sentiment data were sourced from a prior study [12], which classified tweets related to Safaricom PLC into positive, negative, or neutral categories.

In the referenced study, tweets collected from X (formerly Twitter) were preprocessed and analyzed using sentiment classification techniques. The VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon was employed to score each tweet. The compound score was particularly critical for classification. The compound score aggregates the overall sentiment expressed in a tweet into a single value ranging from approximately -1 (most negative) to +1 (most positive). In the dataset used, compound scores ranged from -0.9595 to 0.9657.

A common and intuitive thresholding approach was adopted: tweets with compound scores greater than 0.1 were categorized as positive, those below -0.1 as negative, and those in between as neutral. This classification framework helped to assign categorical sentiment labels to each tweet, thereby classifying social media sentiments into positive, neutral, and negative categories.

The resulting sentiment labels were then aggregated on a daily basis to create a time-indexed sentiment dataset. This dataset reflects public opinion trends over the study period and was aligned with daily historical stock price data for Safaricom PLC to form a unified dataset. The combined dataset includes sentiment scores (compound, positive, negative, neutral) and key market indicators (open, high, low, and close prices) serving as input features for the stock price prediction model.

**Table 1 Sample Merged Dataset with Sentiment Scores and Market Prices**

| Date | Compound | Positive | Negative | Neutral | Open | High | Low | Close | Class Type |
|---|---|---|---|---|---|---|---|---|---|
| 30/12/2022 | 0.0929 | 0.1012 | 0.0336 | 0.8652 | 24.00 | 24.50 | 24.00 | 24.15 | POSITIVE |
| 29/12/2022 | 0.0830 | 0.1019 | 0.0447 | 0.8534 | 24.50 | 24.50 | 23.90 | 24.00 | POSITIVE |
| 28/12/2022 | 0.0357 | 0.1030 | 0.0647 | 0.8322 | 24.70 | 24.70 | 24.30 | 24.45 | POSITIVE |
| 23/12/2022 | 0.0637 | 0.1218 | 0.0627 | 0.8155 | 24.75 | 24.80 | 24.30 | 24.60 | POSITIVE |
| 22/12/2022 | 0.0611 | 0.1230 | 0.0531 | 0.8239 | 25.00 | 25.00 | 24.50 | 24.70 | POSITIVE |

Table 1 presents a sample of the merged dataset combining social media sentiment scores and historical stock prices for Safaricom PLC, with the class type, sentiment scores, and market indicators serving as inputs for the predictive model.

### *B. Proposed Extreme Gradient Boosting Model*

To predict the stock prices of Safaricom PLC, this study employed the Extreme Gradient Boosting (XGBoost) algorithm, a scalable and efficient implementation of gradient boosted decision trees. The model integrates social media sentiment features and historical stock attributes to forecast future stock performance.

Given a training set $\{(X_i, Y_i)\}_{i=1}^{N}$, $X_i$denotes the input features (i.e., sentiment class, open, high, low prices) and $Y_i$ is the corresponding target (close price), the XGBoost model optimizes a differentiable loss function $L(y, \hat{f}(x))$ through a series of additive functions. A learning rate $\alpha$ and a number of boosting rounds $M$, typically determined via early stopping to avoid overfitting, guide the learning process.

**Step 1: Initialize the model with a constant value**

$$\hat{f}_{(0)}(x) = arg\ \min_{\theta} \sum_{i=1}^{N} L\left(y_i, \hat{f}(x)\right) \quad (1)$$

**Step 2: For *m* = 1 to M, perform the following:**

**a) Compute gradients and hessians:**

$$\hat{g}_{(m)}(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f = \hat{f}_{(m-1)}(x)} \quad (2)$$

$$\hat{h}_{(m)}(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f=\hat{f}_{(m-1)}(x)} \quad (3)$$

**b) Fit a base learner $\hat{\phi}_m$ by minimizing:**

$$\hat{\phi}_m = arg\min_{\phi\in\Phi} \sum_{i=1}^{N} \frac{1}{2}\hat{h}_m(x_i)\left[-\frac{\hat{g}_{(m)}(x_i)}{\hat{h}_{(m)}(x_i)} - \phi(x_i)\right]^2 \quad (4)$$

**c) Update the model:**

$$\hat{f}_{(m)}(x) = \hat{f}_{(m-1)}(x) + \alpha\hat{\phi}_m(x) \quad (5)$$

**Step 3: Final model**

$$\hat{f}(x) = \sum_{m=0}^{M} \hat{\phi}_m(x) \quad (6)$$

This iterative process builds a sequence of decision trees, each correcting the residual errors of its predecessor, as illustrated in Figure 1. The final model aggregates the output of all trees, weighted by the learning rate $\alpha$.

