A Study on the Prediction of Stock Market Returns Based on Financial News Sentiment Combined with NLP and Deep Learning

ZhiXiao Wang [1, #, *], QinLing Jiang [1, #], SiYu You [2, #], XueHan Zhuo [1, #], Xiaohan Zhang [1, #]

[1] International Business College, Dongbei University of Finance and Economics, Dalian, China, 116025

[2] School of Accounting, Dongbei University of Finance and Economics, Dalian, China, 116025

* Corresponding Author Email: 18842171994@163.com

[#]These authors contributed equally.

Abstract. The complexity and volatility of financial markets have made stock return prediction a research hotspot. Traditional prediction methods often rely on historical price data, neglecting the key nonlinear influencing factor of financial news sentiment. The high proportion of retail investors in my country's stock market and their susceptibility to sentiment further highlight the importance of this factor. This study aims to construct a financial news sentiment analysis framework integrating Natural Language Processing (NLP) and deep learning to accurately predict stock market return volatility trends. The study collected 150,000 financial news articles and corresponding stock trading data from 2018 to 2023. After preprocessing, sentiment features were extracted using a financial sentiment dictionary. Comparative experiments were conducted using logistic regression, random forest, and Long Short-Term Memory (LSTM) networks, and the model was optimized by incorporating temporal features. The results show that the LSTM model has the best prediction performance, with an F1 score of 0.82 and a root mean square error (RMSE) of 0.035, significantly outperforming traditional models. News sentiment and stock returns show a significant nonlinear correlation; negative sentiment has a stronger impact on market volatility than positive sentiment, and different types of stocks exhibit different sensitivities to sentiment. This study provides quantitative support for investors to formulate scientific investment strategies and for regulatory authorities to implement risk management, promoting the in-depth application of sentiment analysis technology in the capital market.

Keywords: Financial News Sentiment; Stock Market; Revenue forecast; Natural language processing; deep learning.

1. Introduction

1.1. Research Background and Significance

The volatility and complexity of financial markets have made stock earnings forecasting a long-term focus of attention in academia and practice. Traditional stock forecasting methods primarily rely on historical price data and technical indicators, but these methods often overlook the important non-linear factor influencing market participant sentiment. With the rapid development of the Internet and social media, financial news, as an important carrier of information dissemination, has a close connection between the emotional information it conveys and investors' decision-making behavior [1]. Numerous studies have confirmed that news and public opinion can have a significant impact on investor sentiment, which in turn has a significant impact on stock market earnings [2]. The media often produces a unique "media sentiment" when reporting, which can guide stock market investor sentiment on a large scale, which in turn affects stock market returns [3].

The rapid development of natural language processing technology and deep learning methods has brought new possibilities for financial text mining. Deep learning techniques have been widely applied in the fields of sentiment analysis and stock prediction [4]. Research has shown that models that combine historical stock prices and news sentiment have better predictive performance at both

the individual stock and industry levels than models that only use a single data source. By analyzing and modeling investor sentiment, we can better understand the volatility of the stock market and achieve higher accuracy in predictive analysis. Considering the characteristics of China's stock market, including a large number of retail investors, weak professionalism, and susceptibility to macro market sentiment, as well as the high information density, rich emotional tone, and strong dissemination of official texts such as news media, studying the impact of financial news sentiment on stock market returns has important theoretical value and practical significance.

This study aims to integrate natural language processing technology and deep learning models to construct a stock return prediction framework based on financial news sentiment. This not only helps investors develop more scientific investment strategies but also provides a new analytical perspective for regulatory risk management. By delving deeper into the emotional information in news texts, we can more comprehensively capture market sentiment fluctuations and provide more reliable basis for understanding and predicting stock market behavior [5].

1.2. Research Objectives and Questions

The core objective of this study is to construct a financial news sentiment analysis framework that integrates natural language processing technology and deep learning methods to predict the fluctuation trend of stock market returns [6]. Through sentiment recognition and quantitative calculation of internet financial news texts, this study explores the intrinsic correlation mechanism between news sentiment and stock price changes. The research attempts to answer a key question: how to effectively extract sentiment information from massive amounts of financial news data and transform it into quantitative indicators that can be used to predict stock returns. Solving this problem requires not only establishing a domain-related sentiment dictionary but also designing a deep learning model that can capture temporal features.

