electronics

MDPI

Article

A Deep Learning Approach with Extensive Sentiment Analysis for Quantitative Investment

Wang Li, Chaozhu Hu and Youxi Luo *

School of Science, Hubei University of Technology, Wuhan 430068, China; 102112268@hbut.edu.cn (W.L.); 20140035@hbut.edu.cn (C.H.)
* Correspondence: 20051038@hbut.edu.cn

Abstract: Recently, deep-learning-based quantitative investment is playing an increasingly important role in the field of finance. However, due to the complexity of the stock market, establishing effective quantitative investment methods is facing challenges from various aspects because of the complexity of the stock market. Existing research has inadequately utilized stock news information, overlooking significant details within news content. By constructing a deep hybrid model for comprehensive analysis of historical trading data and news information, complemented by momentum trading strategies, this paper introduces a novel quantitative investment approach. For the first time, we fully consider two dimensions of news, including headlines and contents, and further explore their combined impact on modeling stock price. Our approach initially employs fundamental analysis to screen valuable stocks. Subsequently, we built technical factors based on historical trading data. We then integrated news headlines and content summarized through language models to extract semantic information and representations. Lastly, we constructed a deep neural model to capture global features by combining technical factors with semantic representations, enabling stock prediction and trading decisions. Empirical results conducted on over 4000 stocks from the Chinese stock market demonstrated that incorporating news content enriched semantic information and enhanced objectivity in sentiment analysis. Our proposed method achieved an annualized return rate of 32.06% with a maximum drawdown rate of 5.14%. It significantly outperformed the CSI 300 index, indicating its applicability to guiding investors in making more effective investment strategies and realizing considerable returns.

Keywords: quantitative investment; sentiment analysis; stock prediction; deep learning; LSTM; Transformer

Citation: Li, W.; Hu, C.; Luo, Y. A Deep Learning Approach with Extensive Sentiment Analysis for Quantitative Investment. *Electronics* **2023**, *12*, 3960. https://doi.org/10.3390/electronics12183960

Academic Editor: Ioannis Hatzilygeroudis

Received: 28 August 2023
Revised: 18 September 2023
Accepted: 18 September 2023
Published: 20 September 2023

1. Introduction

With the development of data science and deep learning, quantitative investment, as a data-driven and model-based investment approach, has been playing an increasingly important role in the investment field. Research on using computational methods for predicting financial markets and stock prices has gained growing attention, where quantitative investment methods based on computer technology have taken on a more significant role. Developing effective quantitative investment methods holds critical significance for guiding investors in their strategic decisions. However, accurately quantifying investment has remained a challenging problem because of the myriad complex factors influencing the stock market.

Stock prediction is a key component of quantitative investment. Early prediction research primarily applied statistical methods, feature engineering techniques [1], and fuzzy theory [2] to model historical prices. Subsequently, various machine learning methods were widely applied for this task [3–7]. However, these methods exhibit limitations such as sensitivity to data, poor prediction stability, and weak multi-source data fusion capabilities. With the rapid growth of stock trading data volume, traditional machine learning methods

began revealing shortcomings in modeling capability, failing to efficiently capture information within vast datasets. Deep-learning-based methods were then developed [8–26]. Ahmed et al. [19] summarized the application of deep learning methods in stock price prediction. Among them, long short-term memory (LSTM) [27] and convolutional neural network (CNN) [28] have been used to model long sequence patterns [12,13]. In terms of input data sources for models, in addition to price information, text sentiment has also been proven to provide predictive information on stock prices [29–47]. For example, Shynkevich et al. [29] used different categories of financial news to predict price changes of heavyweight stocks in the healthcare sector. However, most of these methods used text sentiment as the sole input, neglecting other direct information such as technical indicators, thereby failing to capture the strong association between textual sentiment and price information and making the models susceptible to noise. Recently, some research has focused on fusing stock price and text sentiment, such as Zhang et al. [32] combining technical indicators and online news to develop a stock price prediction framework and Jing et al. [38] proposing a hybrid model for stock price prediction combining deep learning methods and sentiment analysis. In these studies, user commentary on social media platforms, such as Twitter and StockTwits, has been widely leveraged for stock prediction. Herrera et al. [43] extracted investor sentiment from Twitter via natural language processing and incorporated deep learning models to forecast stock returns. Despite the proven ability of such textual content to improve stock market prediction, the user base across social media platforms may vary, and the posted information could be misleading. Therefore, the predictive quality of such information requires consideration. Additionally, the prediction results may differ when using data gathered from various platforms. As noted by Ashtiani et al. [45], while numerous studies have analyzed social media to predict the stock market, less attention has been paid to utilizing news content. Additionally, advanced natural language processing methods like Transformer have not been fully exploited for stock market prediction. With the progression of research in this domain, there is increasing acknowledgment regarding the critical role of news in stock market prediction. Ma, Y et al. [47] proposed a model that incorporated numerical features and market-driven news sentiments of target stocks, together with news sentiments of related stocks. While these methods simultaneously consider textual sentiment and historical stock prices, many of them overlook the temporal delay between textual sentiment and stock price changes, leading to a temporal matching issue. Additionally, they only used news titles to extract sentiment from text. To our knowledge, we are the first to propose extracting semantic sentiment from both news titles and summarized news content, addressing the issue of prior methods omitting rich, detailed sentiment information from the textual content.

By fully considering both the title and content textual aspects and leveraging deep learning, we propose a novel stock investment approach based on a deep hybrid model that processes and integrates multi-source heterogeneous data (stock technical factors and news semantic representations). The framework of our proposed method is illustrated in Figure 1. Our approach first employs fundamental analysis to rank and filter stocks with investment analysis (Figure 1d). Subsequently, we construct technical factors from historical trading data and utilize a language model to extract semantic information containing both headlines and content from news articles (Figure 1c). These two types of features are fused to create global features, forming the foundation for a deep learning model driven by momentum trading strategies (Figure 1b). Benefiting from the accurate and stable stock price prediction of our model, our method can offer investors rational and comprehensive investment guidance, assisting them in formulating investment strategies. Our contributions can be summarized in two aspects.

The first is enhanced utilization of news information. Diverging from other studies that solely use text headlines, to our knowledge, we are the first to comprehensively consider both news titles and summarized contents. This approach constructs more accurate, rich, and comprehensive semantic features from two dimensions, facilitating the thorough utilization of news information.

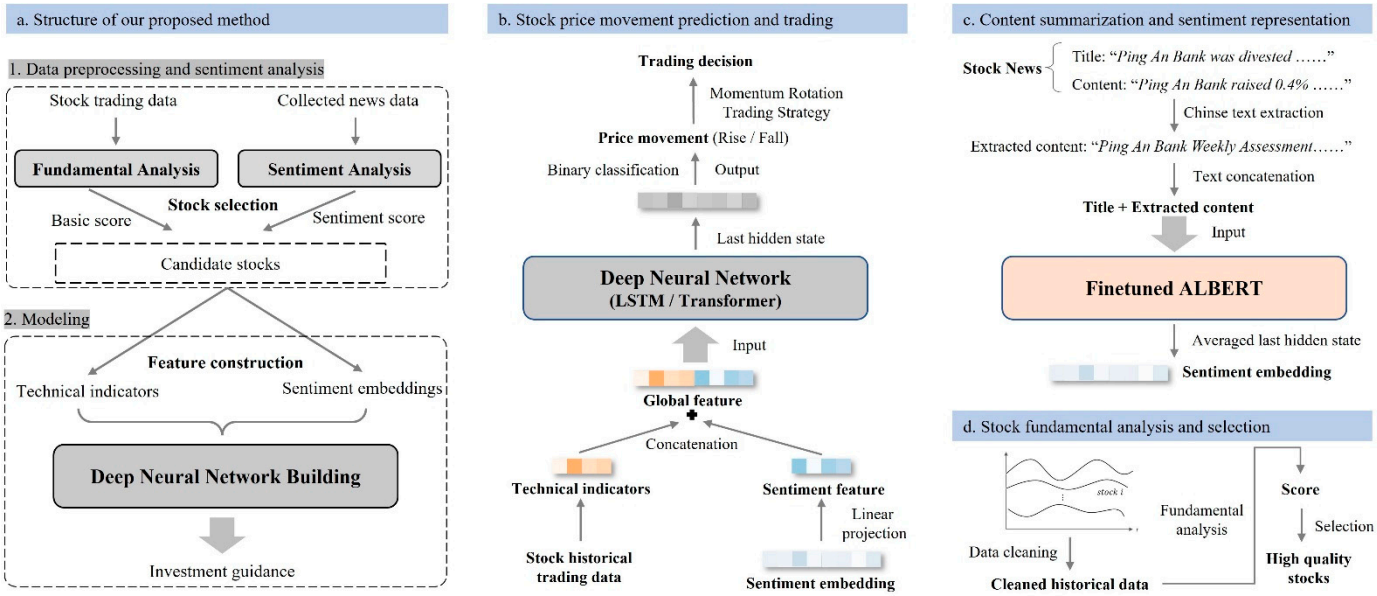

Figure 1. The framework of our proposed approach. (**a**) The structure and workflow of our proposed method. (**b**) The architecture of our proposed hybrid deep neural model for stock price movement prediction and trading decision. (**c**) The generation of sentiment embedding from the news. First, a large pretrained language model was utilized to extract summarized content, which was then concatenated with title and input to a fine-tuned ALBERT model to obtain sentiment embedding. (**d**) Stock fundamental analysis and selection. With the scores obtained from fundamental analysis, high-quality stocks can be preliminarily screened.