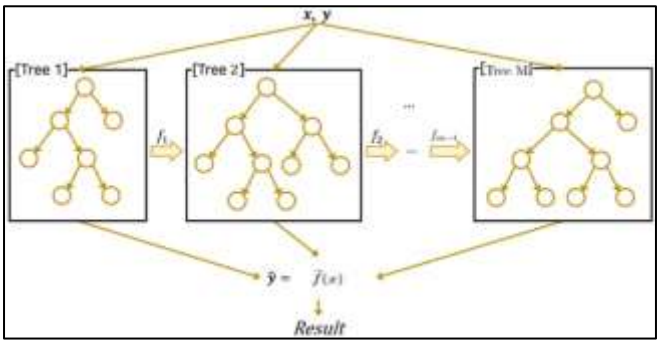



**Figure 1 Conceptual XGBoost Model**

During training, XGBoost optimizes the tree structure and feature splits while applying regularization techniques to prevent overfitting. Once trained, the model makes predictions on unseen data by applying the learned function $\hat{f}(x)$ to new input features. The implementation was conducted using the xgboost library in the Python programming language [31].

*C. Model Testing and Validation*

To assess the reliability and performance of the Extreme Gradient Boosting (XGBoost) model, evaluation was conducted using Mean Squared Error (MSE) and *K*-fold cross-validation.

The Mean Squared Error (MSE) measures the average squared difference between actual and predicted stock prices. A lower MSE value indicates better predictive accuracy. It is computed as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \quad (7)$$

where $n$ is the number of observations, $Y_i$ is the actual value, and $\hat{Y}_i$ is the predicted value.

To further evaluate model generalization, $K$-fold cross-validation was applied. The training dataset was partitioned into $K$ equal subsets $D_1, D_2, \ldots, D_k$. In each iteration, the model was trained on $K-1$ subsets and validated on the remaining one. This process was repeated $K$ times such that each fold served as the validation set exactly once.

The procedure is described as follows:

1. Partition the dataset:
   $D_1 \cup D_2 \cup \ldots \cup D_k$.
2. Model training and validation:
   For each $i = 1, 2, \ldots, K$, train the model on $D/Di$ and validate it on $Di$.
3. Performance evaluation:
   Calculate the MSE for each fold $Mi$, where $i = 1, 2, \ldots, K$.
4. Average performance metric:
   Compute the average MSE:

$$\bar{M} = \frac{1}{K}\sum_{i=1}^{K} M_i \quad (8)$$

## IV. RESULTS & DISCUSSIONS

*D. Model Training for Stock Price Prediction*

This section presents the training and evaluation of the Extreme Gradient Boosting (XGBoost) regression model for stock market price prediction. The training process involved analyzing various combinations of input features to assess their effect on the model's predictive accuracy. Three cases were considered, each with a different configuration of input features (X), while the target variable (y) remained the stock's closing price.

Sentiment categories—positive, negative, and neutral—were encoded as numeric values: positive as 0, negative as 1, and neutral as 2. These encodings enabled integration into the model alongside numerical stock data.

In the first case, the input features comprised the encoded sentiment classes along with the stock's Open, High, and Low prices. The goal was to evaluate the additional contribution of sentiment information to the model's performance. A sample of the input features for this case is provided in Table 2.

**Table 2 Sample of Case 1 X Features**

| Class Scores | Open | High | Low |
|---|---|---|---|
| 0 | 24.50 | 24.50 | 23.90 |
| 0 | 30.50 | 30.60 | 29.95 |
| 0 | 25.95 | 26.00 | 25.15 |
| 0 | 38.75 | 39.50 | 38.65 |
| 0 | 26.80 | 27.00 | 26.40 |
| 0 | 26.45 | 26.85 | 26.05 |
| 0 | 26.80 | 26.80 | 26.00 |
| 0 | 24.65 | 25.10 | 24.65 |
| 0 | 37.50 | 37.70 | 37.40 |
| 1 | 31.50 | 32.50 | 31.10 |

The dataset was split into training and testing subsets in an 80:20 ratio, resulting in 196 observations for training and 49 for testing. Table 3 displays a sample of the actual and

predicted values from the testing set, along with the corresponding residuals computed as the difference between actual and predicted values.