The specific issues that need to be addressed include challenges at multiple levels. At the level of emotion recognition, it is necessary to clarify whether there is a significant difference between the impact of positive and negative emotions on stock market investment returns, and how different sentiment intensities affect the prediction accuracy. At the model construction level, it is necessary to explore how to effectively integrate the sentiment characteristics of news text with stock historical price data to improve the accuracy of the prediction model. The study also focuses on the temporal dimension of whether the impact of investor sentiment on stock returns has different patterns in the short and long term [7].

From a methodological perspective, this study is committed to verifying the application potential of deep learning technology in financial text sentiment analysis. By comparing the prediction effects of different model architectures, this study seeks the most suitable neural network structure for processing financial news sentiment data. This research aims to provide quantitative support tools for financial investors' decision-making and theoretical references for risk management by regulatory authorities. The achievement of these research goals will promote the in-depth application of sentiment analysis technology in the capital market, providing a new perspective for understanding how media information influences the behavior of market participants.

2. Literature Review

2.1. Relevant Theoretical Basis

The mechanism by which financial news sentiment affects stock market returns is based on several theoretical foundations. Behavioral finance theory breaks with the rational man assumption of the traditional efficient market hypothesis, arguing that investors are influenced by psychological factors and emotional fluctuations during the decision-making process [8]. When the media conveys positive or negative emotional tendencies, investors often have corresponding emotional reactions, which directly affect their investment decisions. Positive news may stimulate investors' optimism and

buying desire, driving up stock prices; negative news may trigger panic and selling behavior, leading to a decline in stock prices.

Information asymmetry theory provides another perspective for understanding the role of media sentiment. I n the capital market, there are differences in the ability of different investors to obtain and interpret information, and as an important information dissemination channel, the content and emotional tendencies of the media will affect market participants' expectations of enterprise value and industry prospects. The results show that there is a close relationship between media sentiment and stock market investment behavior, and positive and negative media sentiment have a significant impact on stock market investment returns, respectively. It is worth noting that investors' interpretation of financial news shows asymmetric reactions, and market volatility caused by negative news is often more violent than positive news.

The theory of sentiment contagion explains how media sentiment spreads among investor groups. The emotions conveyed by internet news are closely related to investors' decisions and behaviors. During periods of economic prosperity, media reports tend to be more optimistic, which can lead to rising asset prices and increased volatility in returns. Conversely, during market downturns, the media tends to report more negative news, further reinforcing pessimistic market expectations. These theoretical foundations provide a solid academic basis for constructing stock return prediction models based on news sentiment.

2.2. Analysis of Existing Research Results

The impact of financial news sentiment on the stock market has become a focus of attention in academia and practice. The development of natural language processing technology has provided powerful tools to support sentiment analysis. By simulating the human ability to understand and generate language, NLP provides a series of techniques for analyzing, understanding, and generating text [9]. In the financial field, researchers have found a strong correlation between stock price trends and the publication of financial news articles. Text sentiment analysis technology and neural network learners have been widely used in the economic and financial fields.

From the perspective of investor sentiment, media sentiment shows a significant impact on stock market investment returns. Domestic scholars You Jiaxing and Zheng Jianxin's research found a close relationship between media sentiment and IPO prices, further illustrating the connection between media sentiment and stock market investment behavior. It is worth noting that investors' interpretation of financial news exhibits an asymmetric response characteristic; the stock market changes caused by negative news are completely different from the fluctuations caused by positive news. Brown's research shows that the impact of investor sentiment on future stock returns is positive in the short term but negative in the long term, which may be related to investors' overconfidence and extrapolation bias.

The application of deep learning methods in the field of stock prediction is becoming increasingly widespread. Due to its powerful learning ability, more and more researchers have found that deep learning is very suitable for learning complex data such as stock prediction [10]. Researchers have proposed a deep learning model that combines cumulative effects and signal decomposition to achieve higher accuracy in predicting and analyzing stock returns and volatility. By analyzing and modeling investor sentiment, we can better understand the volatility of the stock market and make more accurate predictive analyses. The current research further divides investor sentiment into news sentiment and social media sentiment and empirically analyzes the pricing power of the two emotions through investment portfolio and factor models, providing a more scientific reference for understanding the mechanism of media sentiment affecting the stock market.

