*VNUHCM Journal of Economics, Business and Law 2026, 10(1):*6410-6424

Research article

Open Access Full Text Article

Sentiment-driven forecasting of the stock index in Vietnam: A machine learning perspective

Phong Anh Nguyen*, Tam Phan Huy, Thanh Ngo Phu

Use your smartphone to scan this QR code and download this article

ABSTRACT

The increasing influence of news sentiment on financial markets has drawn significant attention in recent years, yet its predictive potential in emerging markets remains underexplored. This study investigates whether multi-dimensional sentiment signals, derived from a large corpus of Vietnamnet news articles, can enhance the prediction of daily VN-Index directional movements when combined with technical indicators. Drawing on behavioral finance, information asymmetry, and media framing theories, the research posits that sentiment-laden narratives, alongside price-based signals, provide complementary insights into investor behavior. The dataset comprises 6,480 news articles (2019–2025) and daily VN-Index historical data, from which sentiment features—including polarity, subjectivity, compound scores, and sentiment proportions—are extracted at title, excerpt, and content levels. Technical indicators such as Moving Averages, RSI, MACD, Bollinger Bands, and Volatility are also constructed.

The predictive framework is modeled as a binary classification task (``Up'' vs. ``Unchanged/Down'') and evaluated using multiple machine learning algorithms, including Naive Bayes, Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting, AdaBoost, and CatBoost. Ten-fold cross-validation with metrics such as accuracy, precision, recall, F1-score, and ROC-AUC ensures robust performance assessment. Results reveal that ensemble models, particularly CatBoost, Gradient Boosting, and Random Forest, consistently outperform linear and probabilistic baselines regarding accuracy, recall, and F1-score. Logistic Regression and AdaBoost also show competitive ROC-AUC values, while Naive Bayes underperforms in distinguishing market movements.

These findings underscore the incremental predictive power of sentiment features in an emerging market setting, challenging the semi-strong form of the Efficient Market Hypothesis and reinforcing behavioral finance perspectives on bounded rationality and sentiment-driven trading. Practically, the study offers implications for portfolio managers, policymakers, and media organizations by demonstrating that hybrid sentiment-technical models can improve market forecasting, regulatory monitoring, and responsible financial reporting. Limitations regarding data sources and frequency are acknowledged, and avenues for future research include multi-source sentiment integration, deep learning approaches, and real-time deployment.

Key words: News sentiment, Stock index prediction, Machine learning, Vietnam

University of Economics and Law and Vietnam National University, Ho Chi Minh City, Vietnam

Correspondence

Phong Anh Nguyen, University of Economics and Law and Vietnam National University, Ho Chi Minh City, Vietnam

Email: :phongna@uel.edu.vn

History

- Received: 14-7-2025
- Revised: 26-8-2025
- Accepted: 03-12-2025
- Published Online: 27-03-2026

DOI :

https://doi.org/10.32508/stdjelm.v10i1.1687

Check for updates

INTRODUCTION

In recent years, the increasing influence of news sentiment on financial markets has attracted substantial attention from academics and practitioners. As financial markets become more interconnected and information diffuses instantaneously, investor behavior is increasingly driven by quantitative fundamentals and qualitative perceptions shaped by media coverage and sentiment-laden narratives [1,2]. The rapid proliferation of online financial news and advances in natural language processing have provided new tools to quantify sentiment and incorporate it into predictive models of asset price movements. Despite this progress, accurately predicting stock index direction remains challenging, particularly in emerging markets where investor sentiment may be more reactive to media signals due to less mature institutional frameworks [3,4]. In the context of Vietnam, the relationship between news sentiment and stock market performance has received relatively limited empirical attention, and existing studies often rely on narrow sentiment measures or overlook the integration of sentiment with technical market indicators.

The research problem addressed by this study is how to effectively leverage multi-dimensional sentiment scores extracted from a large corpus of news articles, combined with selected technical indicators, to predict the directional movement (up versus unchanged or down) of the Vietnamese VN-Index. Previous studies have demonstrated that sentiment can carry incremental predictive value beyond traditional price and volume data [5,6]. However, in emerging markets

Cite this article : Phong N A, Tam P H, Thanh N P. **Sentiment-driven forecasting of the stock indexin Vietnam: A machine learning perspective**.*VNUHCM J. Econ. Bus. Law.* 2026; 10(1):6410-6424.

This article was published during the journal's renaming from Journal of Science & Technology Development Journal – Economics-Lawand Management (ISSN: 2588-1051) to Journal of VNUHCM Journal of Economics, Business and Law; the new ISSN is currentlypending assignment.

such as Vietnam, where investor psychology may be particularly sensitive to news flow, the predictive potential of systematically measured sentiment remains underexplored. This research, therefore, aims to fill this gap by constructing an extensive dataset of sentiment signals derived from over 6,000 news articles and integrating these with well-established technical indicators to assess their combined effectiveness in forecasting index trends.

The significance of this research lies in its potential contributions to both academic understanding and practical forecasting applications. Academically, the study expands the literature on behavioral finance and machine learning by empirically examining whether the informational content of news sentiment can be harnessed to improve prediction accuracy in an emerging market setting[7]. The findings could offer valuable insights for portfolio managers, traders, and policymakers interested in anticipating market shifts triggered by sentiment dynamics and market momentum. Moreover, by evaluating multiple machine learning classifiers, including probabilistic, linear, and ensemble approaches, this research provides a comparative perspective on which algorithms are most suitable for sentiment-driven prediction tasks under volatile conditions.

This study aims to develop, implement, and evaluate a predictive framework that combines sentiment scores and technical indicators to classify daily VN-Index movement as either Up, Unchanged, or Unchange/Down. The research follows a structured approach: first, it collects and processes sentiment data from Vietnamnet news articles; second, it derives technical features from historical price data; third, it applies and compares the performance of multiple machine learning classifiers; and finally, it assesses predictive accuracy using cross-validation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The remainder of this paper is organized as follows: Section 2 reviews relevant background theories and empirical studies; Section 3 describes the data collection process and variable measurement and details the machine learning methodologies and evaluation procedures; Section 4 presents empirical results and discussion; and Section 5 concludes with implications and suggestions for future research.

LITERATURE REVIEW

Background theories

The dynamic relationship between news sentiment and stock index movements has been a central focus in financial economics, drawing upon several foundational theories. One of the most influential is the Efficient Market Hypothesis, which posits that financial markets quickly and accurately incorporate all available information into asset prices[8]. Under the semi-strong form of the Efficient Market Hypothesis, public disclosures, such as news articles, should be instantaneously reflected in stock prices, leaving no room for investors to earn abnormal returns by processing publicly available sentiment. However, empirical evidence has often challenged this notion, suggesting that market participants do not process information homogeneously or rationally[1]. This divergence has motivated research into how qualitative textual data, including media sentiment, can influence short-term price fluctuations and predictive models of index performance.

Behavioral finance offers critical insights into why sentiment can systematically affect market outcomes, even if it is publicly observable. According to Barberis, Shleifer, and Vishny[9], investor sentiment and bounded rationality often lead to overreaction or underreaction to news, generating predictable price patterns inconsistent with the Efficient Market Hypothesis. Theories of representativeness and conservatism bias explain how investors might overweigh vivid and recent news stories, leading to momentum or reversal in returns. Shiller[10] further emphasizes the role of ”narratives” and collective psychological factors, which can amplify sentiment-driven trading. This body of work underpins the hypothesis that aggregated measures of news sentiment can have explanatory and predictive power over index movements, particularly during periods of heightened uncertainty or volatility.