**Table 3 Actual, Predicted, and Residuals Sample from Case 1**

| Actual | Predicated | Residuals |
|---|---|---|
| 25.00 | 24.970978 | 0.029022 |
| 24.80 | 24.707245 | 0.092755 |
| 30.45 | 30.187998 | 0.262002 |
| 37.65 | 37.573555 | 0.076445 |
| 36.70 | 36.899155 | -0.289155 |
| 35.10 | 35.091743 | 0.008257 |
| 35.65 | 36.123512 | -0.473512 |
| 23.35 | 23.413219 | -0.063219 |
| 24.80 | 24.938412 | -0.138412 |

Model training was conducted for 100 boosting rounds (M), with corresponding Root Mean Squared Error (RMSE) values recorded as shown in Table 4. Initially, at boosting round 0, the RMSE was 21.29, reflecting high prediction error. The error progressively declined with each boosting round, ultimately reaching 0.01149 at round 99, indicating high model precision.

**Table 4 Boosting Rounds and Corresponding RMSE – Case 1**

| Boosting Round | MSE |
|---|---|
| 0 | 21.28987 |
| 1 | 15.03395 |
| 2 | 10.61997 |
| 3 | 7.51272 |
| 4 | 5.32226 |
| . | . |
| . | . |
| . | . |
| 97 | 0.01199 |
| 98 | 0.01181 |
| 99 | 0.01149 |

The relationship between actual and predicted values is depicted in Figure 2. Most points align closely with the diagonal line, demonstrating the model's accuracy. However, a deviation is noticeable beyond the KES 33 price level.

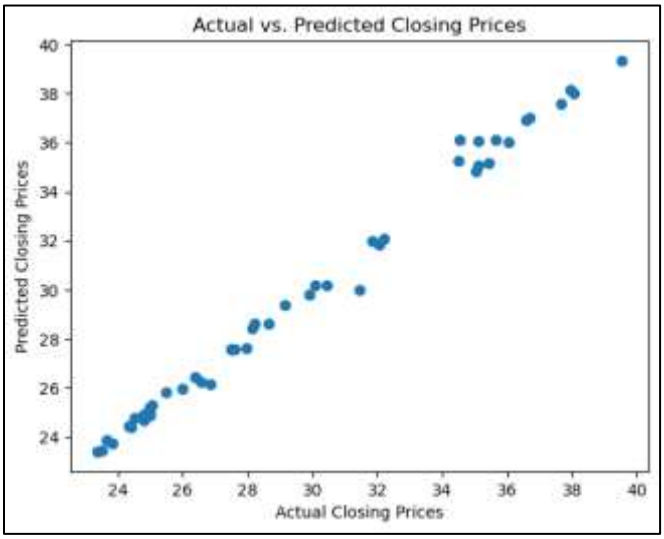



**Figure 2 Relationship between Actual and Predicted Closing Prices in Case 1**

Figure 3 presents the residuals plot, revealing a positive trend in residuals as predicted prices increase. This indicates a consistent underestimation by the model for prices exceeding KES 33. The trend suggests that key influencing factors at higher price levels may have been omitted from the feature set.

