ISSN 2723-6471

Applied Random Forest Algorithm for News and Article Features on The Stock Price Movement: An Empirical Study of The Banking Sector in Vietnam

Nguyen Minh Nhat[1] *

[1]*Faculty of Banking, Ho Chi Minh University of Banking (HUB), Ho Chi Minh City, Vietnam*

(Received: June 19, 2024; Revised: July 08, 2024; Accepted: August 30, 2024; Available online: September 4, 2024)

Abstract

In 2023, in the context of the world economic and political situation continuing to experience many difficulties and challenges, the global stock market has suffered many unfavorable impacts. In that general context, Vietnam's stock market faces many problems, challenges, and strong fluctuations due to unexpected changes in the world's macro economy and geopolitics. Therefore, the study's goal is to investigate the impact of news articles on the stock price movement of commercial banks in Vietnam. Using a dataset of 94,784 news articles from January 2023 to April 2024 and applying the Random Forest algorithm, the author analyzes the significance of various news features. The study identifies that the proportion of news sources with positive evaluations and the proportion of news sources mentioning commercial banks are the most influential features of the stock price movement. The findings reveal that positive news boosts investor confidence, increasing stock prices, while high media attention significantly influences trading activity. Other notable features include the number of news sources and the total sentiment score of articles, which also play crucial roles. This research provides valuable insights for investors and analysts to understand the effect of news articles on stock prices, enhancing their decision-making process in the banking sector. Finally, the research results are scientific proof that helps the Vietnamese stock market to have more positive and robust changes, continue to be an attractive destination for domestic and foreign investment capital flows, and a channel for medium and long-term capital important term for the economy, making an increasingly more outstanding contribution to the country's socio-economic development in the new era.

Keywords: News Articles, Stock Price Movement, Banking Sector, Random Forest, Investor Sentiment, Media Attention

1. Introduction

In the context of modern finance, news articles play an important role in shaping investor perceptions and decision-making [1], which in turn impacts the price movements of stocks in the market. News spreads quickly through various media channels, prompting continuous reactions from investors and other stakeholders. This is particularly true for the banking sector, where information about interest rates, regulatory changes, financial performance, and economic news tends to have a swift and significant impact on the value and price of stocks. Study demonstrated that the sentiment of news can predict stock price movement. Positive news tends to drive stock prices up, while negative news leads to a decline in stock prices [2]. However, the other study emphasized that there are two challenges in prediction: (i) determining the extent of the impact of news articles; (ii) identifying which features of these articles are likely to have a significant influence on stock price movements [3]. The study also recognized that traditional analysis methods are often insufficient to handle the complexity and large volume of news data, necessitating the application of more sophisticated and modern techniques.

The others indicated that machine learning models such as random forest are among the powerful solutions to this problem. Random forest, a type of ensemble learning method, can handle large datasets and recognize nonlinear relationships between variables, thereby determining the importance of news article features in forecasting stock price directions, particularly for news-sensitive stocks such as those in the banking sector [4]. Therefore, the main research objective of this study is to identify which features of news articles have the greatest impact on stock price movement in the banking sector. The author conducted an analysis of data from 94,784 news articles collected during the period from January 1, 2023, to April 1, 2024. Additionally, the stock price data of commercial banks were studied during

***Corresponding author: Nguyen Minh Nhat (nhatnm@hub.edu.vn)**
DOI: https://doi.org/10.47738/jads.v5i3.338

this period. The features considered include the mention rate, the mention rate one day before, the number of news sources mentioning the bank, the number of news articles mentioning the bank, the proportion of news sources with positive evaluations, the proportion of news sources with positive evaluations one day before, the total sentiment score of all news articles mentioning the bank, and the total sentiment score of all news articles mentioning the bank one day before.

This research seeks to enhance the current literature by empirically examining the connection between the characteristics of news articles and stock price movements through the application of the Random Forest algorithm. By identifying the most important features, the author aims to enhance the understanding of how news articles impact stock prices and provide practical insights for investors and analysts in the banking sector. In the following sections, the author will provide a detailed presentation of the literature review, research methodology, analysis results, and a discussion of the significance of these findings.

2. Literature Review

The research demonstrated that integrating different types of news articles can enhance the effectiveness of predicting stock price movements. Using multi-kernel learning techniques, they combined information from five different types of articles classified based on their relevance to the target stock, industry group and sector [5]. The results showed that simultaneously using these types of news articles improved prediction performance compared to using fewer types. Similarly, the study found comparable results when conducting a study on stocks in the healthcare sector. Combining various types of articles can support investors' decision-making process and improve financial predictions [6].

The impact of financial articles on stock prices immediately after they were published. They found that articles from WSJ, Reuters - UK Focus, NYT, and FT had a positive impact on stock prices, while those from Barrons, MarketWatch, Forbes, and Bloomberg had a negative impact [7]. The timing of the article's release also plays a crucial role in influencing stock prices, indicating that both the content and the temporal context of the articles need to be considered when predicting stock prices. The other study used data mining techniques to predict stock prices from financial articles, analyzing the relationship between article content and stock prices to forecast the future [8]. At the same time, other research also designed and implemented an automated system to predict stock price trends immediately after the publication of an article. Both studies demonstrated that automated analysis of financial articles can provide useful information for predicting stock price trends, from data collection, preparation, and classification to applying trading strategies, showcasing the potential of technology and big data in finance [9].

Another research trend explored the impact of news and public sentiment on stock price fluctuations, thereby proposing a trading strategy based on media. This study emphasizes the importance of fundamental information from company articles and the impact of public sentiment on investors' trading decisions [10]. The impact of media on companies varies according to the characteristics and content of the articles, indicating the need to combine both fundamental information and public sentiment in stock price predictions. This opens up a research direction on the interaction between fundamental information and sentiment in financial analysis [11]. Continuing to explore this research direction developed stock price prediction models based on sentiment analysis of financial articles. They used sentiment dictionaries to measure and analyze articles in the sentiment space. The results showed that sentiment analysis models outperformed bag-of-words models in predicting stock price trends [12]. The other study also proposed an automated system for collecting and predicting future stock price movements, utilizing a feature selection process and a sentiment analysis model to assess sentence-level sentiment in financial articles [13]. These studies emphasize the importance of using sentiment analysis and advanced technologies to achieve the highest accuracy in predicting stock price changes.