The second aspect is a novel hybrid model for accurate stock prediction and trading decisions. On the one hand, we are the first to employ large-scale pretrained language models to extract semantic features, enabling the retrieval of fine-grained and highly accurate semantic information from news. On the other hand, LSTM and Transformer, adept at handling long-series data, are employed to model the multi-source heterogeneous features. The hybrid of these two models results in a novel deep hybrid model that can guide investment decisions and achieve stable returns.

The structure of our paper is as follows. Section 2 introduces our proposed method in detail. Section 3 presents extensive experiments demonstrating the performance of our method. Section 4 concludes the paper and provides discussions. The workflow of our proposed method is presented in Figure 1a, which consists of two stages. The first stage is data preprocessing and sentiment analysis (Figure 1c), and the second stage is model building and evaluation (Figure 1b).

2. Method

2.1. Stock Trading and News Data

The historical trading data of over 4000 stocks listed on the Chinese stock market were used for quantitative learning research in this paper. These stocks cover various major industries, including IT, electronics, etc. These stocks can be categorized into four segments: main board, ChiNext board, small and medium-sized enterprises (SME) board, and STAR market, with stock codes starting with 00, 30, 60, and 68, respectively. Among them, the main board and ChiNext board belong to the Shenzhen Stock Exchange (SZSE), while the SME board and STAT market belong to the Shanghai Stock Exchange (SSE). Specifically, we collected three years of stock trading data from 1 January 2020 to 30 December 2022. The data was divided using 30 June 2022 as the cutoff date. Trading data before that date was used as training and validation sets, and the remaining was used as the test set. In total, there were 4565 stocks, and their statistical information is presented in Supplementary Figure S1a.

Historical trading data primarily reflect market supply and demand dynamics, while news information is one of the significant influencing factors in the financial market. News information contains rich information about market trends, public sentiment, and more, thus providing a comprehensive reflection of the financial market. It serves as a valuable supplement to historical stock trading data. The news data for this paper were sourced from East Money (www.eastmoney.com) (accessed on 15 May 2023), one of China's most popular financial information providers. This platform offers comprehensive and accurate stock news. Corresponding to the dates of historical trading data, news reports from those days were collected if available, including the news date, title, and content. In total, 3,889,380 news articles were collected, with statistical results shown in Supplementary Figure S1b. An example of the collected news data is provided in Supplementary Table S1.

2.2. Stock Selection via Fundamental Analysis

When making investment decisions, stocks with undervalued stocks should be prioritized. These stocks have limited downside potential and can generate higher profits with lower risks. The value of a stock can be assessed from the valuation dimension and quality dimension. Valuation indicators are directly calculated from financial metrics, which are frequently used by investors to assess valuation levels. Unlike valuation indicators, quality dimension metrics measure the market valuation of a company from a market perspective and derive indicators from the company's financial statements. For fundamental analysis of stocks, we select indicators from the two dimensions, including price-to-earnings ratio (PE), price-to-book ratio (PB), price-to-sales ratio (PS), return on equity (ROE), profit margin (PM), and earnings quality (IN). These indicators form the foundation of the indicator system. Among them, PE, PB, and PS measure a company's value and are negative indicators. Lower values for these three indicators signify an undervalued company. On the other hand, ROE reflects profitability and, together with PM and IN, measures company quality and potential. These three are positive indicators, with higher values indicating better company quality. This paper obtained and organized the above indicators from the RESSET and CSMAR databases. Based on these indicators, the scores for each stock are calculated in two steps. First, the data are standardized. For positive indicators, the standardization formula is as follows:

$$x_{i,j}^{*} = \frac{x_{i,j} - x_j^{min}}{x_j^{max} - x_j^{min}} \tag{1}$$

For negative indicators, the standardization formula is as follows:

$$x_{i,j}^{*} = \frac{x_j^{max} - x_{i,j}}{x_j^{max} - x_j^{min}} \tag{2}$$

After undergoing the standardization process, all six standardized indicators have a positive impact on stock financial characteristics. In the second step, a weighted average approach is used to calculate the stock scores. The indicators PE, PB, PS, ROE, PM, and IN are assigned weights of 0.15, 0.15, 0.15, 0.3, 0.15, and 0.1, respectively. The standardized indicator values and scores for some stocks are illustrated in Table 1.

Table 1. Indicators and Scores for a Subset of Stocks.

CODE	PE	PB	PS	ROE	PM	IN	SCORE
000820	0.4385	0.9855	0.9970	1.0000	1.0000	0.0133	0.8145
300226	0.4337	0.9983	1.0000	0.7974	0.3132	0.3378	0.6848
600755	0.4362	0.9999	1.0000	0.7973	0.3133	0.2616	0.6777

2.3. Data Preprocessing

The raw stock historical trading data features are high-dimensional, which could potentially contain noise and redundant information. Not all features are equally important for stock prediction. Therefore, this paper extracts stock technical factors to select important

features with predictive capability. This aids in enhancing the model's efficiency and generalization ability. We computed a series of 16 technical factors for each stock, as summarized in Table 2. Details of these factors can be found in the Supplementary Information.

Table 2. Summary of technical indicators.

Technical Indicators	Abbreviation
Moving average (5)	MA (5)
Moving average (30)	MA (30)
Moving average (60)	MA (60)
Exponential moving average (5)	EMA (5)
Exponential moving average (30)	EMA (30)
Exponential moving average (60)	EMA (60)
Moving average convergence/divergence (6, 15, 6)	MACD (6, 15, 6)
Moving average convergence/divergence (12, 26, 9)	MACD (12, 26, 9)
Moving average convergence/divergence (30, 60, 30)	MACD (30, 60, 30)
Relative strength index (14)	RSI (14)
Williams' %R (14)	WILLR (14)
Momentum index (14)	MOM (14)
Chande momentum oscillator (14)	CMO (14)
Ultimate oscillator (7, 14, 28)	ULTOSC (7, 14, 28)
On balance volume	OBV
Chaikin A/D oscillator (3, 10)	ADOSC (3, 10)

For the collected news data, a data cleaning process is carried out, which involves the removal of special characters. Text sequences need to be tokenized before being input to the model. We build a corpus and vocab with the sequences. Each sequence is denoted as an embedding $E_i = \{a_1, a_2, \ldots, a_L\}$, where $a_i \in R^d$ is the d-dimensional embedding of the $i-th$ token, and L is the length of the sequence.

In the above section, after conducting fundamental analysis, we obtained fundamental scores for each stock. Using a threshold of 0.65, we preliminarily filtered high-quality stocks, resulting in 4129 stocks. Technical factors are used as basic features and stock price changes as prediction targets. A time-series dataset of stock technical factors is created using a sliding window approach. A time window T_w is set so that for day t ($t > T_w$), and the technical factor features from the past days equal to the window size are taken as the time-series features for day t. The prediction target was whether the stock would rise on the $t + 2$ day, i.e., calculating the difference between day $t + 2$ and $t - 1$. If the difference is greater than 0, it indicates a rise, and the label is set to 1. Otherwise, it indicates a fall, and the label is set to 0. Using the stock's rise or fall on the $t + 2$ day instead of the t day as the target takes into account considerations such as trading execution delay, data availability, and avoiding short-term market noise.