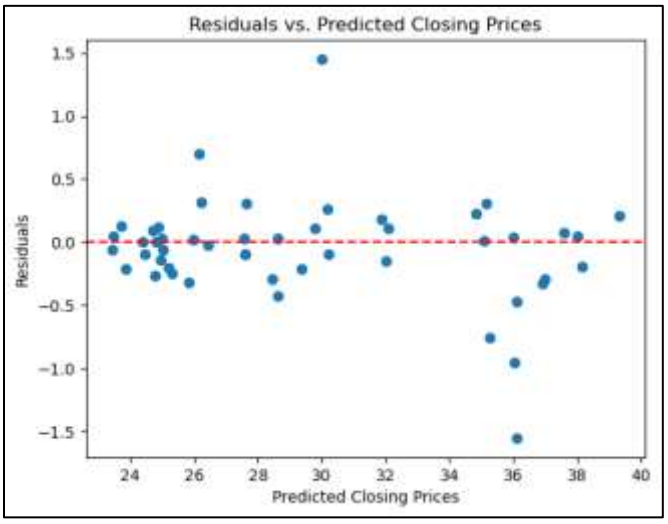



**Figure 3 Residuals Plot for Case 1**

In the second case, only historical stock features—Open, High, and Low prices—were used as input variables, excluding sentiment data. The results, shown in Table 5, reveal slightly different performance characteristics but follow a similar pattern to Case 1.

**Table 5 Actual, Predicted, and Residuals Sample from Case 1**

| Actual | Predicated | Residuals |
|---|---|---|
| 25.00 | 24.979778 | 0.020222 |
| 24.80 | 24.742350 | 0.057650 |
| 30.45 | 30.207863 | 0.242137 |

| | | |
|---|---|---|
| 37.65 | 37.598160 | 0.051840 |
| 36.70 | 37.001274 | -0.301274 |
| 35.10 | 35.103302 | -0.003302 |
| 35.65 | 36.184277 | -0.534277 |
| 23.35 | 23.413839 | -0.063839 |
| 24.80 | 24.926394 | -0.126394 |

RMSE values for Case 2, as shown in Table 6, followed a similar downward trend. Starting from 21.29, the RMSE reduced to 0.01167 by the 99th boosting round, affirming strong model performance.

**Table 6 Boosting Rounds and Corresponding RMSE – Case 2**

| Boosting Round | MSE |
|---|---|
| 0 | 21.28987 |
| 1 | 15.03395 |
| 2 | 10.61997 |
| 3 | 7.51272 |
| 4 | 5.32226 |
| . | . |
| . | . |
| . | . |
| 97 | 0.01198 |
| 98 | 0.01183 |
| 99 | 0.01167 |

The scatter plot in Figure 4 further confirms the model's accuracy, with predicted values closely matching actual prices. Similar to Case 1, a gap appeared beyond the KES 33 price level, again suggesting potential external influences not captured in the features.

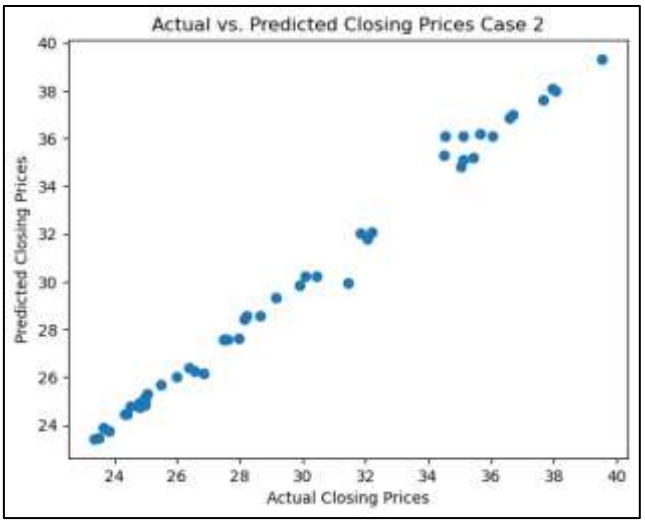



**Figure 4 Relationship between Actual and Predicted Closing Prices in Case 2**

In the third case, scaled sentiment class scores were used as the sole input features for predicting stock closing prices. This case was designed to assess the standalone predictive capability of social media sentiment, independent of historical market data.

The model was trained using the same 80:20 data split. Table 7 presents a sample of the actual, predicted, and residual values for the test dataset under this configuration.

