

AI-Based Stock Market Forecasting: Integrating Financial Indicators, Investor Sentiment, and Attention

Nguyen Quoc Thinh
Academy of Finance (Ha Noi, Viet Nam)
nguyenquocthinh140907@gmail.com

Abstract

This study analyzes the applicability of artificial intelligence (AI) in forecasting the Vietnamese stock market by integrating financial indicators with investor sentiment. The research methodology encompasses a literature review, secondary data analysis, and cross-referencing of official domestic and international sources. The findings indicate that price and trading volume data do not fully reflect investor expectations, attention, and reactions to new information. Consequently, the study proposes an AI data framework structured sequentially: macroeconomics - financial indicators - market information - attention - sentiment - trading behavior - price movements. Random Forest and XGBoost are identified as suitable for tabular data; whereas LSTM and Transformer architectures can effectively process time-series and textual data. The research underscores that data quality, temporal synchronization, interpretability, and data leakage control are critical determinants of forecasting efficacy. The results establish a foundation for developing a multi-source AI forecasting system tailored to the distinct characteristics of the Vietnamese stock market.

Keywords: *artificial intelligence; stock market forecasting; financial indicators; investor sentiment; machine learning.*

Table 1. *Selected Key Indicators of the Vietnamese Stock Market, 2024–2025*

Indicator	2024	2025	Change/Notes
VN-Index at year-end	—	1,784.49 points	+40.9% from end-2024
HNX-Index at year-end	—	248.77 points	+9.4%
Market capitalization: HOSE + HNX + UPCoM	VND 7,173.32 trillion	VND 9,975.36 trillion	+39.1%
Market capitalization/GDP	62.3%	77.6%	+15.3 percentage points
Average daily equity trading value	VND 20.985 trillion/session	VND 29.202 trillion/session	Approximately +39%
Securities accounts	Approximately 9.3 million	More than 11.6 million	Strong increase
GDP growth	7.09%	8.02%	Accelerating economic growth

Source: *State Securities Commission of Vietnam (SSC), 2024 Annual Report and 2025 market-end review; General Statistics Office.*

I. INTRODUCTION

The stock market is increasingly recognized as a vital component of the financial system and the capital allocation process within the economy. For Vietnam, the market expansion in recent years has unfolded simultaneously across multiple dimensions: increased market capitalization, amplified liquidity, a surge in the number of participating investors, modernized trading infrastructure, and escalating interest from international investors.

Data from the State Securities Commission reveals that by the end of 2025, the equity market capitalization on the HOSE, HNX, and UPCoM exchanges reached 9,975.36 trillion VND, representing a 39.1% increase compared to the end of 2024. The VN-Index closed 2025 at 1,784.49 points, up 40.9% year-over-year; concurrently, the average equity trading value attained 29,202 billion VND per session, marking an approximate 39% surge.

Contextualizing these figures within the real economy, the market capitalization at the end of 2025 was equivalent to approximately 77.6% of Vietnam's estimated GDP. Meanwhile, the 2025 GDP grew by 8.02%, the industrial and construction sector by 8.95%, and the service sector by 8.62%. Financial, banking, and insurance activities experienced a 7.82% growth rate.

Consequently, the stock market can no longer be treated as a small dataset confined to a few fundamental variables. The market scale is now sufficiently large to formulate a multidimensional data ecosystem, encompassing price, volume, market breadth, cash flow, foreign investor transactions, corporate disclosures, macroeconomic data, and manifestations of investor attention (Table 1 clearly illustrates this shift).

This escalation in scale engenders a notable research problem. As the market expands, the number of variables potentially correlated with price movements correspondingly increases. Stock prices reflect the outcomes of transactions; however, prior to execution, a sequence of information and reactions typically occurs: investors receive news, form expectations, shift their attention levels, adjust their trading decisions, and ultimately induce fluctuations in price and liquidity.

From this perspective, the author contends that AI-based market forecasting should not be confined to the binary question, "Will the price rise or fall tomorrow?". A more pertinent inquiry is: which information clusters emerge prior to price fluctuations, which clusters provide complementary predictive value, and how can AI process

them within a unified system? This is precisely the research gap this study aims to bridge.

Traditional models conventionally focus on price, returns, volume, or specific technical indicators. This approach offers the advantage of clear, quantifiable data with extensive historical records. However, the stock market also reacts to information that remains inadequately captured by technical variables at the time of observation.

Baker and Wurgler (2006) posited that investor sentiment can induce significant cross-sectional differences in stock returns. Tetlock (2007), through media content analysis, demonstrated that media pessimism correlates with downward price pressure, followed by a reversion to fundamentals; furthermore, exceptionally high or low levels of pessimism are linked to elevated trading liquidity.

