Empirical Study on Stock Market Fluctuation Prediction Using NLP-Based Sentiment Analysis

NGUYEN MINH TUAN[1,a,*], LAM DUY QUY[1,b], NGUYEN HUYNH MINH LUAN[2,c]
PHAYUNG MEESAD[3,d], NGUYEN HONG SON[1,e]

[1]Department of Information Security, Faculty of Information Technology
Posts and Telecommunications Institute of Technology
11 Nguyen Dinh Chieu, Sai Gon ward, Ho Chi Minh city,
VIETNAM

[2]Department of Secondary Education, Faculty of High School Program
Rick Hansen Secondary School
1150 Dream Crest Road, Mississauga, ON, L5V 1N6,
CANADA

[3]Department of Information Technology and Management
King Mongkut's University of Technology North Bangkok
Pracharat 1 Road, Wongsawang, Bangsue, 10800, Bangkok,
THAILAND

[a]ORCID: 0000-0002-4035-1759
[b]ORCID: 0009-0008-7157-5238
[c]ORCID: 0009-0002-2448-2983
[d]ORCID: 0000-0002-3742-6457
[e]ORCID: 0009-0001-4466-7160

**Corresponding Author*

Abstract: This paper investigates the relationship between stock market fluctuations and natural language sentiment extracted from social media. Twitter data are analyzed to derive sentiment signals associated with major stocks, which are then aligned with historical price data collected from Yahoo Finance. The study focuses on three main tasks: (i) synchronizing sentiment information with stock price movements, (ii) analyzing the impact of sentiment fluctuations on individual stock volatility, and (iii) forecasting the Vietnamese stock market index (VN-Index) based on sentiment-driven price predictions. Machine learning and deep learning models, including Long Short-Term Memory (LSTM) networks and hybrid architectures such as CNN-BiLSTM with attention mechanisms, are employed to improve temporal variables between sentiment signals and market dynamics. Experimental results on data from 2021 to 2025 demonstrate a noticeable correlation between sentiment scores and stock price movements for companies such as Apple (AAPL) and Amazon (AMZN). The proposed models show competitive performance in predicting future stock trends and VN-Index movements in 2026 and 2027. The obtained results demonstrate that integrating sentiment analysis with time-series modeling provides valuable insights into market volatility and improves predictive accuracy. This work illustrates the potential of natural language-driven approaches as effective tools for stock market analysis and decision support.

Key-Words: LSTM, Stock market, Vader, Sentiment, Fluctuation market, VN-index.

Received: March 18, 2026. Revised: June 21, 2026. Accepted: July 23, 2026. Available online: November 18, 2026.

1 Introduction

One of the most important factors influencing sentiment expression on social platforms is natural language processing (NLP), which plays a crucial role in analyzing how media sentiment affects stock market behavior. In this context, techniques such as Matthew's Correlation Coefficient (MCC) and the Valence Aware Dictionary and Sentiment Reasoner (VADER) have been employed to examine price fluctuations and media influence, [1]. Moreover, sentiment feedback derived from Simple Syndication (RSS) feeds has been shown to positively impact stock market values through peer-based price prediction mechanisms, [2]. In addition, financial asset values are significantly influenced by sentiments expressed in news headlines, where time series analysis of price data contributes to forecasting market trends. Consequently, combining sentiment analysis of headlines with time series models has emerged as a novel and effective approach to

improving stock price prediction accuracy, [3], [4]. Although sentiment extracted from tweets can predict stock returns, it has been observed that volatility prediction remains challenging, [5]. Furthermore, sentiment data has attracted considerable attention for enhancing stock market decision-making, as models can project index movements using positive or negative sentiment strategies, [6]. To achieve this, a wide range of predictive models including ARIMA, SARIMA, SVM, Random Forest, GRU, RNN, and LSTM have been widely adopted, [7], [8], [9], [10]. These models, along with Logistic Regression, Boosted Regression Trees, Multilayer Perceptron Neural Networks, and Gradient Boosting Machines, have also been applied to forecast trends of major stocks such as AAPL, AMZN, and NFLX, [11], [12].

Additionally, sentiment analysis has been extensively applied to platforms like Twitter and StockTwits, where massive user-generated opinions are processed using machine learning techniques, [13], [14]. By incorporating environmental, social, and governance (ESG) data, machine learning enables deeper understanding of investor sentiment and its influence on stock valuation, [15], [16]. However, sentiment integration does not always guarantee favorable outcomes for individual investment decisions, [17], [18]. Through correlations between internal and external sentiments, factors such as political climate, financial policies, and historical stock data collectively influence market volatility, [19], [20], [21]. As a result, integrating social media sentiment into traditional stock prediction frameworks provides valuable insights for investors and financial institutions, [22], [23]. Moreover, advanced modeling approaches involving clustering techniques such as K-means and DBSCAN, combined with deep learning models like LSTM, CNN, and Multilayer Perceptrons optimized using MAE, MSE, and RMSE, have enhanced predictive flexibility, [24], [25], [26]. By constructing NLP-based text datasets and modifying classification methods such as Support Vector Machines, sentiment polarity has been shown to predict stock price movements several days ahead using Granger causality analysis, [27], [28], [29], [30]. Beyond finance, machine learning has demonstrated remarkable success in diverse domains including physics, applied mathematics, natural language translation, and epidemic prediction, [31].

Finally, machine learning models have been developed to analyze sentiment-driven targets across various applications, such as forecasting Vietnamese sentiment to enhance library service quality and evaluating customer intentions through natural language emotion analysis, [32]. In parallel, chatbot systems further bridge the gap between humans and machines, while computational approaches continue to support social analysis and prediction tasks, [33], [34].

In order to provide more insight into stock market price movements, this paper illustrates the association between sentiment collected from Twitter and stock volatility, [35], [36]. The stock market price is influenced by a variety of factors, and this study examines the connection between social media tweeting and the price of stocks. The VN index will be consulted in response to the historical data based on the prediction. The research is carried out in three steps: first, the stock market price is predicted; second, the relationship between sentiment on stocks and the stock market price is shown; and third, the effect of stock price fluctuations on the Vietnamese stock market. The prediction will be based on the stock market price derived from 2021 to 2022 that is compatible with the sentiment recovered from Twitter from 2021 to 2022. LSTM has been demonstrated to be effective in analysis using machine learning models. The Vietnamese stock database records the country's stock market prices from 2021 to 2022, and forecasts are made using the data.

