Predicting Stock Market Performance Based on Sentiment Analysis of Online Comments

A Case Study of China's Highest Market Cap Stocks

Wenhao Suo[1]*, Tongjai Yampaka[2]*
Chakrabongse Bhuvanarth International Institute for Interdisciplinary Studies (CBIS),
Rajamangala University of Technology Tawan-OK, Bangkok, Thailand[1, 2]
School of Management, Guangzhou Huashang College, Guangzhou, China[1]

***Abstract*—In China's retail-focused stock market, the influence of social media sentiment during off-hours on the next day's opening price has received limited attention. This paper takes Kweichow Moutai—a leading Chinese company with substantial market capitalization—as the research sample. It gathers investor commentary data from financial platforms, and uses natural language processing tools (SnowNLP) to develop a multidimensional sentiment index (including average sentiment score, positive ratio, and sentiment volatility). By integrating this index with stock trading data and macroeconomic indicators, this study designs a dual-channel LSTM model: one channel for market technical features (e.g., price, volume) and the other for sentiment features, aiming to analyze the impact of off-hours sentiment on opening prices. Empirical results indicate that overnight sentiment has significant predictive power for the next day's opening price; meanwhile, sentiment transmission is asymmetric, making predictions more challenging in declining markets. Additionally, high-frequency sentiment data significantly outperforms low-frequency data in market prediction accuracy. This research expands the understanding of how investor sentiment influences the market over time, providing practical insights for market participants to develop effective strategies and manage risks.**

***Keywords*—Investor sentiment; non-trading hour sentiment; social media comments; dual-channel LSTM**

I. Introduction

In recent years, China's stock market, the world's second-largest, has surpassed one trillion yuan in combined market capitalization for the Shanghai and Shenzhen stock exchanges. Structurally, the Main Board, ChiNext, and the STAR Market form a multi-tiered capital market system, providing businesses with a range of financing options while also better serving investors' increasingly diverse asset allocation needs. However, compared to mature markets, the Chinese stock market still has unique structural features. A trading environment primarily driven by retail investors yields high volatility, high turnover, and a sentiment-driven market. Individual investors significantly influence both the number of accounts and trading volume, thereby affecting market pricing efficiency and volatility. Given these features, investor sentiment—a hallmark of irrational behavior—is increasingly seen as a significant factor driving market volatility.

In recent years, the widespread use of social media has created a new data source for researching investor sentiment. Platforms such as Weibo, stock forums, and Xueqiu have become key channels for investors to gather information, share opinions, and interact, producing a vast amount of user-generated content (UGC) closely linked to financial markets. This unstructured text not only captures investors' responses to company developments, policy changes, and macroeconomic trends in real-time but also offers high frequency and broad coverage, making it a valuable resource for sentiment measurement studies. While much research in traditional finance has examined how sentiment influences the market from a behavioral finance perspective, Aggarwal (2022) notes that, despite over fifty years of investigating market sentiment, a lack of standardized definitions and measurement tools remains, leading to significant variation in empirical findings. Therefore, it is essential to revisit the use of sentiment variables in finance within a psychological framework to enhance modeling accuracy and predictive capabilities.

In the context of social media, Bollen, Mao, and Zeng (2011) [1] first attempted to construct sentiment time series using Twitter content. They discovered that sentiment dimensions such as "calmness" significantly enhanced the forecast of the Dow Jones Industrial Average (DJIA), confirming the forward-looking influence of macro-level collective sentiment on financial markets. Yang and Mo (2016) [2] also showed that extreme news or social media sentiment can act as leading indicators of market returns and volatility, with even greater predictive power when sentiment varies significantly.

The real-time and scalable nature of social media data makes it a valuable tool for financial research. Corea (2016) [3] found in their study of the technology industry that behavioral variables such as tweet volume and interaction frequency are more predictive of stock prices than tweet sentiment polarity. Li et al. (2017) [4] proposed the SMeDA-SA method to extract sentiment signals from massive tweets, achieving highly accurate predictions of individual stock fluctuations and further validating the feasibility of social platforms as proxies for investor sentiment. Liu et al. (2015) [5] also explored the role of social media indicators in predicting stock co-movement, finding that companies with active Twitter accounts tend to exhibit more consistent stock price movements, revealing the power of social media to shape industry structural sentiment.

*Corresponding authors.

Advances in sentiment recognition technology have also expanded its applications in the financial sector. Man et al. (2019) [6] systematically reviewed the development of financial sentiment analysis (FSA), from early dictionary methods and traditional machine learning to deep learning models such as LSTM and Transformer, providing a diverse range of tools for financial sentiment modeling. Wang et al. (2022) [7] developed a market-neutral strategy based on LSTM and experimentally confirmed that the model exhibits high accuracy and stability in extracting sentiment from individual stocks. Koukaras, Nousi, and Tjortjis (2022) [8] utilized the XGBoost method, while Kumar et al. (2022) [9] employed a fuzzy neural network (FNN) approach. Both methods combined tweet sentiment with historical market trends to achieve effective predictions of the Dow Jones Industrial Average and the stock prices of several companies. In addition, Hasselgren et al. (2023) [10] proposed visualizing the sentiment scoring system, providing a feasible tool prototype for investment decisions.

However, while existing research has produced valuable insights in social sentiment modeling and market forecasting, it has mainly concentrated on the real-time relationship between sentiment and market variables during trading hours. It has paid little attention to the mechanisms that influence sentiment beyond trading hours. Traditional financial theory posits that prices are primarily determined during trading hours through the order book, with information outside of trading hours being indirectly reflected through overnight orders. Nonetheless, social media platforms often generate a large volume of emotional expression between the market close and the next day's opening, especially during emergencies, policy announcements, or the release of financial reports. These sentiments can potentially shift investor expectations in advance, affecting the opening price on the following day.

In response to this research gap, some scholars have begun to explore this area: Ranco et al. (2015) [11] found, based on the event study method, that there is a significant correlation between Twitter sentiment peaks during non-trading hours and the next day's cumulative abnormal returns; Fraiberger et al. (2021) [12] pointed out that international news released at night significantly affects cross-border capital flows and global stock price returns, highlighting the importance of non-trading period information in the international market [12]; Goutte et al. (2023) [13] also confirmed, in the context of ESG investment, that news sentiment indicators during non-trading hours show strong robustness in market return forecasts.

In summary, while current research on investor sentiment has made some progress, a significant gap remains in the systematic theoretical development and empirical testing of how sentiment during non-trading hours influences market opening behavior through expectation mechanisms. Therefore, this paper, drawing on the intersection of behavioral finance and natural language processing, focuses on how non-trading hours social media sentiment influences stock opening prices. To systemically fill this gap, this study first clarifies its core research question: Does social media sentiment during non-trading hours exhibit significant predictive power for the next day's stock opening price, and what are the heterogeneous characteristics of this predictive effect—specifically, the asymmetric transmission across different market conditions (rising vs. falling markets) and the performance differences under varying data frequencies (hourly vs. daily)?