In recent studies have continued to develop news sentiment analysis models to predict stock price directions, using machine learning models to calculate the subjectivity of the news. The results showed that their models achieved high accuracy in predicting stock price directions, even when the data contained outliers [14]. Although the current studies have demonstrated that combining various types of articles and using machine learning techniques can improve the accuracy of stock price direction predictions, there are still many gaps that need further research and expansion. In particular, the interaction between fundamental and sentiment factors in financial articles has not been fully studied [15]. Moreover, the impact of the timing of release and news sources has not been comprehensively analyzed, which

can significantly affect stock price fluctuations. Additionally, the researches mainly focus on major markets, lacking applications in frontier or emerging markets. Therefore, deeper studies are needed to explore how to combine these factors and apply them to different markets to enhance prediction effectiveness, while also determining the importance of news article features on stock price movements.

3. Research Methodology

In this section, the article will introduce the methods and techniques applied to identify the important characteristics of news articles that influence stock price trends. The content includes research design, features and target variables, random forest algorithm, feature interpretation techniques, and model evaluation metrics [16].

3.1. Research design

The research design shown above outlines the key steps in gathering, processing, and analyzing news articles and stock price data. Initially, news articles are collected from the internet using a Semantic Crawler and stored in a News Storage system. Subsequently, these articles are classified by the News Classifier, and the sentiment of the news is extracted through the Stock News Sentiment Extraction process. The next step involves merging the news articles and stock price data in the News and Price Merger before applying Feature Engineering and developing the model. Lastly, the model undergoes testing in the Testing phase, and the results are presented in the Reporting phase using metrics such as Accuracy, AUC, Confusion Matrix, Feature Importance, and Feature SHAP in Figure 1.

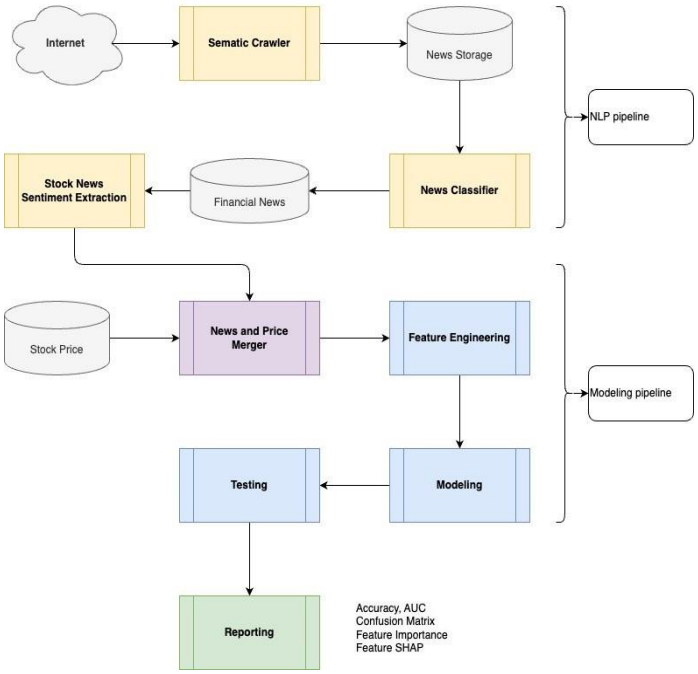

Figure 1. Description of research design

3.2. Features and target variables

There are 8 features extracted from the characteristics of news articles, including: mention_rate, mention_rate_1d, n_mention_publisher, n_mention_article, positive_rate, positive_rate_1d, sentiment_sum, sentiment_sum_1d, and 01 target variable, up_down_signal, used to measure stock price movement. The formulas and characteristics of each variable will be detailed in Table 1.

Table 1. Description of the features and target variable

	Symbol	Formula	Description
Features			
01	mention_rate	(Number of news sources mentioning commercial banks)/(Total number of news sources)	The proportion of news sources mentioning commercial banks

 ISSN 2723-6471

02	mention_rate_1d	(Number of news sources mentioning commercial banks one day before)/(Total number of news sources one day before)	The proportion of news sources mentioning commercial banks one day before
03	n_mention_publisher		The number of news sources mentioning commercial banks
04	n_mention_article		The number of news articles mentioning commercial banks
05	Positive_rate	(Number of news sources with positive evaluations) / (Total number of news sources)	The proportion of news sources with positive evaluations
06	Positive_rate_1d	(Number of news sources with positive evaluations one day before)/ (Total number of news sources one day before)	The proportion of news sources with positive evaluations one day before
07	Sentiment_sum		The total sentiment score of all news articles mentioning commercial banks
08	Sentiment_sum_1d		The total sentiment score of all news articles mentioning commercial banks one day before
Target Variable			
09	up_down_signal	Let up_down_signal be the stock price movement for a given day: $\begin{cases} 0 \ if \ closing \ price \ \leq \text{opening price} \ (no \ increas \\ 1 \ if \ closing \ price > \text{opening price} \ (increase \end{cases}$	The up_down_signal is a binary variable, reflecting the change in stock prices and indicating whether the stock price has increased or not over a specific period.

Table 1 showed that this feature set is designed to capture both the volume and sentiment of news coverage, which are critical factors in predicting stock price movements. By analyzing these features, the model attempts to forecast whether the stock price of commercial banks will increase or not on a given day.

3.3. Random Forest algorithm

The Random Forest algorithm is a powerful ensemble learning method used for classification tasks. In this article, the author employs the Random Forest classifier to predict the stock price movement of Vietnamese commercial banks using eight features derived from news articles.

Let X = $\{x_1, x_2, \dots, x_n\}$ be the set of features and y be the target variable, where y ∈ {0,1}.