The ability of Long Short-Term Memory (LSTM) networks to fully capture complex temporal patterns and heterogeneous sentiment signals is still limited, despite their widespread use and proven efficacy in modeling sequential financial data combined with sentiment information. Specifically, sentiment extracted from social media sites like Twitter frequently exhibits short-term fluctuations, is noisy, and depends on context, making it difficult for a single recurrent structure to fully capture. Consequently, a growing number of recent studies have investigated hybrid deep learning architectures to improve temporal modeling and feature representation in sentiment-driven stock prediction tasks.

When working with high-dimensional or multivariate inputs, Convolutional Neural Networks (CNNs), [37], [38], have demonstrated significant promise in identifying short-term and local patterns in sequential data. The resulting architecture can capture richer contextual information across time when paired with Bidirectional LSTM (BiLSTM) networks, which process temporal dependencies in both forward and backward directions. Additionally, the model can now focus on the most instructive temporal segments when making predictions thanks to the introduction of attention mechanisms that dynamically weight hidden states.

This study investigates the use of a CNN–BiLSTM–Attention (CNN–BiLSTM–AM) model as an advanced baseline for sentiment-driven stock price prediction, motivated by recent advancements in hybrid deep learning architectures. The CNN–BiLSTM–AM framework is used to investigate whether the combination of convolutional feature extraction, bidirectional temporal modeling,

and attention-based weighting can offer advantages beyond a traditional LSTM model, rather than assuming that higher model complexity always results in better performance. Short-term local patterns are captured by convolutional layers in this architecture, and both forward and backward temporal dependencies are modeled by the BiLSTM component. The interaction between past stock prices and sentiment signals obtained from Twitter data can then be more flexibly modeled by using the attention mechanism to highlight informative temporal representations (Table 3).

2 Methodology

This paper investigates the impact of natural language sentiment on stock prices and examines its relationship with fluctuations in the VN-Index. As illustrated in Figure 10, this section consists of two main components: data collection and model construction. In the original framework, textual information is represented using a lexicon-based sentiment analysis approach, in which each social media post is assigned a single polarity score. These sentiment scores are aggregated over time and combined with historical stock price features. The resulting feature sequence is then processed by a Long Short-Term Memory (LSTM) network to capture temporal dependencies and predict future stock movements. In the proposed framework, textual information is represented in terms of sentiment scores, which are extracted from Twitter data using a VADER sentiment analysis method. Each tweet is assigned a sentiment polarity score and daily sentiment values are calculated by aggregating individual scores collected within the same trading day. The sentiment features are then fused with historical stock price features (Open, High, Low, Close, and Volume) to generate the input sequence for the prediction models. The extracted feature set is then fed into deep learning architectures such as LSTM, CNN-LSTM, SimpleRNN-LSTM and CNN-BiLSTM-Attention networks to learn the temporal dependencies and predict the future movements of stock prices.

2.1 Data collection

This study illustrates how stock market prices fluctuate in relation to semantic language gathered from Twitter. The information taken from Twitter includes company names, stock names, daytime, and tweets. Day, open, high, low, last, adj close, volume, and stock name are all included in the stock price dataset procured from Kaggle.

Between 2021 and 2025, there were over 80,793 tweets, 6,300 tweets about the stock market price, and 1,288 tweets about the VN-index. On the same day, the sentiment is converted into a sentiment score and added to the mean function. To ensure consistency in forecasting, Sentiment Days and Stock Market Days will be interchanged. Additionally, the days on Twitter sentiments will be used in place of the days in the VN index.

2.2 Model construction

The Valence Aware Dictionary for Sentiment Reasoning (VADER) is utilized in the current work to assess sentiments and determine the new score for a single day because the sentiment dataset gathered from Twitter is updated at various times of the day. The VADER model, [39], is a rule-based and lexical database-based sentiment analysis tool designed specifically for social media sentiments. It uses a variety of tactics. A collection of lexical components classified as positive or negative according to the polarity of the sentiment they are linked with is known as a sentiment lexicon. In addition to the positivity and negativity scores, it shows how much a sensation is positive or bad. VADER is a vocabulary and rule-based sentiment analysis tool that focuses on sentiments posted on social media. The MIT License governs its full open-sourcing status. All of the lexicon's word valence ratings are added up to determine the compound score, which is then normalized to lie between the most extreme positive $(+1)$ and negative (-1) values after accounting for rule compliance. The most useful metric for a given statement is a unidimensional sentiment measure. In application, the sentiments are used below normalized by the function as Eq. (1):

$$f(x) = \frac{x}{\sqrt{x^2 + \alpha}}, \tag{1}$$

where α is the normalization constant (default value: 15) and x is sum of valence scores of constituent words.

The data collected for stock market prediction consists of Day, Open, High, Low, Last, Adj-Close, Volume, and Stock Name. The most popular program used for evaluation is Generative Adversarial Networks (GANs), [40], [41], [42] in predicting stock market price. The efficacy of utilizing GANs for stock price forecasting is still up for debate. However, these networks have shown promise in handling several challenging tasks, such as producing realistic images and videos and converting text to images. In the past, this kind of model was mostly employed to create new images, texts, or videos but not time series, particularly ones with as much volatility as stock prices. The issue is that because of market volatility and social influences on individual stock trends, stock prices are typically more difficult to predict. Data scientists have recently tested GAN models for stock price prediction, and some of the findings are encouraging. Additionally, the use of news and social media analysis for this purpose is

growing.

In this study, Long Short-Term Memory (LSTM), a widely used recurrent neural network (RNN) architecture in deep learning, is employed for sentiment-driven stock market prediction. It is excellent at identifying long-term dependencies, which makes sequence prediction jobs a perfect fit for it. Because LSTM has feedback connections, as opposed to standard neural networks, it can handle complete data sequences instead of single data points. Because of this, it is very good at identifying and forecasting patterns in sequential data, such as time series, text, and voice. Because it can extract meaningful insights from sequential data, LSTM has developed into a potent tool in deep learning and artificial intelligence that is facilitating advances across a wide range of domains.

Table 1 presents the architecture of the proposed CNN–LSTM model. The Conv1D and MaxPooling1D layers are used to extract local temporal features from the input sequence, while the LSTM layer captures long-term dependencies. A Dropout layer is employed to reduce overfitting, and a Dense layer with a single neuron produces the final prediction. The model contains a total of 33,345 trainable parameters.

Table 4 presents the architecture of the proposed CNN–BiLSTM–Attention model. The Conv1D layer is employed to extract local temporal features from the input sequences, while the Bidirectional LSTM layer is used to capture long-term temporal dependencies in both forward and backward directions. Dropout layers are incorporated to reduce overfitting during training. An attention mechanism is then applied to highlight informative temporal representations derived from the BiLSTM outputs. A Global Average Pooling layer is utilized to aggregate temporal features, followed by a fully connected Dense layer that performs feature transformation. Finally, a Dense layer with a single neuron produces the final prediction. The model contains 71,809 trainable parameters, ensuring sufficient capacity for sentiment-driven stock market forecasting tasks.