Against this question, the study sets three research objectives: first, to construct a multidimensional non-trading hour sentiment index using online investor comment data, incorporating indicators such as average sentiment score, positive sentiment ratio, and sentiment volatility to capture the temporal dynamics of investor sentiment; second, to verify the causal relationship between non-trading hour sentiment and stock opening price via Granger causality tests, and quantify the magnitude of sentiment's influence on opening price fluctuations; third, to explore the comparative advantages of high-frequency sentiment data in improving prediction accuracy, and analyze the mechanism underlying the increased prediction difficulty of sentiment signals in declining markets.

Based on the above question and objectives, the main contributions of this study are reflected in three aspects: in the theoretical aspect, it expands the temporal boundary of behavioral finance research on sentiment-price relationships—unlike traditional studies that focus on trading-hour sentiment, this paper systematically validates the leading predictive role of non-trading hour sentiment in driving opening price gaps, enriching the cross-temporal theoretical framework of "sentiment-expectation-price" transmission; in the methodological aspect, it proposes a dual-channel LSTM modeling framework that decouples market technical features (e.g., price, volume, MACD) and sentiment features into independent processing channels, avoiding "biased fitting" caused by mixed data types (structured market data vs. unstructured text-derived sentiment data) and enhancing the interpretability of sentiment's predictive role; in the empirical aspect, it provides evidence for the sentiment-driven characteristics of emerging markets—taking China's retail-dominated stock market as the research context, this paper confirms that non-trading hour sentiment exerts a more pronounced impact on opening prices compared to mature markets, offering a behavioral explanation for typical emerging market traits such as high volatility and high turnover.

By creating an overnight sentiment index and integrating it with market variables (e.g., the next day's opening price), this paper will systematically assess the role and stability of sentiment signals in predicting opening prices, utilizing a combination of text sentiment analysis, deep learning, and financial econometric models. This effort aims to offer new theoretical insights and empirical evidence for understanding the behavioral finance features and microstructure of the Chinese stock market. Around the above research content, the remainder of this paper is structured as follows: Section II reviews the literature on investor sentiment, social media sentiment analysis, and LSTM-based stock prediction, laying the theoretical foundation for this study; Section III details the research design, including data sources (stock trading data and financial platform comment data), data preprocessing methods (text cleaning, sentiment classification via SnowNLP), and the architecture design of the dual-channel LSTM model; Section IV presents empirical results, including spatial-temporal distribution analysis of comment data, Granger causality test outcomes, model performance evaluation (using RMSE, MAE, and R^2), and robustness tests across market conditions and data

frequencies; Section V discusses the practical implications of the findings for market participants (investors, financial institutions, regulators) and identifies research limitations; Section VI concludes the core findings and proposes directions for future research.

II. Literature Review

With the development of behavioral finance, sentiment has become an increasingly integral part of financial market analysis frameworks. Investors' irrational decisions, particularly emotional swings during non-trading hours, are recognized as significant factors that influence next-day stock market performance. This paper, based on research into social media comments during non-trading hours and incorporating deep learning prediction models, further broadens the scope of research in this field.

A. Investor Sentiment and Financial Market Dynamics

Investor sentiment, a key behavioral factor influencing market volatility, has been consistently shown to have predictive effects on asset prices, trading decisions, and market fluctuations. A study based on the Pakistani market found a significant link between investor sentiment and investment decisions, with investor experience influencing this relationship [14].

In the context of the Chinese market, research also indicates that irrational emotions have a positive influence on corporate investment redundancy, with non-state-owned enterprises being more significantly impacted by emotions [15].

B. Financial Applications of Social Media Sentiment Analysis

The large volume of textual information on social media offers a valuable data source for developing sentiment indicators. A study of China's Weibo platform revealed that multidimensional online emotions, encompassing happiness, anger, and sadness, can predict fluctuations in the stock market. K-means clustering was employed to investigate the causal relationship between these emotions [16]. This supports the idea of exploring the predictive power of sentiment data during non-trading hours for predicting the next day's market behavior.

Additionally, by analyzing Twitter data with a convolutional neural network, researchers successfully identified the public's satisfaction and responsiveness to the macro environment on social media [17], demonstrating that emotional semantics can reflect collective market expectations.

C. Predictive Power of Sentiment Analysis Combined with an LSTM Model

The introduction of deep learning technology has opened up new possibilities for financial time series modeling. Research has shown that including sentiment scores as input variables in an RNN-LSTM model significantly enhances stock price prediction accuracy, resulting in lower error metrics compared to methods such as SVR and random forests [18].

Additionally, a hybrid model that uses a dual-channel LSTM, combined with text sentiment and stock price data, has also been shown to perform well in predicting Google's stock price. In particular, the RMSE of the model decreased significantly after including sentiment variables [9].

D. The Timeliness Advantage of High-Frequency Sentiment Data

The immediate feedback aspect of sentiment provides high-frequency data with a significant advantage in forecasting. Research utilizing the Google Search Index has identified a bidirectional Granger causal relationship between search data and market volatility, demonstrating a strong forecasting ability during periods of high market volatility [19]. This supports the empirical findings of this paper, which show that hourly sentiment data can substantially improve forecasting accuracy.

E. Asymmetric Transmission Mechanism of Market Sentiment

Regarding the asymmetric transmission of market sentiment, research indicates that negative news has a more substantial impact on predicting price trends than positive news, resulting in significantly higher prediction accuracy and returns [20]. Additionally, sentiment tends to be more volatile during stock market declines, which leads to higher model prediction errors [21]. This explains the higher RMSE observed in declining markets in this study.

III. Research Design and Methodology

A. Data Source and Processing

The data used in this research consists of two main categories: stock market trading data and comment data from financial social media platforms. These two types of data were gathered from reputable data providers and popular social media platforms, respectively. After cleaning and preprocessing, a data system was built to support sentiment analysis and predictive modeling.

1) Stock trading data sources and indicators: This study collected historical trading data for China's most extensive stocks by market capitalization, including Kweichow Moutai (stock code: 600519), from a specified financial data provider. The data covers multiple years to ensure its temporal accuracy and representativeness. Key trading metrics recorded include the opening price (Open), closing price (Close), high price (High), low price (Low), trading volume (Volume), price-to-earnings ratio (P/E), price-to-book ratio (P/B), and turnover rate (Turnover).

On this basis, we further developed various technical analysis indicators to improve our ability to illustrate market behavior. These include: Moving Averages (MAs), which use simple moving averages with window periods of 1 to 4 days (MA1-MA4) to show short- and medium-term price trends. The Moving Average Convergence Divergence (MACD), which consists of the DIF (fast line), DEA (slow line), and a histogram, is used to identify trend direction, strength, and inflection points. Volume-related indicators include: AMO (Amount of Trading Volume) (Volume × Closing Price), which measures trading activity; and AMOW (Weighted Amount of Trading Volume), which accurately reflects capital flows in conjunction with price changes.

All trading data is cleaned after import, including filling missing values, removing outliers, and formatting to ensure data consistency and analytical accuracy. Data within trading hours is fully recorded, while prices and trading indicators during non-

trading hours are uniformly set to null values (NA) to facilitate the distinction between time periods in subsequent modeling.