Random Forest builds multiple decision trees using different subsets of the training data created through bootstrapping. Each subset is generated by sampling with replacement from the original dataset. This method ensures that each tree is trained on a slightly different dataset, enhancing the model's robustness [17].

D_i = BootstrapSample (D) where D is the training data and D_i is the i-th bootstrap sample.

At each split in a tree, Random Forest chooses a random subset of features to find the optimal split. This random feature selection creates diversity among the trees and decreases the correlation between them, thereby enhancing the model's overall performance. For classification tasks, each tree in the forest makes a prediction based on the input data. The final prediction is made by taking the majority vote among all the trees, which enhances the model's overall accuracy and stability.

$\hat{y}$ = model$\{\widehat{y_1}, \widehat{y_2}, \dots, \widehat{y_t}\}$ where $\widehat{y_i}$ is the prediction of the i-th tree.

We can understand the Random Forest Classifier Algorithm through the following pseudocode in Python (figure 2):

 ISSN 2723-6471

Random Forest Classifier Algorithm
Pseudocode Code in Python

```
def RandomForestClassifier(D):
  forest = []
  for i in range(B):
    D_i = BootstrapSample(D)
    T_i = DecisionTree()
    features_i = RandomFeatureSelection(D_i)
    T_i.train(D_i, features_i)
    forest.append(T_i)
  return forest
```

```
1: procedure RandomForestClassifier(D) ◂ D is the labeled training data
2:   forest = new Array() ◂ Initialize an empty array for the forest
3:   for i = 0 to B ◂ Iterate B times to create B trees
4:     D_i = Bagging(D) ◂ Create a bootstrap sample of the training data
5:     T_i = new DecisionTree() ◂ Initialize a new decision tree
6:     Feature_i = RandomFeatureSelection (D_i) ◂ Select a random subset of features
7:     T_i.train(D_i, Feature_i) ◂ Train the decision tree with the bootstrap sample and selected features
8:     forest.add(T_i) ◂ Add the trained tree to the forest
9:   end for
10:  return forest ◂ Return the random forest
11: end procedure
```

Figure 2. Pseudocode Random Forest

3.4. Feature interpretation techniques

In this study, to determine which features of news articles have the greatest impact on stock price movement in the banking sector, the author uses two techniques: Feature Importance Values (FIV) and SHapley Additive exPlanations (SHAP). Feature importance values provide a detailed insight into the impact of each feature on the prediction outcome, helping us better understand how the model makes decisions. In the Random Forest model, Feature Importance Values are calculated in two main ways: the degree of impurity reduction and the decrease in accuracy when the feature's values are permuted. The degree of impurity reduction measures the total decrease in Gini impurity or entropy caused by each feature when splitting the data at the nodes in the decision trees. The decrease in accuracy, also known as Permutation Importance, measures the change in the model's accuracy when the values of a feature are randomly shuffled. The higher the value of these features, the more important they are to the model [18]. Identifying the importance of features helps to detect the most crucial factors in the model, support the elimination of unnecessary features, and optimize the performance and accuracy of the predictions.

The SHapley Additive exPlanations (SHAP) is a technique for interpreting machine-learning models based on the principles of game theory. This method assigns the model's prediction value to the input features by evaluating the contribution of each feature to the outcome. SHAP uses Shapley values to represent the influence of each feature on the model's prediction. The strength of SHAP lies in providing consistent and fair explanations of feature importance [19]. Applying SHAP helps us identify the most important features by elucidating how the model works and makes decisions. This method also helps to identify the features that have the greatest impact on the prediction outcome, thereby optimizing the model. Additionally, SHAP addresses potential issues in the model, ensuring the transparency and reliability of the predictions.

3.5. Model evaluation metrics

To assess the performance of model, several metrics were used as follows (table 2): Confusion Matrix: A table that outlines the performance of the classification model by detailing the counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN)

Table 2. Confusion matrix

		Predicted Model	
		No increase	*Increase*
		True Negatives (TN)	*False Positives (FP)*
No increase		when the model predicts that the stock price will not increase (up_down_signal = 0), and the stock price indeed does not increase.	When the model predicts that the stock price will increase (up_down_signal = 1), but the stock price actually does not increase. This is referred to as a Type I error.

 ISSN 2723-6471

		False Negatives (FN)	*True Positives (TP)*
Actual data	*Increase*	When the model predicts that the stock price will not increase (up_down_signal = 0), but the stock price actually does increase. This error is also referred to as a Type II error.	When the model forecasts that the stock price will increase (up_down_signal = 1) and the actual stock price does increase.

Table 2 showed that the confusion matrix is a table used to evaluate the performance of a classification model. It compares the actual values with the predicted values to provide a comprehensive view of the model's accuracy. Besides, Understanding the confusion matrix and these derived metrics allows researchers to evaluate and improve their stock price prediction model by focusing on minimizing these errors and maximizing true positive and true negative predictions. Based on the results from the confusion matrix in Table 2, several metrics such as Accuracy, Recall, Precision, and F1-Score will be calculated to gain a clearer understanding of the model's prediction performance. The meanings and formulas for these metrics are described in Table 3.

Table 3. Description of Model Evaluation Metrics

No.	Metrics	Formula	Meaning
01	*Accuracy*	$\frac{TP + TN}{TP + TN + FP + FN}$	Accuracy is a measure of performance that reflects the ratio of correct predictions made by the model to the total number of predictions. A higher accuracy value signifies a more reliable model in making accurate predictions.
02	Recall	$\frac{TP}{TP + FN}$	Recall assesses the model's effectiveness in accurately detecting cases where the stock price increase. A higher recall value means the model is effective in detecting stock price increases, minimizing the number of missed opportunities for identifying true positive movements.
03	Precision	$\frac{TP}{TP + FP}$	Precision reflects the proportion of predicted stock price increases that are actually correct. A higher precision value indicates that the model is reliable in its positive predictions, reducing the number of false alarms for stock price increases.
04	F1 Score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	The F1-score evaluates the model's ability to predict stock price increases by balancing precision and recall. A higher F1-score suggests that the model effectively balances correct predictions of stock price increases with minimizing both false positives and false negatives, offering a more comprehensive measure of the model's overall performance.