Table 5 presents the architecture of the hybrid SimpleRNN–LSTM model. The network begins with a SimpleRNN layer to capture short-term temporal dependencies, followed by stacked LSTM layers to model long-term sequential patterns. Dropout layers are employed to mitigate overfitting, and a fully connected Dense layer produces the final prediction. The model contains 49,697 trainable parameters, ensuring sufficient capacity for time-series forecasting tasks.

To facilitate a clearer comparison, the experimental results are first analyzed from a quantitative perspective. The performance of the proposed CNN–BiLSTM–Attention model is compared against the conventional LSTM baseline using multiple regression metrics that reflect both prediction accuracy and explanatory power shown in Table 6.

As shown in Figure 1 and Figure 2, the models present two deep learning designs for time-series prediction. Figure 1 follows a hybrid CNN–LSTM strategy, where convolution and pooling act as an automatic feature extractor that highlights local trends and reduces sequence length before the LSTM performs temporal modeling, often improving efficiency and robustness to noise. In contrast, Figure 2 relies purely on recurrent processing with a SimpleRNN + stacked LSTMs, aiming to learn temporal structure directly from the full-length sequence; this can be expressive but may require stronger regularization and careful tuning to avoid overfitting. Both architectures end with a single-neuron dense layer to generate one continuous output value.

In addition to the primary LSTM-based forecasting models, a GAN-based implementation was explored during the experimental phase to investigate its capability in learning stock price dynamics. The GAN experiment was not included in the final comparative evaluation because the primary objective of this study is to assess the forecasting performance of LSTM-based architectures. The corresponding generator and discriminator training losses are reported in Figure 3 to illustrate the convergence behavior of the adversarial learning process.

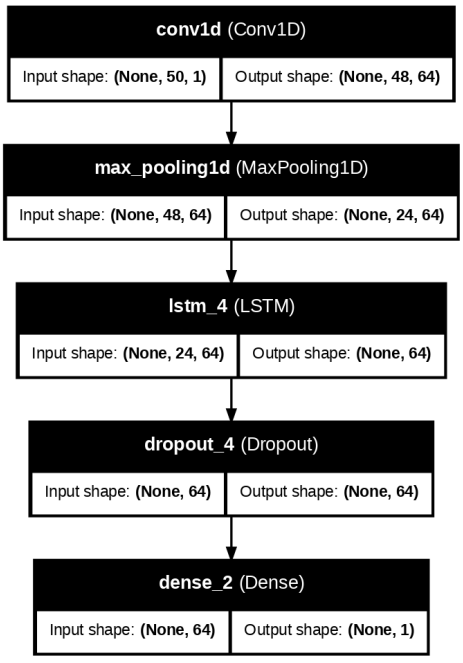


Fig. 1. CNN–LSTM forecasting architecture. Source: created by the authors.

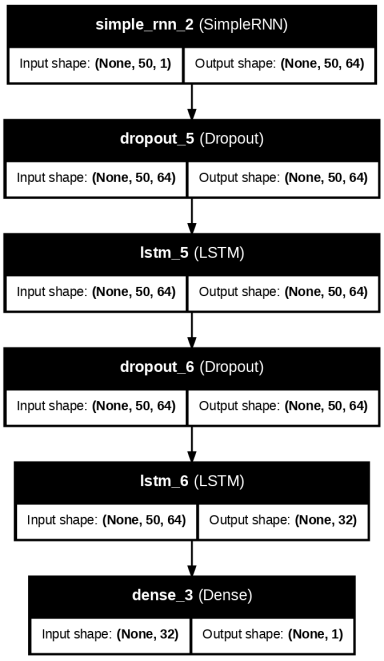


Fig. 2. RNN–LSTM stacked recurrent architecture. Source: created by the authors.

Table 1. Architecture of the CNN–LSTM Model. Source: created by the authors.

Layer	Output	Params.
Conv1D	(None, 48, 64)	256
MaxPooling1D	(None, 24, 64)	0
LSTM	(None, 64)	33,024
Dropout	(None, 64)	0
Dense	(None, 1)	65
Total		**33,345**
Trainable		33,345
Non-trainable		0

3 Numerical results

Based on the specific data, the metrics for the evaluation of prediction results are calculated by mean squared error (MSE). This function estimates the risk metric consistent with the expected value of the given error value or loss value. Given $\hat{y}_i$ is the predicted value of the $i$the sample, and y_i is the corresponding true value, the MSE is calculated over n samples N as Eq. (2).

$$MSE(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^{N-1} (y_i - \hat{y}_i)^2. \tag{2}$$

Eq. (3) assigned a score of 0 to negative feelings and a score of 1 to positive sentiments in the data. Sentiment measurements (Sc) were computed as the difference between the number of positive and negative posts in a given dataset as follows

$$Sc = \frac{M_{pt} - M_{nt}}{M_{pt} + M_{nt}}. \tag{3}$$

M_{nt} denotes the number of negative news articles on day t and M_{pt} is the number of positive news articles. The sentiment index has a range of values from -1 to 1. A negative tone in the news for that particular date is indicated if the sentiment index number is close to zero. On the other hand, if it gets close to 1, the news is generally positive. A min-max scaler was used before using the chosen features as the framework's input.

To quantitatively evaluate the performance and accuracy of the proposed model, several widely used error and goodness-of-fit metrics are employed, including the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). RMSE measures the square root of the average squared difference between the predicted and actual values, placing greater emphasis on larger errors, while MAE provides the average magnitude of the prediction errors without considering their direction. MAPE expresses the prediction error as a percentage, allowing for intuitive interpretation of relative accuracy. In addition, the R^2 metric evaluates the proportion of variance in the observed data explained by the model, thereby indicating the overall goodness of fit. Together, these metrics offer a comprehensive assessment of both prediction accuracy and model reliability shown as follows: Root Mean Squared Error (RMSE) is shown at Eq. (4):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}. \tag{4}$$

Mean Absolute Error (MAE)is shown at Eq. (5):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| . \tag{5}$$

Mean Absolute Percentage Error (MAPE %) is shown at Eq. (6):

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| . \tag{6}$$

Coefficient of Determination (R^2) is shown at Eq. (7):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}. \tag{7}$$