2.4. Deep Hybrid Model

2.4.1. News Representation and Sentiment Analysis

The objective of sentiment analysis on news is to obtain sentiment scores and semantic representations. The sentiment scores will be further utilized for stock screening, while the semantic representations will serve as input features for constructing the subsequent deep-learning-based quantitative investment model. Unlike previous methods that focused solely on sentiment analysis of news headlines, we comprehensively consider both the headlines and the content. In general, sentiment analysis on news consists of two steps: first, news content summarization; second, news representation and sentiment analysis.

Stock news headlines usually provide a brief overview, while news content can offer more comprehensive information. Therefore, in the first step, we utilize state-of-the-art natural language processing (NLP) methods to summarize news content. We employ Randeng-Pegasus [48], which is a large-scale pretrained semantic model specialized in generating text summaries. It is fine-tuned on a dataset comprising around 4 million sam-

ples from seven Chinese domains for text summarization based on the Pegasus-large [49] model. Compared to traditional extractive summarization methods, Randeng-Pegasus can generate novel and coherent summary content. The model's architecture is illustrated in Figure 2.

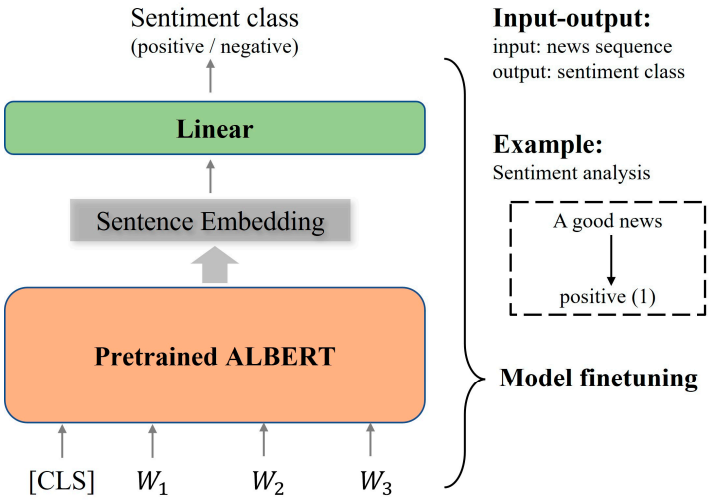

Figure 2. Fine-tuning process and examples of sentiment analysis. The model takes preprocessed text sequences as input and outputs probabilities of positive/negative sentiment. If the probability of positive sentiment is greater than 0.5, it is considered positive.

Transformer has exhibited outstanding performance in many research fields. Its encoder consists of multiple layers, each of which is further composed of two sub-layers: multi-head self-attention mechanism, and position-wise fully connected feed-forward network. Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions [50]:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_1)W^O \tag{3}$$

where single-head scaled dot-product attention is computed as:

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) = softmax\left(\frac{QK^T}{\sqrt{d_i}}\right)V \tag{4}$$

where the projections are parameter matrices $W_i^Q \in R^{d_1 \times d_2}$, $W_i^K \in R^{d_1 \times d_2}$, $W_i^V \in R^{d_1 \times d_2}$, and $W^Q \in R^{d_1 \times d_2}$. In addition to the same two sub-layers in the encoder layer, the Transformer decoder layer inserts a third sub-layer, which performs masked multi-head attention. In this work, similarly to [48], we employ $N = 16$ identical layers and $h = 16$ heads; for each of these, we use $d_q = d_k = d_v = d_{model}/h = 64$ to build Transformer encoder and decoder. After tokenization, the news contents are input into the Randeng-Pegasus model to obtain summarized contents.

In the second step, we consider both news titles and summarized content to represent news content and predict news sentiment polarity. Traditional word vector methods [38,51] have been widely used in sentiment analysis on news in previous research. However, these methods have limitations in weak contextual recognition and limited ability to represent semantic relationships. In this study, we follow a new paradigm in the field of NLP known as transfer learning, where a large language model is pretrained on a large corpus and then fine-tuned on specific tasks to achieve better performance. Specifically, we use a pretrained ALBERT (A Lite BERT) [52], a Chinese sentiment prediction model. It is then fine-tuned on the ChnSentiCorp (Chinese Sentiment Corpus) dataset to enhance its ability to predict sentiment in Chinese news. Finally, the concatenated texts consisting of news

headlines and content summaries are input into the fine-tuned ALBERT model to obtain news representations and predict their sentiment polarities.

ALBERT is a lightweight version of the BERT [53] model. The pretrained ALBERT model used in this study is built based on an open-source model available on Huggingface. ALBERT utilizes a Transformer encoder to extract information, with parameters set as $N = 12$, $h = 12$, $d_q = d_k = d_v = 64$. ChnSentiCorp contains a large number of Chinese text samples with corresponding binary sentiment labels. The dataset encompasses texts from various sources like news, covering different domains and topics. The process and an example of fine-tuning the ALBERT model on the ChnSentiCorp dataset are illustrated in Figure 2. Tokenized text sequences are input into the pretrained ALBERT model to obtain sentence representations. After passing through a fully connected layer with a dimension of 384, the model outputs binary sentiment categories. Once the ALBERT model is fine-tuned, it is subjected to end-to-end transfer learning on the collected stock news text dataset from the Chinese stock market to perform sentiment analysis. This process is consistent with the flow illustrated in Figure 2.

2.4.2. Stock Prediction Model

The two main characteristics of stock time-series signals are temporality and sequence nature, similar to text sequence data in NLP. This paper selects representative sequence models, LSTM [27] and Transformer, to model stock signals. LSTM models are adept at learning temporal trends and perform well on moderate-sized datasets. They have a smaller number of parameters and better generalization capability. On the other hand, the Transformer model employs self-attention mechanisms to capture correlations within input sequences and it is capable of capturing long-range dependencies in sequences. The choice of these two models is based on the following considerations: first, both of them excel at handling time-series signals, making them well suited for stock signal modeling. LSTM is effective at capturing and modeling long-term dependencies in stock time series data, whereas Transformer's self-attention mechanism allows the model to simultaneously consider the entire sequence while flexibly focusing on patterns and relationships at different time scales. Second, previous related studies [25,41,43] have shown that these models demonstrate great performance in finance or stock market prediction tasks, making them suitable for quantitative investment. The two models have their own strengths and weaknesses. By contrasting them, we aim to explore the effectiveness of sequence models in quantitative investment.

LSTM units consist of three types of gates: input gate, forget gate, and output gate. Figure 3 illustrates the operation of an LSTM unit at time t. Here, X_t and Y_t represent the current input and output, h_{t-1} and h_t are the previous and current hidden states, and C_{t-1} and C_t are the previous and current cell states, respectively. The hidden state serves as an encoding of information learned from previous input sequences, while the cell state is a crucial component responsible for storing information extracted from past inputs. The forget gate determines how much information from the previous time step's cell state is retained in the current time step, The formula of the forget gate is as follows:

$$f_t = \sigma\Big(W_f \cdot (h_{t-1}, x_t) + b_f\Big) \tag{5}$$

The input gate determines how much of the current model's input vector is stored in the current state of the unit. The calculation involves three steps:

$$i_t = \sigma(W_i \cdot (h_{t-1}, x_t) + b_i) \tag{6}$$

$$\widetilde{C}_t = tanh(W_c \cdot (h_{t-1}, x_t) + b_c) \tag{7}$$

$$C_t = f_t C_{t-1} + i_t \widetilde{C}_t \tag{8}$$

The output gate determines how much information from the current unit is output to the model's output. The calculation formula for the output gate is as follows:

$$o_t = \sigma(W_o \cdot (h_{t-1}, x_t) + b_o) \tag{9}$$

$$h_t = o_t \cdot tanh(C_t) \tag{10}$$

where σ is a sigmoid function, W_f, W_i, W_c, W_o are weight matrices in neurons, and b_f, b_i, b_c, b_o are bias terms.