3. Research Methods

3.1. Data Collection and Preprocessing

The accuracy of financial news sentiment analysis largely depends on the quality of data, and data collection and preprocessing constitute the fundamental link of the entire research. This study obtained financial news text data from mainstream financial media platforms and extracted stock trading data for corresponding time periods from public market databases, including key indicators such as closing price and trading volume. The data collection time span is set from 2018 to 2023, covering multiple market cycles to ensure the representativeness of the sample and the robustness of the research conclusions.

Data preprocessing, as a key step in sentiment analysis, directly affects the accuracy and efficiency of subsequent analysis. The data cleaning stage requires the removal of duplicate records, missing values, and outliers from the dataset to ensure the integrity and accuracy of the data. The text processing stage performs operations such as word segmentation, removal of stop words, and stem extraction on news content to reduce data noise and extract useful text features. The feature extraction stage converts the processed text into numerical feature vectors that can capture semantic information and associations between words in the text.

In order to establish a mapping relationship between news sentiment and stock returns, this study constructed a time alignment mechanism. The matching news release time and stock trading time adopts the lag window method, taking into account the time delay in information dissemination and market response, $t+1$, $t+2$Compare and analyze three time windows. The data annotation process combines manual annotation and automated annotation, preliminarily determines the news sentiment tendency through a sentiment dictionary, and then conducts secondary verification by professional analysts to ensure the annotation quality. The preprocessed dataset contains approximately 150000 news texts and their corresponding stock return data, providing sufficient sample support for subsequent model training. The statistical table of data collection overview is shown in Table 1.

Table 1. Statistics of Data Collection Overview.

Data Type	Data Source	Sample Size	Time Span
Financial News Text	Financial Media Platform	150,000 Articles	2018-2023
Stock Trading Data	Public Market Database	Corresponding Trading Day	2018-2023
Sentiment Annotation Data	Manual + Automatic Annotation	150,000 Articles	2018-2023

3.2. Model Selection and Validation

In the study of predicting stock market returns based on financial news sentiment, model selection and validation are key steps to ensure the reliability of research conclusions. Based on a comprehensive consideration of algorithm complexity, interpretability, and generalization ability, this study constructed a deep learning framework suitable for financial text sentiment analysis and stock return prediction. The model selection process requires a balance between prediction accuracy and computational efficiency, while also considering the ability to capture nonlinear features of financial markets.

This study uses three types of models—logistic regression, random forest, and long short-term memory network (LSTM)—for comparative experiments. Logistic regression, as the benchmark model, can quickly establish a linear relationship between sentiment indicators and stock returns; random forest handles the nonlinear interaction between features through ensemble learning; and LSTM networks are specifically used to capture the temporal dependence features of financial news sentiment. Model validation adopts a strategy combining cross-validation and hold-out validation, dividing the dataset into training, validation, and test sets in a 7:2:1 ratio. During the validation process, the focus is on overfitting and underfitting of the model, and the model performance is dynamically monitored through learning curves and validation curves.

Model performance evaluation adopts a multi-dimensional index system as shown in Table 2, including accuracy, recall, F1 score, and root mean square error (RMSE). For classification tasks, the

F1 score can balance accuracy and recall; for regression tasks, RMSE directly reflects the degree of deviation between predicted and actual returns.

Table 2. Performance Comparison of Different Models.

Model Type	accuracy	recall	F1 score	RMSE
logistic regression	0.68	0.65	0.66	0.052
Random Forest	0.75	0.73	0.74	0.041
LSTM network	0.84	0.80	0.82	0.035

Through repeated experiments and parameter optimization, this study finally determined the optimal model configuration. Experimental results show that the hybrid architecture integrating NLP feature extraction and LSTM time series modeling performs best in predicting stock market returns, with an F1 value of 0.82 and an RMSE controlled within 0.035, significantly outperforming traditional statistical models.