In this study, we apply the LSTM model for analyzing semantics and predicting stock market

prices. Figure 3 shows the evolution of the discriminator loss (top) and generator loss (bottom) over training epochs. Early in training, both losses fluctuate sharply, indicating the initial competition as the discriminator rapidly learns to distinguish real from generated samples while the generator adjusts to improve its outputs. After the early phase, the losses settle into relatively stable ranges—approximately 1.38–1.40 for the discriminator and 0.68–0.71 for the generator—suggesting the adversarial game reaches a near-equilibrium where neither network consistently dominates. The occasional spikes near later epochs reflect brief instability or shifts in one network temporarily outperforming the other, but the losses quickly return to their steady levels, indicating overall stable GAN training without sustained divergence or collapse. The values of close will be chosen for prediction. Figure 4 and Figure 5 illustrate the label general stock market price and VN index prediction based on the historical data using LSTM. Figure 4 illustrates the future forecasting results obtained from the trained model. The predicted trajectory reflects a coherent continuation of historical patterns while incorporating recent market dynamics. The figure highlights the model's capability to generalize learned temporal features and provide meaningful long-term projections. These results suggest that the proposed approach can serve as an effective analytical tool for anticipating potential future market movements. Figure 5 presents an overall visualization of the model's predictive behavior, emphasizing its ability to track market trends under varying conditions. The close alignment between predicted and observed values confirms the effectiveness of the model in learning nonlinear dependencies within the data. This figure further supports the reliability of the proposed framework for stock price analysis, while acknowledging that prediction accuracy may vary during highly volatile market phases.

Based on the history, Figure 4 and Figure 5 provide the prediction using the LSTM model including the AAPL, AMD, AMZN, BA, BX, COST, CRM, DIS, ENPH, F, GOOG, INTC, KO, META, MSFT, NFLX, NIO, NOC, PG, PYPL, TSLA, TSM, VZ, XPEV, ZS, and VN index under the sentiments to 2024 and 2025. The movement of AMZN is shown in Figure 6 under the changing of sentiment score has been combined with the mean function shown in Figure 7. Figure 6 illustrates the architecture of the proposed SimpleRNN–LSTM model. The network begins with a SimpleRNN layer that captures short-term temporal patterns from the input sequence. The extracted representations are subsequently processed by stacked LSTM layers to learn long-term dependencies and sequential relationships in the data. Dropout layers are incorporated to mitigate overfitting during training, while a fully connected Dense layer generates the final stock price prediction. This hybrid architecture combines the computational efficiency of SimpleRNN with the memory capability of LSTM, enabling effective modeling of complex financial time-series data. Figure 7 illustrates the architecture of the proposed CNN–LSTM model. The network first employs a one-dimensional convolutional layer (Conv1D) to extract local temporal features from the input sequence, followed by a MaxPooling layer to reduce dimensionality and suppress noise. The extracted feature maps are then processed by an LSTM layer to capture long-term temporal dependencies and sequential relationships within the stock market data. A Dropout layer is incorporated to improve generalization and reduce overfitting during training, while the final Dense layer generates the stock price prediction. By combining convolutional feature extraction with recurrent temporal modeling, the CNN–LSTM architecture is capable of learning both local patterns and long-term trends in financial time-series data.

Figure 8 and Figure 9 perform the estimation and fluctuation of the AAPL under the comments related to the AAPL label. Figure 8 actual versus predicted stock prices for AAPL generated by the proposed model over the considered time period. The figure shows that the predicted curve closely follows the trend of the observed data, indicating that the model is able to capture the underlying dynamics of stock price movements. Minor deviations are observed during periods of high volatility; however, the overall trend alignment suggests strong predictive capability. This result demonstrates the effectiveness of the model in learning temporal patterns and responding to changes in market behavior. Figure 9 presents the forecasted stock price trend for AMZN for the future period based on the trained model. The predicted trajectory reflects a smooth continuation of historical trends while accounting for recent market fluctuations. The figure highlights the model's ability to extrapolate meaningful future patterns from past data, providing insights into potential price movements. These projections serve as a valuable reference for analyzing long-term trends, although they should be interpreted with caution due to the inherent uncertainty of financial markets.

To improve reproducibility and provide a comprehensive description of the proposed forecasting framework, the experimental settings adopted in this study are summarized in this subsection. Historical stock prices and sentiment scores extracted from Twitter data were first normalized using the Min–Max scaling technique before being used as input features. The dataset was divided into training and testing subsets using an 80:20 ratio. A sliding window of 30 trading days was employed to construct the input sequences, enabling the models to capture both short-term and long-term

temporal dependencies. All deep learning models were implemented using TensorFlow/Keras. The Adam optimizer was adopted due to its effectiveness in training non-stationary time-series forecasting models. The initial learning rate was set to 0.001, the batch size was fixed at 32, and all models were trained for 100 epochs (Table 2). To ensure consistency and reproducibility, the random seed was fixed throughout all experiments. The stock price forecasting task was formulated as a regression problem. Therefore, the Mean Squared Error (MSE) was employed as the loss function shown at Eq. (8) and is defined as

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\,, \tag{8}$$

where y_i denotes the actual stock price, $\hat{y}_i$ represents the predicted stock price, and N is the total number of samples.

For the proposed CNN–BiLSTM–Attention model, the convolutional layer first extracts local temporal features from the fused sentiment and stock market inputs. The extracted representations are subsequently processed by a Bidirectional Long Short-Term Memory (BiLSTM) network to capture temporal dependencies in both forward and backward directions. Let h_t denote the hidden state generated by the BiLSTM layer at time step t. The attention score is calculated as shown as Eq. (9)

$$e_t = \mathbf{v}^T \tanh\left(\mathbf{W}_h h_t + \mathbf{b}_h\right), \tag{9}$$

where $\mathbf{W}_h$, $\mathbf{b}_h$, and $\mathbf{v}$ are trainable parameters.

The attention weights are obtained through the Softmax normalization function as shown at Eq. (10):

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T}\exp(e_k)}, \tag{10}$$

where T denotes the sequence length.

The resulting context vector is computed as shown as Eq. (11)

$$c = \sum_{t=1}^{T}\alpha_t h_t. \tag{11}$$

The context vector c is subsequently passed to the fully connected layers to generate the final stock price prediction. By assigning higher weights to informative temporal representations, the attention mechanism enables the model to focus on the most relevant sentiment-driven patterns and stock market dynamics. All experiments were conducted on a workstation equipped with an Intel Core i5 processor, 32 GB RAM, and an NVIDIA RTX-series GPU.