These findings do not imply that investor sentiment can substitute for rigorous financial analysis. Conversely, the author argues that substantially greater value lies in the capacity to integrate both informational layers.

A corporation exhibiting growing profitability, improving cash flows, and a reasonable valuation establishes a robust financial foundation. Nevertheless, if potent negative information simultaneously permeates the market, prompting heightened attention and a synchronized pessimistic response from investors, short-term stock prices may deviate significantly from fundamental signals. AI possesses a distinct advantage in its ability to concurrently process heterogeneous variables.

Therefore, this study focuses on three primary research questions:

First: How have the current scale and data structure of the Vietnamese stock market facilitated the application of AI in forecasting?

Second: What complementary informational value can be generated by incorporating investor sentiment data into financial indicators?

Third: In what direction should an AI forecasting framework tailored to Vietnam be developed to concurrently ensure predictive capacity, interpretability, and practical applicability?

II. THEORETICAL FRAMEWORK

A. Stock Market Forecasting and the Limitations of Price Data

Stock prices reflect the supply-demand equilibrium at a specific point in time and are driven by information, expectations, risk, liquidity, and investor behavior. According to the Efficient Market Hypothesis (Fama, 1970), prices tend to fully reflect all available information, thereby constraining the predictability of future price movements based solely on public data. Nevertheless, behavioral biases, cognitive limitations in information processing, and transaction costs can generate supplementary predictive signals. Baker and Wurgler (2006) demonstrated that investor sentiment is correlated with the cross-sectional returns of various stock portfolios. Consequently, AI can extend forecasting capabilities beyond historical price data by incorporating signals derived from information flows, market expectations, and investor behavior.

B. Financial Indicators

Input variables can be categorized into four primary groups: (i) price and liquidity metrics, including price, returns, trading volume, and turnover value; (ii) technical indicators, such as moving averages, RSI, MACD, and historical volatility; (iii) fundamental indicators, encompassing revenue, profitability, EPS, ROE, ROA, cash flows, and valuation multiples; and (iv) macroeconomic variables, including GDP, inflation, interest rates, credit growth, exchange rates, and global financial conditions. The integration of these variable clusters enables the model to simultaneously capture market dynamics, corporate financial health, and the broader macroeconomic context.

C. Investor Sentiment

Investor sentiment can be encapsulated through news polarity, investor attention levels, trading behaviors, and event-driven market reactions. Baker and Wurgler (2006) established a correlation between sentiment and stock returns, whereas Tetlock (2007) indicated that media content can furnish predictive information regarding price volatility and liquidity. Therefore, sentiment need not be quantified by a singular composite index; AI methodologies can amalgamate multiple proxy variables and evaluate the predictive contribution of each individual feature.

D. Distinguishing Investor Sentiment from Attention

Attention reflects the intensity of investors' focus on specific information or topics, whereas sentiment captures the directional polarity (positive, negative, or neutral) of that focus. These two constructs are distinct; for instance, adverse news can still precipitate exceptionally high levels of attention. Consequently, data should be stratified into three layers: Attention → Sentiment → Behavior, thereby formulating the informational sequence: Information → Attention → Sentiment → Behavior → Price. This sequential approach prevents the erroneous conflation of all news or social media data directly with sentiment.

III. THE VIETNAMESE MARKET CONTEXT AND THE GROWING VOLUME OF DATA

A. Rapid Expansion in Market Scale

The year 2024 represents a significant milestone. According to the State Securities Commission, the combined equity market capitalization on the HOSE, HNX, and UPCoM exchanges reached 7,173.32 trillion VND by year-end, reflecting a 20.8% year-over-year increase and equating to approximately 62.3% of the GDP. The average equity trading value attained 20,985 billion VND per session. The number of securities accounts reached approximately 9.3 million, with an influx of roughly 2 million new retail investor accounts during the year.

In 2025, the market experienced even more robust growth. Market capitalization peaked at 9,975.36 trillion VND; the VN-Index surged by 40.9%; and the average trading value reached 29,202 billion VND per session.

The absolute increase in market capitalization of approximately 2,802 trillion VND within a single year is a remarkable figure. However, it is imperative to avoid

equating increased market capitalization with enhanced forecasting quality. A larger scale inherently generates a greater volume of data; yet, big data yields empirical value only when it is characterized by high quality, temporal relevance, and the capacity to accurately reflect actual market behaviors.

B. Growth in the Number of Investors

By the end of May 2026, the number of securities accounts culminated at 13.16 million, representing a nearly 11% increase relative to the end of 2025 and equivalent to approximately 13% of the national population.