Information asymmetry theory also contributes to understanding why news sentiment can play a decisive role in price discovery. Grossman and Stiglitz[11] argue that markets can never be perfectly efficient because information acquisition is costly, creating incentives for informed traders to exploit informational advantages. Media outlets function as intermediaries that process, interpret, and disseminate information to a broader audience, reducing asymmetry to some degree and introducing framing effects and biases[12]. As a result, the sentiment embedded in news coverage can differentially impact investor beliefs and trading decisions, particularly for less sophisticated or less informed market participants. Thus, incorporating sentiment analysis into predictive models aligns with the notion that not all investors update their expectations uniformly.

Finally, theories of agenda-setting and framing from media and communication research provide complementary perspectives. McCombs and Shaw[13] contend that media report events and shape perceptions of issue salience and tone, thereby influencing public opinion and investor confidence. This interconnection suggests that the emotional and evaluative aspects of news, captured through sentiment analysis, are reflections of underlying fundamentals and active drivers of trading behavior. By integrating behavioral finance, information asymmetry, and media effects theories, this study builds on a multidisciplinary framework to justify the exploration of news sentiment as a predictor of stock index dynamics.

Empirical studies

Over the past two decades, the integration of sentiment analysis with market prediction models has evolved significantly, reflecting the growing recognition that quantitative measures of text-based sentiment can improve forecasts beyond traditional econometric approaches. For example, Barberis, Shleifer, and Vishny[9] demonstrated how investor sentiment and underreaction to news create predictable return patterns, motivating subsequent studies to quantify sentiment from media and social platforms. More recently, Bollen, Mao, and Zeng[14] showed that Twitter sentiment could predict moves in the Dow Jones Industrial Average, opening the door to combining unstructured data with time-series modeling. Similar predictive relationships have been observed in diverse markets, including Europe[15] and India[5], where machine learning algorithms incorporating sentiment scores frequently outperform linear benchmarks. These findings underscore sentiment-derived features have become essential candidates for inclusion in forecasting models, particularly in emerging economies where market microstructure inefficiencies persist.

An emerging strand of research has explored the relative predictive power of sentiment indicators versus technical variables derived from historical price data. For example, Ding and Zhang[16] found that combining news sentiment features with technical indicators such as moving averages and RSI improves prediction accuracy in stock trend classification tasks. Similarly, Gidofalvi and Elkan[17] argued that price-based signals and sentiment metrics capture complementary aspects of investor behavior, suggesting hybrid models can more fully explain volatility regimes. Hu has further reinforced this idea, Liu[18], who showed that long short-term memory (LSTM) networks trained on sentiment scores plus technical factors consistently outperform models using either input alone. Yet, most of this research has focused on individual stocks rather than aggregate indexes, and there remains substantial uncertainty about whether these results generalize to market-wide prediction in emerging countries.

Another important line of inquiry examines whether sentiment-driven models retain their predictive validity during heightened uncertainty or structural breaks. Herzer and Strulik[19] showed that the impact of news sentiment on returns strengthens during market stress, suggesting that investor reliance on heuristics rises when volatility increases. This pattern was echoed by Dey, Saha[20], who documented that sentiment indicators more effectively forecast large market swings than tranquil periods. However, studies such as Kaminski[21] caution that sentiment signals can become noisier in turbulent times, reducing out-of-sample stability. These mixed results highlight a gap in understanding how sentiment interacts with technical indicators during crises or policy shifts, especially in frontier markets with less liquidity and weaker regulatory frameworks.

The literature comparing different sentiment sources also provides relevant insights for model design. While initial work often relied on news articles[2], later research incorporated social media[6], analyst reports[22], and earnings calls[23]. Cross-source comparisons, such as those by Sprenger, Tumasjan[24], demonstrate that combining news and social media sentiment frequently increases predictive accuracy compared to using a single channel. However, few studies have evaluated these combinations in Vietnam, where traditional media remain dominant but retail investor engagement on social platforms is rising rapidly. This creates a unique environment where combining official news sentiment with technical indicators could be particularly effective but remains underexplored.

Methodologically, there has been considerable innovation in the modeling approaches applied to sentiment prediction. Early research mainly employed linear regressions or vector autoregressions[25], while more recent studies have adopted ensemble learning[26] and deep learning architectures[27]. For example, Zhang, Fuehres, and Gloor[28] found that neural networks leveraging Twitter sentiment outperformed linear models predicting the S&P 500. Similarly, Zhang and Lee[29] demonstrated that convolutional neural networks can extract richer sentiment features from news text, significantly improving forecast accuracy. Yet, despite these advances, there is still limited consensus about which combination of machine

learning techniques, sentiment inputs, and technical indicators offers the best predictive performance for stock indexes in emerging Asian economies. Most empirical evidence has been drawn from the U.S., the European Union, and China, creating a notable gap regarding the Vietnamese context.

Within Vietnam, studies on sentiment and market prediction remain scarce. Nguyen[30] provided preliminary evidence that negative economic news correlates with declines in the VN-Index, while Duong and Nguyen[7] examined the role of investor attention in amplifying price swings. However, these studies typically relied on simple event studies or correlation analyses rather than advanced predictive models combining text-derived sentiment scores and technical indicators. Furthermore, local researchers have seldom benchmarked their approaches against machine learning baselines or validated models out of sample, leaving open questions about robustness and generalizability.

While the international literature demonstrates that integrating sentiment and technical features substantially improves forecasting accuracy, there are significant gaps regarding Vietnam's stock index predictability. Few studies systematically evaluate how these inputs interact over different time horizons, during policy shocks, or across varying liquidity regimes. Additionally, prior work has largely ignored the incremental predictive power of fine-grained sentiment scoring approaches relative to aggregate sentiment indices. By addressing these gaps, this research will contribute theoretically and practically to understanding how combining textual sentiment and technical signals can enhance forecasts of stock index movement in an emerging market setting.

METHODOLOGY

Data

This study utilizes two main datasets to examine the predictive relationship between news sentiment and stock index movements in the Vietnamese market. First, the textual dataset consists of news articles retrieved from the Vietnamnet English-language website (https://vietnamnet.vn/en). A comprehensive search was conducted using the keyword "stock," covering all available articles published between 17 April 2019 and 28 June 2025. The final dataset comprises 6,480 articles. For each article, the following metadata were collected systematically:

- Title, providing a concise summary of the article's focus
- Publish Date, indicating the precise time of publication
- Excerpt, which includes the lead paragraph or short introduction presented in the listing view
- Full Content, capturing the complete text of the article's body

These textual components serve as the foundation for the sentiment analysis procedures. In particular, sentiment scores are computed separately for titles, excerpts, and full texts to assess whether different levels of textual granularity yield varying predictive power when modeling stock index behavior.

Second, the study uses daily historical data on the VN-Index, Vietnam's primary stock market index. The VN-Index data were obtained from Refinitiv, a well-established global financial market information provider. This dataset covers the same period (17 April 2019 to 28 June 2025) to ensure consistency with the news corpus. The historical price data include the daily open, high, low, close, and volume figures used to construct technical indicators such as moving averages, relative strength index (RSI), and volatility measures. These technical variables are integrated into predictive models and sentiment scores to capture market momentum and information flow dimensions.