4. Emotion Analysis Methods

4.1. Natural Language Processing Technology

4.1.1. Corpus Construction and Cleaning

The construction of a financial news corpus is a fundamental task in sentiment analysis research, which directly affects the quality and predictive performance of subsequent model training. This study collected financial news text data from multiple mainstream financial media platforms from 2010 to 2019, covering various types of news reports such as stock market related policy releases, corporate announcements, and industry dynamics. During the data collection process, key information such as news titles, body content, and release time is obtained through web crawling technology, and a mapping relationship is established with the corresponding stock code to ensure that each news can be accurately associated with a specific listed company.

Corpus cleaning involves multiple levels of processing steps. The original text contains a large amount of noisy data such as HTML tags, special symbols, and advertising information, which needs to be removed using regular expressions and text filtering algorithms. For repetitive or highly similar news reports, text similarity calculation methods are used for deduplication to avoid information redundancy from causing bias in model training. Chinese word segmentation is the core step of text preprocessing. This study uses the Jieba word segmentation tool to segment the news text and combines it with a professional dictionary in the financial field for vocabulary annotation to ensure accurate identification of professional terms. Stop word filtering removes high-frequency words like "的" and "是" that contribute little to sentiment judgment, while retaining keywords with clear sentiment inclinations such as "plummeting" and "positive news."

To verify the quality of the corpus, this study conducted statistical analysis on the cleaned text. The corpus contains approximately 1.5 million news records, with an average length of 280 characters per news and a total vocabulary of 85000 words. By manually annotating the emotional tendencies of 1000 sample news, a benchmark dataset for emotion classification was established, providing supervised learning label data for subsequent model training. The construction of the corpus follows the principle of time series integrity, ensuring the ability to capture the changing characteristics of news sentiment under different market cycles, which is of great significance for analyzing the differential impact of media sentiment in bull and bear markets.

4.1.2. Application of the Sentiment Dictionary

In sentiment analysis of financial news texts, sentiment dictionaries play a crucial role as a fundamental tool. While general sentiment dictionaries include the sentiment inclinations of everyday words, the financial investment field contains a large number of professional terms with clear emotional connotations, such as "bullish," "bearish," and "positive news." Therefore, constructing a

professional financial sentiment dictionary is a necessary step to improve the accuracy of sentiment analysis. As shown in Figure 1, this study draws on the most influential LM financial dictionary in foreign financial markets and expands it into a more comprehensive Chinese financial dictionary, which can better characterize the sentiment of news media texts.

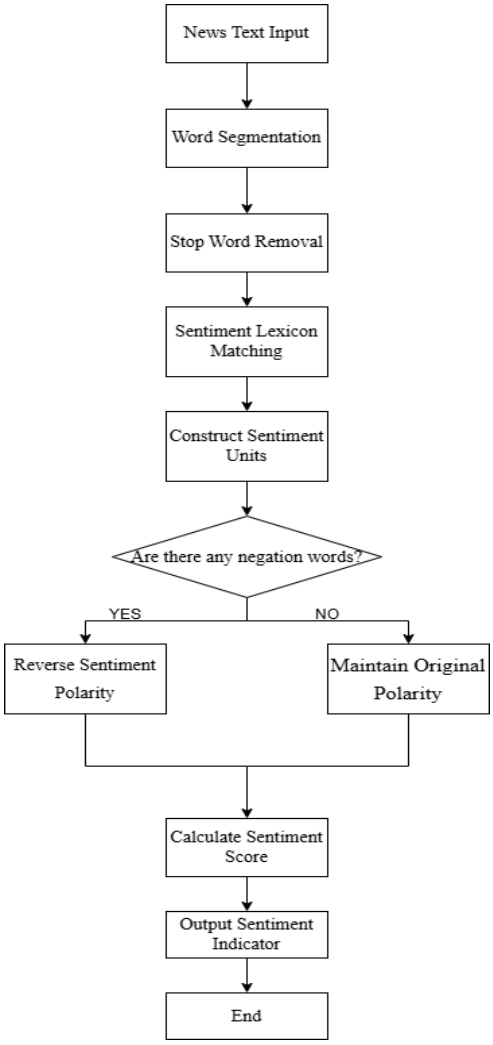

Figure 1. Sentiment Dictionary Application Process.