**Table 7 Actual, Predicted, and Residuals Sample from Case 3**

| Actual | Predicated | Residuals |
|---|---|---|
| 25.00 | 24.894268 | 0.105732 |
| 24.80 | 24.688568 | 0.111432 |
| 30.45 | 30.089899 | 0.360101 |
| 37.65 | 37.395192 | 0.254808 |
| 36.70 | 36.823539 | -0.123539 |
| 35.10 | 35.124204 | -0.024204 |
| 35.65 | 35.864593 | -0.214593 |
| 23.35 | 23.294840 | 0.055160 |
| 24.80 | 24.899769 | -0.099769 |

Despite relying solely on sentiment data as input, the model in Case 3 demonstrated a gradual improvement in performance across boosting rounds, as shown in Table 8. The RMSE decreased from 1.01382 at round 0 to 0.16571 by round 99. However, this was the least accurate outcome among the three cases. The results indicate that while sentiment data alone offers some predictive value, it lacks the robustness required for high-precision forecasting when compared to models trained on more comprehensive feature sets.

**Table 8 Boosting Rounds and Corresponding RMSE – Case 3**

| Boosting Round | MSE |
|---|---|
| 0 | 21.28987 |
| 1 | 15.03395 |
| 2 | 10.61997 |
| 3 | 7.51272 |
| 4 | 5.32226 |
| . | . |
| . | . |
| . | . |
| 97 | 0.01228 |
| 98 | 0.01216 |
| 99 | 0.01205 |

The actual versus predicted plot in Figure 5 shows how compound sentiment scores relate to actual and predicted prices in Case 3. Dotted points represent actual prices, while crossed points show predictions. Overlaps indicate accurate

predictions, whereas visible gaps highlight instances of reduced model precision.

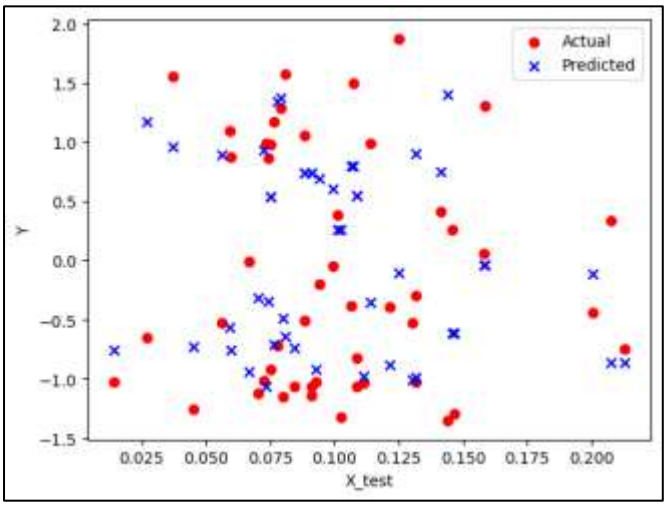



**Figure 5 Relationship between Actual and Predicted Closing Prices in Case 3**

*E. Model Evaluation and Validation*

This phase assessed the performance of the XGBoost regression model, enabling informed adjustments to improve accuracy and robustness.

Mean squared error quantified the average squared difference between predicted and actual closing prices (Equation 7). The results for the three cases were:

- Case 1 MSE: 0.1681
- Case 2 MSE: 0.1675
- Case 3 MSE: 1.6193

Case 3, which relied solely on the compound sentiment score as input, exhibited the highest error by a significant margin. Cases 1 and 2 showed nearly identical MSE values, with Case 2 marginally outperforming Case 1 by 0.0006. This minimal difference suggests that including sentiment features alongside traditional predictors provided negligible improvement. Overall, models incorporating multiple features (Cases 1 and 2) consistently outperformed the single feature model in Case 3.

A 10-fold cross-validation was conducted to evaluate model stability and generalization:

1. The dataset was split into ten equal folds.
2. For each iteration, one-fold was used for validation while the remaining nine formed the training set.
3. The model was trained and evaluated iteratively across all folds.
4. The average MSE across folds was computed.

Cross-validation scores were:

- Case 1: 0.9949
- Case 2: 0.9947
- Case 3: -1.001

Cases 1 and 2 achieved similarly high accuracy, confirming the reliability of models with multiple input features. In contrast, Case 3's negative score indicated poor predictive performance, reinforcing that relying solely on sentiment scores drastically reduces model effectiveness.

## V. SUMMARY & RECOMMENDATIONS

This study explored the application of an Extreme Gradient Boosting (XGBoost) regression model for predicting the stock price of Safaricom PLC using both financial indicators and sentiment data extracted from social media. The modelling process involved three distinct cases: the first combined sentiment classes with traditional financial variables such as opening, high, and low prices; the second used only these financial indicators; and the third relied solely on the compound sentiment score. The results indicated that the model performed well in Cases 1 and 2, achieving mean squared errors of 0.1681 and 0.1675 and cross-validation scores of 0.9949 and 0.9947, respectively. However, Case 3 underperformed, highlighting the limitations of using only the compound sentiment score as a predictor.