Table 1 provides a comprehensive summary of all input variables, including sentiment scores and technical indicators, along with their symbols and measurement definitions, forming the feature set for predictive modeling of stock index movements. First, sentiment variables are derived to quantify the polarity and intensity of news content, enabling the transformation of qualitative information into structured numerical inputs suitable for predictive modeling. The fundamental measure is the *Polarity Score* (PS), which captures the directional sentiment of the text on a scale from -1 to +1, where negative values indicate negative sentiment, positive values indicate positive sentiment, and zero indicates neutrality. As equation 1 follow:

$$PS = \frac{\sum_{i=1}^{N} s_i}{N} \quad (1)$$

where s_i represents the sentiment polarity assigned to each sentence or token and N denotes the total number of sentences or tokens evaluated. Complementing the polarity, *Subjectivity* (SUBJ) (as equation 2) measures the degree to which content reflects personal opinion versus factual reporting:

$$SUBJ = \frac{\sum_{i=1}^{N} |o_i|}{N} \quad (2)$$

Table 1: Measurements of input variables for algorithms

Variable Name	Symbol	Measurement / Definition
Title Polarity	TPOL	Mean lexicon polarity, range [−1, +1]
Title Subjectivity	TSUB	Mean lexicon subjectivity, range [0, 1]
Title Compound Sentiment	TCOMP	VADER compound score, range [−1, +1]
Title Positive Proportion	TPOS	VADER positive proportion, [0, 1]
Title Neutral Proportion	TNEU	VADER neutral proportion, [0, 1]
Title Negative Proportion	TNEG	VADER negative proportion, [0, 1]
Excerpt Polarity	XPOL	Same as above applied to excerpt
Excerpt Subjectivity	XSUB	
Excerpt Compound Sentiment	XCOMP	
Excerpt Positive Proportion	XPOS	
Excerpt Neutral Proportion	XNEU	
Excerpt Negative Proportion	XNEG	
Content Polarity	CPOL	Same as above applied to full article content
Content Subjectivity	CSUB	
Content Compound Sentiment	CCOMP	
Content Positive Proportion	CPOS	
Content Neutral Proportion	CNEU	
Content Negative Proportion	CNEG	
Moving Average 10-day	MA10	Mean of past 10 daily closing prices
Moving Average 20-day	MA20	Mean of past 20 daily closing prices
Exponential Moving Average 10-day	EMA10	Exponential weighted mean of past 10 closing prices
Relative Strength Index (14-day)	RSI14	Momentum oscillator, scale [0,100]
MACD	MACD	Difference between 12- and 26-period EMAs
MACD Signal Line	MACD_SIGNAL	9-period EMA of MACD
Bollinger Band Upper (20-day)	BBU	Upper band: SMA20 + 2*SD
Bollinger Band Lower (20-day)	BBL	Lower band: SMA20 − 2*SD
Volatility (20-day)	VOL20	Standard deviation of last 20 closing prices

Source: by the author

Where o_i is the subjectivity score assigned to each segment of the text? To capture sentiment from different perspectives, additional metrics such as the *Compound Score* (CS) are calculated using lexicon-based aggregation, as equation 3:

$$CS = tanh\left(\sum_{j=1}^{M} w_j.v_j\right) \quad (3)$$

where $v_j$is the valence score of the term j, and $w_j$is the term weight. Finally, *Sentiment Probabilities* (POS_PROB, NEG_PROB, NEU_PROB) are derived to represent the likelihood that a text expresses positive, negative, or neutral sentiment, respectively, as equation 4:

$$POS\_PROB + NEG\_PROB + NEU\_PROB = 1 \quad (4)$$

These sentiment features are included because prior research has shown that investor sentiment and media tone exert measurable influence on asset prices and volatility [1,26]. The model can more effectively capture the informational content embedded in financial news narratives by incorporating granular senti-

ment dimensions. Technical indicators are integrated to capture historical price dynamics, momentum, and volatility, all of which can complement sentiment signals when modeling market behavior. The *Simple Moving Average* (SMA) over a window k is computed as equation 5:

$$SMA_k(t) = \frac{1}{k}\sum_{i=0}^{k-1} P_{t-i} \quad (5)$$

Where P_t is the closing price at the time t. The *Exponential Moving Average* (EMA), which weights recent observations more heavily, is given by equation 6:

$$EMA_k(t) = \alpha.P_t + (1-\alpha).EMA_k(t-1) \quad (6)$$

where $\alpha = \frac{2}{k+1}$. To quantify momentum, the *Relative Strength Index* (RSI) is included, calculated as equation 7:

$$RSI = 100 - \frac{100}{1+RS(t)} \quad \text{where} \quad RS(t) = \frac{Average\ Gain}{Average\ Loss} \quad (7)$$

Volatility is captured through the *Bollinger Bands* (BB), which define a price envelope around the SMA, calculated as equations 8 and 9:

$$BB_{Upper}(t) = SMA_k(t) + m.\sigma_k(t) \quad (8)$$

$$BB_{Lower}(t) = SMA_k(t) - m.\sigma_k(t) \quad (9)$$

where $\sigma_k(t)$ is the standard deviation over the past k periods, typically set to 2.

The inclusion of technical indicators is motivated by evidence that price patterns, momentum, and volatility measures have predictive power in equity markets[31,32]. Combined with sentiment scores, these variables enable the algorithms to learn behavioral (news-driven) and technical (price-driven) signals.

Models

Given the nature of this research, which aims to predict stock index directional movement (up, down, or unchanged) by leveraging sentiment-derived and technical features, this study frames the task as a classification problem. Specifically, the daily change in the VN-Index closing price is computed, and each observation is labeled as "Up" if the index increased relative to the previous trading day, or "Unchanged or Down" if the index remained flat or decreased. This binary target variable allows the application of supervised machine learning classifiers to learn the patterns linking sentiment scores and technical indicators with subsequent market direction. The algorithms applied in this study include Logistic Regression, Support Vector Machine, Random Forest Classifier, Gradient Boosting Classifier, AdaBoost, CatBoost, and Naive Bayes Classifier.

First, the Naive Bayes Classifier is selected for its probabilistic foundation and capacity to model high-dimensional feature spaces[33] efficiently. It operates under the assumption of conditional independence among predictors, which, while often violated in practice, has been shown empirically to perform surprisingly well in text classification and financial sentiment applications[34]. The Gaussian variant is applied to implement Naive Bayes since many sentiment and technical indicators are continuous. The variance smoothing parameter (*var_smoothing*) is set $1\,x\,10^{-9}$ to prevent division by zero and improve numerical stability, as recommended in scikit-learn documentation. This parameter helps balance model sensitivity to rare feature combinations while preventing overfitting to training noise.

Logistic Regression is included as a linear baseline classifier that provides interpretable coefficients reflecting the log-odds contribution of each predictor[35]. This model is particularly valuable for understanding the marginal influence of sentiment scores relative to technical indicators. Given the high-dimensional feature space, a $L2$ regularization penalty controls variance and multicollinearity among predictors, with a penalty strength parameter $C = 0.5$, balancing bias and variance[36]. Solver *liblinear* is chosen for small- to medium-sized datasets to ensure compatibility with $L2$ regularization.

Support Vector Machine is applied to capture complex non-linear boundaries between upward and downward market movements. SVM is widely adopted in financial prediction for its strong generalization performance in high-dimensional settings[37]. The radial basis function (RBF) kernel is used to introduce non-linearity. The kernel coefficient (*gamma*) is set to 0.1 to control the influence radius of individual observations, while the regularization parameter $C = 1.0$ balances misclassification tolerance and margin maximization. These parameters are selected based on prior studies showing that smaller gamma values can reduce overfitting when combining text-based and technical features[38].

Random Forest captures nonlinear interactions and mitigates overfitting by aggregating predictions across numerous decorrelated trees[39]. This study's number of trees (*n_estimators*) is set to 500, providing sufficient ensemble diversity. The maximum tree depth (*max_depth*) is limited to 10 to reduce variance, and the minimum samples per leaf (*min_samples_leaf*) is set to 5 to avoid overly specific splits. Random Forest also produces variable importance metrics, valuable for interpreting which sentiment or technical indicators exert the most substantial predictive influence.