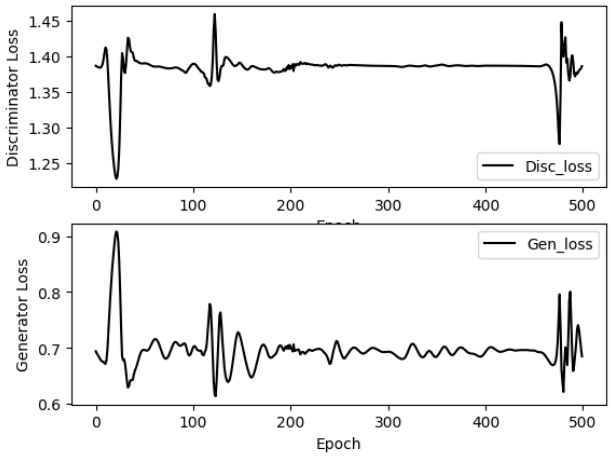


Fig. 3. Training loss curves of the Conv-LSTM models. Source: created by the authors.

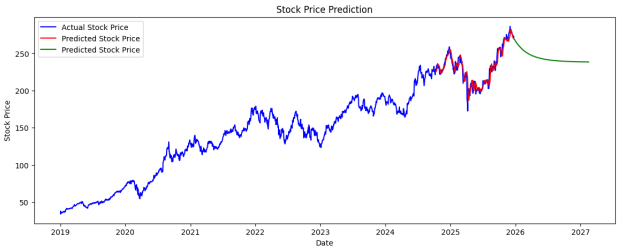


Fig. 4. General stock market price prediction under the sentiments. Source: created by the authors.

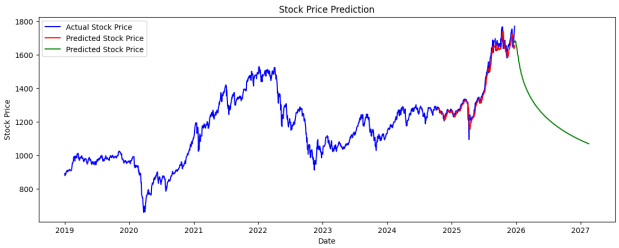


Fig. 5. VN index prediction affected by sentiments. Source: created by the authors.

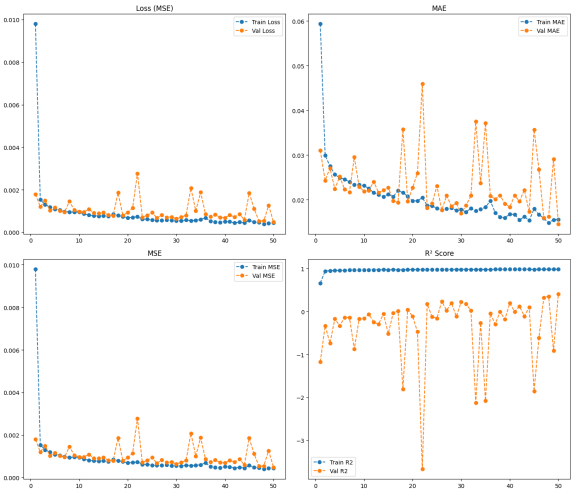


Fig. 6. Construction of the model RNN-LSTM. Source: created by the authors.

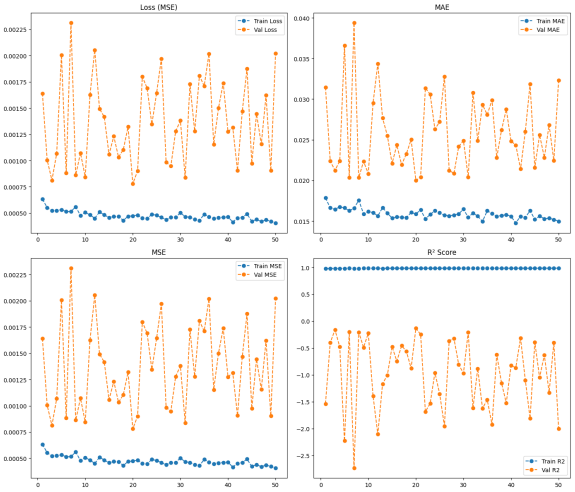


Fig. 7. Construction of the model CNN-LSTM. Source: created by the authors.

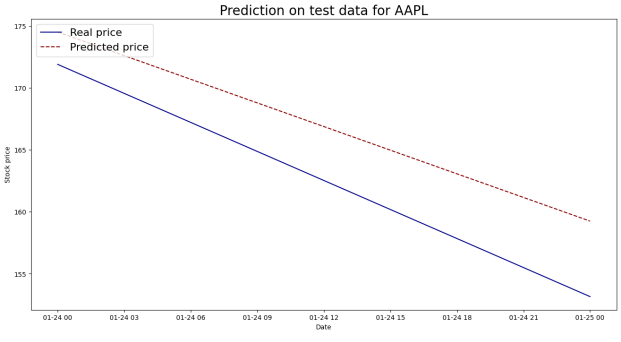


Fig. 8. Comparison between actual and predicted stock prices for AAPL. Source: created by the authors.

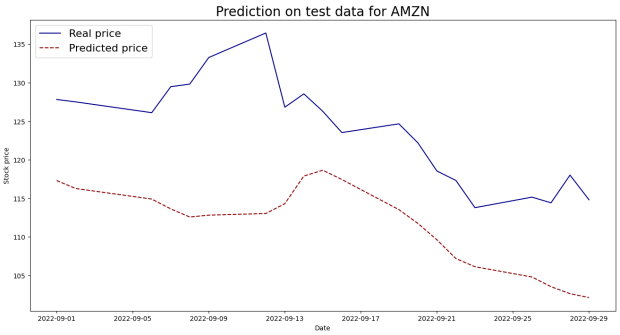


Fig. 9. Forecasted stock price trend for AMZN. Source: created by the authors.

Table 2. Experimental configuration used in this study. Source: created by the authors.