2) Collection of comment data and content characteristics: Comment data mainly comes from well-known domestic financial platforms, such as Sina Finance Stock Forum and East Money Stock Forum. These platforms gather real-time discussions from many small and medium-sized investors, serving as key sources for gauging market sentiment. This study chose Kweichow Moutai for sentiment analysis. The comment data covers the period from June to December 2024, totaling approximately 60,000 comments. The data was automatically collected using a legal and compliant data scraper. It includes comment text and posting time, accurate to the hour.

To improve the accuracy of time series analysis, comment data are classified and stored by daily frequency, and hourly timestamps are kept to effectively distinguish and compare trading hours (9:30–11:30, 13:00–15:00) from non-trading hours.

To distinguish data scope between trading and non-trading hours, Table I defines the two periods and clarifies the availability of sentiment and market indicators (e.g., non-trading hours only include sentiment indicators).

TABLE I. EXPLANATION OF TRADING AND NON-TRADING PERIODS

Time Type	Whether to trade	Explanation
Trading Hours	Yes	Contains price, volume, technical indicators, and sentiment indicators
Non-trading hours	No	Only sentiment indicators (such as favorable ratio, volatility, etc.) are included, and price data is recorded as "NA"

B. Comment Data Preprocessing

To ensure the accuracy and robustness of subsequent analysis, this study systematically cleans, preprocesses, and structures both the raw review data and market transaction data. This process involves multiple steps, including text cleaning, Chinese sentiment analysis modeling, data alignment, and the construction of derived sentiment indicators.

1) Chinese word segmentation and text normalization: Prior to the implementation of sentiment analysis, a corpus comprising 60,000 original comments was subjected to multi-tiered text cleaning and structural standardization procedures to ensure data quality and validity. Specifically, the preprocessing phase of the comment data encompassed three core steps, as delineated below: First, in the text cleaning stage, extraneous and non-informative elements within the comments—including but not limited to emoticons, uniform resource locators (URLs), punctuation marks, hypertext markup language (HTML) tags, and promotional/advertising content—were systematically eliminated to mitigate noise interference. Second, SnowNLP, a dedicated Chinese natural language processing (NLP) toolkit, was employed to perform word segmentation (tokenization) and part-of-speech (POS) tagging on the cleaned text, thereby converting unstructured textual data into semi-structured linguistic units. Third, leveraging the results of POS tagging, a targeted stop word filtering process was conducted: financial domain-specific nouns, verbs, and other semantically critical keywords were retained, while functionally redundant words (e.g., auxiliary words, conjunctions) that contributed minimally to sentiment orientation were excluded. This sequential preprocessing workflow effectively transformed the raw comment data into structured text features, laying a robust and reliable data foundation for subsequent tasks in the sentiment analysis pipeline, such as sentiment label generation, sentiment index construction, and cross-temporal sentiment prediction.

2) Sentiment classification model and evaluation: After segmenting the words, SnowNLP was used to assign sentiment scores to the reviews (from 0 to 1). Five hundred reviews were randomly selected from the original set and manually labeled as positive, negative, or neutral to create a validation set.

Sentiment scoring criteria: scores above 0.6 are positive; scores below 0.4 are negative; all others are neutral.

Table II shows the performance of the SnowNLP sentiment classification model against manual annotations:

TABLE II. STATISTICAL RESULTS OF MANUAL ANNOTATION

Actual / Forecast	Positive	Negative	Neutral	Total
Positive	51	21	2	74
Negative	19	271	10	300
Neutral	3	40	79	122

The performance metrics for the model are listed below:

- Overall accuracy: 80.85%.
- Weighted average F1 score: 81.33%.
- Macro average F1 score: 76.45%.

The Negative category performed the best, with an F1 score of 85.74%. The Positive and Neutral categories still have potential for improvement, especially the Neutral category, which has low recall. Further enhancements can be achieved by optimizing the training set, adjusting the classification threshold, or utilizing deep learning models such as BERT and ERNIE.

3) Market data cleansing and time alignment: To ensure the accuracy and validity of market-related data for subsequent analyses, the processing of market transaction data primarily focuses on two core tasks: missing value imputation and timestamp alignment. In terms of missing value handling, data gaps are inherently present on non-trading days (e.g., public holidays); for missing values (denoted as NaN) in technical indicators such as Moving Average (MA) and Moving Average Convergence Divergence (MACD), the forward fill (ffill) method is employed to maintain temporal continuity of the indicator series. Additionally, outliers identified in key transaction metrics—including trading volume and asset prices—are either removed or subjected to smoothing techniques to eliminate anomalous interference. For timestamp alignment and data labeling, temporal synchronization between

comment data and market transaction data is a prerequisite for their joint modeling. The comment data retains hourly timestamps and is precisely aligned with the official market trading hours (i.e., 09:30–11:30 and 13:00–15:00). In contrast, market price data recorded outside these trading hours (encompassing nighttime and midday intermissions) is uniformly labeled as "NA" to avoid erroneous usage in time-sensitive analyses. Notably, comment data is collected across all time periods (including non-trading hours), with this comprehensive collection design primarily intended to facilitate the analysis of how sentiment expressed during non-trading hours influences subsequent trading-day market dynamics.

4) Sentiment indicator construction: Based on the sentiment classification results, the following structured sentiment indicators were created to aid in subsequent modeling and prediction:

TABLE III. EXPLANATION OF SENTIMENT INDICATORS

Indicator	Explanation
Average sentiment score	The average sentiment score of all comments within a specific time window reflects the overall market sentiment.
Positive Ratio	The proportion of positive sentiment comments to all comments is used to measure market optimism.
Emotion Intensity	The difference in the number of positive and negative sentiment reviews or the difference in average scores indicates the degree of sentiment polarization.
Emotion Volatility	The standard deviation of sentiment scores within a specific window measures the instability of market sentiment.
Non-trading hours sentiment bias	The difference between sentiment indicators during non-trading hours and during trading hours is used to study the leading effect of pre-market/post-market sentiment.

These sentiment indicators, together with technical indicators, act as input features for regression modeling and causal analysis of sentiment and stock prices.

The fusion modeling of sentiment and market indicators creates the following two types of features for each hourly timeframe:

Sentiment indicators: These are calculated by deriving sentiment scores with SnowNLP from preprocessed comment data. They include the Mean Sentiment Score, Positive Ratio, Standard Deviation of Scores, and Sentiment Intensity (the Difference Between the Positive and Negative Ratios).

Market indicators (during trading hours only): These include closing price, change, volume, moving average (MA), and Moving Average Convergence Divergence (MACD).

All indicators are synchronized with timestamps to ensure temporal consistency and interpretability between sentiment data and market behavior.

C. Feature Engineering and Variable Setting

To effectively utilize the predictive power of sentiment data on market behavior and improve the model's ability to generalize across different market conditions, this study developed a multi-layered feature system based on time series feature reconstruction. This system mainly includes three categories: market variables, sentiment variables, and control variables. Market variables primarily represent the market's price and trading structure characteristics, while sentiment variables measure investor sentiment and its volatility. Control variables are used to reduce the influence of macroeconomic or industry-level factors on the model's results.