The findings suggest that while social media sentiment can be integrated into predictive models, its marginal impact on improving predictive performance, at least for Safaricom PLC, was limited. This may be attributed to the predominance of neutral sentiment in the analyzed tweets and the possibility that stock price movements for the company are less sensitive to public sentiment. As such, sentiment features did not provide a substantial advantage over traditional financial indicators in forecasting stock prices in this context.

Despite the limited effect of sentiment data, the study demonstrates the feasibility of incorporating sentiment analysis into financial modeling. It is recommended that future research apply this framework across other companies and industries within the Nairobi Securities Exchange and other markets to evaluate whether sentiment exerts a stronger influence elsewhere.

This approach could be particularly beneficial for companies with higher public engagement or exposure to sentiment-driven volatility. Investors, analysts, and researchers could benefit from models that integrate both sentiment and financial indicators, especially in more sentiment-sensitive markets.

Several limitations were observed during the study. The analysis was restricted to data from the year 2022, which limits the ability to capture long-term market patterns or seasonal effects. The focus on a single organization also constrains the generalizability of the findings. Furthermore, the sentiment data was collected solely from X (formerly Twitter), potentially excluding relevant sentiment from other platforms. The model also did not account for macroeconomic, geopolitical, or regulatory factors, all of which may significantly influence stock prices.

To enhance and build upon this study, future research should expand the scope to include multiple organizations across different sectors of the Nairobi Securities Exchange and other

markets. Integrating fundamental indicators such as price-to-earnings ratios, earnings per share, and book value could offer a more comprehensive view of the drivers of stock prices.

Incorporating time-series forecasting techniques such as ARIMA and GARCH may help in modelling the temporal dynamics of both sentiment and price data. Comparative studies involving machine learning models like LSTM and traditional statistical models such as linear regression and autoregressive approaches could offer deeper insight into the relative strengths and limitations of these methods. Moreover, aggregating sentiment data from multiple platforms would ensure a more complete sentiment profile, further improving the robustness and predictive power of the models.