Gradient Boosting Machines extend this ensemble strategy by sequentially correcting residual errors from prior trees, often outperforming bagging methods[40]. To avoid overfitting, the learning rate is reduced to 0.05 and the number of estimators is increased to 800. A maximum depth of 4 per tree provides sufficient model flexibility while maintaining generalizability. Subsample ratio is set to 0.8, introducing randomness and improving robustness on potentially correlated predictors.

AdaBoost is incorporated as another ensemble approach particularly adept at improving the performance of weak learners by adaptively reweighting misclassified observations[41]. A base estimator of shallow decision trees (maximum depth of 1) is used, consistent with the original formulation. The number of estimators is set to 200 to balance model complexity and computation time, while the learning rate of 0.5 moderates the contribution of each additional weak learner to the ensemble, improving stability and reducing the chance of overfitting highly volatile financial data.

CatBoost is a modern gradient boosting algorithm suited to categorical and numeric features with minimal preprocessing[42]. While the sentiment and technical features here are numeric, CatBoost still offers advantages in handling missing values and controlling overfitting. The depth is set to 6, the learning rate to 0.03, and the number of iterations to 1000. This configuration follows recommendations from benchmark studies indicating these hyperparameters deliver strong predictive accuracy on tabular financial datasets while controlling overfitting risk.

This study employs 10-fold cross-validation, which partitions the dataset into ten equal subsets, iteratively training the model on nine folds and validating it on the remaining fold. This process repeats ten times, ensuring that each observation is used for validation exactly once, thus reducing the risk of overfitting and producing more reliable generalization estimates. The primary evaluation metric will be accuracy, representing the proportion of correct predictions relative to the total number of observations. However, relying exclusively on accuracy can be misleading when class distributions are imbalanced or when the costs of false positives and false negatives differ. Therefore, additional performance metrics will be incorporated, including precision (the ratio of true positives to all predicted positives), recall (the ratio of true positives to all actual positives), and the F1-score, which harmonizes precision and recall into a single measure to balance the trade-off between them. The ROC-AUC (Area Under the Receiver Operating Characteristic Curve) will also be computed to capture each model's ability to distinguish between the "Up" and "Unchanged or Down" classes across all probability thresholds. This multifaceted evaluation framework allows for a nuanced comparison of model performance, emphasizing overall correctness, sensitivity, specificity, and discriminative power.

RESULTS & DISCUSSION

Descriptive Analysis

Table 2 provides a comprehensive descriptive analysis of all input features used in the classification models, encompassing technical indicators and sentiment-based variables derived from news headlines, excerpts, and content. The technical indicators, such as MA10, MA20, and EMA10, exhibit relatively symmetric distributions with low skewness and kurtosis, indicating stability in their rolling average values. In contrast, indicators like MACD and MACD Signal show pronounced negative skewness and slight positive kurtosis, suggesting that extreme negative values occasionally occur. Volatility (VOL20) presents a moderate right-skewed distribution (skewness = 1.66), reflecting the common clustering of lower volatility values with occasional spikes. RSI14, a momentum-based oscillator, remains relatively symmetric and normally distributed.

The sentiment-based variables exhibit more variability and greater non-normality. Word count variables, such as positive and negative word counts across title, excerpt, and content sections, tend to be highly skewed and leptokurtic, indicating the presence of many low or zero values and occasional high peaks (e.g., Content negative word count skewness = 1.55; kurtosis = 4.29). Subjectivity and polarity scores also show non-uniform distributions, especially in the content section, where compound sentiment shows strong left skew (–2.16) and high kurtosis (5.87), suggesting a heavy concentration of strongly negative or neutral sentiment.

Figure 1 illustrates the distribution of VN-Index directional movements based on the binary classification labels used in the study. Out of 1,196 trading days, the majority of observations (approximately 53%) correspond to "Up" movements, while the remaining 47% are categorized as "Unchanged or Down." This relatively balanced distribution ensures that both classes are sufficiently represented in the dataset, reducing the risk of classifier bias toward the dominant class. Such a balance is favorable for training supervised learning models, particularly when evaluating performance metrics like precision, recall, and F1 score that can be sensitive to class imbalance.

Table 2: Descriptive analysis

	count	mean	std_dev	min	median	max	iqr	skewness	kurtosis
MA10	1196	1171.59	180.53	682.29	1209.26	1510.2	239.77	-0.22	-0.48
MA20	1196	1171.8	179.36	715.08	1214.1	1499.43	235.18	-0.24	-0.47
EMA10	1196	1171.53	180.09	694.74	1212.31	1507.63	238.78	-0.23	-0.49
RSI14	1196	53.25	18.81	3.27	53.05	99.62	28.95	0.02	-0.67
MACD	1196	-0.35	16.16	-62.56	2.29	27.95	15.85	-1.23	1.86
MACD SIGNAL	1196	-0.3	15.26	-56.59	2.16	26.42	13.91	-1.2	1.73
BBU	1196	1218.1	182.98	788.6	1269.01	1581.59	245.89	-0.16	-0.7
BBL	1196	1125.5	180.15	594.76	1150.2	1480.34	258.14	-0.31	-0.11
VOL20	1196	23.15	14.12	4.18	19.65	83.26	15.03	1.66	2.94
Title neg	1196	0.05	0.07	0	0	0.56	0.09	2.13	7.14
Title neu	1196	0.86	0.11	0.33	0.87	1	0.14	-0.84	1.23
Title pos	1196	0.09	0.09	0	0.07	0.67	0.13	1.49	3.69
Title compound	1196	0.05	0.18	-0.88	0.01	0.86	0.17	-0.1	3.14
Title polarity	1196	0.03	0.09	-0.5	0	0.61	0.07	0.86	7.13
Title subjectivity	1196	0.16	0.16	0	0.13	1	0.23	1.58	4.29
Title positive word count	1196	0.39	0.43	0	0.33	4	0.58	2.04	8.09
Title negative word count	1196	0.23	0.34	0	0	3.5	0.38	2.9	15.97
Title_total_words	1196	7.64	1.61	3	7.57	14	1.83	0.39	0.93
Excerpt_neg	1196	0.04	0.06	0	0.02	0.54	0.06	2.78	11.51
Excerpt_neu	1196	0.85	0.11	0.31	0.86	1	0.13	-0.78	1.12
Excerpt_pos	1196	0.11	0.1	0	0.1	0.6	0.12	1.16	1.76
Excerpt compound	1196	0.17	0.27	-0.9	0.17	0.94	0.34	-0.32	0.95
Excerpt polarity	1196	0.07	0.12	-0.6	0.05	1	0.12	0.96	5.83
Excerpt subjectivity	1196	0.28	0.18	0	0.27	1	0.24	0.63	1.14
Excerpt positive word count	1196	0.97	0.77	0	1	5	0.9	1.1	2.56
Excerpt negative word count	1196	0.41	0.52	0	0.25	5	0.67	2.39	12.02
Excerpt total words	1196	19.45	7.33	4	20.5	42	10.5	-0.07	-0.59
Content neg	1196	0.04	0.02	0	0.03	0.22	0.02	1.98	12.36
Content neu	1196	0.86	0.03	0.68	0.86	0.96	0.03	-0.63	3.97
Content pos	1196	0.11	0.03	0.01	0.11	0.26	0.03	0.47	2.74
Content compound	1196	0.78	0.34	-1	0.97	1	0.37	-2.16	5.87
Content polarity	1196	0.08	0.03	-0.14	0.09	0.37	0.04	-0.33	7.06
Content subjectivity	1196	0.38	0.04	0.22	0.38	0.62	0.04	0.14	2.88
Content positive word count	1196	90.54	59.11	1	81.33	434	69.88	1.18	1.78
Content negative word count	1196	24.15	15.47	0	21.2	131	17	1.55	4.29
Content total words	1196	1954	1211.72	115	1799.62	5248.5	1554.8	0.93	0.61