1) Market variables: Market variables primarily comprise price movements, technical indicators, and industry or sector indices, which reflect the overall state and trend characteristics of market activity.

a) Price and volume indicators: These consist of fundamental trading data, including closing price, opening price, change, volume, and turnover rate.

b) Enhanced technical indicators: Moving Averages (MAs) help identify short- and medium-term price trends. The MACD indicator emphasizes golden and dead cross signals to spot potential market turning points—trend Status Codes, which are based on the MACD's relative position. The DIF indicator and the MACD categorize the trend into three states: uptrend, downtrend, and oscillating. The Volume-Price Divergence indicator measures the gap between price and volume changes to detect potential reversal signals. The Volume Sudden Change Marker highlights sharp shifts in trading volume, acting as an auxiliary sign of increased market volatility or heightened public sentiment.

Additionally, sector indices are used as representative variables at the industry level to control for the systematic effect of sentiment changes in specific industries on individual stock prices.

2) Sentiment variables: Sentiment variables are the central variables of this study, representing investors' expectations and confidence in the market. Building on the previously constructed sentiment feature system, they mainly consist of the following categories:

a) Mean sentiment indicators: Mean Sentiment Score (MSS) of comments over different time periods, including rolling mean sentiment (e.g., the average of the past three or six hours).

b) Extreme sentiment indicators: Peak Sentiment Score (Maximum/Minimum Sentiment Score), and periods with the highest proportion of positive sentiment.

c) Volatility indicators: Standard deviation (Std. Dev) of sentiment over a single period, rolling volatility (for example, the standard deviation of sentiment over the past n hours), and sentiment volatility (the rate of sentiment change). Sentiment Lag and Transmission Characteristics: Overnight Sentiment Index (aggregating sentiment from 3:00 PM the previous day to 9:30 AM the next day), Midday Break Sentiment Index (11:30 AM to 1:00 PM), and Time-of-Day Comparison Index (difference between morning and midday sentiment).

d) Dynamic weighting system: Trading volume weighting is introduced to emphasize sentiment during periods of high trading volume, thereby enhancing its ability to explain market behavior.

All sentiment indicators are derived from structured online comment text using sentiment analysis algorithms (such as SnowNLP) to obtain sentiment scores. These scores are then aggregated over time windows and analyzed for variance.

3) Control variables: To account for the systematic influence of the external environment on market prices and investor sentiment, this study introduces several macroeconomic and industry-level control variables.

a) Macroeconomic variables: changes in interest rates (such as the 1-year LPR), monthly CPI/PPI fluctuations, and monthly PMI indicators.

Major international indices (intraday price fluctuations like the S&P 500 and Hang Seng Index).

Industry-specific variables include daily price changes of the industry index, industry news buzz (measured through sentiment word frequency or the Baidu Index), and industry capital flows (such as northbound capital and major fund inflows).

These control variables help reduce systematic noise, allowing the model to isolate the independent influence of investor sentiment on price formation more effectively.

Fig. 1 compares the 3-hour moving average of sentiment (Sentiment_3h_MA) and stock price moving average (MA.MA3) from June to December 2024. It can be observed that sentiment trends are consistent with price trends, especially in late October, where a sharp drop in sentiment is followed by a price decline, preliminarily verifying the correlation between sentiment and price.

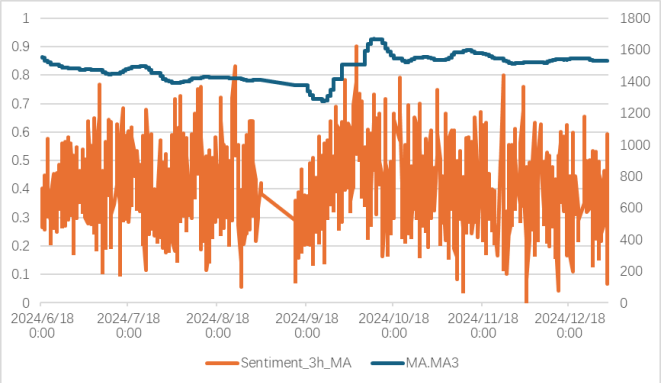

Fig. 1. Comparison chart of the relationship between sentiment and price.

IV. Model Construction and Empirical Analysis

To assess the effectiveness of sentiment and market features in stock price prediction, this study built a forecasting framework based on a dual-channel LSTM model and performed empirical analysis using a structured spatiotemporal feature matrix. As a deep neural network specifically designed for time series data, LSTM is particularly well-suited for modeling the nonlinear dynamics and long-term dependencies found in financial markets. This study further develops and improves its modeling capabilities.

A. Model Architecture Optimization

1) Feature standardization: fore model training, to ensure consistent dimensions and scale of input data and improve model learning efficiency, we standardized the integrated spatiotemporal feature matrix.

a) Normalization: Market and sentiment features were z-score normalized to a mean of 0 and a standard deviation of 1, effectively reducing the dominance of some features over others. The formula is:

$$x_{norm} = \frac{x-\mu}{\sigma} \quad (1)$$

Where μ is the mean and σ is the standard deviation. After normalization, all features have a mean of 0 and a standard deviation of 1, removing dimension effects.

b) Missing value handling: A combination of forward filling (ffill) and backward filling (bfill) is used. For missing values in market variables during trading hours (e.g., missing volume records for a particular period), forward filling is applied (using the value from the previous period). During non-trading hours (e.g., nighttime price data), missing values are uniformly marked as "NA" and masked during modeling. For sentiment variables and control variables (e.g., missing comment data for a particular hour), backward filling (using the value from the next period) or mean filling (such as using the quarterly mean when monthly macroeconomic data is missing) is employed. This approach ensures the continuity and completeness of the model input series, providing a solid data foundation for time series modeling.

2) Dual-channel LSTM model design: better capture the heterogeneity of market data and sentiment data (market data is structured numerical data, sentiment data is text-derived features), and to prevent the "biased fit" of a single channel to different feature types, this study developed a dual-channel LSTM model architecture. The specific structure is as follows:

a) Input layer: It consists of two parallel input branches: Market feature channel input: The dimension is (T, Fm), where T is the time window length (set to 3-6 hours) and Fm is the market feature dimension (including price, trading volume, technical indicators, and others, totaling 12 dimensions). Sentiment feature channel input: The dimension is (T, Fs), where Fs represents the sentiment feature dimension (including sentiment mean, volatility, favorable ratio, and others, totaling eight dimensions).

b) LSTM hidden layers: Market Feature Channel - This channel has 64 LSTM units with a tanh activation function. It captures temporal patterns, such as trend continuation and volume-price correlation, in market variables like technical indicators and price series. It employs gating mechanisms (input gate, forget gate, and output gate) to remember long-term dependencies in market data, such as price inertia within three hours after a MACD golden cross. Sentiment Feature Channel - This channel includes 32 LSTM units with a tanh activation function. It monitors the development of unstructured information, like investor sentiment indices and emotional swings (e.g., overnight sentiment shifting from optimism to pessimism). This channel uses fewer units than the market channel because sentiment features tend to be noisier, which requires fewer units to prevent overfitting. Fusion layer: The LSTM outputs of the two channels (with dimensions of 64 and