4. Result and Discussion

4.1. Research results

The author compiled an extensive dataset containing 94,784 news articles published from the beginning of 2023 until April 1, 2024. These articles were carefully selected from the stock and economy sections of leading financial websites such as vnEconomy, Cafef, Vietstock, and vietnambiz, among others. Approximately 1,595 news articles were specifically related to Vietnamese publishers and news about bank tickers, with extracted sentiment. Figure 3 provides a detailed breakdown of the distribution of articles across these sources.

 ISSN 2723-6471

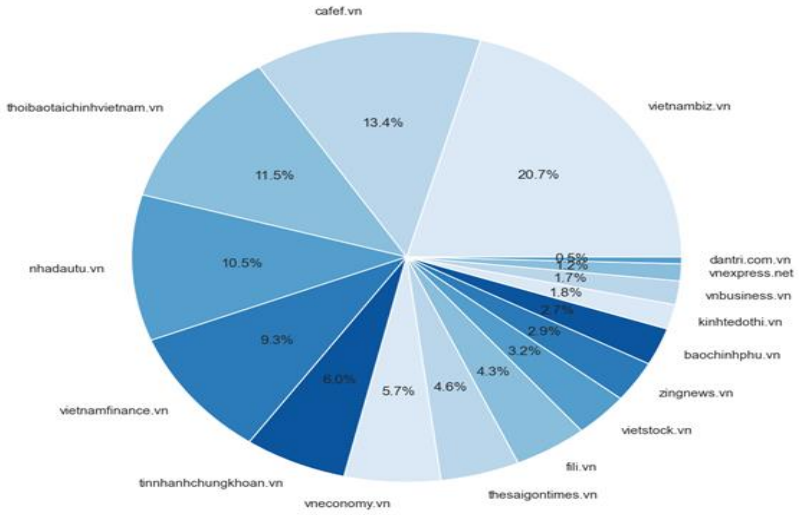

Figure 3. Description of data sources

The distribution of 1595 news articles about bank sector stocks primarily focuses on mentioning major banks and is specifically presented in Figure 4.

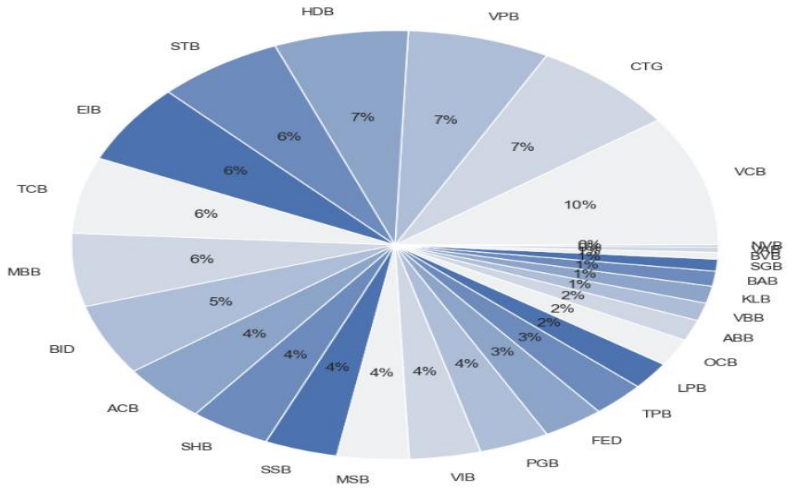

Figure 4. The distribution of news articles about bank sector stocks

Moreover, the provided dataset consists of 1,595 entries, each with eight news articles features relevant to stock price movement analysis and one target variable, up_down_signal. Statistical summaries for each feature provide details on the count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum values. Notably, the up_down_signal has a mean of -0.40495 and a standard deviation of 0.999, reflecting variability in stock price movements. This summary offers a thorough overview of the dataset's characteristics, which is essential for understanding and preparing the data for predictive modeling [20]. Table 4 showed the Random Forest classifier was implemented using Python's scikit-learn library. The dataset was split into training and testing sets, with 90% used for training and 10% for testing. Key hyperparameters, such as the number of trees (n_estimators), maximum depth (max_depth), and the number of features considered at each split (max_features), were fine-tuned to optimize the model's performance. Moreover, a more balanced approach, such as an 80/20 split, or even better, employing cross-validation techniques like k-fold cross-validation, could indeed provide a more thorough evaluation of the model's generalizability. Cross-validation, in particular, ensures that every data point gets a chance to be in the training and test sets, thereby providing a more accurate estimate of the model's performance.

 ISSN 2723-6471

Table 4. Statistical data description

	Up_down_signal	n_mention_article	n_mention_publisher	mention_rate	positive_rate	Sentiment_sum	Mention_rate_1d	Positive_rate_1d	Sentiment_sum_1d
Count	1595	1595	1595	1595	1595	1595	1595	1595	1595
Mean	-0.0495	12.5755	6.7216	0.1704	0.7829	-1.0169	0.0811	0.3155	-0.4244
Std	0.9990	6.5462	2.4183	0.1270	0.1400	2.441	0.1417	0.3977	1.577
Min	-1.0000	1.0000	1.00000	0.0270	0.2500	10.0000	0.0000	0.0000	-10.0000
25%	-1.0000	8.0000	5.0000	0.0833	0.7000	-2.0000	0.0000	0.0000	0.0000
50%	-1.0000	11.0000	7.0000	0.1250	0.8095	-1.0000	0.0000	0.0000	0.0000
75%	1.0000	16.0000	8.0000	0.2222	0.8888	0.0000	0.1250	0.7500	0.0000
max	1.0000	37.0000	13.0000	1.0000	1.0000	5.0000	1.0000	1.0000	5.0000

4.2. The confusion matrix results for the random forest model on the test dataset

The Confusion Matrix results for the Random Forest model on the test dataset indicate that the model correctly identified 76 instances where the stock price did not increase (True Negatives - TN). Furthermore, the model did not produce any false positives, meaning there were no cases where the model incorrectly predicted a rise in stock price when there was none [21]. However, there were 31 instances where the model predicted that the stock price would not increase, but it actually did (False Negatives - FN), indicating that the model missed some signals of rising stock prices. Additionally, the model correctly predicted 38 instances where the stock price increased (True Positives - TP), demonstrating its ability to identify upward trends. Overall, the Random Forest model shows high accuracy in recognizing non-increase signals but needs improvement in detecting increase signals to reduce the number of False Negatives in Figure 5.