By the end of June 2026, the aggregate equity market capitalization on HOSE, HNX, and UPCoM surpassed 10.57 quadrillion VND, an increase of 6% compared to the end of 2025, equating to 82.3% of the estimated 2025 GDP. The average trading value approached 29,300 billion VND per session.

Table 2. The Scale of the Vietnamese Stock Market at Key Milestones, 2024–2026

Time Point	Market Capitalization	Average Daily Trading Value	Securities Accounts	Notable Indicator
End-2024	VND 7,173.32 trillion	VND 20.985 trillion/session	~9.3 million	Market capitalization ~62.3% of GDP
End-2025	VND 9,975.36 trillion	VND 29.202 trillion/session	>11.6 million	Market capitalization ~77.6% of GDP
End-May 2026	—	—	13.16 million	~13% of the population
End-June 2026	>VND 10.57 quadrillion	~VND 29.3 trillion/session	—	~82.3% of 2025 GDP

Source: State Securities Commission of Vietnam; General Statistics Office.

The aforementioned statistics hold direct implications for AI research. As the number of accounts escalated from approximately 9.3 million to over 13 million within an 18-month span, the market has generated a substantially larger repository of behavioral data. While not all data is publicly accessible at the granular individual level, the aggregated manifestations of these behaviors have materialized through liquidity, cash flows, price volatility, and event-driven reactions.

The author posits that this phenomenon underscores the necessity for sentiment data to be evaluated with greater academic rigor in predictive studies of the Vietnamese market.

IV. MACROECONOMIC CONDITIONS AND FORECASTING CAPACITY

Stock market analysis cannot be conducted in isolation from the broader macroeconomic environment. Vietnam's GDP in 2025 recorded a growth rate of 8.02%. Regarding sectoral contributions, the service sector expanded by 8.62%, contributing 51.08% to the total increase in value added; the industrial and construction sector grew by 8.95%, contributing 43.62%. Notably, financial, banking, and insurance activities exhibited a growth rate of 7.82%.

These metrics illustrate that the real economy is expanding at a robust trajectory. However, the stock market inherently reflects forward-looking expectations.

Consequently, a forecasting model reliant solely on low-frequency GDP data faces intrinsic difficulties in capturing short-term market fluctuations.

AI can partially resolve this limitation through the integration of mixed-frequency variables. GDP data can serve as a macroeconomic baseline; data on interest rates, exchange rates, or credit growth can furnish higher-frequency signals; price and liquidity data offer daily granular information; and news data can be processed on a daily or even hourly basis.

Table 3. Selected Economic Indicators Relevant to the Forecasting Model

Indicator	Latest Data Used	Relevance to the Market
Vietnam GDP, 2025	+8.02%	Reflects the health of the economy
Q4 2025 GDP	+8.46%	Indicates the momentum of growth toward year-end
Industry and construction, 2025	+8.95%	Related to production and corporate profitability
Services, 2025	+8.62%	Reflects domestic demand
Finance, banking, and insurance, 2025	+7.82%	Directly related to the financial system
World Bank 2026 growth forecast	6.8%	Macroeconomic expectations
IMF 2026 growth forecast in the 2025 Article IV	5.6%	A more cautious scenario

Source: General Statistics Office, World Bank, and IMF.

The divergence among international forecasts constitutes another noteworthy data point. In its May 2026 update, the World Bank projected Vietnam's GDP growth at 6.8% for 2026, succeeding the approximate 8% expansion in 2025. Conversely, the IMF's 2025 Article IV Consultation forecasted a 5.6% growth rate for 2026, premised on the assumed impacts of trade barriers and the gradual phasing out of 2025 stimulus measures.

In the context of forecasting research, disparities among official projections should not merely be dismissed as the "margin of error of a single entity". Rather, they signify the coexistence of heterogeneous expectational states within the market. Investors may react to realized empirical data or to projections of future economic conditions. This juncture is precisely where sentiment and informational data can effectively complement traditional financial datasets.

V. CREDIT, FINANCIAL CONDITIONS, AND THE STOCK MARKET

Another critical variable warranting inclusion in the analytical framework is credit. Credit influences the broader economy through its capacity to finance production, consumption, and investment. Fluctuations in credit conditions can consequently alter the cost of capital and corporate profit expectations. These dynamics may subsequently be transmitted to the equity market.

In 2026, the State Bank of Vietnam and the Government have continued to closely monitor credit growth. According to data cited by the Vietnam Government Portal, as of July 29, 2026, the aggregate outstanding credit across the banking system exceeded 20.15 quadrillion VND, representing an 8.38% increase relative to the end of 2025. This constitutes a salient observable variable for constructing a forecasting dataset, provided that the temporal context is rigorously maintained, given that this represents mid-year rather than year-end data.