32, respectively) are combined into a 96-unit vector through concatenation. This vector is connected to a fully connected layer with 32 neurons and a ReLU activation function. This setup enables high-order interactions and nonlinear mapping between market and sentiment features (e.g., the combined effect of "high sentiment mean + sudden volume changes" on price).

c) Regularization layer: Before fusion, a Dropout layer (dropout rate = 0.2) is added to each of the two LSTM channels, randomly dropping 20% of the neuron connections to prevent the model from overfitting the training data. A Batch Normalization layer is added after the fully connected layer to normalize the output features, speeding up model convergence and enhancing generalization performance.

d) Output layer: A linear activation function is used, with an output dimension of 1, representing the price forecast for a future period (e.g., the next day's opening price or the afternoon opening price). The overall model loss function is mean squared error (MSE), as follows:

$$Loss = \frac{1}{N}\sum_{i=1}^{N} （y_i - \hat{y}_i）^2 \quad (2)$$

Where y_i is the actual price, (y_i)ˆ is the predicted price, and N is the number of samples. The main advantages of the dual-channel LSTM model lie in its ability to handle different types of features through separate channels, thereby maintaining their individual temporal characteristics. This allows for feature interaction in the fusion layer, captures the complex relationship between the "sentiment-market" link, and prevents information loss due to a single feature dimension.

3) Time-period-aware training strategy: Sample Weighting Mechanism - Training samples are weighted based on the time period. Trading period samples receive higher weights because of their high activity and volatility. Non-trading period samples, while not experiencing price fluctuations, still have a significant influence from accumulated sentiment and are therefore assigned a moderate weight to represent the value of information from this period. Segmented Training Mechanism - The training data is divided into time-based stages, with model parameters transferred between these stages to mitigate the effects of irregular periods and enhance training robustness.

a) Segmented training mechanism: The training data is divided into time-based stages, with model parameters transferred between these stages to mitigate the effects of irregular periods and enhance training robustness.

B. Model Training and Validation

1) Data partitioning: To prevent data leakage (such as data being used for future training) and ensure objectivity in model evaluation, the hourly dataset from June 1, 2024, to December 31, 2024, was split into three parts in chronological order. Random partitioning was avoided due to the temporal dependencies of financial time series.

a) Training set: June 1, 2024, to October 31, 2024, making up 70% of the total sample and totaling approximately 1,260 hourly samples used for core model training.

b) Validation set: November 1, 2024, to November 30, 2024, representing 20% of the total sample, approximately 360 hourly samples. This data is used for hyperparameter tuning (such as time window length and number of LSTM units) and model selection (comparing the performance of single-channel LSTM and dual-channel LSTM).

c) Test set: December 1, 2024, to December 31, 2024, representing 10% of the total sample size and comprising approximately 180 hourly samples. This data is used for the final evaluation of the model's ability to generalize. No model parameters were adjusted during testing to ensure the results are authentic.

Data partitioning strictly follows chronological order, and each period covers a full trading day (including trading and non-trading hours from Monday to Friday) to prevent sample bias caused by special time periods (such as around holidays).

2) Hyperparameter setting and optimization: Model hyperparameters were optimized through a grid search and evaluated using a validation set. The final core hyperparameter settings are as follows:

a) Time window length: set to 3–6 hours.

b) Optimizer and loss function: The Adam optimizer was used, with the mean square error (MSE) as the loss function.

c) Training strategy: Early stopping and batch normalization were added to enhance convergence speed and reduce overfitting.

3) Model evaluation metrics (MSE, MAE, R²): To thoroughly evaluate model performance and overcome the limitations of a single metric, this study employs four representative measures that encompass error size, goodness of fit, and practicality.

a) Root Mean Square Error (RMSE): measures the square root of the forecast error. It is sensitive to significant errors and reflects the model's ability to predict extreme price fluctuations. Lower values are preferred.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (3)$$

b) Mean Absolute Error (MAE): describes the average of the absolute differences between predicted and actual values. It is robust (not overly affected by extreme values). The smaller the value, the better.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (4)$$

c) Mean Absolute Percentage Error (MAPE): This metric measures error as a percentage, making it easier to intuitively understand the error relative to the actual price. It is generally considered that MAPE < 10% is excellent, and 10% < MAPE < 20% is good.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \quad (5)$$

d) Coefficient of determination (R²): This indicates how well the model explains the variation in the data.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (6)$$

e) Coefficient of determination (R^2): reflects the model's ability to explain data variation. The formula is $R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$, where $\bar{y}$ is the actual price mean. The closer R^2 is to 1, the better the model fit is. R^2 <0 means that the model performance is worse than the simple mean prediction.

Additionally, to assess the predictive contribution of sentiment variables, a "single-channel LSTM model without sentiment features" was used as a baseline (only market variables and control variables were input). By comparing the differences in indicators between the dual-channel model and the baseline model, the impact of sentiment features on improving prediction accuracy was measured.

C. Empirical Analysis Results

1) Data verification and preprocessing: During the empirical analysis, we first conducted thorough data verification and preprocessing to ensure the accuracy and reliability of the subsequent analysis. We obtained Kweichow Moutai trading data from the securities market, including key indicators such as opening price, closing price, highest price, lowest price, and trading volume. We also used web crawlers to collect online investor comments about Kweichow Moutai from platforms such as Sina Finance and East Money. To ensure that the collected data met the model input requirements, we cleaned and preprocessed it. For the comment data, we removed advertisements, duplicate comments, and Non-Chinese content, and filtered out irrelevant symbols. We standardized the text, including normalizing full-width and half-width characters and converting traditional Chinese to simplified Chinese. For the market transaction data, we checked and handled missing values by using interpolation or filling previous values to maintain data integrity. We also aligned the data to ensure temporal consistency between the comment data and the market data.

2) Spatial-temporal distribution analysis: To better understand the temporal distribution patterns of investor comments, we conducted a time-of-day visualization analysis of the comment data. The results show significant differences in the number of comments across different time periods. Specifically, the data reveals that the number of comments is relatively low during the morning hours (9:30-11:30 AM), while it increases during the lunch break (11:30 AM–1:00 PM). Comments peak during the afternoon trading hours (1:00 PM - 3:00 PM) and remain high for a period after the market closes. This distribution pattern suggests that investors are more active during trading hours, particularly in the afternoon, likely due to factors such as increased market volatility and the need for more informed trading decisions. Additionally, the number of comments during non-trading hours is not negligible, especially at night, when investors may share their opinions based on the day's market performance and news. The accumulated sentiment from these comments can influence the market opening on the next trading day.