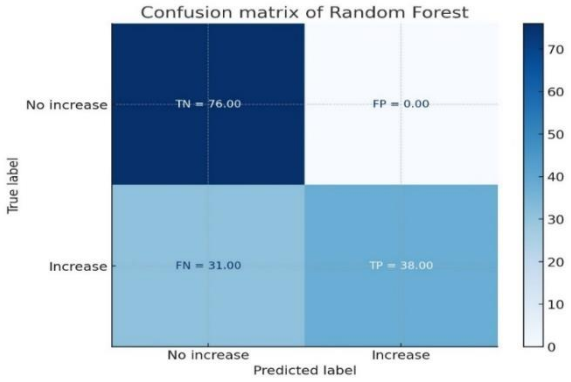

Figure 5. Confusion matrix of random forest

In terms of model performance metrics, the accuracy of the Random Forest model was 78.62%, demonstrating the model's capability to correctly predict the majority of instances. The recall metric, however, was 55.07%, indicating that the model lacks sensitivity in identifying cases where stock prices increased [22]. On the other hand, the precision was 100%, suggesting that all instances where the model predicted an increase in stock prices were accurate. The F1 Score was 71.01%, highlighting a balance between precision and recall, yet underscoring the need for further enhancement in detecting price increase signals. These results suggest that the Random Forest model may need adjustments or to be combined with other methods to enhance its prediction performance (Table 5).

Table 5. The results of model on the tested data

	Performance metrics			
	Accuracy	Recall	Precision	F1 Score
Random Forest Model	78.62%	55.07%	100%	71.01%

The results from the Feature Importance Values method show that mention_rate is the most important feature, with the highest score, followed by positive_rate. This indicates that the proportion of news sources mentioning the bank and the proportion of positive evaluations have the greatest impact on stock price predictions. Other features such as n_mention_publisher, sentiment_sum and n_mention_article also have high importance, indicating that the number of sources and the total sentiment score of news articles play significant roles as well [23]. Besides, the ROC curve

 ISSN 2723-6471

(Receiver Operating Characteristic curve) plots the true positive rate (recall) against the false positive rate and helps in visualizing the performance of the classification model across different threshold values. The AUC (Area Under the ROC Curve) provides a single scalar value that summarizes the overall ability of the model to discriminate between classes. An AUC closer to 1 indicates a better performing model, as it suggests the model has a good measure of separability between the classes. The model highly values feature related to the level and sentiment of news sources when predicting stock price movements. However, features like mention_rate_1d, positive_rate_1d and sentiment_sum_1d have lower importance, indicating that their impact is less significant (Figure 6).

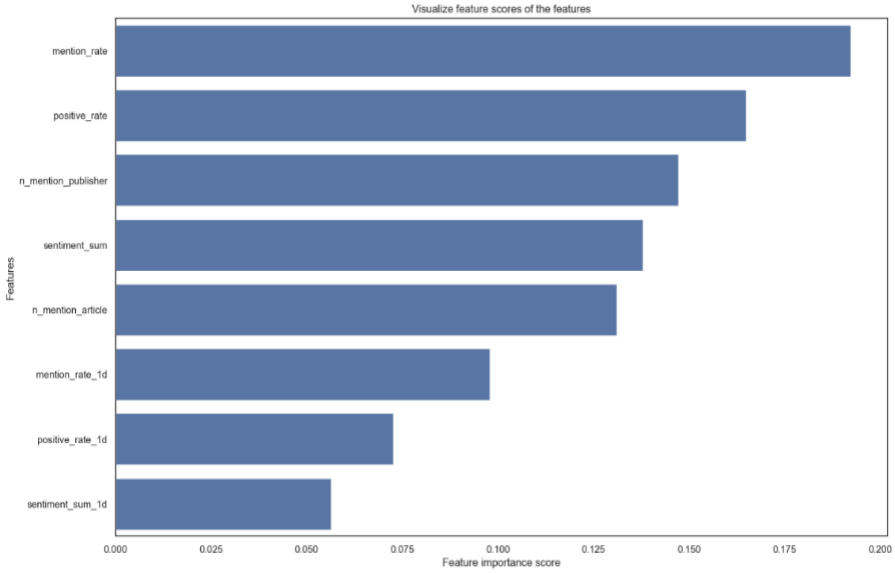

Figure 6. Feature Importance Values

The results from the SHapley Additive exPlanations (SHAP) method support the findings from the Feature Importance Values method, with positive_rate and mention_rate having the highest SHAP values, highlighting their importance for the model's output. Features like n_mention_publisher, sentiment_sum and n_mention_article also show high SHAP values, indicating a significant influence of the number of sources and the total sentiment score on stock price predictions. The color distribution in the SHAP plot shows that these features can have either positive or negative effects depending on the specific context. This understanding helps to see how each feature influences the prediction model and assists in model optimization [24]. Combining both methods provides a comprehensive understanding of the importance and influence of each feature, enhancing the accuracy of the predictions (Figure 7).

