Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 18, No. 1, 2025

Stock Trend Prediction for Energy Sector Stocks Based on Investor Sentiment and Deep Learning

Keyu Long *

School of Business, Lehigh University, Bethlehem, Pennsylvania, 18015, USA
* Corresponding author: Keyu Long (Email: kel524@lehigh.edu)

Abstract: The energy industry has always been one of the focuses of attention for investors and the media, among others, due to its special characteristics. Energy companies can be sensitively driven by a variety of factors such as policy changes, international situation and investor sentiment, etc. Traditional forecasting models are unable to make accurate predictions based on factors such as industry characteristics, and how to effectively capture these complex factors as well as more accurately predict stock prices in this industry based on these factors has become one of the important topics in financial research. In this study, we use historical trading data of energy stocks, take advantage of deep learning to process time series data and capture the dependencies between features, construct a stock prediction model with energy industry characteristics, and experimentally evaluate the model. The experimental results show that the prediction results are superior with the addition of industry characteristic indicators as well as indicators generated from texts related to the industry, which suggests that compared to generalized prediction models, models based on sentiment text analysis and deep learning are able to effectively identify and capture industry characteristics and make more accurate predictions.

Keywords: Investor Sentiment, CNN, LSTM, Energy Stocks.

1. Introduction

As a raw material for human economic development, social progress and personal life, energy is an indispensable part of people's life and production. Traditional energy sources include fossil fuels such as coal, oil and natural gas, which have driven economic and social development since the Industrial Revolution and are still widely used today. Renewable energy, as an emerging industry, is an important direction for future energy development. In the financial market, small and medium-sized investors are one of the main forces of investment, they lack professional judgment on the market, will be greatly affected by public opinion and market fluctuations, most of them are difficult to hold the same financial product for a long time. As for enterprises, they need to obtain long-term stable investment from stocks for expanding production, researching and developing new products, etc. Stocks are an important financing channel for them. The prediction of stock trends is influenced by a variety of factors, among which investor sentiment plays an important role. For energy stocks, some keywords in the media related to the politics and economy of such industries also affect the prediction of trends. Traditional stock market analysis methods, which often rely on historical data and technical indicators, are hardly applicable anymore. Based on this, this paper introduces deep learning techniques to enable the model to handle huge data sets more rationally, capture complex patterns and potential relationships in the data more efficiently, and improve the accuracy and reliability of the prediction with a richer number of features.

2. Literature Review

The study of a single type of stock often provides insights into industry characteristics and allows for more accurate forecasting. The use of technical analysis in energy stocks provides better guidance to energy investors when making decisions [1]. With the gradual and wide application of machine learning, more algorithms are applied in stock analysis. BP neural networks have been used, demonstrating the superiority of this method in dealing with complex nonlinear data [2]. A stock trading model based on the CNN algorithm has been designed, which performs excellently and demonstrates the potential of CNN in capturing complex market trends and making accurate predictions [3]. Due to the high degree of uncertainty in stock market data, many researchers have begun to utilize a combination of algorithms to construct models. A hybrid model that combines CNN and Reinforcement-LSTM has been designed, and this hybrid model in Big Data performs better than many traditional prediction methods [4]. Many scholars have studied investor sentiment since a long time ago. For example, 11 significant pricing anomalies of financial assets have been selected and the impact of market sentiment has been evaluated through a long-short strategy [5]. With the popularity of social media, a large amount of information related to the market continues to emerge. Sentiment analysis, combined with a variety of machine learning models, has been conducted to analyze the impact of sentiment keywords in social media on stock prices [6]. In summary, applying sentiment analysis and deep learning to stock trend prediction is a popular trend nowadays. Meanwhile, conducting detailed research on specific types of stocks based on industry characteristics and building models specifically for predicting single stock categories is also a research hotspot.