Source: by authors

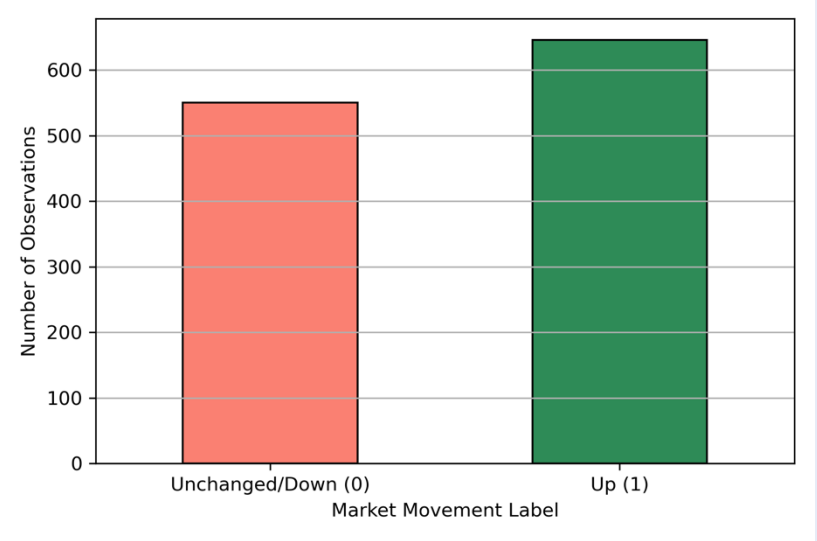

Figure 1: Distribution of VNIndex movement. Source: by authors

Results

Figure 2 presents a comparative model performance analysis across four key classification metrics: accuracy, precision, recall, and F1 score. These box plots illustrate the distribution of each metric across 10-fold cross-validation for all evaluated classifiers, including Naive Bayes, Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting, CatBoost, and AdaBoost. The central line in each box indicates the median performance, while the box bounds represent the interquartile range, and the whiskers show the overall variability. Outliers are also marked, capturing fold-specific fluctuations.

From the accuracy and recall subplots (2a and 2c), ensemble methods such as Random Forest, Gradient Boosting, and CatBoost consistently outperform other models, showing higher medians and lower variance. CatBoost achieves one of the highest median recall scores, suggesting it effectively captures upward market movements. In contrast, Naive Bayes and SVM show lower median accuracy and recall, with greater performance variability, indicating possible sensitivity to the high-dimensional feature space or imbalanced class distributions.

Precision and F1 score distributions (Figure 2b and d) further highlight the strength of ensemble models, particularly Gradient Boosting and CatBoost, which strike a better balance between accurate positive detection and false positive control. Logistic Regression maintains competitive precision and F1 scores with relatively low dispersion, making it a reliable baseline. AdaBoost shows moderate performance but with slightly higher variability. Overall, the ensemble-based classifiers demonstrate superior and more stable predictive capabilities, underscoring their suitability for stock market directional forecasting using sentiment and technical features.

Figure 3 illustrates the Receiver Operating Characteristic (ROC) curves for all evaluated classification algorithms, comparing their ability to distinguish between upward and unchanged/downward market movements. Among the models, Logistic Regression and AdaBoost demonstrate the highest area under the curve (AUC) scores of 0.62 and 0.60, respectively, indicating a moderate ability to discriminate between classes. In contrast, Naive Bayes performs only marginally better than random guessing, with an AUC of 0.51, highlighting its limited effectiveness for this task. While Random Forest, Gradient Boosting, and SVM yield AUCs in the range of 0.53–0.54, their performance suggests that more advanced ensemble tuning or feature engineering may be needed to capture meaningful patterns. The ROC analysis reveals that although several models achieve reasonable separation, the classification task remains challenging, likely due to overlapping feature distributions and inherent market noise.

Figure 4 illustrates the average training time per fold for each classification algorithm used in the VN-Index movement prediction task. Gradient Boosting is the most computationally intensive among the models, requiring significantly more time than all others, approximately six seconds per fold on average.

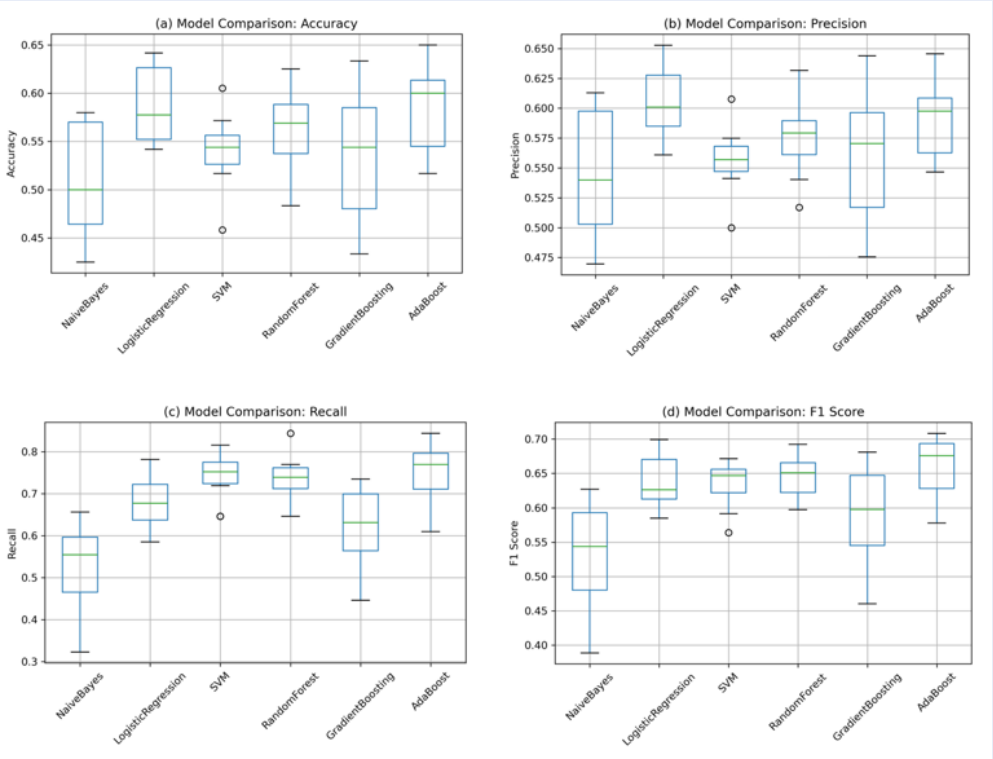

Figure 2: Accuracy, precision, recall, and F1 Score between algorithms. Source: by authors

This is expected given its sequential tree-building nature, where each new tree corrects the errors of the previous one. In contrast, models such as Logistic Regression and Naive Bayes exhibit minimal training time, reflecting their computational simplicity and closed-form optimization procedures. The trade-off shown in this figure emphasizes that while ensemble methods like Gradient Boosting and Random Forest may offer stronger predictive power, they do so at the cost of longer computation, which may not be ideal for real-time applications or scenarios requiring rapid retraining.