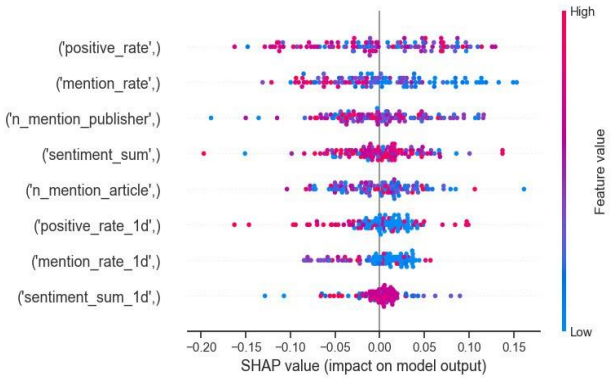

Figure 7. SHapley Additive exPlanations (SHAP)

4.3. Discussions

Based on the research results, we can observe the significance of news article features on stock price movement in the banking sector, specifically as follows:

Firstly, the positive_rate feature, which indicates the proportion of news sources with positive evaluations, significantly affects stock price movement as it reflects market sentiment and investor confidence in the business operations of commercial banks [25]. When there is a high volume of positive news, investors tend to view the banks more optimistically, leading to increased stock purchases and consequently driving up the stock price. Positive news often indicates good financial results, successful strategies, or favorable market conditions, all contributing to investor optimism. This optimism boosts the demand for stocks, thus pushing up the stock prices. Therefore, the positive_rate is a crucial indicator showing how market sentiment, influenced by news, has a strong impact on stock prices. Moreover, the paper limitations should also discuss the accuracy of the sentiment model, including any validation metrics e.g., precision, recall, F1 score for the sentiment model itself that were used to assess its performance. It's also important to acknowledge any potential limitations of the sentiment analysis approach and how these might affect the study's overall results.

Secondly, the mention_rate feature, representing the proportion of news sources mentioning commercial banks, indicates the level of media and public interest in these financial institutions. A high mention rate signifies that the banking system is drawing substantial media attention, possibly due to significant events or outstanding business performance during that period [26]. This increased attention often leads to higher stock trading activity as investors respond to new information in the banking sector. If the mentioned news is positive, it can create optimism and drive stock prices up. Conversely, if the news is negative, it can cause concern and lower stock prices. Therefore, this feature is the next indicator showing its impact on stock price movement in the market [27]. Besides, the author had recommendations for addressing timing in the study based on the discussion on timing impact. The paper should include a discussion on how the timing of news releases might affect investor behavior and stock price movements. This could involve analyzing whether certain times of the day or week are more sensitive to news or if market conditions e.g., volatility influence the strength of the reaction.

Thirdly, the n_mention_publisher feature indicates the number of news sources mentioning the commercial banks. This feature reflects the prevalence and dissemination of information in the investment market. When many news sources mention a particular bank, it shows that information about the bank is being widely circulated, attracting the attention of many investors [28]. This increased attention can stimulate stock trading activities of that bank, depending on the nature of the news being spread during that period. The study focuses exclusively on the banking sector in Vietnam but does not compare the results with other sectors. Therefore, the author had recommendations for the Study is to expand the scope, which is to consider expanding the study to include other sectors such as technology, healthcare, or consumer goods. This could involve applying the same methodology to datasets from these sectors and comparing the results with those from the banking sector.

Fourthly, the sentiment_sum feature represents the total sentiment score of news articles mentioning commercial banks. It also affects stock price movements as it aggregates both positive and negative sentiments from various sources. When the sentiment score is high, it indicates that most of the news is positive, creating optimism and encouraging investors to buy stocks, thereby driving up the stock price [29]. Conversely, if the sentiment score is low or negative, it shows that the news is mostly negative, causing concern and potentially leading to a sell-off, lowering the stock price. Combining all related news into a single score provides investors with an overall view of market sentiment, so the sentiment_sum is also an important indicator reflecting the overall impact of news on stock prices. The study should include an evaluation of the credibility of the news sources used. This could involve discussing the reputation of vnEconomy, Cafef, Vietstock, and others within the Vietnamese market, as well as any known biases or inclinations these outlets may have. In this case, the study results should explicitly acknowledge any limitations regarding the news sources used, such as potential biases or lack of diversity in viewpoints. Thus, the author proposed future research to encourage future studies to explore the impact of news source credibility and bias on sentiment analysis in more detail, possibly by comparing the results from different sources or by using more advanced techniques to adjust for potential biases.

Fifthly, the n_mention_article feature indicates the number of news articles mentioning commercial banks. This feature reflects the level of public interest and attention towards these financial institutions. When the number of news articles about the bank increases, it indicates that the bank is receiving significant media attention, which can lead to increased

ISSN 2723-6471

stock trading activity [30]. However, when comparing the impact of n_mention_article with mention_rate as discussed earlier, we can see that the n_mention_article feature has less influence on stock price volatility. This is because the mention_rate feature not only measures the number of mentions but also the proportion of news sources mentioning the bank compared to the total number of news sources, thus reflecting a broader and stronger dissemination. SHAP values provide a way to explain the output of complex machine learning models by attributing a prediction to individual features. However, without a detailed explanation, the significance of these values might not be clear to the readers, especially those less familiar with SHAP based on the summary plot can show the overall impact of each feature on the model's output across all samples. It helps in identifying the most important features and understanding their distribution of impact.

Finally, the features "mention_rate_1d," "positive_rate_1d," and "sentiment_sum_1d" reflect the previous day's metrics and impact stock price movements because they show the lasting influence of information on the market. "Mention_rate_1d" indicates the proportion of news sources mentioning the bank the day before, demonstrating continued attention that can influence investment decisions. "Positive_rate_1d" measures the proportion of news sources with positive evaluations from the previous day, reflecting sustained optimism and its effect on investor sentiment. "Sentiment_sum_1d" aggregates the sentiment scores of all news articles mentioning the bank the previous day, showing prolonged sentiment trends and their impact on investor behavior [31]. However, these features have less impact on stock price movements compared to current day metrics because fresh information typically has a stronger influence on investor decisions. Investors often react quickly to the latest news, so the previous day's metrics may not fully capture the current market situation. This lag in information can reduce the predictive accuracy of these features, making them less influential on stock price movements compared to same-day metrics.