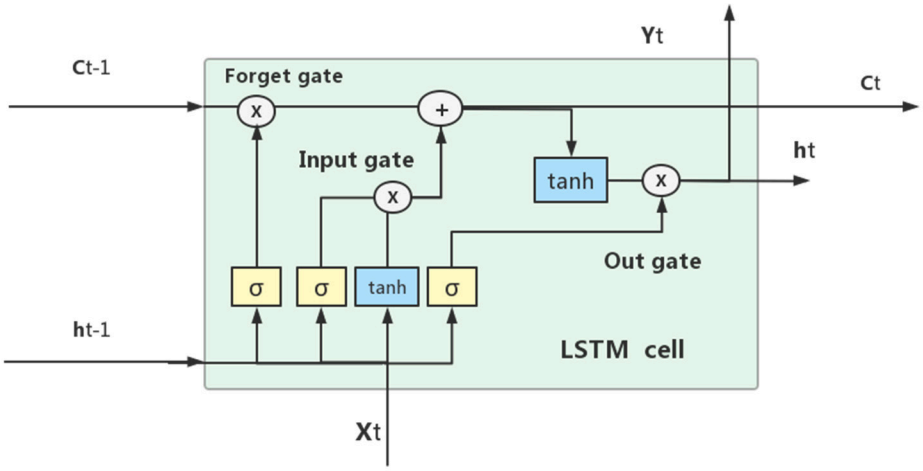

Figure 3. The structure of an LSTM unit.

We employ a modified LSTM with two layers to model the stock time-series signals, with the hidden dimension of each layer set to 32. The fused global features of technical factors and semantic representations are used as model inputs. To balance the importance of the two types of features, their dimensions are unified. First, the 384-dimensional semantic representations are projected to the same 16 dimensions as the technical factors through a fully connected layer. Then, the two types of features are concatenated in the feature dimension to obtain 32-dimensional global features. The framework of using LSTM to predict stock changes is illustrated in Figure 1b.

In addition, as a comparison, a modified Transformer is also utilized, with parameters set as $T = 64, N = 6, h = 8, d_q = d_k = d_v = d_{model}/d_k = 64$. The parameter choices for the two deep learning models are shown in Supplementary Table S2.

2.4.3. Trading Strategy

We propose a simple and effective trading strategy consisting of two main components: a momentum rotation trading strategy for capturing short-term price trends and a profit-taking and stop-loss strategy for risk management and preserving gains. The trading strategy can be outlined in five steps:

1. Select the top 50 stocks with the highest prediction accuracy from the validation set.
2. Calculate the momentum of these selected stocks for momentum rotation trading. Here, momentum is defined as the slope of the 20-day closing price series. The slope for each day is computed by fitting a linear regression to the 20-day closing price sequence using the formula:
$$close = \beta * x + \varepsilon \tag{11}$$
3. Sort the slopes and select the top 60% of stocks to form a portfolio. The deep learning model is then used to predict whether the price will rise on the $T_w + 2$ day. If a stock is predicted to rise, it is bought.

4. For stocks held in the portfolio, if the deep learning model predicts a price decrease, the stock is sold.
5. Implement profit-taking and stop-loss strategies. Stocks are sold if their returns exceed 22% or if their prices decline by 8%.

2.4.4. Backtesting Evaluation Metrics

This study employs the annualized return rate and the maximum drawdown rate as evaluation metrics for backtesting the trading strategy. The annualized return rate (ARR) quantifies the actual gains of investing in an asset into an annualized measure of returns. It reflects the realized returns obtained from investing in an asset over the course of one year. Assuming an investor holds the asset for a period of T_p terms, with an achieved or expected return rate of R_{T_p}, and there are m individual periods within a year, the formula to calculate the ARR for the asset is as follows:

$$ARR = \left[\left(1 + R_{T_p}\right)^{\frac{1}{T_p}} - 1 \right] \times m \tag{12}$$

In practical applications, maximum drawdown (MDD) is frequently utilized to assess the performance of investments. It signifies the magnitude of decline from the peak to the trough of an asset's value within the time interval $(0,\ T_p)$, representing the maximum drop. The mathematical formula for calculating the MDD is as follows:

$$MDD(T_p) = \underset{\tau \in (0,T_p)}{max}\ D(\tau) = \underset{\tau \in (0,T_p)}{max} \left[\underset{\tau \in (0,T_p)}{max}\ P_t - P_\tau \right] \tag{13}$$

The corresponding maximum drawdown rate (MDR) is defined as:

$$MDR(T_p) = \underset{\tau \in (0,T_p)}{max}\ d(\tau) = \frac{MDD(T_p)}{max_{t \in (0,T_p)} P_t} \tag{14}$$

3. Results and Discussion

3.1. News Sentiment Analysis

After fine-tuning, the ALBERT model exhibits a significant enhancement in its predictive ability for sentiment polarity, resulting in precise sentiment prediction and comprehensive semantic representations for stock news. The process of model fine-tuning and its performance are detailed in the Supplementary Information. Utilizing the fine-tuned ALBERT model, sentiment analysis is performed end to end on the collected stock news data. Among the 4129 stocks selected after fundamental analysis, the sentiment polarity of each news article in their training and validation sets is predicted. The proportion of positive news is then calculated, constituting the sentiment score. By ranking stocks based on their sentiment scores and setting a threshold of 0.7, stocks with a proportion of positive news exceeding the threshold are retained. Ultimately, after the threshold-based selection, 577 stocks remain.

To investigate the impact of introducing news content summaries on sentiment polarity prediction, two control groups are constructed: the Title group and the Title + Content group. Table 3 presents the proportions of positive news under different threshold values for these two groups. Notably, the Title + Content group exhibits a lower proportion than the Title group. One possible explanation is that stock news headlines often employ attention-grabbing language and tend to be sensationalized, whereas news content is generally more objective in sentiment. Consequently, the integration of both title and content summary results in a decrease in the proportion of positive news. This reflects the supplementary role of content to title information, and its importance for objective evaluation of polarity. Statistics by stock board are shown in Figure 4. For all four boards, the Title + Content groups have lower positive news ratios than the Title groups, with SME board having the highest positive news ratios overall.

Table 3. Statistics of the proportion of predicted positive sentiment news for different groups.

Threshold	Title	Title + Content
0.9	9.7%	1.6%
0.8	22.9%	6.2%
0.7	39.7%	15.7%

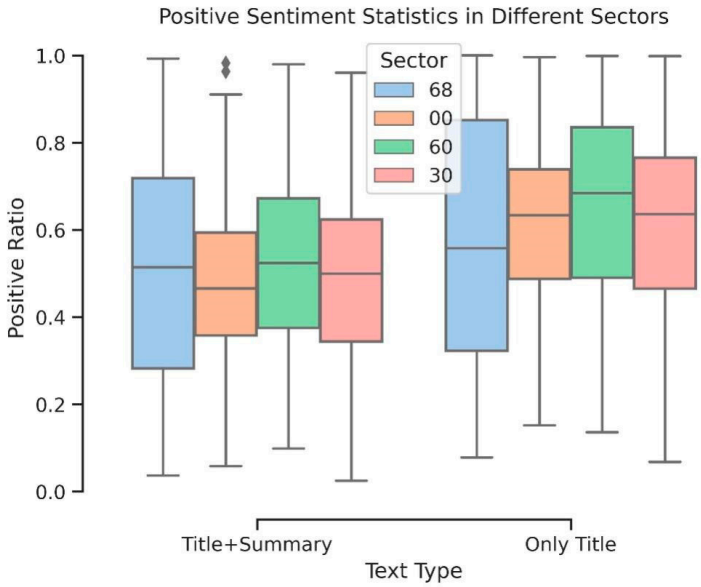

Figure 4. Statistics of the proportion of positive sentiment news for stocks in different boards.

3.2. Stock Prediction Performance

The above results analyze the importance of news content in fully reflecting news information from a statistical perspective. To further investigate the role of news information and the impact of incorporating news content, we set up three controlled experiments: Vanilla, Title, and Title + Content. Accordingly, the features for the three groups are technical factors only, technical factors + news title semantic representations, and technical factors + fused semantic representations from news titles and summarized content. For both the LSTM and Transformer models, stock rise/fall prediction is performed with models trained on the three groups separately. Figure 5 shows the average loss and accuracy curves on the training and validation sets of the two models across the three controlled groups. It can be observed that after 30 rounds of training, the models have mostly converged. There are slight performance improvements in subsequent rounds before oscillating within a range. This indicates the models are not overfitting and have generalization capability. Complete convergence is achieved at around 150 rounds.

Evaluation metrics AUC and Recall are used to assess model performance. Table 4 shows the results of the two models on the 50 stocks with the highest validation accuracy. Under both metrics, LSTM consistently demonstrates significantly better performance over Transformer. Specifically, after introducing semantic representations of news titles and summaries to the technical factor features, the AUC of LSTM and Transformer improved from 83.19% and 70.79% to 85.43% and 78.16%, representing increases of 2.70% and 10.41%, respectively. Their Recall metrics also had significant improvements of 23.95% and 16.57%, reaching 80.94% and 65.02%, respectively. It can also be observed that the Title groups generally outperform the Vanilla groups, except for a minor decrease in LSTM's AUC from 83.19% to 82.69%. This indicates that incorporating news title information alone can also improve prediction accuracy to some extent.