**REFERENCES**


1. De Carosia, A. E. O., Coelho, G. P., & da Silva, A. E. A. (2021). Investment strategies applied to the Brazilian stock market: A methodology based on sentiment analysis with deep learning. *Expert Systems with Applications*, 184, 115470.
2. Alqahtani, A. F., et al. (2021). Stock market prediction using social media sentiment analysis. *Journal of Financial Data Science*, 6(1), 45-67.
3. Smith, J. D. (2019). Financial news, market behavior, and stock price movements. *Journal of Economic Behavior*, 102, 77-94.
4. Gilbert, E., et al. (2010). Emotion and sentiment in social media: A survey of sentiment analysis techniques. *Journal of Computational Linguistics*, 42(2), 101-130.
5. Antweiler,W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of Internet stock message boards. *The Journal of Finance*, 59(3), 1259-1294.
6. Gilbert, E., & Karaholios, K. (2010). Predicting the mood of the crowd on social media. *Journal of Social Computing*, 12(4), 50-65.
7. Zhang, W., & Kim, Y. (2019). News events and stock returns: A new methodology for market prediction. *Journal of Financial Markets*, 22(3), 221-245.
8. Hutto,C., et al. (2014).VADER:Aparsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the Eighth International Conference on Weblogs and Social Media*, 216-225.
9. Hutto, C. J., & Gilbert, E. E. (2014). Vader: A parsimonious and effective sentiment analysis tool. *Proceedings of the Eighth International Conference on Weblogs and Social Media*, 216-225.
10. Tetlock, P. C., et al. (2007). All the news that's fit to reprint: Do investors react to stale information? *Review of Financial Studies*, 20(3), 929-957.
11. Xing, Z., et al. (2017). Financial market prediction with sentiment analysis. *International Journal of Financial Engineering*, 4(2), 125-139.
12. Kimani, J., Karanjah, A., & Kihara, P. (2024). Sentiment classification of Safaricom PLC social media sentiments on X (formerly Twitter). *Asian Journal of Probability and Statistics*, 26(6), 31–40. https://doi.org/10.9734/ajpas/2024/v26i6622
13. Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep learning for event-driven stock prediction. In *Twenty-Fourth International Joint Conference on Artificial Intelligence; AAAI Publications: Menlo Park, CA, USA*.
14. Di Persio, L., & Honchar, O. Artificial neural networks architectures for stock price prediction: Comparisons and applications. *International Journal of Circuits, Systems and Signal Processing*, 10, 403–413.
15. Sousa, M. G., Sakiyama, K., Rodrigues, L. D. S., Moraes, P. H., Fernandes, E. R., &Matsubara, E. T. (2019). BERT for stock market sentiment analysis. In *Proceedings of the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, 4–6 November*, pp. 1597-1601.
16. Chen, J., Song, Y., Weng, Y., Zhang, Y., & Hsu, C. (2020). Stock price prediction with a hybrid model. *Expert Systems with Applications.*
17. Zhang, Y., & Wu, L. (2020). Attention-based deep learning model for stock price prediction. *Expert Systems with Applications.*
18. Gitonga, J., et al. (2019). Long short-term memory (LSTM) recurrent neural networks for stock price prediction: A case study of Nairobi Securities Exchange. *International Journal of Scientific and Technology Research*, 8(9), 1839–1845.
19. Ngugi, M., Mwangi, W., & Kamau, E. (2019). Recurrent neural networks for stock price prediction: A case study of the Kenyan financial market. *International Journal of Computer Applications.*
20. Mwangi, K., et al. (2021). Stock price prediction using recurrent neural networks: A case study of the Nairobi Securities Exchange. *Journal of Business and Finance Research*, 9(1), 33-41.
21. Al-Shayea, Q., & AbdelHameed, A. (2019). Stock prices prediction using machine learning algorithms. In 2019 *International Conference on Intelligent Computing and Its Applications* (pp. 12-17). IEEE.
22. Mwangi, E. M.,&Wafula, R.N. (2018).Amachine learning-based stock price prediction model for the Nairobi Securities Exchange. *Journal of Business and Management*, 20(7), 30-38.
23. Njoroge, A., Mwangi, W., & Kamau, E. (2020). Enhancing the interpretability of support vector machines in predicting stock prices: A case study of the Nairobi Securities Exchange. *In 2020*

*International Conference on Information and Communication Technology and Systems (ICTS)*.

24. Yu, L., & Yuan, Y. (2021). A logistic regression-based model for stock price prediction integrating financial ratios and technical indicators. *Economic Research-Ekonomska Istraživanja*, 34(1), 1819-1837.
25. Wanjohi, R., Mwangi, W., & Kamau, E. (2019). Forecasting stock prices using logistic regression: A case of the Nairobi Securities Exchange. In *2019 International Conference on Information and Communication Technology and Systems (ICTS)*.
26. Wongkaew, S., Chantaruk, T., & Khongprasert, N. (2018). Stock price prediction using dynamic k-nearest neighbors. *Applied Artificial Intelligence*, 32(8), 958-984.
27. Mwangi, J., Gichuki, J.,&Maina, D. (2021). Application of K-nearest neighbor algorithm in predicting stock prices of companies listed in the Nairobi Securities Exchange. *Journal of Data Analysis and Information Processing*, 9(1), 1-10. doi: 10.4236/jdaip.2021.91001
28. Tsai, C. F., Lu, L. C., & Chou,W. C. (2012). A hybrid model of support vector regression and fuzzy time series for stock price prediction. *Expert Systems with Applications*, 39(1), 139-147.
29. Akbas, S., Ozturk, I., & Arik, S. (2018). Stock price prediction using machine learning techniques: A comparative study on linear regression, support vector machines, and random forest algorithms. *Expert Systems with Applications*, 109, 84-91.
30. Kiptoo, R., Ndirangu, L., & Kavoi, M. (2021). Predicting volatility in the Nairobi Securities Exchange using random forest. *Journal of Finance and Investment Analysis*, 10(3), 20-36.
31. XGBoost Developers. (2022). XGBoost parameters. *XGBoost Documentation*. Retrieved from https://xgboost.readthedocs.io/en/latest/parameter.html