3. Relevant Theoretical Foundations

3.1. Deep Learning Model Theory

3.1.1. Long and Short-term Memory Network Theory

In 1997, the theory of Long Short-Term Memory Networks (LSTM) was proposed, introducing a gating mechanism that is able to establish a balance between long-term and short-term information, effectively capturing dependencies over

longtime spans [7]. Memory unit is the core of LSTM, which controls the flow of information through forgetting gates, input gates and output gates, and the update formula is:

$$C_t = f_t * C_{t-1} + i_t \quad (1)$$

Where C_t is the memory cell at the current moment t, C_{t-1} is the memory cell at the moment before t, f_t and i_t are the outputs of the forgetting gate and the input gate, respectively, and is the candidate memory cell, which is the new information generated based on the x_t of the current input and the hidden state h_{t-1} of the previous moment. The forgetting gate f_t controls the extent to which memories of the previous moment are retained or forgotten,

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

Where σ is the Sigmoid activation function, W_f is the weight matrix of the forgetting gate, b_f is the bias vector of the forgetting gate, h_{t-1} is the hidden state of the previous moment, and x_t is the current input data. On the basis of the forgetting gate, the input gate i_t determine show much of the current candidate information is updated into the memory cell, and can adjust the proportion of information in the candidate memory cell that is updated into the memory cell, with a formula similar to the forgetting gate,

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

Where W_i and b_i are the weight matrix and bias vector of the input gate, respectively. Similarly, h_{t-1} is the hidden state of the previous moment and x_t is the current input data. However, the input gate does not directly introduce the input information into the memory unit, it will combine with the candidate memory unit to calculate the result, the candidate memory unit will generate a new potential information candidate value based on the current input and the hidden state of the previous moment, and it will decide whether to be updated into the memory unit or not under the action of the input gate, which is calculated by the formula:

$$\widetilde{C_t} = \tan h(W_C \cdot [h_{t-1}, X_t] + b_C \quad (4)$$

Where is the candidate memory cell at the current moment t, W_c and b_C are the weight matrix and bias vector associated with the candidate memory cell, respectively, h_{t-1} is the hidden state at the previous moment, x_t is the current input data, and tanh restricts the output value to between -1 and 1 by means of a hyperbolic tangent function. The hidden state h_t of the LSTM combines the current memory cell information and the role of the output gate, which is a representation of the output at the current time step t. It can be viewed as an intermediate result of the external output of the LSTM at each time step, which can be used as an input to the next layer of the network,

$$h_t = o_t \cdot tanh(C_t) \quad (5)$$

Where o_t is the output of the output gate that determines the effect of the memory cell C_t on the hidden state h_t at the current moment t. C_t is the current memory cell. The output gate determine show much of the information in the memory cell is output to the hidden state,

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (6)$$

Where W_o and b_o are the weight matrix and bias vector of the output gate, respectively. In summary, LSTM captures short-term and long-term dependencies in time series through the collaborative work of memory cells and hidden states, overcoming the limitations of traditional recurrent neural networks in handling long-term dependencies.

3.1.2. Convolutional Neural Network Theory

Convolutional neural networks were first proposed by Yann LeCun in the 1980s for image processing tasks. The core advantage of CNN is its local connectivity and parameter sharing mechanism. The local connectivity and dynamic propagation characteristics enable it to process multidimensional signals quickly, as described in [8]. In time series analysis, convolutional neural networks can efficiently extract local temporal features through convolutional operations, pooling operations, etc., reducing the number of parameters and computation. A convolutional layer can locally scan the input data with a small convolutional kernel to generate a feature map for extracting local features, the mathematical expression of the convolutional operation is as follows:

$$s(i,j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} x\,(i+m, j+n) \cdot w(m,n) + b \quad (7)$$