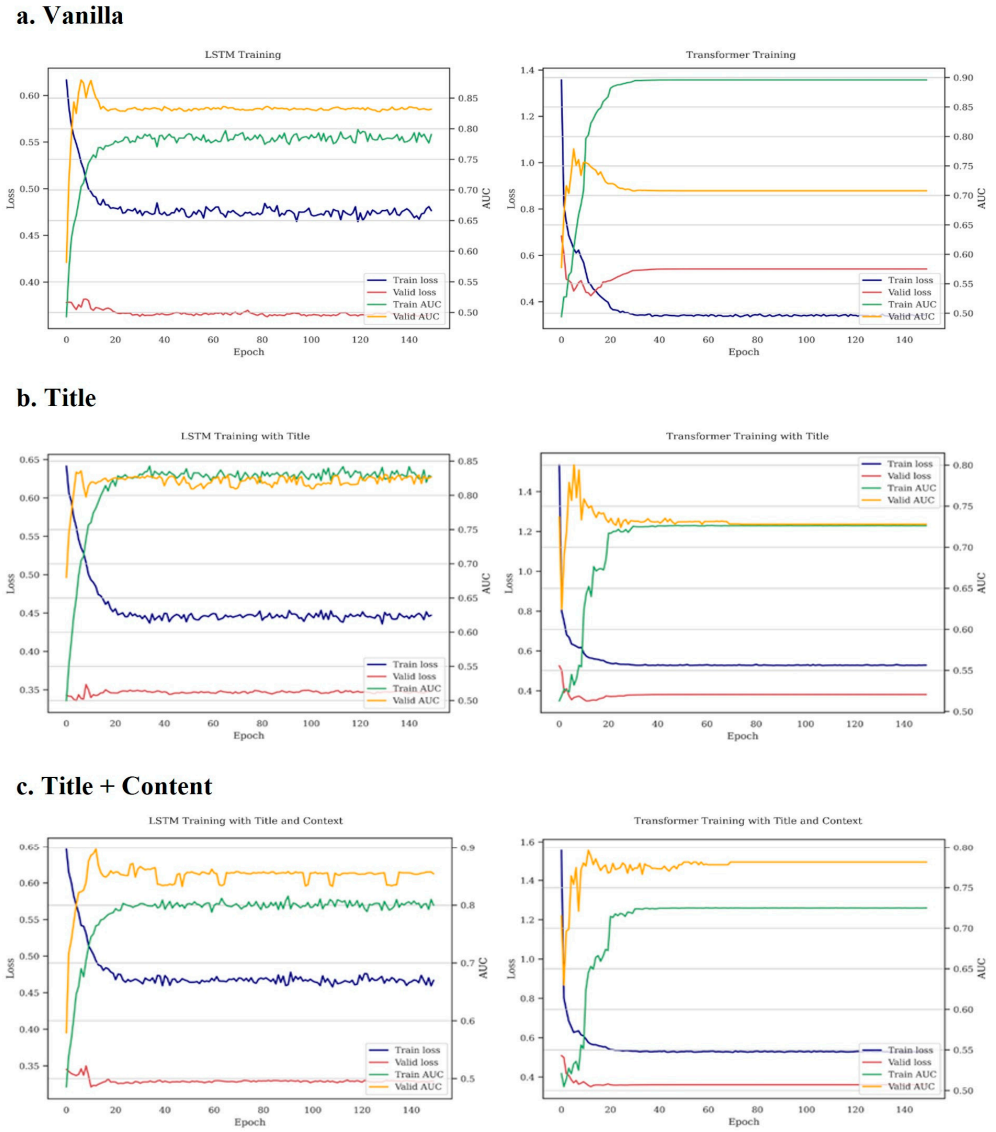

Figure 5. Average loss and accuracy curves of LSTM and Transformer on the train and validation sets across three groups.

Table 4. Evaluation results of stock changes prediction on the validation set.

Model / Group	LSTM		Transformer	
	AUC	**Recall**	**AUC**	**Recall**
Vanilla	83.19%	65.30%	70.79%	55.78%
Title	82.69%	77.67%	72.80%	59.64%
Title + Content	85.43%	80.94%	78.16%	65.02%

3.3. Backtesting Evaluation

To evaluate the effectiveness of the investment strategy, we applied the strategy to the test set to simulate its past performance. In this study, we utilized an initial capital of CNY 100,000 and incurred a transaction fee of 0.025% for backtesting. The backtesting results for the LSTM and Transformer models under three control groups are presented in Figure 6. The results indicate that for both models, when incorporating news titles and summarized contents as knowledge enhancement, the strategy achieved the maximum returns. In this scenario, the ARR of both models significantly increased, from 16.1% and 13.59% to 32.06% and 26.98%, respectively. Furthermore, the LSTM model yielded higher overall returns, reaching 32.06%, whereas the Transformer model exhibited lower maximum drawdown,

indicating more stable returns and reduced risk. This highlights the distinct strengths and weaknesses of the two models. The backtesting evaluation results align with the models' ability to predict stock price trends. Notably, the inclusion of only news title sentiment representation also led to a significant increase in the ARR. However, it was accompanied by an increase in the MDR. This suggests that while news titles can provide Supplementary Information, their content might be limited and biased because of their subjective emotional tone, potentially leading to model instability. In contrast, the comprehensive improvement brought by incorporating both news titles and content summaries reflects the objective information provided by content summaries. This not only further supplements various aspects of stock information but also enhances model robustness and noise resistance through its objective sentiment.

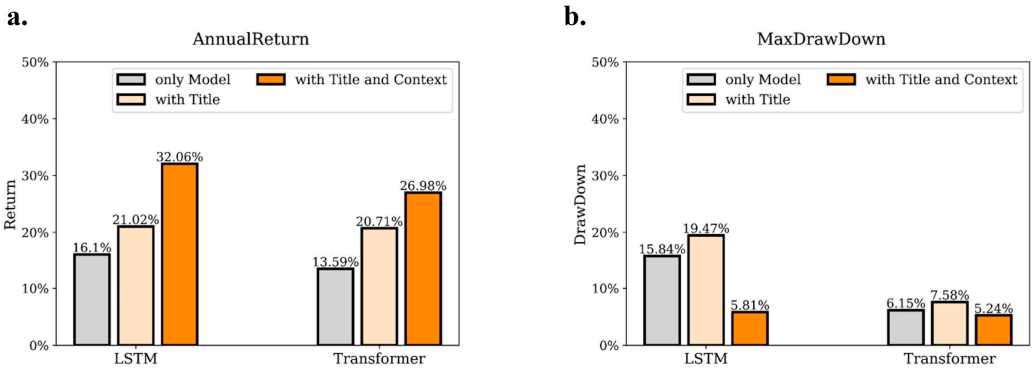

Figure 6. Comparison of backtesting metrics. (**a**) Evaluation with annual return rate; (**b**) evaluation with max drawdown rate.

Figure 7 shows the dynamic changes of annualized return and maximum drawdown over trading time during backtesting in more detail. In each subfigure, the blue, red, and yellow curves at the top, middle, and bottom positions represent the cumulative return curve, drawdown curve, and daily return curve, respectively. Among them, the cumulative return curve reflects the dynamic changes in the long-term returns of the investment strategy, while the drawdown curve and daily return curve depict the short-term performance on the current trading day. It can be observed that the profits and losses of stock investment are constantly fluctuating. The annualized return may start to decrease after reaching a historical peak at some point, then enter fluctuation again. Furthermore, our quantitative investment approach is compared against the CSI 300 index, with the results shown in Figure 8. In each subfigure, the blue and yellow curves represent the cumulative returns of our approach and the baseline, respectively. The results indicate that the proposed approach significantly outperforms the baseline, proving its practical value. This affirms its potential to provide guidance for investors' investments and to yield actual returns.