The Bank for International Settlements (BIS) provides a more comprehensive data framework regarding credit to the non-financial sector, differentiating total credit provided by all sectors from bank credit, and facilitating the monitoring of credit-to-GDP ratios. The BIS updates this dataset quarterly and employs standardized conceptual definitions to enable cross-country comparability.

Furthermore, the BIS highlights that private credit, the credit-to-GDP gap, and debt service ratios serve as valuable indicator clusters for assessing financial conditions and systemic risks.

Within the scope of this study, BIS data is utilized to establish international benchmarks and macroeconomic context, whereas data from the State Bank of Vietnam is prioritized for the granular analysis of domestic credit dynamics. This stratified methodological approach mitigates the risk of conflating datasets governed by disparate statistical compilation methods.

A practical AI model may incorporate the following financial variables:

- credit growth;
- policy interest rates;
- deposit interest rates;
- lending interest rates;
- exchange rates;
- GDP growth;
- CPI;
- industrial production growth;
- stock market indices;
- trading value.

When these variables are combined with sentiment measures, the model may distinguish between two otherwise similar states. For example, the same increase in the VN-Index may have different forecasting implications if, in one case, it is accompanied by strong credit growth, higher liquidity, and positive sentiment, whereas in another it coincides with declining liquidity and weakening sentiment.

VI. LIQUIDITY AND CAPITAL FLOWS: IMPORTANT BEHAVIORAL DATA

A prominent advantage of the Vietnamese stock market is the increasing abundance and granularity of trading data.

In 2025, the average equity trading value attained 29,202 billion VND per session, representing an approximate 39% increase compared to 2024. Nevertheless, aggregate liquidity is insufficient to fully decipher cash flow dynamics. An AI system must evaluate liquidity across multiple dimensions: whether liquidity expansions are positively or negatively correlated with price movements, whether capital is concentrated in large-cap equities or distributed across the broader market, whether liquidity spikes precede or succeed informational events, and whether they are correlated with foreign investor trading activities.

Empirical data from December 2025 indicates that foreign investor transactions aggregated to over 139,514

billion VND, constituting 12.83% of the total market trading value; within this month, foreign investors registered a net buying value of over 2,098 billion VND. Conversely, assessing the entirety of 2025 reveals that foreign investors maintained a net selling position of approximately 124,114 billion VND, comprising roughly 691,276 billion VND in purchasing value against 815,390 billion VND in selling value.

This serves as an invaluable exemplar for forecasting research. An isolated observation of December could lead investors to deduce that foreign capital inflows are structurally improving. However, a longitudinal analysis over the full calendar year contradicts this, exposing a substantial net outflow.

Table 4. *Selected Indicators of Foreign Investor Transactions in 2025*

Period	Trading Value	Share of Market	Net Capital Flow
September 2025	>VND 152.021 trillion	12.23%	Net selling: >VND 24.742 trillion
December 2025	>VND 139.514 trillion	12.83%	Net buying: >VND 2.098 trillion
Full year 2025	Purchases: VND 691.276 trillion; sales: VND 815.390 trillion	—	Net selling: VND 124.114 trillion

Source: *State Securities Commission of Vietnam.*

Another case is even more noteworthy. At the end of September 2025, the VN-Index stood at 1,661.7 points, up 29.02% year-on-year, while foreign investors recorded net selling of more than VND 24.742 trillion during the month.

From a research perspective, these data demonstrate why a model should not be built around a simple logic such as “foreign selling → market decline.” The relationship between the capital flows of a particular investor group and an aggregate market index may depend on numerous other factors. AI can address this issue by incorporating capital flows alongside:

- index trends;
- liquidity;
- market breadth;
- news sentiment;
- exchange-rate movements;
- macroeconomic data;
- sector-specific characteristics.

The fundamental efficacy of AI resides in its capacity to identify complex, multi-dimensional signal combinations that remain imperceptible through conventional univariate observational paradigms.

VII. MARKET CONCENTRATION AND ITS IMPLICATIONS FOR AI

Another frequently overlooked dimension is the concentration of market capitalization. According to the State Securities Commission, as of late December 2025, the Ho Chi Minh Stock Exchange (HOSE) hosted 681 listed and traded instruments, comprising 400 equities, 4 closed-end fund certificates, 18 ETF certificates, and 259 covered

warrants. Aggregate stock market capitalization on HOSE surpassed 8.31 quadrillion VND, representing approximately 95.1% of the total capitalization of listed equities market-wide and equivalent to 72.18% of the 2024 GDP. Fifty enterprises listed on HOSE held a market capitalization exceeding 1 billion USD, with 5 companies surpassing 10 billion USD.