Parameter	Value
Train/Test Split	80% / 20%
Sequence Length	30
Optimizer	Adam
Learning Rate	0.001
Batch Size	32
Epochs	100
Loss Function	MSE
Random Seed	42
Feature Scaling	Min-Max

4 Limitation and Discussion

Machine learning has effectively captured stock market volatility by analyzing natural language data from Twitter alongside stock price movements. Sentiment data is collected each time relevant comments appear on Twitter, and these sentiments are aggregated into daily sentiment scores. While this approach can introduce potential biases, as sentiment from different times of the day is condensed into a single score, the overall trend remains useful for observing the influence of natural language on stock market behavior. The Long Short-Term Memory (LSTM) model, in particular, demonstrates strong predictive capability by analyzing historical data, providing valuable insights that can support informed investment decisions, [43]. While this study offers promising results in using sentiment analysis to predict stock market movements, several limitations must be acknowledged. First, relying on social media data, particularly Twitter, may introduce noise and biases, as not all posts provide reliable information regarding stock price changes. Additionally, the VADER sentiment model, primarily designed for English-language text, may not fully capture the complexities of Vietnamese sentiment or financial jargon. Moreover, the temporal alignment of sentiment scores and stock market data might not accurately reflect real-time market reactions,

as external factors such as political events or economic policies can impact prices independently of sentiment, [44],[45]. Finally, the LSTM model's predictions are based on historical data and may struggle with sudden market shifts, highlighting the need for further improvements to address these challenges. The primary objective of this study is to investigate the effectiveness of sentiment-enhanced deep learning models for stock market prediction rather than to introduce a new state-of-the-art forecasting architecture. Therefore, we selected widely adopted recurrent and hybrid architectures, including LSTM, CNN-LSTM, SimpleRNN-LSTM, and CNN-BiLSTM-Attention, as representative baselines for comparison. Nevertheless, we agree that Transformer-based approaches represent an important research direction.

5 Conclusion

Stock market prices play a crucial role in guiding investors' strategic decision-making. This study investigates the relationship between sentiment scores and stock price movements, with a particular focus on major equities such as AMZN and AAPL, as well as the VN index. By employing machine learning techniques, the results reveal a strong correlation between sentiment dynamics and stock price fluctuations. Notably, variations in sentiment are shown to closely align with changes in the VN index, enabling reliable forecasts of individual stock prices and index trends for the 2026-2027 period. The findings highlight the effectiveness of sentiment-driven machine learning models, especially the Long Short-Term Memory (LSTM) network, in capturing long-term dependencies in financial time-series data. Nevertheless, the results should be interpreted with caution due to limitations associated with the reliability of social media data and language-dependent constraints of sentiment analysis models. Future research may benefit from incorporating more diverse data sources and enhancing multilingual sentiment analysis techniques to further improve predictive accuracy and support more informed investment decisions.

Declarations

Authors' contributions

Nguyen Minh Tuan, Lam Duy Quy, Nguyen Huynh Minh Luan: includes conceptualization, data curation preparation, research, methodology composition, software, visualization, original draft writing, and validation. Phayung Meesad, Nguyen Hong Son: vision, writing-original version, checking writing and final editing

Availability of data and materials

We declare no conflict of interest. Data shown at [46].

Competing interests

We confirm no conflict of interest.

Funding

This work was supported by the Posts and Telecommunications Institute of Technology (PTIT). Number: 999/QĐ-HV

Ethics declaration

not applicable.

Acknowledgments

We sincerely appreciate the financial research support of the Posts and Telecommunications Institute of Technology in Ho Chi Minh City, Viet Nam, for providing the necessary resources, a conducive research environment that contributed significantly to the success of this study.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work, the author(s) utilized artificial intelligence (AI) technologies solely for the purpose of linguistic refinement and grammatical correction. Following the application of these tools, the author(s) comprehensively reviewed, verified, and edited the resulting text. The author(s) maintain full accountability and responsibility for the accuracy, integrity, and final content of the publication.

References:

[1] Abdelfattah, B. A., Darwish, S. M., & Elkaffas, S. M. (2024). Enhancing the Prediction of Stock Market Movement Using Neutrosophic-Logic-Based Sentiment Analysis. Journal of Theoretical and Applied Electronic Commerce Research, 19(1), 116–134. https://doi.org/10.3390/jtaer19010007

[2] Bharathi, S., Geetha, A., & Department of Computer Science and Engineering, B.S.Abdur Rahman University, Vandalur, Chennai-600 048, Tamil Nadu, India. (2017). Sentiment Analysis for Effective Stock Market Prediction. International Journal of Intelligent Engineering and Systems, 10(3), 146–154. https://doi.org/10.22266/ijies2017.0630.16

[3] Feng, L., Zheng, Y., Wang, X., Guo, C., & Xue, R. (2025). Global stock market forecasting: Insights from series and parallel combination of machine learning models. Pacific-Basin Finance Journal, 93, 102905. https://doi.org/10.1016/j.pacfin.2025.102905

[4] Hu, Z., Zhao, Y., & Khushi, M. (2021). A

survey of forex and stock price prediction using deep learning. Applied System Innovation, 4(1), 9. https://doi.org/10.3390/asi4010009

[5] Deveikyte, J., Geman, H., Piccari, C., & Provetti, A. (2022). A sentiment analysis approach to the prediction of market volatility. Frontiers in Artificial Intelligence, 5, 836809. https://doi.org/10.3389/frai.2022.836809

[6] Fazlija, B., & Harder, P. (2022). Using Financial News Sentiment for Stock Price Direction Prediction. Mathematics, 10(13), 2156. https://doi.org/10.3390/math10132156

[7] Nabipour, M., Nayyeri, P., Jabani, H., S, S., & Mosavi, A. (2020). Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. IEEE Access, 8, 150199–150212. https://doi.org/10.1109/access.2020.3015966

[8] John, A., & Latha, T. (2023). Stock market prediction based on deep hybrid RNN model and sentiment analysis. Automatika, 64(4), 981–995. https://doi.org/10.1080/00051144.2023.2217602

[9] Khedr, A. E., S.E.Salama, & Yaseen, N. (2017). Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis. International Journal of Intelligent Systems and Applications, 9(7), 22–30. https://doi.org/10.5815/ijisa.2017.07.03

[10] Joshi, K., H. N, B., & Rao, J. (2016). Stock Trend Prediction Using News Sentiment Analysis. International Journal of Computer Science and Information Technology, 8(3), 67–76. https://doi.org/10.5121/ijcsit.2016.8306

[11] Davidovic, M., & McCleary, J. (2025). News Sentiment and Stock Market Dynamics: A Machine Learning investigation. Journal of Risk and Financial Management, 18(8), 412. https://doi.org/10.3390/jrfm18080412

[12] Kolasani, S. V., & Assaf, R. (2020). Predicting Stock Movement Using Sentiment Analysis of Twitter Feed with Neural Networks. Journal of Data Analysis and Information Processing, 08(04), 309–319. https://doi.org/10.4236/jdaip.2020.84018

[13] Koukaras, P., Nousi, C., & Tjortjis, C. (2022). Stock Market Prediction Using Microblogging Sentiment Analysis and Machine Learning. Telecom, 3(2), 358–378. https://doi.org/10.3390/telecom3020019