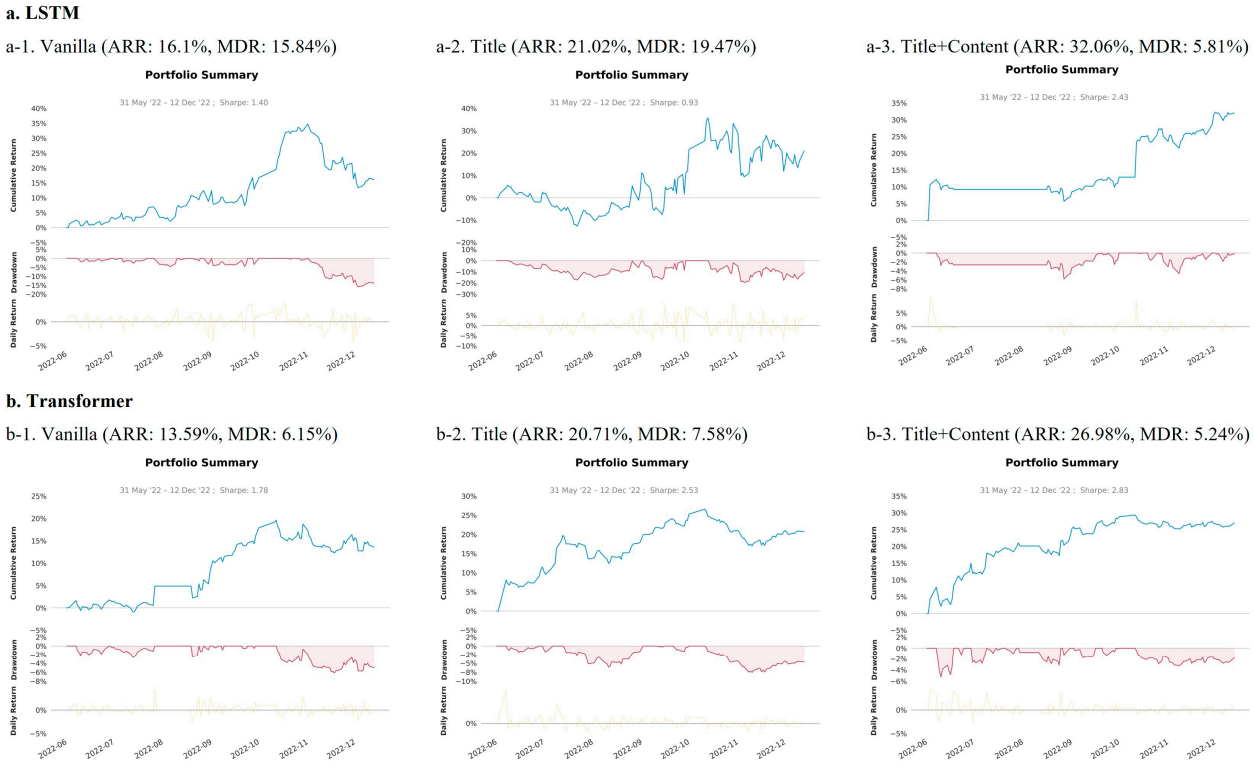

Figure 7. Variation of backtesting metrics over trading time. (**a**,**b**) present the dynamic change of metrics for the LSTM and Transformer models, respectively, across three control groups. Note that each subfigure consists of three parts. The upper part is the cumulative return curve, the middle part is the drawdown curve, and the lower part is the daily return curve. The x-axis denotes the trading time, spanning from June 2022 to December 2022, within the test dataset.

Figure 8. Comparison of ARR between our proposed investment method and the CSI 300 index. Here, (**a**,**b**) illustrate the dynamic comparative results for the LSTM and Transformer models, respectively. Note that for each subfigure, the x-axis represents trading time, and the y-axis denotes cumulative returns for our proposed method and baseline benchmark CSI 300 index. The cumulative returns curves for the corresponding methods are marked in blue and yellow, respectively.

3.4. Backtesting on Separate Markets

The above backtesting evaluation results are estimated across the entire Chinese stock market, which includes two separate stock exchanges, SZSE and SSE. To test the robustness and generalizability of our model, we divided the collected stock data into two subsets, applied our methodology to guide quantitative investments, and conducted backtesting evaluations separately. The results of the backtesting on these two independent stock exchanges are presented in Tables 5 and 6. Our model consistently delivered significant returns for investors with relatively low maximum drawdown rates, reaching 32.49% and 29.92% of ARR on the two markets, respectively. It is noticeable that for both the LSTM and Transformer models, the ARR of the Title control group is higher than that of the Vanilla group, which utilizes only technical indicators. However, the maximum drawdown rate is also higher. In contrast, the Title + Content control group achieved the best performance on both evaluation metrics. This indicates that the integration of news titles alone is insufficient for modeling, and summarized news contents can provide a more objective and comprehensive view of news, helping to capture price patterns more effectively. These findings are consistent with the results above on the entire Chinese stock market. The results from the independent SZSE and SSE markets collectively demonstrate that our model can perform effectively in different stock markets, confirming its robustness and generalizability.

Table 5. Backtesting performance on Shenzhen Stock Exchange (SZSE).

Group	LSTM		Transformer	
	ARR	**MDR**	**ARR**	**MDR**
Vanilla	17.56%	11.19%	15.02%	5.95%
Title	21.66%	7.29%	21.33%	6.27%
Title + Content	32.49%	4.04%	27.29%	3.63%

Table 6. Backtesting performance on Shanghai Stock Exchange (SSE).

Group	LSTM		Transformer	
	ARR	**MDR**	**ARR**	**MDR**
Vanilla	18.95%	21.18%	15.41%	6.45%
Title	24.04%	22.55%	19.89%	7.36%
Title + Content	29.92%	5.46%	26.24%	4.14%

3.5. Performance Comparison with Baselines

To further evaluate the robustness and generalizability of our proposed methods, we conducted an extensive empirical and analytical comparison with established baselines. Specifically, a deep-learning-based strategy (RNN [28]), a machine-learning-based strategy (XGBoost [54]), and four traditional quantitative methods are included. RNN (recurrent neural network) is a neural network well suited for processing sequential data, while XGBoost is an efficient gradient-boosting decision tree model. For both methods, we adopted identical trading strategies as our proposed approach for a fair comparison. The four traditional quantitative methods are mean reversion strategy [55], linear regression on fundamental analysis, linear regression on technical indicators, and AIRMA. Among them, the mean reversion strategy assumes that asset prices tend to revert to their long-term mean and utilizes the moving average of the logarithmic returns over the past 60 days as a trading signal. The linear regression models fit regressions on the five fundamental indicators and the 16 technical indicators described in the Methods section, respectively. As for ARIMA, the autoregressive integrated moving average (ARIMA) model, a commonly used time-series forecasting model, is employed to predict stock price movements based on technical indicators. The backtesting performance comparison between our proposed method and the baselines is presented in Table 7. Notably, our strategy significantly outperformed all

baselines, demonstrating the effectiveness of our approach and its applicability in a broader spectrum of quantitative investment scenarios.

Table 7. Backtesting performance comparison.

Strategy		ARR	MDR
Our strategy		32.06%	5.81%
Deep-learning-based strategy (RNN)		29.11%	11.68%
Machine-learning-based strategy (XGBoost)		12.92%	5.92%
Traditional quant strategy	Mean reversion [55]	2.58%	85.07%
	Fundamental analysis (LR) [18]	6.26%	36.65%
	Technical indicators (LR) [18]	7.65%	4.01%
	ARIMA [44]	7.80%	6.79%

From our perspective, the superior performance of our method primarily stems from the following differences and enhancements compared to other approaches. The first is objective and thorough stock screening and preprocessing. Unlike other methods that determine target stocks for quantitative investment via manually subjective selection, our method screens candidate stocks through ranking based on scores from fundamental analysis and sentiment analysis. This objective manner simultaneously leverages the domain knowledge from economics theory and the practical sentiments information from stock news. The second is sentiment analysis and representation using state-of-the-art NLP techniques. A large pretrained generative language model is utilized to obtain meaningful extractions from raw news contents, and a fine-tuned variant of ALBERT is employed as the sentiment encoder. The third is comprehensive extraction and utilization of news information. In contrast to the majority of other methods that only consider news titles, our approach innovatively integrates both news titles and summarized news content from the summarization model, effectively capturing extensive information within the news. The fourth is integration of extensive heterogeneous stock information. Different from traditional quantitative strategies that mainly rely on historical data and technical indicators, our model integrates both technical indicators and comprehensive sentiment embeddings. The fifth is a novel effective quantitative framework based on deep learning models. Our hybrid framework effectively incorporates a news content summarization model (Pegasus), a news sentiment analysis model (ALBERT), and a stock movement prediction model (LSTM/Transformer). These advanced deep-learning-based models enable our framework with outstanding capabilities to accurately guide trading decisions in quantitative investment. The above discussion collectively illustrates how our model fits into the broader landscape of quantitative investment strategies.