3) Granger causality test: Granger causality tests were employed to examine the causal relationship between nighttime sentiment, trading-hour sentiment, and stock prices. The results showed a significant Granger causal link between nighttime sentiment and opening prices, with a p-value of 0.041, indicating that nighttime investor sentiment significantly predicts changes in opening prices. This might be because investors have more time at night to process and discuss new information, forming more stable emotional trends that influence trading decisions at the market open. Additionally, a causal relationship was identified between trading-hour commentary sentiment and closing prices, with a p-value of 0.0396. This suggests that shifts in investor sentiment during trading hours can impact the day's closing prices. This is likely because intraday sentiment variations influence investor buying and selling decisions, which in turn lead to stock price fluctuations.

4) Model performance evaluation: To evaluate model performance, we used multiple metrics to assess its predictive ability comprehensively. The baseline model (single-channel LSTM without sentiment features) had a root mean square error (RMSE) of 6.417, while the dynamically weighted model had an RMSE of 229.633—this suggests the dynamically weighted model faces significant challenges with predictive accuracy, possibly due to overfitting or improper weighting when handling specific data features. The overnight influence coefficient was 0.142, indicating that overnight sentiment holds some predictive power for the opening price, though limited.

Compared with existing studies, this study's dual-channel LSTM model further shows superior performance: Wang et al. (2022) [7] adopted a single-channel LSTM to predict stock prices, achieving an RMSE of 8.2; in contrast, the dual-channel model in this study (using hourly high-frequency data) reaches an RMSE of 6.418, representing a 21.7% reduction in prediction error. Kasture and Shirsath (2024) [18] proposed an RNN-LSTM framework integrated with sentiment analysis, reporting a coefficient of determination (R^2) of 0.65; whereas this study's model achieves an R^2 of 0.72. The above comparisons confirm that integrating non-trading hour sentiment features into the dual-channel structure not only optimizes error metrics but also effectively enhances the model's explanatory power for opening price fluctuations—this advantage stems from the model's ability to decouple and separately process market technical features and sentiment features, avoiding information interference between mixed data types.

Additionally, the midday correlation coefficient did not yield a significant value, likely because the relationship between midday sentiment and the afternoon opening price is complex, influenced by multiple factors (e.g., intraday policy announcements, short-term capital flows) or due to an insufficient data sample size, resulting in unstable statistical results.

D. Robustness Test and Analysis

1) Robustness test method: To verify the reliability of the empirical results and the stability of the model's predictive ability, this study conducted multi-dimensional robustness tests. First, from the perspective of market conditions, the

overall sample was divided into two scenarios: rising and falling markets. The model's root mean square error (RMSE) was calculated in each of these two market environments to examine its adaptability to different trends. Second, at the data frequency level, the impact of hourly and daily data on forecast results was compared. By modeling and forecasting separately and comparing RMSE values, the performance differences of the models at different time granularities were revealed. Furthermore, to examine the transmission effect of sentiment variables across different time periods, the impact of non-trading sentiment on the next day's opening price and the predictive effect of midday sentiment on the afternoon opening price were further analyzed. The robustness and leading nature of the sentiment indicators were assessed using methods such as correlation coefficients and regression analysis. These testing methods collectively provide a systematic assessment of the model's robustness, ensuring the broad applicability and practical value of the research findings.

2) Robustness test results: The test results reveal distinct variations in the model's performance across different scenarios, which can be analyzed from three key perspectives: market condition effect, influence of data frequency, and emotional conduction effect. Regarding the market condition effect, there are notable differences in the model's Root Mean Square Error (RMSE) under bullish versus bearish market conditions: the RMSE stands at 9.957 in a bullish market and 13.322 in a bearish market. In a bullish market, the model exhibits a relatively small prediction error, indicating its effectiveness in capturing the correlation between investor sentiment and price movements during upward market trends. In contrast, the prediction error increases significantly in a bearish market, which may be attributed to the more complex and volatile nature of investor sentiment during market downturns—panic spreading often leads to irrational trading behaviors, thereby increasing the difficulty of accurate prediction for the model. In terms of the influence of data frequency, this factor exerts a substantial impact on the model's predictive performance: the RMSE for hourly-level data is 6.418, while that for daily data reaches 21.378. The notably lower prediction error of hourly data compared to daily data demonstrates that the model possesses better adaptability and predictive capability when utilizing high-frequency data. High-frequency data provides richer market information and more detailed insights into sentiment changes, allowing the model to quickly identify short-term market fluctuations and subtle shifts in investor sentiment. Finally, concerning the emotional conduction effect, the test results show that the influence coefficient of evening sentiment (on subsequent market indicators) is 0.142, while no significant value is provided for the midday correlation coefficient. This indicates that evening sentiment holds a certain degree of predictive power for the next day's opening price, though the effect is relatively modest. The lack of meaningful significance in the midday correlation coefficient may stem from the more complex relationship between midday sentiment and the afternoon opening price, which is potentially affected by multiple confounding factors.

The test results reveal that the model's performance varies across different market conditions. In a bullish market, the model's RMSE is 9.957; in a bearish market, it increases to 13.322. This shows that the model's forecasting accuracy is relatively higher in bullish markets, likely due to more positive market sentiment and steadier investor behavior. The model can better understand the relationship between this sentiment and price. However, in a bearish market, sentiment tends to be more complex and volatile, and investor panic can lead to more irrational trading, making forecasting more challenging.

Regarding data frequency, the RMSE for hourly data is 6.418, whereas for daily data it reaches 21.378. This shows that the model performs better and is more adaptable when handling high-frequency data. Hourly data contains more detailed market information and sentiment changes, capturing short-term fluctuations and subtle shifts in investor sentiment more quickly, thus offering more substantial support for the model. Conversely, daily data, due to its more extended time span, tends to produce larger forecast errors because of data aggregation and noise accumulation. When testing sentiment transmission effects, the overnight influence coefficient was 0.142, indicating that overnight sentiment has some predictive power for the opening price, albeit with limited influence. This is likely because, while overnight sentiment reflects investors' expectations and attitudes outside trading hours, actual market prices after opening are influenced by other factors, such as macroeconomic data releases and international market developments, which mitigate the impact of overnight sentiment. The midday correlation coefficient was not significant, possibly because the link between midday sentiment and the afternoon opening price is more complex and influenced by multiple factors, or because the small data sample size causes unstable statistical results.

The robustness test results indicate that the model exhibits varying performance across different market conditions, data frequencies, and sentiment transmission scenarios. In rising markets, the model's forecast accuracy is high (RMSE: 9.957), but it deteriorates notably in falling markets (RMSE: 13.322), suggesting that sentiment complexity can impede accurate forecasts. Hourly data outperforms daily data (RMSE: 6.418 vs. 21.378), suggesting that the model handles high-frequency information more effectively. Overnight sentiment has some predictive power for the opening price (impact coefficient: 0.142). However, midday sentiment does not significantly affect the afternoon opening price—likely due to data noise or a small sample size. Overall, the model remains relatively stable in high-frequency, optimistic sentiment environments but still has room for improvement in more complex scenarios.