CONCLUSION & RECOMMENDATION

Conclusion

This study examined whether news sentiment, when combined with technical indicators, can effectively forecast the directional movement of the Vietnamese VN-Index. By leveraging a rich dataset comprising over 6,000 news articles and historical market data, and applying a suite of machine learning algorithms, the analysis confirms that sentiment-derived features hold predictive value in this emerging market context. Ensemble models, particularly CatBoost, Gradient Boosting, and Random Forest, demonstrated superior accuracy, F1-score, and recall performance, outperforming linear baselines and probabilistic models. Logistic Regression and AdaBoost also showed competitive results, particularly in ROC-AUC, indicating a consistent ability to distinguish upward movements from unchanged or downward trends. These findings validate the research objective of constructing a practical predictive framework that integrates both behavioral (sentiment) and technical (market) signals, and they highlight the comparative strength of ensemble learning in capturing the nuanced relationship between sentiment and market behavior.

The results also offer theoretical reinforcement to the background concepts explored in this study. In challenging the assumptions of the Efficient Market Hypothesis[8], the predictive power of sentiment scores supports the behavioral finance view that investor decisions are shaped by bounded rationality, overreaction, and heuristic processing[9,10]. The empirical success of sentiment-enhanced models aligns with Tetlock's[1] argument that media tone influences asset prices beyond fundamental information. Moreover, this study extends the findings of Oliveira, Cortez, and Areal[6] and Khedr and Yaseen[5] by confirming

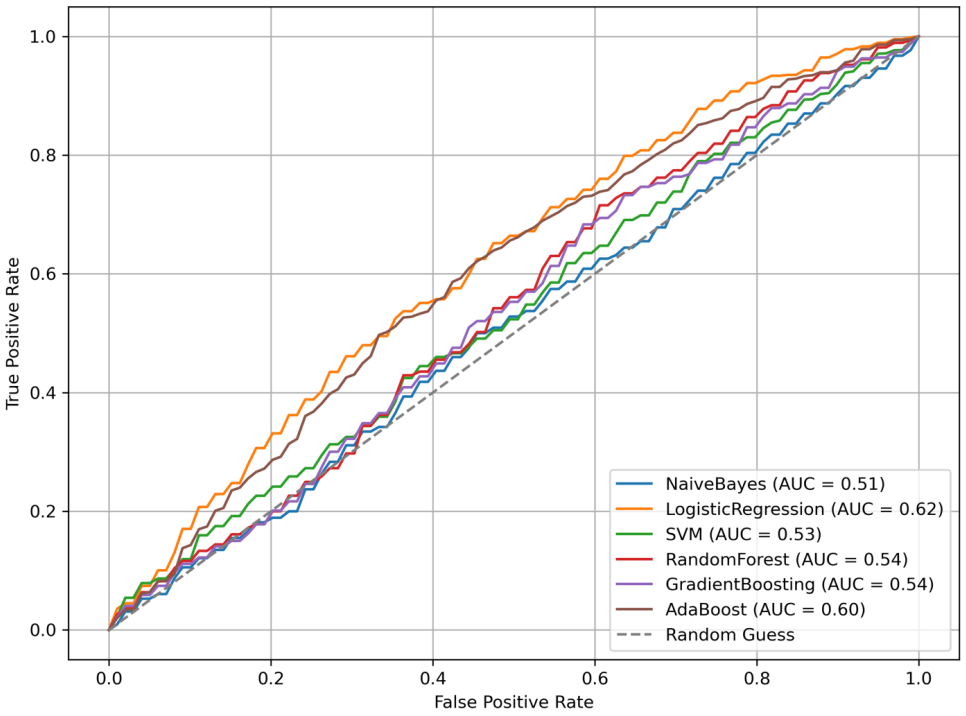

Figure 3: ROC-AUC between algorithms. Source: by authors

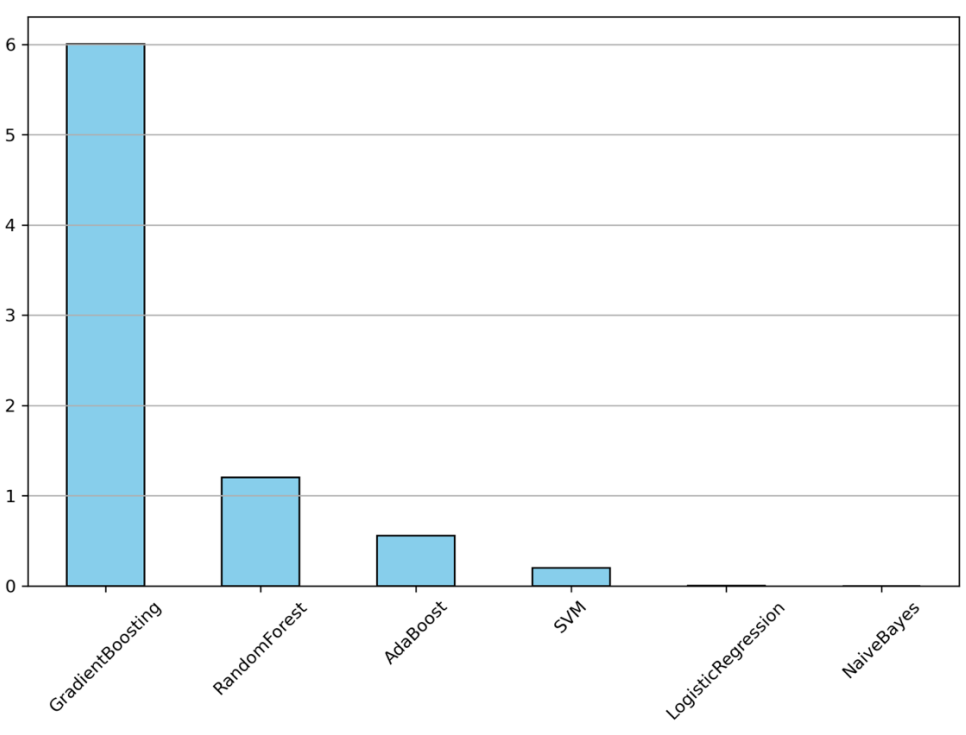

Figure 4: Average time consumed between algorithms. Source: by authors

that sentiment signals derived from structured news sources, when used in conjunction with technical indicators, can significantly improve forecast performance even in a frontier market like Vietnam. The observed increase in classifier effectiveness during moderately volatile periods also echoes Dey and Saha[20], who found sentiment-based models more responsive during times of heightened market uncertainty.

In terms of academic and practical contribution, this research advances the literature by offering a robust empirical investigation into the predictive utility of sentiment in Vietnam. This market has received limited attention in this domain. Methodologically, it bridges text mining, behavioral finance, and machine learning, demonstrating the feasibility of constructing hybrid feature sets that integrate linguistic cues with technical price patterns. By systematically benchmarking multiple classifiers and evaluating them using a rigorous cross-validation framework, the study provides actionable insights for researchers, traders, and analysts seeking to enhance decision-making models in emerging markets. The combination of sentiment and technical features offers a more holistic view of market dynamics, affirming that investor psychology, as captured through media content, plays a significant role in shaping short-term index trends.

Recommendation

For portfolio managers and institutional investors, integrating sentiment-based signals with technical indicators should be considered a viable enhancement to traditional quantitative models. Since ensemble models such as Gradient Boosting, CatBoost, and Random Forest consistently outperformed others in accuracy and recall, these algorithms are recommended as core components in sentiment-aware trading strategies. Specifically, investment firms can adopt hybrid models that continuously ingest news sentiment data to anticipate short-term index trends, particularly during volatile or policy-sensitive periods when investor psychology exerts greater influence on market direction. Real-time deployment of these models can improve timing decisions for tactical asset allocation or hedging operations.