[14] Kraaijeveld, O., & De Smedt, J. (2020). The predictive power of public Twitter sentiment for forecasting cryptocurrency prices. Journal of International Financial Markets Institutions and Money, 65, 101188. https://doi.org/10.1016/j.intfin.2020.101188

[15] Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., García, Á. L., Heredia, I., Malík, P., & Hluchý, L. (2019). Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. Artificial Intelligence Review, 52(1), 77–124. https://doi.org/10.1007/s10462-018-09679-z

[16] Lee, H., Kim, J. H., & Jung, H. S. (2024). Deep-learning-based stock market prediction incorporating ESG sentiment and technical indicators. Scientific Reports, 14(1), 10262. https://doi.org/10.1038/s41598-024-61106-2

[17] Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

[18] Shah, D., Isah, H., & Zulkernine, F. (2019). Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. International Journal of Financial Studies, 7(2), 26. https://doi.org/10.3390/ijfs7020026

[19] Maqbool, J., Aggarwal, P., Kaur, R., Mittal, A., & Ganaie, I. A. (2023). Stock Prediction by Integrating Sentiment Scores of Financial News and MLP-Regressor: A Machine Learning Approach. Procedia Computer Science, 218, 1067–1078. https://doi.org/10.1016/j.procs.2023.01.086

[20] Mittal, A., & Goel, A. (2011). Stock Prediction Using Twitter Sentiment Analysis. Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019). Stock Price Prediction Using News Sentiment Analysis. 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), 205–208. https://doi.org/10.1109/BigDataService.2019.00035

[21] Sarker, I. H. (2021). Deep Learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Computer Science, 2(6), 420. https://doi.org/10.1007/s42979-021-00815-1

[22] Ding, G., & Qin, L. (2019). Study on the prediction of stock price based on the associated network model of LSTM. International Journal of Machine Learning and Cybernetics, 11(6), 1307–1317. https://doi.org/10.1007/s13042-019-01041-1

[23] Chandrasekaran, R., Mehta, V., Valkunde,

T., & Moustakas, E. (2020). Topics, trends, and sentiments of tweets about the COVID-19 pandemic: Temporal Infoveillance Study. Journal of Medical Internet Research, 22(10), e22624. https://doi.org/10.2196/22624

[24] Parvatha, L. S., Naga Veera Tarun, D., Yeswanth, M., & Kiran, Jonnalagadda. S. (2023). Stock Market Prediction Using Sentiment Analysis and Incremental Clustering Approaches. 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), 888–893. https://doi.org/10.1109/ICACCS57279.2023.10112768

[25] Puh, K., & Bagic Babac, M. (2023). Predicting stock market using natural language processing. American Journal of Business, 38(2), 41–61. https://doi.org/10.1108/AJB-08-2022-0124

[26] Mohammad, A. R. (2021). Stock market prediction through sentiment analysis of Social-Media and financial stock data using machine learning. In uO Research (University of Ottawa). https://doi.org/10.20381/ruor-27045

[27] Saxena, A., Jain, A., Sharma, P., Singla, S., & Ticku, A. (2023). Sentiment Analysis of Stocks Based on News Headlines Using NLP. In V. E. Balas, R. C. Bansal, S. K. Mangipudi, & S. Dawn (Eds.), Proceedings of the International Conference on Artificial Intelligence Techniques for Electrical Engineering Systems (AITEES 2022) (Vol. 3, pp. 124–135). Atlantis Press International BV. https://doi.org/10.2991/978-94-6463-074-9_12

[28] Shi, Y., Zheng, Y., Guo, K., & Ren, X. (2021). Stock movement prediction with sentiment analysis based on deep learning networks. Concurrency and Computation: Practice and Experience, 33(6), e6076. https://doi.org/10.1002/cpe.6076

[29] Smailovic, J., Grcar, M., Lavrac, N., & Žnidarcic, M. (2013). Predictive Sentiment Analysis of Tweets: A Stock Market Application. In A. Holzinger & G. Pasi (Eds.), Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data (Vol. 7947, pp. 77–88). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-39146-0_8

[30] Xiao, Q., & Ihnaini, B. (2023). Stock trend prediction using sentiment analysis. PeerJ Computer Science, 9, 1293. https://doi.org/10.7717/peerj-cs.1293

[31] Tuan, N.M., Meesad, P. and Son. N. H. 2024. On a Stock Prediction Aligned to Natural Language Sentiments. In 2024 8th International Conference on Natural Language Processing and Information Retrieval (NLPIR 2024), December 13–15, 2024, Okayama, Japan. ACM, New York, NY, USA, Article 111, 5. https://doi.org/10.1145/3711542.3711597

[32] Tuan, N. M., Luan, N. H. M., Hien, N. T., Son, N. H., Thua, H. T., & Tinh, V. V. (2026). Metaheuristic feature optimization using GA, PSO, and HHO for AWVs data classification. An International Journal of Optimization and Control: Theories & Applications (IJOCTA), 16(4), 1467. https://doi.org/10.36922/IJOCTA026220101

[33] Tuan, N.M., Meesad, P. & Nguyen, H.H.C. English–Vietnamese Machine Translation Using Deep Learning for Chatbot Applications. SN COMPUT. SCI. 5, 5, 2024. https://doi.org/10.1007/s42979-023-02339-2

[34] Tuan, N. M., Duy, T. T., Meesad, P., Son, N. H., Hien, N. T., Thua, H. T., Tinh, V. V., Van, C. V., & Trang, N. L. N. (2026). Optimizing predictive accuracy in general medical exams using hybrid machine learning and metaheuristic optimization methods. An International Journal of Optimization and Control: Theories & Applications (IJOCTA), 16(3), 888–907. https://doi.org/10.36922/IJOCTA026020008

[35] Anupama B S, Rakshith D B, Rahul, K. M., & Navaneeth M. (2020). Real Time Twitter Sentiment Analysis using Natural Language Processing. International Journal of Engineering Research And, V9(07), IJERTV9IS070406. https://doi.org/10.17577/IJERTV9IS070406

[36] Jain, R., Rai, R. S., Jain, S., Ahluwalia, R., & Gupta, J. (2023). Real time sentiment analysis of natural language using multimedia input. Multimedia Tools and Applications, 82(26), 41021–41036. https://doi.org/10.1007/s11042-023-15213-3

[37] Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner. “Gradient-Based Learning Applied to Document Recognition.” Proceedings of the IEEE 86, no. 11 (1998): 2278–324. https://doi.org/10.1109/5.726791

[38] Kim, Yoon. “Convolutional Neural Networks for Sentence Classification.” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2014, 1746–51. https://doi.org/10.3115/v1/D14-1181.