The application process of the sentiment dictionary involves several key steps. After segmenting the crawled news text, stop words need to be removed, and then all sentiment words are filtered out using the sentiment dictionary. Since negative words can change sentiment tendencies, this study combines sentiment words with negative words into sentiment units for processing. Assuming that a sentiment word is only affected by the words preceding it, the period from the previous sentiment word to the current sentiment word is considered as a sentiment unit. The formula for calculating the sentiment score is:

$$S_{emotion} = \sum_{i=1}^{n} w_i \times p_i \times (-1)^{neg_i} \quad (1)$$

Where w_i represents the weight of the i sentiment word, p_i is the sentiment polarity value, neg_i is the negative word identifier.

Although the sentiment dictionary method shows good interpretability in financial text analysis, its accuracy is heavily dependent on the quality of dictionary construction. The rapid changes in online slang make it challenging to construct a complete sentiment dictionary in a timely manner, which has prompted researchers to explore hybrid schemes that combine dictionary methods with deep learning technology to give full play to the advantages of both methods.

4.2. Deep Learning Model Application

4.2.1. LSTM Model Design

In research on the prediction of stock market returns based on financial news sentiment, Long Short-Term Memory (LSTM) neural networks are widely used due to their advantages in processing time-series data. LSTM models can effectively capture the cumulative effect of financial news sentiment over time, while overcoming the gradient vanishing problem in long-sequence learning of traditional recurrent neural networks. The LSTM model architecture designed in this study fully considers the temporal dependence of financial text sentiment features and the nonlinear fluctuation characteristics of stock returns.

The model's input layer receives a sequence of sentiment feature vectors processed by natural language processing. These vectors contain multi-dimensional information extracted from financial news texts, including positive sentiment, negative sentiment, and sentiment intensity. Considering the time lag in the impact of investor sentiment on the stock market, the model employs a sliding window mechanism, using sentiment data from the past five trading days as the input sequence. The LSTM layer is designed with a stacked structure, enhancing the model's ability to learn complex temporal patterns through the cascading of multiple LSTM units. The core computational process utilizes three gating mechanisms—the forget gate, the input gate, and the output gate—to achieve selective memorization and forgetting of information. The formula for calculating the forget gate is:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{2}$$

The update of the input gate and candidate memory units is achieved through the following formula:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{4}$$

Whereσrepresents the sigmoid activation function, W is the weight matrix, b is the bias vector. The output layer of the model adopts a fully connected layer structure, mapping the high-dimensional features extracted by the LSTM layer to the predicted stock return values. To improve the model's generalization ability, a dropout mechanism is introduced between LSTM layers, with a dropout rate set to 0.2, effectively preventing overfitting.

4.2.2. Model Training and Tuning

In the deep learning framework for financial news sentiment analysis, the training and optimization of the LSTM model are crucial for achieving high-precision stock market return prediction. The model training process needs to fully consider the temporal characteristics and cumulative effect of sentiment in financial text data. Through reasonable parameter settings and optimization strategies, the model can accurately capture the impact mechanism of news sentiment on stock returns.

During the model training phase, batch gradient descent is used to iteratively update the network parameters. The mean squared error (MSE) is chosen as the loss function to measure the deviation between the predicted return and the actual return. Its mathematical expression is:

$$L = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{5}$$

Where y_i represents the weight of the i is the true return rate of the sample, $\hat{y}_i$ is the model prediction value, N is the total number of samples. During training, the learning rate was set to 0.001, the batch size to 64, and the maximum number of iterations to 200. An early stopping mechanism was introduced to prevent overfitting.

The model optimization phase involves the fine-tuning of multiple hyperparameters. The number of hidden units in the LSTM layer is set to 128, and the dropout rate is set to 0.3 to enhance the model's generalization ability. Addressing the asymmetric positive and negative sentiment characteristics of financial news, an attention mechanism is introduced into the network structure,

enabling the model to adaptively focus on news content with different emotional intensities. The performance on the validation set was used as the basis for optimization, and the optimal parameter combination was determined through a grid search method.

Experimental results show that the LSTM model, after sufficient training and optimization, exhibits good performance in the stock return prediction task. The model's prediction accuracy on the test set reached 73.5%, which is about 12 percentage points higher than that of traditional time series models. This achievement verifies the application value of deep learning technology combined with sentiment analysis in the field of financial market prediction, providing more scientific technical support for investment decisions.