Table 5. *Selected Structural Characteristics of HOSE at Year-End 2025*

Indicator	Value
Total number of securities on HOSE	681
Number of stocks	400
Closed-end fund certificates	4
ETF certificates	18
Covered warrants	259
HOSE equity market capitalization	>VND 8.31 quadrillion
Share of total listed-equity market capitalization	>95.1%
Companies with market capitalization >USD 1 billion	50
Companies with market capitalization >USD 10 billion	5

Source: *State Securities Commission of Vietnam, December 2025 data.*

This characteristic has two implications.

On one hand, the heavy concentration of liquidity and capitalization on HOSE renders benchmark indices like the VN-Index significantly sensitive to large-cap equities. Consequently, an AI index forecasting model must explicitly account for index weighting structures and the relative contributions of these bellwether constituents.

On the other hand, incorporating social media sentiment data risks introducing bias toward heavily discussed tickers. A small-cap stock experiencing a spike in discussion volume may generate a stronger sentiment signal than a mega-cap enterprise with lower public engagement. This discrepancy highlights the necessity of distinguishing "attention intensity" from "economic significance." A robust AI system should be able to answer two separate questions:

1. Which stocks or topics are attracting the greatest attention?
2. Is the topic sufficiently large or economically significant to affect the market index or a particular sector?

Without distinguishing between these two questions, the model may learn from the "noise" of the data rather than from information of genuine predictive value.

VIII. THE ROLE OF FTSE RUSSELL AND MARKET INTEGRATION

A pivotal structural catalyst influencing Vietnamese market data is the ongoing reclassification process. FTSE Russell confirmed the reclassification of Vietnam from a Frontier Market to a Secondary Emerging Market, with an effective implementation date of September 21, 2026. The August 2026 classification update published by FTSE Russell explicitly details this status modification alongside the associated list of eligible securities.

At the time of this study, this reclassification has not yet taken effect; its official implementation date according to FTSE Russell is September 21, 2026.

The World Bank projects that this market upgrade could catalyze additional international portfolio inflows, estimated at 3 to 5 billion USD during the initial years, with the potential to reach up to 25 billion USD by 2030. It should be emphasized that these figures represent prospective estimates by the World Bank rather than realized capital inflows.

For AI research, the reclassification constitutes a valuable event variable. A model can examine:

- sentiment before and after the announcement;
- liquidity before and after the announcement;
- foreign investor transactions;
- price movements of eligible stocks;
- changes in market breadth;
- changes in investor attention.

Table 6. *Selected Milestones in Market Development and Integration for Inclusion in an AI Dataset*

Milestone	Event	Data Significance
2024	Market capitalization reached VND 7,173.32 trillion	Expansion in market scale
2025	VN-Index increased by 40.9%	Substantial price volatility
2025	Market capitalization reached VND 9,975.36 trillion	Continued expansion in market scale
2025	Securities accounts >11.6 million	Expansion of the investor base
2026	13.16 million accounts at the end of May	Growth in behavioral data
2026	Market capitalization >VND 10.57 quadrillion at the end of June	Continued market expansion
September 21, 2026	FTSE Russell scheduled to apply Secondary Emerging Market classification	International event variable

Source: *State Securities Commission of Vietnam and FTSE Russell.*

Market reclassification carries broader structural implications: as the market attracts a larger footprint of international institutional investors, market microstructure and trading behavior are likely to evolve. Trading volume distributions, information elasticity, and the co-movement between foreign capital flows and large-cap equities may undergo fundamental shifts.

Consequently, AI models trained on pre-upgrade historical data should not assume that statistical relationships will remain stationary post-reclassification. This underscores the critical imperative of managing model drift and data distribution shifts.

IX. RESEARCH METHODOLOGY

This study adopts a qualitative methodology integrated with secondary data analysis. Data sources are structured into four group:

Group 1: Regulatory and trading data. The principal sources include the State Securities Commission of Vietnam, HOSE, and HNX. The data used comprise:

- VN-Index;
- HNX-Index;
- market capitalization;
- trading value;

- number of securities accounts;
- foreign investor transactions;
- number of securities;
- market size.

Group 2: Macroeconomic data

Sources include the General Statistics Office, the State Bank of Vietnam, the Government, the IMF, and the World Bank. The variables of interest include:

- GDP;
- CPI;
- credit;
- interest rates;
- exchange rates;
- industrial production;
- growth forecasts.