Where $s(i,j)$ is the value of the output feature map at position (i,j), $x(i+m, j+n)$ is the value of the input data at position $(i+m, j+n)$, $w(m,n)$ is the weight of the convolution kernel at position (m,n), and b is the bias term. After the convolution operation, in order to enhance the nonlinear representation of the model, the output of the convolutional layer is passed through a nonlinear activation function to ensure that the convolutional neural network can learn the nonlinear relationship, the most commonly used activation function is ReLU, which is formulated as:

$$ReLU(x) = max(0, x) \quad (8)$$

Where x is the output of the convolutional layer, the activation function serves to set the negative values in the output of the convolutional layer to zero and retain only the positive values, which effectively solves the problem of gradient vanishing. After the feature maps extracted by the convolutional layer are activated by the activation function, the pooling layer downsamples the feature maps to reduce the data dimensions and retain the important features, which is used to reduce the information redundancy. Maximum pooling is done by selecting the maximum value in the pooling window with the formula:

$$p(i,j) = max\{s(i,j), s(i+1, j+1), \ldots, s(i+k, j+k)\} \quad (9)$$

Where, $p(i,j)$ is the output value after pooling, which is the pooled value at position (i,j). $s(i,j)$ is the value of the input feature map at position (i,j). k is the size of the pooling window, which defines the range. After being processed by the convolutional layer, activation function and pooling layer, the high-level features of the input data that have been extracted by the convolutional neural network are passed to the fully connected layer, which is responsible for integrating these features that have been compressed and processed and mapping them to the output space for outputting the final classification or regression results. Each neuron in the fully connected layer is connected to all the neurons in the previous layer with the formula:

$$z = W \cdot x + b \tag{10}$$

Where z is the output of the fully connected layer, W is the weight matrix, x is the feature vector of the input and b is the bias term. This hierarchical structure of convolutional neural networks is well suited for processing high-dimensional data, and can efficiently learn and classify complex input data by capturing complex feature representations through a multilayer structure.

3.2. Text Sentiment Analysis Theory

Text Sentiment Analysis is a technique for analysing sentiment tendencies in text through natural language processing techniques for extracting user's emotions or opinions, for example, the overall framework of sentiment analysis was discussed, and a sentiment lexicon was constructed [9]. For text sentiment analysis, constructing a sentiment lexicon is an early and commonly used method, it matches the sentiment words in the text to be analyzed with the emotional vocabulary in the dictionary. Six sentiment lexicons were compared, and researchers can choose the most appropriate lexicon based on the type of text and the specific analysis task [10]. Machine learning-based approaches predict the sentiment tendencies of text by training classifiers. An Apache Spark-based sentiment analysis framework was proposed to classify the sentiment of large amounts of data generated on social media, using machine learning algorithms to handle datasets of different sizes [11]. Methods such as Naive Bayes and Support Vector Machines were used for the task of sentiment classification on Twitter [12]. The results all show that machine learning methods are better in handling the task of sentiment analysis. Deep learning-based approaches can handle long texts and some complex sentiment expressions with better performance through Convolutional Neural Networks and Long and Short-Term Memory Networks, etc. The application of CNN and LSTM in sentiment analysis was studied, and the results indicated that the CNN-LSTM hybrid model and the bidirectional LSTM model performed better in terms of accuracy [13].

4. CNN-LSTM Based Prediction Model Construction

4.1. Data Pre-processing

4.1.1. Data Selection

The study selected performance data from seven listed companies covering both traditional and renewable energy fields from 2018 to 2023. These seven companies include traditional energy companies primarily engaged in oil, natural gas, and coal: PetroChina, Sinopec, and Shenhua. Additionally, it includes representative companies in the renewable energy field, covering solar energy, wind energy, and battery storage: LONGi, Trina Solar, Goldwind, and CATL. In order to construct a more accurate stock price prediction model, this study extracts a number of indicators from multiple dimensions, including stock price and trading volume. All the indicators and their classification are shown below:

Table 1. Indicator Categories and Names

Indicator Category	Indicator Name
Price Indicator	Open Price Close Price High Price Low Price Volume
Technical Indicator	Simple Moving Average (SMA) Exponential Moving Average (EMA) Volume Weighted Average Price (VWAP) Bollinger Band Width Average True Range (ATR) True Range (TR) On-Balance Volume (OBV) Moving Average Convergence Divergence (MACD) Relative Strength Index (RSI)
External Indicator	Overnight Interbank Lending Rate Brent Crude Oil Price USD to CNY Exchange Rate Gold Price CRB Commodity Index Northbound Net Capital Inflow 10-Year Treasury Yield
Energy-Specific Indicator	Energy Price Correlation Policy Sensitivity Index

To further enhance the model's ability to capture the characteristics of energy stocks, this study designs two industry-specific indicators. These two indicators are the 'Energy Price Correlation' and the 'Policy Sensitivity Index'. The Energy Price Correlation is used to quantify the correlation between an energy company's stock price and international energy prices. This study first collects the daily closing prices of energy company stocks and international energy prices, then calculates the daily rate of change for both. Subsequently, a sliding window method is used, selecting 10 days as a fixed-length window period to calculate the correlation between the rate of change in energy company stock prices and international energy prices within this period. The calculation formula is as follows:

$$Correlation = \frac{\sum_{t=1}^{N}\left(\Delta P_{\text{company}}(t) \times \Delta P_{\text{energy}}(t)\right)}{N} \tag{11}$$

Where N is the window length, which can be set to 10, and $\Delta Pcompany\ (t)$ and $\Delta Penergy\ (t)$ represent the daily rate

of change in stock price for the energy company and the international energy price, respectively. The policy sensitivity index is used to measure the sensitivity of a company's stock price to policies. To calculate this indicator, this study firstly screens important policy release events related to the renewable energy industry and selects five days before and after as the time window to quantify the company's sensitivity to policy releases by calculating the relative rate of change in share prices before and after the policy releases.

4.1.2. Text Selection

This study collects discussions related to the energy industry from various finance-related social media platforms and investor forums, such as Guba and Xueqiu, from 2018 to 2023. Keywords such as "oil", "natural gas", and "renewable energy" are used to filter the data, removing irrelevant information. The data is periodically scraped at regular intervals to ensure coverage of major hot events and market fluctuation periods.