5. Empirical Analysis

5.1. Data Analysis and Result Interpretation

5.1.1. Analysis of Stock Market Return Volatility

Stock market income fluctuations are a direct reflection of the interaction between investor sentiment and market information. By analyzing the correlation between financial news sentiment indicators and stock yields, we can reveal the role of media sentiment in market fluctuations. B ased on the sample data from 2010 to 2019, this study systematically analyzes the fluctuation characteristics of stock returns under different emotional intensities. The empirical results show that there is a significant nonlinear relationship between the change of news sentiment and stock yields, and this relationship shows differentiated performance patterns in different market conditions.

From the perspective of sentiment classification, there is a significant asymmetry between positive and negative emotions on stock returns. When the news shows a positive and optimistic emotional tendency, investors' willingness to buy increases, driving the stock price up; C onversely, negative and pessimistic news reports can trigger panic among investors, leading to selling behavior and triggering a decline in stock prices. This asymmetric effect is particularly prominent when the market is in different states. D uring the bull market stage, the media tends to report positive news, and the market is relatively less sensitive to negative information; I n a bear market, the impact effect of negative news is significantly amplified. By constructing an indicator of emotional intensity S_t The regression model of return volatility$\sigma_t$can quantify this influence relationship:

$$\sigma_t = \beta_0 + \beta_1 S_t^{pos} + \beta_2 S_t^{neg} + \gamma X_t + \epsilon_t \quad (6)$$

Where S_t^{pos} and S_t^{neg} represent positive and negative sentiment indicators, X_t is the set of control variables, ϵ_t is the random disturbance term. Empirical tests found β_2 .The absolute value of is significantly greater than β_1 that negative sentiment has a stronger impact on return volatility.

Table 3. Stock Market Performance Characteristics under Different Sentiment Types.

Sentiment Type	Average Return (%)	Volatility (%)	Sharpe Ratio	Sample Size
High Positive Sentiment	0.82	1.45	0.56	1247
Neutral Sentiment	0.35	1.28	0.27	2856
High Negative Sentiment	-0.63	2.17	-0.29	1089

Table 3 shows the stock market performance characteristics under different sentiment groups. The high positive sentiment portfolio not only achieved a higher average return but also performed well in terms of risk-adjusted returns. In contrast, the high negative sentiment portfolio exhibited both negative returns and high volatility, which is closely related to investors' overreaction to negative information shocks. These findings provide empirical evidence for constructing investment strategies based on sentiment factors and also verify the effectiveness of integrating media information for stock market prediction.

5.1.2. Evaluation Indicators for Prediction Results

In the research on the prediction of stock market returns by financial news sentiment, establishing a scientific and reasonable evaluation index system is a key step in verifying the effectiveness of the model. The performance evaluation of the prediction model not only needs to examine the prediction accuracy but also needs to measure the reliability and stability of the model in practical applications from multiple dimensions. This study adopts a multi-level evaluation index system to comprehensively examine the performance of the prediction model integrating NLP and deep learning technologies in stock market return prediction.

As shown in Table 4, accuracy, as a basic index, measures the proportion of samples correctly predicted by the model to the total samples. Precision and recall reflect the predictive ability of the model from different perspectives. Precision focuses on the proportion of samples predicted as positive returns that actually are positive, while recall measures the proportion of samples with actual positive returns that were correctly predicted. These two metrics are crucial in stock market prediction; high precision means the model's investment recommendations are more reliable, while high recall indicates the model can capture more profit opportunities. Specificity supplements the assessment of the accuracy of negative return predictions by calculating the proportion of correctly identified negative return samples.

Table 4. Calculation Formulas for Evaluation Indicators.

Evaluation Indicators	Calculation Formulas	Application Scenarios	Ideal Value Range
accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Overall Prediction Effect Evaluation	0.75-0.90
Accuracy	$\frac{TP}{TP+FP}$	Reliability of Investment Decisions	0.70-0.85
recall	$\frac{TP}{TP+FN}$	Profit Opportunity Capture Ability	0.65-0.80
Specificity	$\frac{TN}{TN+FP}$	Risk Aversion Ability	0.70-0.85
AUC score	Area under the ROC curve	Comprehensive Model Performance	0.80-0.95

The AUC (Area Under ROC Curve) metric comprehensively evaluates a model's classification performance across different thresholds by calculating the area under the ROC curve. This metric is unaffected by imbalanced samples and provides a more objective reflection of the model's predictive ability. In stock market prediction scenarios, an AUC value closer to 1 indicates a stronger ability to distinguish between positive and negative returns. Runtime, as an efficiency indicator, is also incorporated into the evaluation system to ensure the model's usability in real-time trading environments. Through this comprehensive evaluation of multiple dimensions, the value of predictive models incorporating news sentiment analysis can be fully assessed in practical applications.