Group 3: International financial data

The BIS is used to cross-reference credit and financial-condition indicators within the international statistical framework. The BIS provides quarterly data on credit to the non-financial sector, including credit from all sources of financing as well as bank credit.

The IMF and World Bank are used to assess growth, trade, financial conditions, and external risks.

Group 4: Sentiment data

In empirical implementation, this group of data may include:

- financial news;
- corporate reports;
- policy announcements;
- analytical articles;
- investment forums;
- search data;
- publicly available social media engagement data.

Sentiment data must strictly adhere to point-in-time collection constraints. Information items are eligible for model ingestion only if they were publicly available prior to the designated forecast timestamp. Utilizing articles revised or updated post-trading without proper timestamp validation introduces look-ahead bias and future data leakage.

X. PROPOSED AI MODEL FRAMEWORK

Rather than relying on a single isolated algorithm, the proposed framework is architected across three distinct feature layers:

First: The Market Data Layer

This layer includes:

- price;
- returns;
- trading volume;
- trading value;

- volatility;
- market indices;
- market breadth;
- domestic and foreign capital flows.

This constitutes the foundational data layer.

Second: The Financial Data Layer

This layer includes:

- revenue growth;
- earnings;
- EPS;
- ROE;
- ROA;
- P/E;
- P/B;
- leverage;
- cash flows;
- credit growth;
- interest rates;
- exchange rates;
- GDP;
- CPI.

These variables help prevent the model from being driven entirely by short-term fluctuations.

Third: The Sentiment Data Layer

This layer includes:

- positive/negative news scores;
- news volume;
- number of mentions;
- search intensity;
- discussion intensity;
- abrupt changes in attention;
- investor responses to events.

Table 7. Proposed Data Framework for the AI System

Data Layer	Representative Variables	Appropriate Frequency	Objective
Market	Price, returns, trading volume	Daily/hourly	Trend identification
Liquidity	Trading value, breadth, capital flows	Daily	Identification of trend strength
Corporate	Revenue, earnings, EPS	Quarterly/annual	Assessment of fundamentals
Macroeconomic	GDP, CPI, credit, interest rates	Monthly/quarterly	Identification of economic conditions
News	Tone, topics	Daily/hourly	Measurement of new information

Attention	Searches, mentions	Daily/hourly	Measurement of investor interest
Sentiment	Positive/negative/neutral	Daily/hourly	Measurement of sentiment

Crucially, all feature layers must be strictly time-synchronized. A financial disclosure released after market close cannot be treated as information accessible to investors during that day's trading session.

XI. ALGORITHM SELECTION: COMPLEXITY IS NOT NECESSARILY REQUIRED

In artificial intelligence (AI) research within the financial sector, model selection should be predicated on data characteristics and research objectives rather than defaulting to complex algorithms. A suitable approach is to begin with interpretable models, subsequently increasing complexity as necessitated.

Random Forest is appropriate for tabular data with multiple input variables and allows for evaluating feature importance in forecasting outcomes. XGBoost exhibits an advantage when data encapsulate non-linear relationships and interactions among variables; for instance, the impact of liquidity may vary depending on market sentiment states. LSTM (Long Short-Term Memory) networks are suitable for time-series data due to their capacity to capture temporal dependencies. Transformers are more appropriate for large datasets, particularly textual data, owing to their effective contextual processing capabilities in sentiment analysis.

However, a complex model does not necessarily equate to superior results. With small, noisy datasets or inaccurately recorded timestamps, complex models can lead to overfitting, thereby diminishing out-of-sample forecasting performance. Therefore, studies should benchmark models ranging from simple to complex, evaluating them against identical criteria to select the most optimal model aligned with the data and research objectives.

XII. RESULTS OF THE SECONDARY DATA ANALYSIS

A. First Result: Market scale has generated a sufficiently large data foundation

Data from 2024–2026 indicate that the Vietnamese market has undergone rapid scale transformations. Market capitalization increased from VND 7,173.32 trillion at the end of 2024 to VND 9,975.36 trillion at the end of 2025, subsequently exceeding VND 10.57 quadrillion by the end of June 2026.

Average trading value increased from VND 20,985 billion/session to VND 29,202 billion/session in 2025, and reached approximately VND 29,300 billion/session in the first half of 2026.

The number of accounts grew from roughly 9.3 million at the end of 2024 to over 11.6 million by the end of 2025, and 13.16 million by late May 2026.

The author notes that these three indicators constitute a structural shift: a larger market, higher trading density, and an increased number of participants. These provide favorable conditions for developing AI systems based on behavioral data.