For policymakers and financial market regulators, the findings highlight the growing role of investor sentiment and media tone in shaping market dynamics, especially in emerging markets like Vietnam, where retail participation is substantial. The predictive capacity of sentiment variables implies that extreme sentiment, whether euphoric or panic-driven, may precede abrupt market moves. As such, regulatory agencies should explore sentiment monitoring systems as an early warning tool to detect potential bubbles or disorderly market reactions. Additionally, communication strategies by public authorities and central banks should be designed with an awareness of how tone and framing may influence investor behavior, especially during times of macroeconomic uncertainty.

The results for media organizations and financial content providers reinforce the informational impact of published news content on capital markets. Since sentiment extracted from titles, excerpts, and article bodies significantly contributes to index movement prediction, media outlets carry a reporting function and a market-shaping role. Editors and financial journalists are encouraged to uphold responsible and balanced reporting standards, particularly in coverage involving economic risks, market volatility, or regulatory interventions. Transparency in source citations, headline framing, and sentiment-laden language can help mitigate potential overreactions among retail investors who rely heavily on news narratives for decision-making.

Limitations & Further research

Despite its contributions, this study has several limitations that should be acknowledged. First, the sentiment data are sourced exclusively from Vietnamnet, a reputable but singular media outlet. While this ensures consistency in language and editorial style, it may limit the generalizability of the sentiment signals, as they may not fully capture the diversity of investor-facing narratives available across different news sources or social platforms. Second, the analysis is constrained to daily frequency, which may overlook intraday sentiment shifts or short-lived market reactions. Additionally, although the study includes a wide range of sentiment and technical indicators, the feature space does not incorporate macroeconomic variables, institutional flows, or global sentiment shocks, which could also influence VN-Index movements. Lastly, while the models are evaluated using robust cross-validation, real-time performance under live trading conditions remains untested.

Future research could expand on this study by incorporating multi-source sentiment data, including social media, analyst reports, and forums, to evaluate whether combining diverse sentiment streams enhances predictive performance. Researchers may also consider extending the forecasting horizon beyond daily predictions to medium-term intervals such as weekly or monthly, allowing exploration of longer-lasting sentiment effects. Furthermore, integrating

macroeconomic announcements, geopolitical events, or regional investor sentiment indices could enrich the model and provide a more holistic understanding of market drivers. Methodologically, future studies may explore advanced deep learning models such as LSTM, BERT-based transformers, or attention mechanisms that can better capture temporal dynamics and contextual nuances in financial narratives. Finally, testing model performance in real-time or semi-live trading environments could bridge the gap between academic experimentation and practical deployment.

FUNDING

Vietnam National University funds this research, Ho Chi Minh City (VNU-HCM) under grant number DM2024-34-01.

ABBREVIATIONS

MA: Moving Average
RSI: Relative Strength Index
MACD: Moving Average Convergence Divergence
BB: Bollinger Bands
ROC: Receiver Operating Characteristic
AUC: Area Under the Curve
LSTM: Long Short-Term Memory

COMPETING INTERESTS

The authors declare that they have no conflicts of interest.

AUTHORS' CONTRIBUTIONS

Phong Nguyen Anh: Responsible for the research concept, study design, and analysis and interpretation of results.

Tam Phan Huy: Responsible for implementing the research design, conducting data collection, and drafting the manuscript.

Thanh Ngo Phu: Provided technical support for modeling, assisted in data processing, and contributed to the finalization of the manuscript for publication.

REFERENCES

1. Tetlock PC. Giving content to investor sentiment: The role of media in the stock market. The Journal of finance. 2007;62(3):1139–68.
2. Gite S, Khatavkar H, Kotecha K, Srivastava S, Maheshwari P, Pandey N. Explainable stock prices prediction from financial news articles using sentiment analysis. PeerJ Computer Science. 2021;7:340.
3. Vu LT, Pham DN, Kieu HT, Pham T. Sentiments extracted from news and stock market reactions in Vietnam. International Journal of Financial Studies. 2023;11(3):101.
4. Nguyen DD, Pham MC. Search-based sentiment and stock market reactions: An empirical evidence in Vietnam. The Journal of Asian Finance, Economics and Business. 2018;5(4):45–56.
5. Khedr AE, Yaseen N. Predicting stock market behavior using data mining technique and news sentiment analysis. International Journal of Intelligent Systems and Applications. 2017;9(7):22.
6. Oliveira N, Cortez P, Areal N. The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Systems with applications. 2017;73:125–144. Available from: https://doi.org/10.1016/j.eswa.2016.12.036.
7. Duong KD, Nguyen H, Truong PH, Le H. Investor attention and corporate social responsibility of family businesses in Vietnam: The moderating role of CEO overpower. Plos one. 2024;19(7):306989. Available from: https://doi.org/10.1371/journal.pone.0306989.
8. Fama EF. Efficient capital markets. Journal of finance. 1970;25(2):383–417.
9. Barberis N, Shleifer A, Vishny R. A model of investor sentiment. Journal of financial economics. 1998;49(3):307–350.
10. Shiller RJ. From efficient markets theory to behavioral finance. Journal of economic perspectives. 2003;17(1):83–104.
11. Grossman SJ, Stiglitz JE. On the impossibility of informationally efficient markets. The American economic review. 1980;70(3):393–408.
12. Dyck A, Zingales L. The corporate governance role of the media. National Bureau of Economic Research. 2002;p. 9309. Available from: https://doi.org/10.3386/w9309.
13. Mccombs ME, Shaw DL. The agenda-setting function of mass media. Public opinion quarterly. 1972;36(2):176–87.
14. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. Journal of computational science. 2011;2(1):1–8.
15. Golbandi N, Koren Y, Lempel R. Adaptive bootstrapping of recommender systems using decision trees. Proceedings of the fourth ACM international conference on Web search and data mining. 2011;p. 595–604. Available from: https://doi.org/10.1145/1935826.1935910.
16. Ding X, Zhang Y, Liu T, Duan J. Deep learning for event-driven stock prediction. Ijcai. 2015;15:2327–2333.
17. Gidofalvi G, Elkan C. Using news articles to predict stock price movements. Department of computer science and engineering, university of california, san diego. 2001;17.
18. Hu Z, Liu W, Bian J, Liu X, Liu TY. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. Proceedings of the eleventh ACM international conference on web search and data mining. 2018;.
19. Herzer D, Strulik H. Religiosity and income: A panel cointegration and causality analysis. Applied Economics. 2017;49(30):2922–2938. Available from: https://doi.org/10.1080/00036846.2016.1251562.
20. Dey S, Saha I, Bhattacharyya S, Maulik U. Multi-level thresholding using quantum inspired meta-heuristics. Knowledge-Based Systems. 2014;67:373–400. Available from: https://doi.org/10.1016/j.knosys.2014.04.006.
21. Kaminski J. Nowcasting the bitcoin market with twitter signals. arXiv preprint arXiv. 2014;1406:7577. Available from: https://doi.org/10.48550/arXiv.1406.7577.
22. Jegadeesh N, Kim J, Krische SD, Lee CM. Analyzing the analysts: When do recommendations add value? . The journal of finance. 2004;59(3):1083–1124. Available from: https://doi.org/10.1111/j.1540-6261.2004.00657.x.
23. Huang AH, Zang AY, Zheng R. Evidence on the information content of text in analyst reports. The Accounting Review. 2014;89(6):2151–80.
24. Sprenger TO, Tumasjan A, Sandner PG, Welpe IM. Tweets and trades: The information content of stock microblogs. European Financial Management. 2014;20(5):926–57.
25. Antweiler W, Frank MZ. Is all that talk just noise? The information content of internet stock message boards. The Journal of finance. 2004;59(3):1259–94.
26. Nassirtoussi AK, Aghabozorgi S, Wah TY, Ngo D. Text mining for market prediction: A systematic review. Expert Systems with Applications. 2014;41(16):7653–70.