[39] Pano, T., & Kashef, R. (2020). A Complete VADER-Based Sentiment

Analysis of Bitcoin (BTC) Tweets during the Era of COVID-19. Big Data and Cognitive Computing, 4(4), 33. https://doi.org/10.3390/bdcc4040033

[40] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative Adversarial Networks: An Overview. IEEE Signal Processing Magazine, 35(1), 53–65. https://doi.org/10.1109/msp.2017.2765202

[41] Imran, A. S., Yang, R., Kastrati, Z., Daudpota, S. M., & Shaikh, S. (2022). The impact of synthetic text generation for sentiment analysis using GAN based models. Egyptian Informatics Journal, 23(3), 547–557. https://doi.org/10.1016/j.eij.2022.05.006

[42] Yu, L., Zhang, W., Wang, J., & Yu, Y. (2017). SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1). https://doi.org/10.1609/aaai.v31i1.10804

[43] Shah, M. H., Wang, N., De Chun, H., Zhang, K., Ullah, I., Ullah, A., & Iqbal, K. (2023). Exploring the interwoven relationship: Property rights, financial freedom, government regulation, and stock market fluctuations in emerging economies - A novel system GMM perspective. Heliyon, 10(1), e23804. https://doi.org/10.1016/j.heliyon.2023.e23804

[44] Darshan, H. K., Shankar, A. R., Harish, B. S., & Kumar, K. H. M. (2019). Exploiting RLPI for sentiment analysis on movie reviews. Journal of Advances in Information Technology, 10(1), 14–19. https://doi.org/10.12720/jait.10.1.14-19

[45] Darwich, M., Noah, S. a. M., Omar, N., Osman, N. A., & Ahmad, I. S. (2020). Quantifying the natural sentiment strength of polar term senses using semantic gloss information and degree adverbs. Journal of Advances in Information Technology, 109–118. https://doi.org/10.12720/jait.11.3.109-118

[46] Equinxx, Stock Prediction GAN - Twitter Sentiment Analysis, Kaggle. [Online]. Available: https://www.kaggle.com/code/-equinxx/stock-prediction-gan-twitter-sentiment-analysis/input.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Nguyen Minh Tuan, Lam Duy Quy, Nguyen Huynh Minh Luan: Conceptualization, data curation, investigation, methodology, software, visualization, writing-original draft and writing-review and editing, validation, visualization, writing-original draft and writing-review and editing. Phayung Meesad, Nguyen Hong Son: writing-review and editing, validation, visualization, writing-review and editing

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This work was supported by the Posts and Telecommunications Institute of Technology (PTIT). Number: 999/QĐ-HV.

Conflicts of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0. https://creativecommons.org/licenses/by/4.0/deed.en_US

APPENDIX

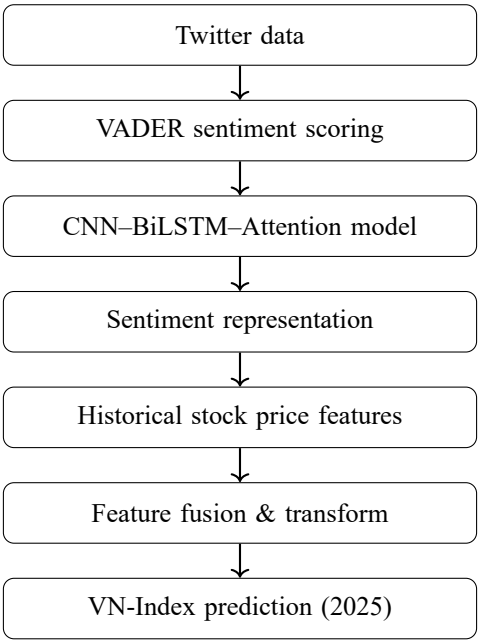


Fig. 10. CNN–BiLSTM–Attention based framework for stock market prediction. Source: created by the authors.

Table 3. Data description for sentiments prediction. Source: created by the authors.

Tweet	Stock Name	Company Name
Mainstream media has done an amazing job at be...	TSLA	Tesla, Inc.
Tesla delivery estimates are at around 364k fr...	TSLA	Tesla, Inc.
Even if I include 63.0M unvested RSUs as of...	TSLA	Tesla, Inc.
@RealDanODowd @WholeMarsBlog @Tesla...	TSLA	Tesla, Inc.
@RealDanODowd @Tesla Stop trying to kill kids,...	TSLA	Tesla, Inc.
A group of lawmakers led by Sen. Elizabeth War...	AMZN	Amazon.com, Inc.
$NIO just because I'm down money doesn't mean ...	AMZN	Amazon.com, Inc.
Today's drop in $SPX is a perfect example of w...	AMZN	Amazon.com, Inc.
Druckenmiller owned $CVNA this year nMunger b...	AMZN	Amazon.com, Inc.
Top 10 $QQQ Holdings n nAnd Credit Rating n n...	AMZN	Amazon.com, Inc.

Table 4. Architecture of the CNN–BiLSTM–Attention model. Source: created by the authors.

Layer	**Output**	**Params.**
Input Layer	(None, 30, 8)	0
Conv1D	(None, 30, 64)	1,600
Dropout	(None, 30, 64)	0
Bidirectional LSTM	(None, 30, 128)	66,048
Dropout	(None, 30, 128)	0
Attention	(None, 30, 128)	0
Global Average Pooling	(None, 128)	0
Dense	(None, 32)	4,128
Dense	(None, 1)	33
Total		**215,429**
Trainable		**71,809**
Non-trainable		**0**
Optimizer		**143,620**

Table 5. Architecture of the SimpleRNN–LSTM Hybrid Model. Source: created by the authors.

Layer	**Output**	**Params.**
SimpleRNN	(None, 50, 64)	4,224
Dropout	(None, 50, 64)	0
LSTM	(None, 50, 64)	33,024
Dropout	(None, 50, 64)	0
LSTM	(None, 32)	12,416
Dense	(None, 1)	33
Total	49,697	
Trainable	49,697	
Non-trainable	0	

Table 6. Performance comparison between LSTM and CNN–BiLSTM–Attention models. Source: created by the authors.

Metric	SimpleRNN LSTM Train	SimpleRNN LSTM Test	CNN BiLSTM–AM Train	CNN BiLSTM–AM Test
MSE	0.1638	0.1689	0.1812	0.1741
RMSE	0.04048	0.04110	4.5014e-04	4.5014e-03
MAE	0.0146	0.3235	0.0152	0.0131
MAPE (%)	0.0219	0.0318	4.7997e-03	4.7997e-04
R^2	0.9823	0.9852	0.9928	0.9917