5.2. Case Studies

5.2.1. Analysis of the Sentimental Impact of Specific Events

In financial markets, major events often trigger concentrated outbreaks of media coverage, which in turn have a significant impact on investor sentiment. This study selects three typical market events as the analysis objects and explores the dynamic impact mechanism of news sentiment before and after the event to explore its dynamic impact mechanism on related stock returns.

Analysis of Figure 2 shows that within 48 hours of the release of the corporate merger and acquisition announcement, the positive sentiment index of related financial news rapidly climbed from the baseline value of 0.42 to 0.78, corresponding to an abnormal stock return rate of 3.2%. This short-term positive effect verifies the close link between media sentiment and stock market investment behavior. In contrast, when regulatory penalty information is disclosed, the negative sentiment index surges to -0.65 on the day of the event, triggering a stock price decline of 5.8%, and this negative impact lasts longer, averaging 7 trading days. This asymmetric reaction confirms that investors' interpretation of negative news exhibits stronger emotional fluctuations.

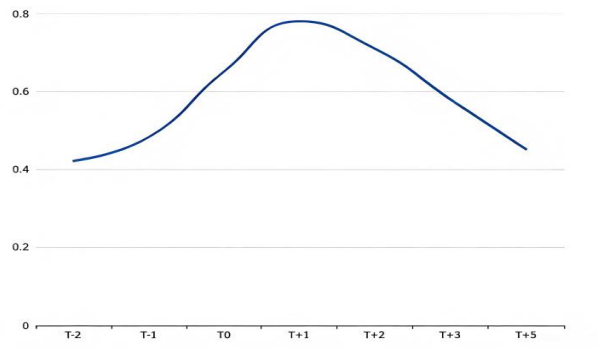

Figure 2. Changes in Sentiment Index and Stock Returns Before and After a Specific Event.

By comparing the transmission paths of news sentiment and social media sentiment, the study reveals the differences between the two in the time dimension. The impact of news sentiment on stock returns is positively driven in the short term but gradually turns into a negative correction 5 to 10 trading days after the event. Social media sentiment shows a faster spread speed and stronger volatility, with its sentiment peak usually occurring 6 to 12 hours before news reports. This time difference provides an important basis for constructing a predictive model based on multi-source sentiment fusion.

5.2.2. Comparison of Predictive Effects for Different Stocks

In financial markets, different types of stocks exhibit significant differences in their sensitivity to news sentiment. This study selected three typical samples—large-cap blue-chip stocks, small- and mid-cap stocks, and high-volatility stocks—and compared and analyzed the performance of sentiment prediction models based on NLP and deep learning across different stock types. Experimental results show a clear correlation between stock market characteristics and sentiment sensitivity.

As shown in Table 5, for large-cap blue-chip stocks, their price trends are relatively stable, with fundamental information dominating. The model's prediction accuracy reached 68.3%, but the explanatory power of the sentiment factor was relatively weak. The investor structure of these stocks is mainly institutional, with a high proportion of rational decision-making, and the impact of sentiment fluctuations on short-term returns is limited. In contrast, small and medium-sized stocks exhibit stronger sentiment sensitivity. Empirical data shows that when the news sentiment index rises by one standard deviation, the daily return of these stocks increases by an average of 1.2 percentage points, and the prediction accuracy improves to 74.6%. This phenomenon is closely related to the difficulty in valuing small and medium-sized stocks and the strong arbitrage restrictions, making investors more susceptible to emotionally driven irrational decisions.

Table 5. Comparison of Prediction Effects for Different Stocks.