3.6. Effectiveness of News Sentiment

In the backtesting results depicted in Figure 6, our proposed method, which incorporates sentiment embeddings from both news titles and summarized content, significantly surpassed its counterpart without using these embeddings. This indicates that the integration of news sentiment facilitates modeling quantitative investment. However, the choice regarding which other forms of data to integrate and how to effectively integrate them are two interesting questions to be investigated.

Social media sentiment has been found useful in some studies [41,43]. Therefore, the representativeness of news sentiment and social media sentiment are compared. We introduced social media sentiment separately into the Indicators (Vanilla) and Indicators + News (Title + Content) control groups, forming the Indicators + Media and Indicators + News + Media groups, respectively. The backtesting evaluation results for these four groups are presented in Figure 9. Compared to the Indicators group, the Indicators + Media group showed a slight improvement in ARR from 16.10% to 16.34%, but also an increase in MDR from 15.84% to 24.45%, suggesting that the introduction of social media sentiment provided relatively limited effective information. After adding social

media sentiment to the Indicators + News group, the backtesting performance degraded, with ARR decreasing from 32.06% to 21.88% and MDR increasing from 5.81% to 17.90%. Both the comparisons collectively indicate that news sentiment better reflects the overall market sentiment trend than social media sentiment, thus validating the effectiveness and reliability of the comprehensive news sentiment integration in our method. Note that these discussions represent a preliminary exploration of integrating other forms of sentiment data. More systematic research is required to answer the questions.

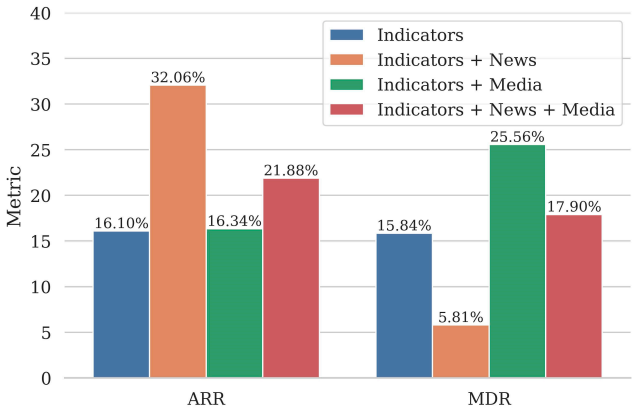

Figure 9. Backtesting performance comparison between the sentiment of news and social media.

4. Conclusions

In recent years, with the advancement of computer technology, especially machine learning and deep learning techniques, quantitative investment has become increasingly significant in the financial markets. However, due to the intricate factors influencing the stock market, achieving accurate quantitative investment has remained a challenging task. The majority of current quantitative investment methods primarily model historical stock trading data, neglecting the rich information embedded in real-world stock market news reports. Moreover, existing methods that consider news merely conduct sentiment analysis on news headlines without thoroughly extracting information from the content.

This paper introduces a novel stock investment approach based on a deep hybrid model. To the best of our knowledge, we are the first to propose simultaneously taking into account both news headlines and news content summaries to extract textual sentiment and representations. The obtained comprehensive sentiment representations, merged with the technical factors extracted from historical trading data, form a global feature. We employ deep learning models, LSTM and Transformer, for modeling stock price changes, allowing for a controlled exploration of the application effectiveness of different series models in stock data. Momentum trading and stop-loss strategies are incorporated to assist the model in predicting stock price directions. Empirical experiments conducted on over 4000 stocks in the entire Chinese stock market demonstrate that our deep hybrid model can accurately and steadily predict stock rises and falls, providing rational and comprehensive investment guidance for investors. Ablation experiments indicate that the incorporation of objective sentiment from news content summaries enhances the model's robustness and noise resistance. With technical factors as basic features, introducing semantic representations of news titles and summaries significantly improves backtesting performance. Specifically, the LSTM model achieves an ARR of 32.06% and an MDR of 5.81%, while the Transformer achieves 26.98% ARR and 5.24% MDR. These results markedly outperform the CSI 300 index, proving the practical value of our proposed quantitative investment approach in providing guidance for investors' strategy making and delivering effective and stable returns. The outstanding backtesting performances of our model on the two separate stock exchanges, SZSE and SSE, demonstrate the robustness and generalizability of our model. Additionally, the comparison with numerous other approaches highlights the differences and enhancements of our method, revealing its applicability in a broader spectrum of

quantitative investment scenarios. Finally, the representativeness and reliability of news sentiment is further validated through comparison with social media sentiment.

The limitation of this study is that the way employed to combine technical indicators and sentiment representations is direct concatenation, which may not be the most efficient way for fusing such heterogeneous information. Future work will focus on studying more effective fusion approaches for technical factors and news semantic representations, which could further enable mutual enhancement and complementarity between the two information sources for better modeling and trading strategy formulation. Improving the selection and creation of technical factors is also an interesting research direction.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/electronics12183960/s1, Figure S1: Statistics of the collected stock and news data; Figure S2: The architecture of Pegasus; Figure S3: The curves of finetuning; Figure S4: A comparison of sentiment prediction; Figure S5: A comparison of model prediction performance on stock market news before and after fine-tuning; Table S1: Example of the collected news; Table S2: Parameters settings for deep learning models; Table S3: Comparison of news and social media; Table S4: Relational analysis between stock price movements and sentiment from different forms of data; Table S5: Backtesting performance of baselines models.

Author Contributions: W.L. conceptualized the idea. W.L. and Y.L. led the research and contributed technical ideas. W.L. developed the proposed method and performed analysis and experiments. W.L. wrote the manuscript. C.H. and Y.L. provided evaluation and suggestions. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Social Science Fund of China (Grant Number 17BJY210), the Key Humanities and Social Science Fund of Hubei Provincial Department of Education (Grant Number 20D043), and the National Natural Science Foundation of China (Grant Number 11701161).

Data Availability Statement: Algorithm code is publicly available at https://github.com/SallyLi0606/Quant (accessed on 29 August 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Weng, B.; Ahmed, M.A.; Megahed, F.M. Stock market one-day ahead movement prediction using disparate data sources. *Expert Syst. Appl.* **2017**, *79*, 153–163. [CrossRef]
2. Chen, M.Y.; Chen, B.T. A hybrid fuzzy time series model based on granular computing for stock price forecasting. *Inf. Sci.* **2015**, *294*, 227–241. [CrossRef]
3. Ballings, M.; Van den Poel, D.; Hespeels, N.; Gryp, R. Evaluating multiple classifiers for stock price direction prediction. *Expert Syst. Appl.* **2015**, *42*, 7046–7056. [CrossRef]
4. Kim, S.; Ku, S.; Chang, W.; Song, J.W. Predicting the direction of US stock prices using effective transfer entropy and machine learning techniques. *IEEE Access* **2020**, *8*, 111660–111682. [CrossRef]
5. Yun, K.K.; Yoon, S.W.; Won, D. Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process. *Expert Syst. Appl.* **2021**, *186*, 115716. [CrossRef]
6. Carta, S.M.; Consoli, S.; Piras, L.; Podda, A.S.; Recupero, D.R. Explainable machine learning exploiting news and domain-specific lexicon for stock market forecasting. *IEEE Access* **2021**, *9*, 30193–30205. [CrossRef]
7. Jiang, M.; Liu, J.; Zhang, L.; Liu, C. An improved Stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms. *Phys. A Stat. Mech. Its Appl.* **2020**, *541*, 122272. [CrossRef]
8. Di Persio, L.; Honchar, O. Artificial neural networks architectures for stock price prediction: Comparisons and applications. *Int. J. Circuits Syst. Signal Process.* **2016**, *10*, 403–413.
9. Chong, E.; Han, C.; Park, F.C. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Syst. Appl.* **2017**, *83*, 187–205. [CrossRef]
10. Singh, R.; Srivastava, S. Stock prediction using deep learning. *Multimed. Tools Appl.* **2017**, *76*, 18569–18584. [CrossRef]
11. Cao, J.; Wang, J. Stock price forecasting model based on modified convolution neural network and financial time series analysis. *Int. J. Commun. Syst.* **2019**, *32*, e3987. [CrossRef]
12. Gunduz, H.; Yaslan, Y.; Cataltepe, Z. Intraday prediction of Borsa Istanbul using convolutional neural networks and feature correlations. *Knowl. Based Syst.* **2017**, *137*, 138–148. [CrossRef]
13. Fischer, T.; Krauss, C. Deep learning with long short-term memory networks for financial market predictions. *Eur. J. Oper. Res.* **2018**, *270*, 654–669. [CrossRef]