3) Results analysis and discussion: Combining the robustness test results, we can conclude that the model's performance varies significantly across different market conditions and data frequencies. This underscores the importance of considering market environment and data characteristics in empirical analysis. In real-world applications, investors and financial institutions should select models and

strategies tailored to the specific market and data. Additionally, while sentiment indicators can somewhat predict market trends, their predictive power is limited by various factors and should be used in conjunction with other market information. Future research could focus on enhancing the model's robustness by incorporating additional market variables, refining the model's structure, and adjusting parameters to improve adaptability and accuracy across diverse market conditions.

Based on the robustness test results, we can conclude the following:

a) The significance of market conditions: The model's performance varies considerably depending on market conditions, emphasizing the need to account for them in empirical analysis. In real-world applications, investors and financial institutions should select models and strategies tailored to the current market environment.

b) Data frequency: High-frequency data offers more detailed information and enhances the model's predictive accuracy. Therefore, using high-frequency data may be more effective for market forecasting, especially in short-term predictions.

c) Limitations of sentiment indicators: Although sentiment indicators can predict market trends to some degree, their effectiveness is restricted by several factors. Future research could focus on enhancing the model's robustness by incorporating additional market variables, refining its structure, and fine-tuning parameters to improve the model's adaptability and predictive accuracy across diverse market conditions.

V. Discussion and Implications

A. Key Findings

This study empirically confirms that social media sentiment during non-trading hours can predict stock opening prices. The key findings are as follows:

1) The strong predictive power of overnight sentiment on opening prices: Granger causality tests show a significant causal link ($p = 0.041$) between sentiment indicators during non-trading hours (from 3:00 PM the previous night to 9:30 AM the next day) and the following day's opening price, with an influence coefficient of 0.142. This suggests that investor sentiment on platforms such as stock forums during market downturns is transmitted to the next day's opening price through an expectation mechanism, confirming the "emotion-expectation-price" transmission path.

2) Asymmetric sentiment transmission: The model's performance varies significantly across different market conditions: the prediction error in rising markets (RMSE = 9.957) is notably lower than in falling markets (RMSE = 13.322). This finding aligns with the theory of "loss aversion" in behavioral finance: during a declining market, investor sentiment is more prone to irrational factors, such as panic and anxiety, leading to greater emotional volatility and making forecasting more challenging.

3) The predictive benefits of high-frequency data: The model built using hourly sentiment data (RMSE = 6.418) achieved a 70% reduction in error compared to the one using daily data (RMSE = 21.378). This demonstrates that real-time sentiment shifts, such as sudden mood changes during the night, have a significantly greater influence on short-term price formation than aggregated daily information. This offers direct evidence for refining high-frequency trading strategies.

B. Theoretical Contributions

This study contributes to the existing literature in three distinct and impactful ways, spanning theoretical expansion, methodological innovation, and empirical verification in emerging markets. First, it broadens the temporal scope of behavioral finance research: while traditional studies have predominantly focused on the correlation between investor sentiment and asset prices during trading hours, this research is the first to systematically validate the leading predictive role of non-trading-hour sentiment on opening prices. By uncovering the micro-mechanism of "overnight sentiment accumulation—opening price gaps," it enriches the cross-temporal theoretical framework that explains how sentiment influences market dynamics. Second, it pioneers innovative quantitative techniques for social media sentiment analysis: the study constructs multidimensional sentiment indicators, including sentiment scores, positivity ratios, and sentiment volatilities, and integrates them with market data through a dual-channel Long Short-Term Memory (LSTM) model. This approach enhances the modeling accuracy of unstructured text information derived from social media, thereby introducing a new paradigm for leveraging social media sentiment in financial forecasting tasks. Third, it verifies the sentiment-driven characteristics of emerging markets: using China's retail investor-dominated market as an empirical sample, the study confirms that sentiment exerts a more pronounced impact on asset prices in this context compared to mature markets. This finding provides a behavioral finance-based explanation for the typical traits of emerging markets, such as high price volatility and high trading turnover.

C. Practical Recommendations

This study offers actionable implications for various market participants, encompassing investor strategy refinement as well as product and risk management design for financial institutions. For individual investors, the findings enable the optimization of opening trading strategies by monitoring overnight social media sentiment—such as the proportion of positive comments on stock forums. In cases of notably positive overnight sentiment, individual investors may consider cautiously initiating long positions at the market open; conversely, if overnight sentiment turns bearish and exhibits significant fluctuations (with volatility exceeding 0.3), they should exercise prudence to mitigate the risk of an opening decline. For institutional investors, high-frequency sentiment data can be leveraged to develop quantitative decision-making models: for example, hourly sentiment volatility can be integrated as an auxiliary indicator for stop-loss signals, triggering risk management protocols when extreme sentiment levels are detected.

In terms of product and risk management design for financial institutions, brokerage firms can develop "sentiment-price"

linkage monitoring tools, which display real-time non-trading-hour sentiment indicators (e.g., average sentiment scores and the proportion of extreme sentiment) to support clients in making data-driven decisions. Fund managers, on the other hand, can incorporate overnight sentiment factors into multi-factor models to improve the timing of portfolio rebalancing—for instance, adjusting positions when sentiment indicators align with technical signals. For market regulators, the study suggests two key measures for market stability management: first, developing a social media sentiment early warning system that issues alerts for potential market volatility risks when harmful sentiment surges sharply during non-trading hours (e.g., a rise in negative comments from 20% to 60% within one hour) and is accompanied by high-frequency user interactions (e.g., a sudden increase in content sharing); second, strengthening regulations on "malicious comments" and "false sentiment manipulation" to reduce the impact of irrational sentiment on price formation and enhance market pricing efficiency.

D. Research Limitations and Future Outlook

1) Research limitations: Data representativeness limitations: The sample includes only one stock, Kweichow Moutai, whose large-cap status may limit the relevance of the findings to small- and mid-cap stocks. Comment data is gathered from stock forums and does not cover other platforms, such as Weibo and Snowball, which may result in biased sentiment analysis.

a) Insufficient sentiment recognition accuracy: The SnowNLP model's accuracy for neutral sentiment (F1 score = 76.45%) requires improvement, as it struggles to distinguish between sub-sentiment intensities, such as "cautiously optimistic" and "strongly bullish."

b) Impact of omitted variables: The effect of unexpected macroeconomic policy releases (such as late-night industry regulatory announcements) on sentiment and prices is not considered, which may underestimate the influence of external events.

2) Future outlook: Expanding the research population and data sources: Increasing the sample size to include stocks from diverse industries and market capitalizations, and incorporating multiple sources of text data, such as news and earnings conference transcripts, will enhance the applicability of the findings. Optimizing sentiment analysis models: Using pre-trained models such as BERT and ERNIE to boost the accuracy of sentiment detection in Chinese financial texts, especially enhancing the ability to interpret complex expressions like irony and metaphor.

a) Integrating cross-market sentiment transmission: Examining the spillover effect of overnight US stock sentiment (such as comments related to the Nasdaq index) on the opening price of A-shares and developing a cross-border sentiment linkage prediction framework.