5. Conclusions and Recommendation

This study explores the impact of news article features on stock price movements in the banking sector using the Random Forest algorithm. Our analysis, based on 94,784 news articles from January 2023 to April 2024, reveals several critical insights. Firstly, the feature positive_rate, indicating the proportion of news sources with positive evaluations, significantly affects stock price movements. Positive news tends to boost investor confidence, leading to increased stock purchases and consequently higher stock prices. This finding underscores the influence of market sentiment, as captured through news evaluations, on investor behavior. Secondly, the feature mention_rate, which measures the proportion of news sources mentioning commercial banks, is another key determinant. A higher mention rate indicates substantial media attention, which can lead to increased stock trading activity. The nature of the news, whether positive or negative, plays a crucial role in determining the direction of stock price movements. Additionally, features like n_mention_publisher and sentiment_sum also significantly influence stock prices by reflecting the breadth of coverage and overall sentiment towards banks. These insights emphasize the importance of timely news analysis for making informed investment decisions.

Based on the results above, 2024 will have many intertwined opportunities and challenges for the Vietnamese stock market. In that context, to develop a safe and sustainable stock market, an effective medium- and long-term capital mobilization channel for the economy, the State Securities Commission will continue to implement many groups of synchronous and specific recommendations following.

First of all, in the immediate future, with the approval of the Ministry of Finance's leaders, the State Securities Commission is urgently preparing for the 2024 Stock Market Development Conference. The conference is expected to have the participation and direction of Government leaders, the participation and sharing of opinions of representatives of ministries, branches, international financial institutions, listed enterprises and market members. Stock prices are influenced by a wide range of factors, including but not limited to news sentiment. Economic indicators, interest rates, and global events can have substantial effects on market movements, and their exclusion may lead to an incomplete understanding of the factors driving stock prices. Thus, the author had recommendations for expanding the scope of the study based on the economic indicators to consider including key economic indicators such as GDP growth rates, unemployment rates, inflation rates, and interest rates in the analysis. These factors are known to affect investor sentiment and stock prices.

Secondly, for regular solutions in 2024, the management agency will continue to perfect the legal framework and policy system for market development, focusing on completing the development of a plan to effectively implement the Stock Market Development Strategy project until 2030. Understanding why specific features were selected enhances the interpretability of the model. It allows stakeholders to see how certain variables impact predictions, which is particularly important in financial modeling where decisions based on these predictions can have significant consequences. Thus, the author had recommendations for justifying feature selection to discuss theoretical and empirical basis to provide a discussion of the theoretical or empirical reasons for selecting each feature. Moreover, explain why sentiment scores were included, possibly by referencing literature that links sentiment analysis with market behavior.

Thirdly, at the same time, the State Securities Commission will strengthen the construction of infrastructure systems, apply information technology, and meet the trend of the 4.0 technology revolution, both creating conditions for management and operation, and creating favorable conditions for businesses, market institutions and investors to participate. Besides, the author had recommendations for the implement and compare multiple algorithms based on the study should implement and compare the performance of alternative algorithms like Gradient Boosting Machines e.g., XGBoost, LightGBM and Support Vector Machines (SVM).

Fourthly, along with that, we will continue to restructure the stock market based on the main pillars that have been proposed, focusing on strengthening the management of securities trading organizations and securities practitioners, diversifying the investor base, developing the institutional investor system, encouraging long-term foreign investment, and training individual investors. Data quality helps to proper preprocessing ensured that the data fed into the model is clean, consistent, and free from noise, which is crucial for accurate predictions. Missing data, unnormalized text, or biases in sentiment extraction can lead to misleading results. Therefore, the next research is improving data quality.

Finally, to ensure market discipline, the management agency will strengthen inspection, examination and supervision to ensure the stock market develops sustainably, openly, transparently, and promptly and strictly handles violations of the law on the stock market. Moreover, the author suggested practical ways for investors to monitor and evaluate the credibility of the news sources used in the study. This could include advice on cross-referencing information from multiple sources or using news aggregation tools.

The study limitations and future research: In this study, the author only focused on analyzing the banking sector; in the future, the author can extend the analysis to other sectors to increase the generalizability of the research results. Additionally, the study primarily uses textual sentiment analysis and does not incorporate multimedia content such as videos or images. Future research can explore the integration of multimedia content to validate and broaden the findings. Finally, studying different time periods should also be considered to gain deeper insights into the impact of news articles on stock price volatility in the market. Recommendations for further investigation is to examine class distribution to check the distribution of the target variable (stock price movement) to see if there is a significant imbalance. If an imbalance exists, consider applying techniques like resampling (oversampling the minority class or undersampling the majority class) or using model adjustments like class weighting. The next research is to adjust classification threshold, such as experiment with different classification thresholds and observe how the confusion matrix changes. This can provide insight into whether the model is making predictions too conservatively.

6. Declaration

6.1. Author Contributions

Conceptualization: N.M.N.; Methodology: N.M.N.; Software: N.M.N.; Validation: N.M.N.; Formal Analysis: N.M.N.; Investigation: N.M.N.; Resources: N.M.N.; Data Curation: N.M.N.; Writing Original Draft Preparation: N.M.N.; Writing Review and Editing: N.M.N.; Visualization: N.M.N.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The author received financial support for the research by Ho Chi Minh University of Banking (HUB), Vietnam.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. S. Chen, M. T. Leung, and H. Daouk, "Application of neural networks to an emerging financial market: Forecasting and trading the Taiwan Stock Index", *Computer and Operations Research*, vol. 30, no. 6, pp. 901-923, 2003.

[2] Y. Shynkevich, T. McGinnity, S. A. Coleman, and A. Belatreche, "Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning", *Decision Support Systems*, vol. 85, no. 5, pp. 74-83, 2016.

[3] A. J. Hussain, A. Knowles, P. J. G. Lisboa, and W. El-Deredy, "Financial time series prediction using polynomial pipelined neural networks", *Expert Systems with Applications*, vol. 35, no. 3, pp. 1186-1199, 2008.

[4] S. Smith, "Comparing traditional news and social media with stock price movements; which comes first, the news or the price change?", *Journal of Big Data*, vol. 9, no. 1, pp. 1-20, 2022.