14. Mukherjee, S.; Sadhukhan, B.; Sarkar, N.; Roy, D.; De, S. Stock market prediction using deep learning algorithms. *CAAI Trans. Intell. Technol.* **2021**, *8*, 82–94. [CrossRef]
15. Agrawal, M.; Shukla, P.K.; Nair, R.; Nayyar, A.; Masud, M. Stock Prediction Based on Technical Indicators Using Deep Learning Model. *Comput. Mater. Contin.* **2022**, *70*, 287–304. [CrossRef]
16. Albahli, S.; Awan, A.; Nazir, T.; Irtaza, A.; Alkhalifah, A.; Albattah, W. A deep learning method DCWR with HANet for stock market prediction using news articles. *Complex Intell. Syst.* **2022**, *8*, 2471–2487. [CrossRef]
17. Yadav, K.; Yadav, M.; Saini, S. Stock values predictions using deep learning based hybrid models. *CAAI Trans. Intell. Technol.* **2022**, *7*, 107–116. [CrossRef]
18. Banik, S.; Sharma, N.; Mangla, M.; Mohanty, S.N.; Shitharth, S. LSTM based decision support system for swing trading in stock market. *Knowl. -Based Syst.* **2022**, *239*, 107994. [CrossRef]
19. Ahmed, S.; Alshater, M.M.; El Ammari, A.; Hammami, H. Artificial intelligence and machine learning in finance: A bibliometric review. *Res. Int. Bus. Financ.* **2022**, *61*, 101646. [CrossRef]
20. Park, H.J.; Kim, Y.; Kim, H.Y. Stock market forecasting using a multi-task approach integrating long short-term memory and the random forest framework. *Appl. Soft Comput.* **2022**, *114*, 108106. [CrossRef]
21. Kanwal, A.; Lau, M.F.; Ng, S.P.; Sim, K.Y.; Chandrasekaran, S. BiCuDNNLSTM-1dCNN—A hybrid deep learning-based predictive model for stock price prediction. *Expert Syst. Appl.* **2022**, *202*, 117123. [CrossRef]
22. Tao, M.; Gao, S.; Mao, D.; Huang, H. Knowledge graph and deep learning combined with a stock price prediction network focusing on related stocks and mutation points. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34*, 4322–4334. [CrossRef]
23. Patil, P.R.; Parasar, D.; Charhate, S. Wrapper-Based Feature Selection and Optimization-Enabled Hybrid Deep Learning Framework for Stock Market Prediction. *Int. J. Inf. Technol. Decis. Mak.* **2023**, 1–26. [CrossRef]
24. Li, M.; Zhu, Y.; Shen, Y.; Angelova, M. Clustering-enhanced stock price prediction using deep learning. *World Wide Web* **2023**, *26*, 207–232. [CrossRef]
25. Zhang, Q.; Qin, C.; Zhang, Y.; Bao, F.; Zhang, C.; Liu, P. Transformer-based attention network for stock movement prediction. *Expert Syst. Appl.* **2022**, *202*, 117239. [CrossRef]
26. Minh, D.L.; Sadeghi-Niaraki, A.; Huy, H.D.; Min, K.; Moon, H. Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. *IEEE Access* **2018**, *6*, 55392–55404. [CrossRef]
27. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
28. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
29. Shynkevich, Y.; McGinnity, T.M.; Coleman, S.A.; Belatreche, A. Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning. *Decis. Support Syst.* **2016**, *85*, 74–83. [CrossRef]
30. Feuerriegel, S.; Gordon, J. Long-term stock index forecasting based on text mining of regulatory disclosures. *Decis. Support Syst.* **2018**, *112*, 88–97. [CrossRef]
31. Shi, L.; Teng, Z.; Wang, L.; Zhang, Y.; Binder, A. DeepClue: Visual interpretation of text-based deep stock prediction. *IEEE Trans. Knowl. Data Eng.* **2018**, *31*, 1094–1108. [CrossRef]
32. Zhang, J.; Cui, S.; Xu, Y.; Li, Q.; Li, T. A novel data-driven stock price trend prediction system. *Expert Syst. Appl.* **2018**, *97*, 60–69. [CrossRef]
33. Carosia, A.E.O.; Coelho, G.P.; Silva, A.E.A. Analyzing the Brazilian financial market through Portuguese sentiment analysis in social media. *Appl. Artif. Intell.* **2020**, *34*, 1–19. [CrossRef]
34. Carta, S.; Consoli, S.; Piras, L.; Podda, A.S.; Recupero, D.R. Event detection in finance using hierarchical clustering algorithms on news and tweets. *PeerJ Comput. Sci.* **2021**, *7*, e438. [CrossRef]
35. Huang, J.Y.; Liu, J.H. Using social media mining technology to improve stock price forecast accuracy. *J. Forecast.* **2020**, *39*, 104–116. [CrossRef]
36. Lin, W.C.; Tsai, C.F.; Chen, H. Factors affecting text mining based stock prediction: Text feature representations, machine learning models, and news platforms. *Appl. Soft Comput.* **2022**, *130*, 109673. [CrossRef]
37. Lin, Y.L.; Lai, C.J.; Pai, P.F. Using deep learning techniques in forecasting stock markets by hybrid data with multilingual sentiment analysis. *Electronics* **2022**, *11*, 3513. [CrossRef]
38. Jing, N.; Wu, Z.; Wang, H. A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. *Expert Syst. Appl.* **2021**, *178*, 115019. [CrossRef]
39. Wu, S.; Liu, Y.; Zou, Z.; Weng, T.H. S_I_LSTM: Stock price prediction based on multiple data sources and sentiment analysis. *Connect. Sci.* **2022**, *34*, 44–62. [CrossRef]
40. Daradkeh, M.K. A hybrid data analytics framework with sentiment convergence and multi-feature fusion for stock trend prediction. *Electronics* **2022**, *11*, 250. [CrossRef]
41. Swathi, T.; Kasiviswanath, N.; Rao, A.A. An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis. *Appl. Intell.* **2022**, *52*, 13675–13688. [CrossRef]
42. Gao, R.; Cui, S.; Xiao, H.; Fan, W.; Zhang, H.; Wang, Y. Integrating the sentiments of multiple news providers for stock market index movement prediction: A deep learning approach based on evidential reasoning rule. *Inf. Sci.* **2022**, *615*, 529–556. [CrossRef]
43. Herrera, G.P.; Constantino, M.; Su, J.J.; Naranpanawa, A. Renewable energy stocks forecast using Twitter investor sentiment and deep learning. *Energy Econ.* **2022**, *114*, 106285. [CrossRef]

44. Zhao, Y.; Yang, G. Deep Learning-based Integrated Framework for stock price movement prediction. *Appl. Soft Comput.* **2023**, *133*, 109921. [CrossRef]
45. Ashtiani, M.N.; Raahemi, B. News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review. *Expert Syst. Appl.* **2023**, *217*, 119509. [CrossRef]
46. Shilpa, B.L.; Shambhavi, B.R. Combined deep learning classifiers for stock market prediction: Integrating stock price and news sentiments. *Kybernetes* **2023**, *52*, 748–773.
47. Ma, Y.; Mao, R.; Lin, Q.; Wu, P.; Cambria, E. Multi-source aggregated classification for stock price movement prediction. *Inf. Fusion* **2023**, *91*, 515–528. [CrossRef]
48. Wang, J.; Zhang, Y.; Zhang, L.; Yang, P.; Gao, X.; Wu, Z.; Zhang, J. Fengshenbang 1.0: Being the foundation of Chinese cognitive intelligence. *arXiv* **2022**, arXiv:2209.02970.
49. Zhang, J.; Zhao, Y.; Saleh, M.; Liu, P. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In Proceedings of the 37th International Conference on Machine Learning, Online, 13 July 2020; pp. 11328–11339.
50. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
51. Akita, R.; Yoshihara, A.; Matsubara, T.; Uehara, K. Deep learning for stock prediction using numerical and textual information. In Proceedings of the 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Okayama, Japan, 26–29 June 2016; pp. 1–6.
52. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942.
53. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
54. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd international Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
55. Lehmann, B.N. Fads, martingales, and market efficiency. *Q. J. Econ.* **1990**, *105*, 128. [CrossRef]