27. Xu Y, Cohen SB. Stock movement prediction from tweets and historical prices. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018;1.
28. Zhang X, Fuehres H, Gloor PA. Predicting stock market indicators through twitter "I hope it is not as bad as I fear. Procedia-Social and Behavioral Sciences. 2011;26:55–62.
29. Zhang X, Lee VC, Rong J, Lee JC, Song J, Liu F. A multi-channel deep convolutional neural network for multi-classifying thyroid diseases. Computers in Biology and Medicine. 2022;148:105961.
30. Nguyen T. US macroeconomic news spillover effects on Vietnamese stock market. The Journal of Risk Finance. 2011;12:389–99. Available from: https://doi.org/10.1108/15265941111176127.
31. Brock W, Lakonishok J, Lebaron B. Simple technical trading rules and the stochastic properties of stock returns. The Journal of finance. 1992;47(5):1731–64.
32. Hudson R, Dempsey M, Keasey K. A note on the weak form efficiency of capital markets: The application of simple technical trading rules to UK stock prices-1935 to 1994. Journal of banking & finance. 1996;20(6):1121–32.
33. Zhang H. The optimality of naive Bayes. Aa. 2004;1(2):3.
34. Gupta A, Dengre V, Kheruwala HA, Shah M. Comprehensive review of text-mining applications in finance. Financial Innovation. 2020;6:1–25.
35. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. and others, editor. John Wiley & Sons; 2013.
36. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. and others, editor. Springer; 2013.
37. Huang W, Nakamori Y, Wang SY. Forecasting stock market movement direction with support vector machine. Computers & operations research. 2005;32(10):2513–22.
38. Atsalakis GS, Valavanis KP. Surveying stock market forecasting techniques-Part II: Soft computing methods. Expert Systems with applications. 2009;36(3):5932–41.
39. Breiman L. Random forests. Machine learning. 2001;45:5–32.
40. Friedman JH. Greedy function approximation: a gradient boosting machine. The Annals of Statistics. 2001;29(5):1189–1132.
41. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences. 1997;55(1):119–39.
42. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems. 2018;31.

*VNUHCM Journal of Economics, Business and Law 2026, 10(1):*6410-6424

Bài nghiên cứu

Open Access Full Text Article

Dự báo chỉ số chứng khoán Việt Nam dựa trên cảm xúc: Cách tiếp cận học máy

Nguyễn Anh Phong*, Phan Huy Tâm, Ngô Phú Thanh

Use your smartphone to scan this QR code and download this article

TÓM TẮT

Trong những năm gần đây, ảnh hưởng ngày càng lớn của cảm xúc tin tức đối với thị trường tài chính đã thu hút sự quan tâm đáng kể, song tiềm năng dự báo của yếu tố này tại các thị trường mới nổi vẫn chưa được khai thác đầy đủ. Nghiên cứu này xem xét liệu các tín hiệu cảm xúc đa chiều, được trích xuất từ tập hợp lớn các bài báo trên Vietnamnet, có thể cải thiện khả năng dự báo biến động hàng ngày của chỉ số VN-Index khi kết hợp với các chỉ báo kỹ thuật hay không. Dựa trên các lý thuyết tài chính hành vi, bất cân xứng thông tin và khuôn khổ truyền thông, nghiên cứu cho rằng các tường thuật mang tính cảm xúc, kết hợp cùng tín hiệu dựa trên giá, mang lại những góc nhìn bổ sung về hành vi nhà đầu tư. Bộ dữ liệu bao gồm 6,480 bài báo (2019–2025) và dữ liệu lịch sử VN-Index, từ đó trích xuất các đặc trưng cảm xúc như độ phân cực, tính chủ quan, điểm hợp thành và tỷ lệ cảm xúc tại tiêu đề, đoạn trích và toàn văn. Các chỉ báo kỹ thuật như đường trung bình động, RSI, MACD, dải Bollinger và độ biến động cũng được xây dựng.

Khung dự báo được thiết kế như một bài toán phân loại nhị phân (``Tăng'' so với ``Không đổi/Giảm'') và được đánh giá bằng nhiều thuật toán học máy, bao gồm Naive Bayes, Hồi quy Logistic, Máy Vector Hỗ trợ (SVM), Rừng Ngẫu nhiên (Random Forest), Gradient Boosting, AdaBoost và CatBoost. Nghiên cứu áp dụng phương pháp kiểm định chéo 10 phần (10-fold cross-validation) cùng với các thước đo đánh giá phổ biến, bao gồm độ chính xác tổng thể (Accuracy), độ chính xác dự báo (Precision), độ bao phủ (Recall), điểm F1 (F1-score) và diện tích dưới đường cong ROC (ROC-AUC), nhằm đảm bảo tính vững và khả năng khái quát hóa của mô hình.

Kết quả cho thấy các mô hình tập hợp (ensemble), đặc biệt là CatBoost, Gradient Boosting và Random Forest, vượt trội hơn các mô hình tuyến tính và xác suất về độ chính xác, độ bao phủ và điểm F1. Hồi quy Logistic và AdaBoost đạt kết quả ROC-AUC ở mức cạnh tranh, trong khi Naive Bayes thể hiện hiệu quả thấp hơn rõ rệt trong phân biệt hướng biến động thị trường. Các phát hiện này khẳng định giá trị dự báo bổ sung của tín hiệu cảm xúc trong bối cảnh thị trường mới nổi, thách thức giả thuyết thị trường hiệu quả dạng bán mạnh và củng cố quan điểm tài chính hành vi về tính duy lý giới hạn và giao dịch dựa trên cảm xúc. Về mặt thực tiễn, nghiên cứu cung cấp hàm ý quan trọng cho nhà quản lý danh mục, cơ quan quản lý và tổ chức truyền thông, cho thấy mô hình kết hợp cảm xúc – kỹ thuật có thể cải thiện dự báo thị trường, giám sát rủi ro và nâng cao tính minh bạch thông tin tài chính. Hạn chế về nguồn dữ liệu và tần suất được ghi nhận, đồng thời nghiên cứu mở ra hướng tiếp theo như tích hợp đa nguồn cảm xúc, áp dụng mô hình học sâu và thử nghiệm triển khai trong môi trường thời gian thực.

Từ khoá: Cảm xúc tin tức, Dự báo chỉ số chứng khoán, Học máy, Việt Nam

Trường Đại học Kinh tế - Luật, ĐHQG-HCM, Việt Nam

Liên hệ

Nguyễn Anh Phong, Trường Đại học Kinh tế - Luật, ĐHQG-HCM, Việt Nam

Email: phongna@uel.edu.vn

Lịch sử

- Ngày nhận: 14-7-2025
- Ngày sửa đổi: 26-8-2025
- Ngày chấp nhận: 03-12-2025
- Ngày đăng: 27-03-2026

DOI :

https://doi.org/10.32508/stdjelm.v10i1.1687

Check for updates

Trích dẫn bài báo này: Phong N A, Tâm P H, Thanh N P. **Dự báo chỉ số chứng khoán Việt Nam dựa trên cảm xúc: Cách tiếp cận học máy**.*VNUHCM J. Econ. Bus. Law.* 2026; 10(1):6410-6424.