[5] B. Shantha Gowri and V. S. Ram, "Influence of news on rational decision making by financial market investors", *Investment Management and Financial Innovations*, vol. 16, no. 3, pp. 142-156, 2019.

[6] Y. Kara, M. A. Boyacioglu, and Ö. K. Baykan, "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul stock exchange", *Expert Systems with Applications*, vol. 38, no. 5, pp. 5311-5319, 2011.

[7] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis", *Knowledge-Based Systems*, vol. 69, no. 10, pp. 14-23, 2014.

[8] L. P. Ni, Z. W. Ni, and Y. Z. Gao, "Stock trend prediction based on fractal feature selection and support vector machine", *Expert Systems with Applications*, vol. 38, no. 5, pp. 5569-5576, 2011.

[9] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "Efficient stock-market prediction using ensemble support vector machine", *Open Computer Science*, vol. 10, no. 1, pp. 153-163, 2020.

[10] B. K. Meher, A. Anand, S. Kumar, R. Birau, and M. Singh, "Effectiveness of Random Forest Model in Predicting Stock Prices of Solar Energy Companies in India", *International Journal of Energy Economics and Policy*, vol. 14, no. 2, pp. 426-434, 2024.

[11] Q. Li, T. Wang, P. Li, L. Liu, Q. Gong, and Y. Chen, "The effect of news and public mood on stock movements", *Information Sciences*, vol. 278, no. 10, pp. 826-840, 2014.

[12] M. Arellano, and O. Bover, "Another look at the instrumental variable estimation of error-components models", Journal of Econometrics, vol. 68, no. 1, pp. 29-51, 1995.

[13] R. Blundell, and S. Bond, "Initial conditions and moment restrictions in dynamic panel data models", *Journal of Econometrics*, vol. 87, no. 1, pp. 115-143, 1998.

[14] A. K. Giri and P. Joshi, "The impact of macroeconomic indicators on Indian stock prices: An empirical analysis", *Studies in Business and Economics*, vol. 12, no. 1, pp. 61-78, 2017.

[15] R. Laduna, and M. Sun'an, "The effect of macroeconomic variables on banking stock price index in Indonesia stock exchange", *Russian Journal of Agricultural and Socio-Economic Sciences*, vol. 73, no. 1, pp. 155-162, 2018.

[16] A. Laporte, and F. Windmeijer, "Estimation of panel data models with binary indicators when treatment effects are not constant over time", *Economics Letters*, vol. 88, no. 3, pp. 389-396, 2005.

[17] M. A. Maharditya, L. Layyinaturrobaniyah, and M. Anwar, "Implication of macroeconomic factors to stock returns of Indonesian property and real estate companies", *Jurnal Dinamika Manajemen*, vol. 9, no. 1, pp. 100-113, 2018.

[18] Y. Wu, "Exchange rates, stock prices, and money markets: Evidence from Singapore", *Journal of Asian Economics*, vol. 12, no. 3, pp. 445-458, 2001.

[19] A. S. Yang, and A. Pangastuti, "Stock market efficiency and liquidity: The Indonesia stock exchange merger", *Research in International Business and Finance*, vol. 36, no. 1, pp. 28-40, 2016.

[20] M. Abdelkarim, and Y. Almumani, "An empirical study on effect of profitability ratios & market value ratios on market capitalization of commercial banks in Jordan", *International Journal of Business and Social Science*, vol. 9, no. 4, pp. 39-45, 2018.

[21] I. Abeysekera, "The influence of board size on intellectual capital disclosure by Kenyan listed firms", *Journal of Intellectual Capital*, vol. 11, no. 4, pp. 504-518, 2010.

[22] H. A. Ahmed, "The role of audit committee attributes in intellectual capital disclosures: Evidence from Malaysia", *Managerial Auditing Journal*, vol. 30, no. 8-9, pp. 756-784, 2015.

[23] Y. G. Lee, J. Y. Oh, D. Kim, and G. Kim, "Shap value-based feature importance analysis for short-term load forecasting", *Journal of Electrical Engineering & Technology*, vol. 18, no. 1, pp. 579-588, 2023.

[24] X. Yan, "Corporate governance and intellectual capital disclosures in CEOs' statements", *Nankai Business Review International*, vol. 8, no. 1, pp. 2-21, 2017.

[25] I. Ulum, R. R. Harviana, S. Zubaidah, and A. W. Jati, "Intellectual capital disclosure and prospective student interest: An Indonesian perspectives", *Cogent Business and Management*, vol. 6, no. 1, pp. 1-13, 2019.

[26] G. J. Staubus, "Ethics failures in corporate financial reporting", *Journal of Business Ethics*, vol. 57, no. 1, pp. 5-15, 2005.

[27] I. Soukhakian and M. Khodakarami, "Working capital management, firm performance and macroeconomic factors: Evidence from Iran", *Cogent Business and Management*, vol. 6, no. 1, pp. 1-24, 2019.

[28] R. Alkebsee, A. Alhebry, G. Tian, and A. Garefalakis, "Audit committee's cash compensation and earnings management: The moderating effects of institutional factors", *Revista Española de Financiación Y Contabilidad*, vol. 51, no. 4, pp. 389-416, 2022.

[29] A. Al-Sartawi and S. M. Reyad, "The relationship between the extent of online financial disclosure and profitability of Islamic banks", *Journal of Financial Reporting and Accounting*, vol. 17, no. 2, pp. 343-362, 2019.

[30] A. T. Ajibade, N. Okutu, F. Akande, J. D. Kwarbai, I. M. Olayinka, and A. Olotu, "IFRS adoption, corporate governance and faithful representation of financial reporting quality in Nigeria's development banks", *Cogent Business and Management*, vol. 9, no. 1, pp. 1-13, 2022.

[31] M. Aria, and C. Cuccurullo, "Bibliometrix: An R-tool for comprehensive science mapping analysis", *Journal of Informetrics*, vol. 11, no. 4, pp. 959-975, 2017.