B. Second Result: Price does not provide comprehensive behavioral information

Foreign investor data in 2025 demonstrate a discrepancy between monthly and annual cash flows. December recorded a net buying value of over VND 2,098 billion, whereas the entire year saw a net selling value of VND 124,114 billion.

The case of September 2025 is even more pronounced. Foreign investors were net sellers of over VND 24,742 billion, yet the VN-Index remained 29.02% higher year-over-year.

This result does not imply that foreign cash flow is insignificant. Conversely, it highlights that a single variable struggles to explain the entire market dynamics. AI can address this by integrating cash flows with multiple other signals.

C. Third Result: Sentiment possesses a theoretical basis to serve as a supplementary variable

Baker and Wurgler (2006) indicated that sentiment can exert heterogeneous effects across stock groups, particularly for small-cap, highly volatile, difficult-to-value, or hard-to-arbitrage stocks.

Tetlock (2007) found a relationship between media pessimism and downward price pressure, fundamental reversals, and liquidity.

Consequently, the author posits that sentiment can be incorporated into models as a supplementary information source; however, there is insufficient evidence to consider sentiment the direct cause of all price fluctuations.

D. Fourth Result: Models must account for structural changes

FTSE Russell confirmed Vietnam's transition from a Frontier to a Secondary Emerging Market, effective September 21, 2026. This event could alter capital flow structures, target stock groups, and international investor behavior.

Thus, models trained on pre-upgrade data require re-validation post-upgrade.

This represents a critical issue in financial AI research: models might accurately learn from historical data but become obsolete once the market structure shifts.

XIII. DISCUSSION

The overall findings suggest that AI in finance does not necessarily have to focus on predicting a specific price level; rather, it can be directed toward identifying market regimes such as uptrend, downtrend, sideways, or high volatility. This approach is more congruent with the continuously changing nature of the market and allows for the assessment of state transition probabilities based on multiple information sources.

Within this context, sentiment should be viewed as a supplementary information source to financial indicators rather than an independent buy or sell signal. The value of sentiment needs empirical validation by comparing the predictive power of models solely utilizing price, liquidity, and financial indicators against models incorporating sentiment. Simultaneously, market attention should be considered as a distinct variable, since an increase in

attention often precedes the formation of an explicitly positive or negative attitude.

An issue even more paramount than algorithm selection is data quality and timeliness. Missing data, incorrect publication timestamps, post-event information updates, or future data leakage (look-ahead bias) can all distort results. Specifically, if data are timestamped prior to the market actually absorbing the information, models might achieve high test accuracy but hold no practical predictive value. Therefore, data processing and control procedures must take precedence over AI model selection.

XIV. IMPLICATIONS

For regulatory authorities, it is imperative to continue standardizing market data, information disclosure timelines, corporate tickers, and corporate events, while concurrently enhancing the connectivity and exploitability of historical data. This also establishes a foundation for developing AI tools that assist in monitoring anomalous fluctuations, irregular trading activities, information diffusion, and abnormal divergences among sentiment, liquidity, and cash flow.

For listed companies, information disclosure provides not only financial data but also impacts market attention, expectations, and sentiment. Consequently, information must be disclosed in a timely, clear, and consistent manner, ensuring it is machine-readable for quantitative processing.

For investors, AI is most suitable as a decision support system, aiding in information synthesis, trend detection, sentiment tracking, and risk alerting. AI-generated signals should be evaluated in conjunction with business performance, valuations, cash flows, industry risks, macroeconomic conditions, and policy changes, rather than functioning as a “black box” that dictates buy or sell decisions.

XV. LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

This study primarily utilizes secondary data and a synthesis methodology without conducting comprehensive empirical experiments on stock-specific micro-level data. Thus, it is not yet possible to directly quantify the improvement in predictive power upon integrating sentiment into the models.

Furthermore, Vietnamese textual data face challenges regarding context, domain-specific terminology, and annotation quality. Social media data may also be subject to sample selection bias, where a highly active user group generates massive interactions but fails to represent the broader investor population. More importantly, the correlation between sentiment and price does not imply causality; sentiment might forecast price movements, but it could also act as a lagging reaction to price changes. Therefore, empirical studies must strictly control temporal ordering and avoid data leakage.

Future research should focus on three primary trajectories. First, constructing a Vietnamese market sentiment dataset labeled with positive, negative, and neutral sentiments, attention levels, sectors, stock tickers, and publication timestamps. Second, conducting comparative experiments among models relying solely on price and

liquidity, models augmented with financial indicators, and models further integrated with sentiment, thereby determining the incremental predictive value of sentiment information. Third, validating models across distinct market regimes and during periods before and after significant structural market changes.