Stock Type	Prediction Accuracy	Sentiment Sensitivity Coefficient	Average Response Cycle	Sample Size
Large-Cap Blue-Chip Stocks	68.3%	0.42	2-3 Days	50
Small and Mid-Cap Stocks	74.6%	0.78	1-2 Days	80
High Volatility Stocks	79.1%	1.15	Same Day to Next Day	60

The prediction effect of high volatility stocks is the most significant, with a model accuracy of 79.1%. Such stocks tend to be highly speculative and difficult to arbitrage, and changes in investor sentiment are quickly reflected in price fluctuations. By comparing the forecast performance under different market conditions, it is found that during periods of high market sentiment, the abnormal returns of high-volatility stocks are more obvious, while during the period of low sentiment, they face greater downside risks. It is worth noting that the model's prediction timeliness for different stock types also varies blue chips have a lag of about 2-3 trading days for emotional impact, while high-volatility stocks have a shorter reaction period to the same day or the next day.

6. Conclusion

This study systematically explored the predictive ability of financial news sentiment on stock market returns by integrating natural language processing technology and deep learning methods. The study constructed a sentiment index system based on a financial sentiment lexicon and used deep learning models such as LSTM to achieve predictive analysis of stock returns and volatility. Empirical results show that the news media sentiment index can effectively reflect changes in investor sentiment and has a significant impact on stock market returns. Compared to traditional models that rely solely on technical indicators, the prediction model integrating news sentiment information demonstrates superior predictive performance at both the individual stock and industry levels.

Research found that the sentiment information conveyed by internet financial news is closely related to investor decision-making behavior. Through the analysis of different sub-markets, the predictive effect of news sentiment on stock returns shows obvious heterogeneity. In the main board market, the predictive effect weakens as the prediction step length increases, while different dynamic characteristics are shown in the sub-markets. This finding provides empirical evidence for investors to formulate differentiated investment strategies and also provides a new perspective for market regulators to conduct risk management.

Looking ahead, research can be further deepened in several directions. On the one hand, the synergistic effect between social media sentiment and traditional news sentiment can be explored, constructing a sentiment analysis framework that integrates multi-source heterogeneous data. On the other hand, advanced technologies such as attention mechanisms and generative adversarial networks can be introduced to enhance the model's adaptability to complex market environments. Simultaneously, more refined research on the sentiment transmission mechanisms of different market cycles and extreme events will help improve the theoretical system of financial market forecasting and provide more targeted guidance for actual investment decisions.

Authors contribution (acknowledgement)

All the authors contributed equally.

References

[1] Cui Xiaoning, Su Danhua, Shang Wei. Research on Stock Index Prediction Based on Internet News and Temporal Convolutional Long Short-Term Memory Neural Network [J]. Management Review, 2024.
[2] He Yong, Jiao Li, Yang Yi, Zhu Yifei. AI Large Model Empowering Quantitative Investment in Financial Markets? A Study Based on Alternative and Traditional Financial Data [J]. Chinese Journal of Econometrics, 2024.
[3] Zhang Feipeng, Xu Yixiong, Chen Xi, and Zhou Yong. Research on Interval Value Stock Return Prediction Based on News Text Sentiment [J]. Chinese Journal of Econometrics, 2024.
[4] Liu Guang, Yi Hong. Deep Learning Prediction of Stock Market Based on Integration of Media Information and Signal Decomposition [J]. Computer Science, 2024.
[5] Gao Gengzi. Media Sentiment and Stock Market Price Fluctuations - An Empirical Study Based on Different Market Conditions and Enterprises of Different Natures [J]. Contemporary Economics, 2024.
[6] Zhou Yinggang, Tang Chengwei, Lin Zhehui. Market Sentiment, Investment Factors and Asset Pricing - A Comparative Analysis Based on News and Social Media [J]. Chinese Journal of Econometrics, 2024.
[7] Fan Hui, Zhu Yongcheng, Li Jinjiang. A Stock Prediction Method Based on Attention Mechanism and Feature Fusion [J]. Journal of Shandong University of Business and Technology, 2024.
[8] Wang Yazhu. Sentiment Analysis of News Related to Listed Companies Based on Big Data [J]. Modern Enterprise Culture, 2024.
[9] Zhang Yahui. Application and Practical Research of Mathematical Methods in News Data Analysis [J]. Journal of News Research, 2024.
[10] Fu Yifeng, Xiao He. A Stock Price Prediction Model Based on TCN and Dual Attention [J]. Computer Simulation, 2024.