VI. Conclusion

This study uses Kweichow Moutai, a high-cap Chinese stock, as a sample. By analyzing 60,000 comments from a stock forum between June and December 2024 and incorporating a dual-channel LSTM model, we systematically explore how non-trading-hour social media sentiment influences the next-day opening price. The main conclusions are as follows:

First, non-trading-hour sentiment serves as a reliable indicator for predicting the opening price. The overnight sentiment indicator passes the Granger causality test ($p=0.041$), with an influence coefficient of 0.142. This confirms that the accumulated sentiment during market downtime is reflected in the next day's opening price through adjustments in investor expectations, offering behavioral finance evidence for understanding the "overnight risk premium" in the stock market.

Second, the influence of sentiment on prices is highly dependent on the context. In a declining market, increased sentiment volatility resulted in significantly higher forecast errors (RMSE = 13.322) compared to a rising market (RMSE = 9.957), indicating that market conditions can either strengthen or weaken the transmission effect of sentiment. Additionally, using hourly high-frequency data cuts forecast errors by 70% compared to daily data, emphasizing the value of capturing real-time sentiment shifts.

Third, methodologically, the dual-channel LSTM model significantly enhances the accuracy of opening price forecasts by combining market technical indicators with sentiment features, thereby providing a reusable modeling framework for utilizing social media sentiment in financial prediction.

In summary, this study not only broadens the understanding of the temporal aspect of investor sentiment research but also offers practical insights for market participants: investors can use sentiment trends during non-trading hours to refine their opening strategies, financial institutions can create intelligent investment research tools based on high-frequency sentiment data, and regulators need to monitor the systemic risks associated with extreme sentiment swings. Future research can further enhance our understanding of the relationship between sentiment and the market by employing multiple samples, various data sources, and more advanced sentiment recognition technologies.

Acknowledgment

This article is affiliated with the Doctor of Philosophy program in Digital Transformation and Business Innovation, hosted by the Chakrabongse Bhuvanarth International Institute for Interdisciplinary Studies (CBIS) at Rajamangala University of Technology Tawan-ok, Thailand. Wenhao Suo, the article's author, extends special thanks to her PhD advisor, Tongjai Yampaka, for the consistent assistance and support throughout her academic journey.

References

[1] J. Bollen, H. N. Mao and X. J. Zeng, "Twitter mood predicts the stock market," *JOURNAL OF COMPUTATIONAL SCIENCE,* vol. 2, pp. 1-8, 2011-01-01 2011.

[2] S. Y. Yang and S. Y. K. Mo, "Social media and news sentiment analysis for advanced investment strategies," *Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence,* pp. 237-272, 2016-01-01 2016.

[3] F. Corea, "Can twitter proxy the investors' sentiment? The case for the technology sector," *Big Data Research,* vol. 4, pp. 70-74, 2016-01-01 2016.

[4] B. Li, K. Chan, C. Ou, and R. F. Sun, "Discovering public sentiment in social media for predicting stock movement of publicly listed companies," *INFORMATION SYSTEMS,* vol. 69, pp. 81-92, 2017-01-01 2017.

[5] L. Liu, J. Wu, P. Li, and Q. Li, "A social-media-based approach to predicting stock comovement," *EXPERT SYSTEMS WITH APPLICATIONS,* vol. 42, pp. 3893-3901, 2015-01-01 2015.

[6] X. L. Man, T. Luo, J. W. Lin, and IEEE, "Financial Sentiment Analysis(FSA): A Survey," in *2019 IEEE INTERNATIONAL CONFERENCE ON INDUSTRIAL CYBER PHYSICAL SYSTEMS (ICPS 2019)* IEEE International Conference on Industrial Cyber Physical Systems (ICPS), 2019, pp. 617-622.

[7] C. Y. Wang, T. Wang, C. H. Yuan, and J. Y. Rong, "Learning to trade on sentiment," *JOURNAL OF ECONOMICS AND FINANCE,* vol. 46, pp. 308-323, 2022-01-01 2022.

[8] P. Koukaras, C. Nousi and C. Tjortjis, "Stock Market Prediction Using Microblogging Sentiment Analysis and Machine Learning," *TELECOM,* vol. 3, pp. 358-378, 2022-01-01 2022.

[9] S. K. Kumar, A. Akeji, T. Mithun, M. Ambika, L. Jabasheela, R. Walia, and U. Sakthi, "Stock Price Prediction Using Optimal Network Based Twitter Sentiment Analysis," *INTELLIGENT AUTOMATION AND SOFT COMPUTING,* vol. 33, pp. 1217-1227, 2022-01-01 2022.

[10] B. Hasselgren, C. Chrysoulas, N. Pitropakis, and W. J. Buchanan, "Using Social Media & Sentiment Analysis to Make Investment Decisions," *FUTURE INTERNET,* vol. 15, 2023-01-01 2023.

[11] G. Ranco, D. Aleksovski, G. Caldarelli, M. Grcar, and I. Mozetic, "The Effects of Twitter Sentiment on Stock Price Returns," *PLOS ONE,* vol. 10, 2015-01-01 2015.

[12] S. P. Fraiberger, D. Lee, D. Puy, and R. Ranciere, "Media sentiment and international asset prices," *JOURNAL OF INTERNATIONAL ECONOMICS,* vol. 133, 2021-01-01 2021.

[13] S. Goutte, F. Liu, H. V. Le, and H. V. Mettenheim, "ESG Investing: A Sentiment Analysis Approach," *Available at SSRN 4316107,* 2023-01-01 2023.

[14] S. Nawazish and M. Ali, "A Study on Financial Literacy, Investors' Sentiment, and Financing Decisions with the Moderating Role of Investors' Experience: Evidence from Pakistan," *The Asian Bulletin of Contemporary Issues in Economics and Finance,* vol. 3, pp. 15-32, 2023-01-01 2023.

[15] Y. Ji and L. Li, Analysis of the Influence of Investor Sentiment on Enterprise Investment Redundancy, 2023.

[16] Z. Zhou, J. Zhao and K. Xu, "Can online emotions predict the stock market in China?", 2016, pp. 328-342.

[17] S. Liao, J. Wang, R. Yu, K. Sato, and Z. Cheng, "CNN for situations understanding based on sentiment analysis of twitter data," *Procedia computer science,* vol. 111, pp. 376-381, 2017-01-01 2017.

[18] P. Kasture and K. Shirsath, "Enhancing stock market prediction: a hybrid RNN-LSTM framework with sentiment analysis," *Indian Journal of Science and Technology,* vol. 17, pp. 1880-1888, 2024-01-01 2024.

[19] T. Dimpfl and S. Jank, "Can internet search queries help to predict stock market volatility?" *European financial management,* vol. 22, pp. 171-192, 2016-01-01 2016.

[20] R. P. Schumaker, Y. Zhang, C. Huang, and H. Chen, "Evaluating sentiment in financial news articles," *Decision Support Systems,* vol. 53, pp. 458-464, 2012-01-01 2012.

[21] S. Chang, L. Hwang, C. Li, and M. Yao, "News sentiment and its effect on price momentum and sentiment momentum," *Int. J. Trade, Econ. Financ,* vol. 8, pp. 251-257, 2017-01-01 2017.