Notably, the FTSE Russell reclassification of the Vietnamese market could provide a natural quasi-experiment to examine shifts in sentiment, liquidity, capital flows, and the predictive power of AI models. This design could help assess whether a model constructed in a pre-shift market environment retains its efficacy once the market structure alters.

XVI. CONCLUSION

The Vietnamese stock market is forming an increasingly expansive and diversified data ecosystem. Surges in market capitalization, liquidity, investor volume, and information sources facilitate the transition of AI from a predominantly price-based forecasting approach to a multimodal data integration model.

Within this evolution, financial indicators such as price, liquidity, profitability, valuation, interest rates, and exchange rates remain fundamental. Market sentiment and attention can be considered supplementary information layers that reflect investor expectations and behaviors. Accordingly, this study proposes an analytical framework: Macroeconomics → Financial Indicators → Information → Attention → Sentiment → Trading Behavior → Price.

The overarching findings demonstrate that AI efficacy does not depend solely on employing complex algorithms, but primarily on data quality, recording timeliness, the capacity to control data leakage, and out-of-sample validation. Hence, efforts should focus on empirically verifying whether sentiment generates incremental predictive value compared to traditional financial variables.

Vietnam's reclassification by FTSE Russell to a Secondary Emerging Market, effective September 21, 2026, establishes an additional critical milestone for investigating shifts in capital flows, liquidity, investor sentiment, and AI model performance.

Consequently, the appropriate research trajectory is not to seek the most complex model, but rather to determine which information sources genuinely enhance AI's capability to understand and forecast the market more accurately.

REFERENCES

- [1] Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4), 1645–1680. <https://doi.org/10.1111/j.1540-6261.2006.00885.x>
- [2] Breiman, L. (2001). *Random forests*. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- [3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>
- [4] Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *The Journal of Finance*, 66(5), 1461–1499. <https://doi.org/10.1111/j.1540-6261.2011.01679.x>
- [5] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417. <https://doi.org/10.1111/j.1540-6261.1970.tb00518.x>

- [6] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [7] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35–65. <https://doi.org/10.1111/j.1540-6261.2010.01625.x>
- [8] Mintarya, L. N., Halim, J. N., Angie, C., Achmad, S., & Kurniawan, A. (2023). Machine learning approaches in stock market prediction: A systematic literature review. *Procedia Computer Science*, 216, 96–102. <https://doi.org/10.1016/j.procs.2022.12.115>
- [9] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), 1139–1168. <https://doi.org/10.1111/j.1540-6261.2007.01232.x>
- [10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, 30. arXiv:1706.03762. <https://doi.org/10.48550/arXiv.1706.03762>
- [11] *Bank for International Settlements*. (2015). *Introduction to BIS statistics*. *Bank for International Settlements*. <https://www.bis.org/publications/qr/2015q3/bisstats.htm>
- [12] International Monetary Fund. (2025). Vietnam: 2025 Article IV consultation—Press release; staff report; and statement by the Executive Director for Vietnam. *IMF Staff Country Reports*, 2025/283. <https://doi.org/10.5089/9798229026949.002.A001>
- [13] International Monetary Fund. (2025). IMF Executive Board concludes 2025 Article IV consultation with Vietnam. *International Monetary Fund*.
- [14] London Stock Exchange Group (FTSE Russell). (2026). FTSE Russell announces results of March 2026 semi-annual country classification review for equities and fixed income.
- [15] World Bank. (2026). Viet Nam's economy remains resilient, but sustained reforms are key to navigating heightened uncertainty. *World Bank*.
- [16] World Bank. (2026). A turning point for Viet Nam's capital markets. *World Bank*.
- [17] General Statistics Office. (2026). Socio-economic situation report for Q4 and 2025. *General Statistics Office*.
- [18] General Statistics Office. (2026). Press release on the socio-economic situation in Q4 and 2025.
- [19] State Securities Commission of Vietnam. (2025). Annual report 2024. *Ministry of Finance*.
- [20] State Securities Commission of Vietnam. (2025). Market information for September 2025. *Ministry of Finance*.
- [21] State Securities Commission of Vietnam. (2025). Market information for December 2025. *Ministry of Finance*.
- [22] State Securities Commission of Vietnam. (2025). 2025 stock market review and summary. *Ministry of Finance*.
- [23] State Securities Commission of Vietnam. (2026). Six-month review of direction and administration and implementation of tasks for the second half of 2026. *Ministry of Finance*.
- [24] Government Office. (2026). Credit outstanding rises by 8.38% as the State Bank of Vietnam works to improve capital access. *Government Portal*