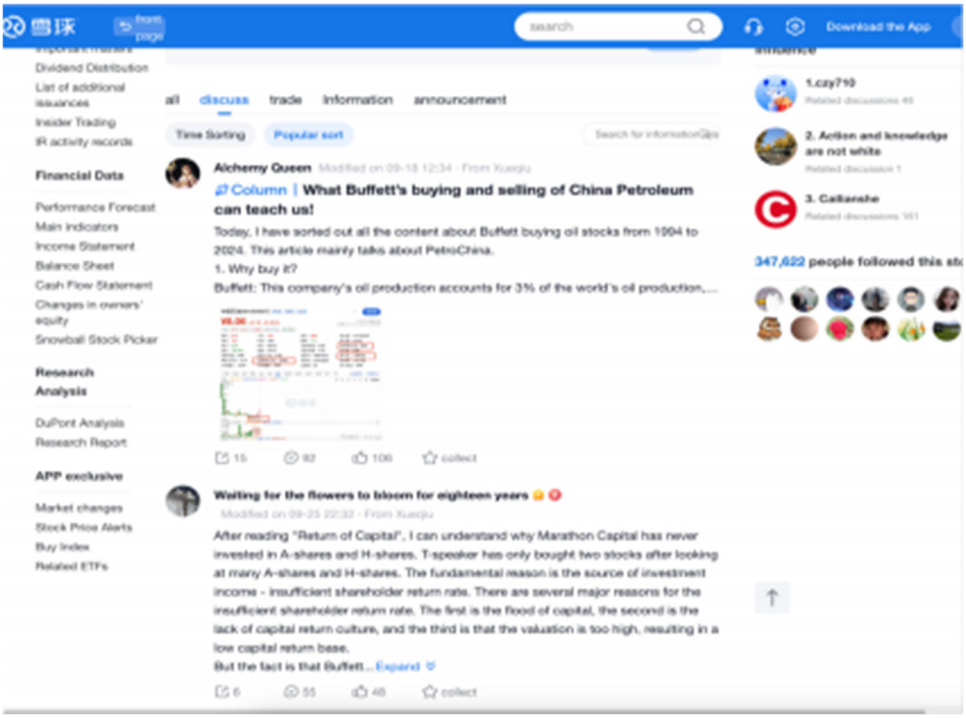

Figure 1. XueQiu Webpage on Investment in PetroChina

In Figure 1, we see a webpage from XueQiu discussing investment in PetroChina. On the basis of keyword screening, this study captures a large number of statements containing emotional words. In addition, the length and interactivity of the post content is also a reflection of investor sentiment. Longer posts tend to contain more in-depth analyses and discussions, and generally have more comments and replies, with higher heat, and the heat and intensity of the discussion is also a reflection of investor sentiment towards the stock. In addition, this study extracts news reports about the energy industry from financial news websites such as Oriental Wealth, using industry-related keywords and selecting this news by timeframe. Then, policy release documents are collected from the official websites of government departments, focusing on policy releases, adjustments, and industry reports that affect the industry.

4.1.3. Data and Text Processing

The data selected for this study is stock data for all trading days of selected energy sector companies for the period from 2018 to 2023, covering a number of dimensions such as opening price, closing price, trading volume, etc. The data is then cleaned, de-weighted, and normalised. For missing values, this study uses the corresponding data of the previous trading day of that missing value to fill in. For duplicate values, these values are directly deleted considering that they maybe redundant data that occurred during the process of performing data capture. For outliers, this study uses statistical methods to detect them, such as by setting upper and lower limits on the data to eliminate certain extreme data that do not conform to market patterns. After completing the data cleaning, Z-score normalisation is used to scale all the data to similar ranges to ensure that there is no excessive difference in magnitude between the data. Principal component analysis is then used to compress multiple features into a small number of principal components to reduce model complexity. In addition to the processing of the data, the text, which contains information relevant to the research object, needs to be converted into a feature quantity that can be used by the model. In the pre-processing stage, for English text, words are separated by natural spaces, while for Chinese text, sentences are divided into separate lexical units using a word separation tool. Next, meaningless stop words are removed to reduce the noise and ensure that the remaining content can express the information effectively. After the basic processing of the text, the sentiment analysis model based on machine learning is used to input the pre-processed text into the pre-training model, and then the text data is subjected to sentiment classification and sentiment score calculation. The sentiment analysis model classifies the text as positive, negative or neutral sentiment by identifying the sentiment information in the text and assigns each text a sentiment score. Also, in order to combine the textual information with stock market data, it is necessary to ensure that the sentiment score is temporally aligned with data such as stock prices and trading volumes. The following metrics were quantified in this study:

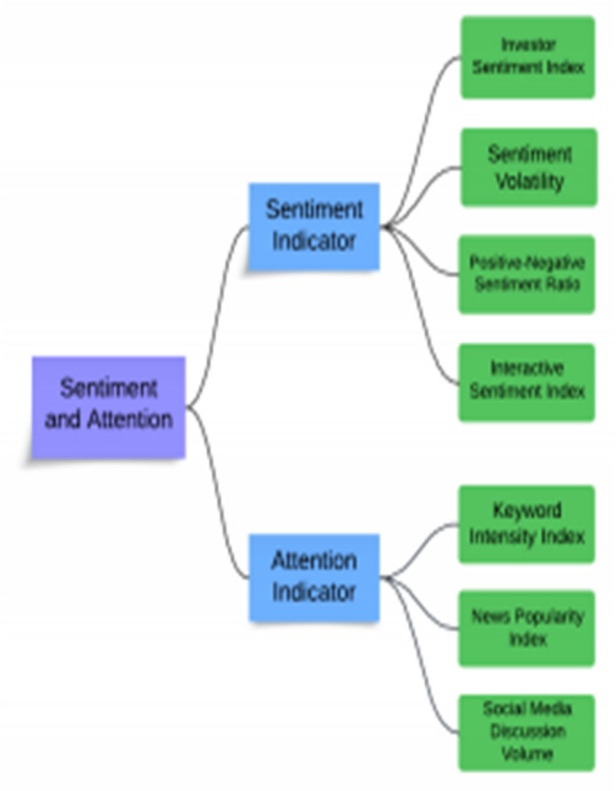

Figure 2. Sentiment and Attention Indicators

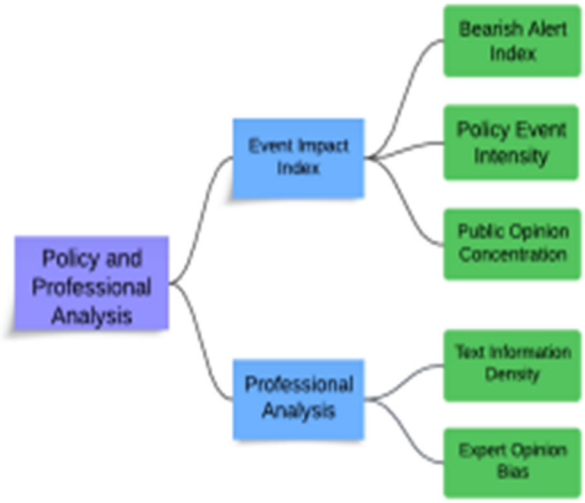

Figure 3. Policy and Professional Analysis Indicators

In addition to this, this study counted the daily search volume of Baidu for keywords related to energy stocks, which also served as indicators for inputting into the model. These metrics are all relevant feature quantities for the characteristics of the energy industry, as the extracted texts were directionally selected to be relevant to the industry.

4.2. Model Architecture Design

4.2.1. Overview of the CNN-LSTM Model

This study combines two deep learning models, CNN and LSTM, to construct a hybrid model. The following flowchart shows the specific operation process of the model:

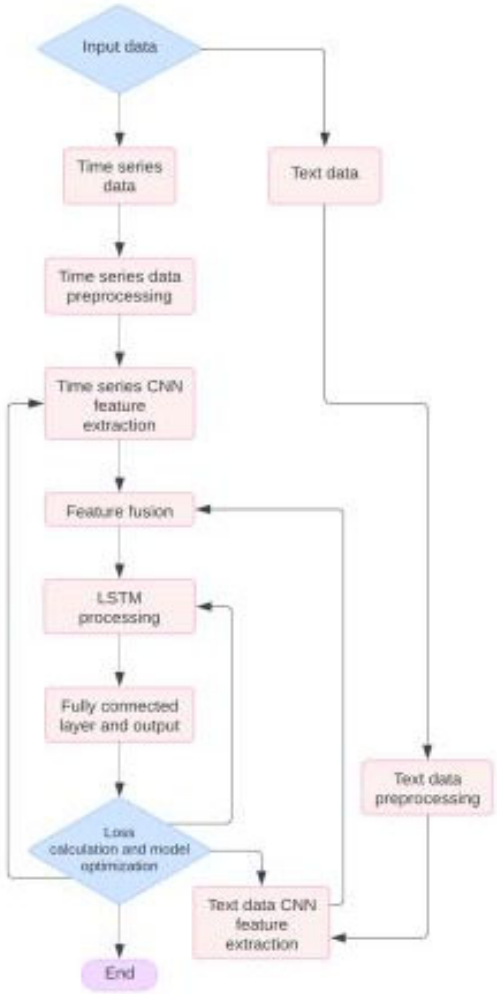

Figure 4. CNN-LSTM Model Flowchart for Stock Trend Prediction

The input data for the model consists of two main categories, time series data such as historical prices and trading volumes of energy stocks, and text data such as investor sentiment, which will be quantified into numerical features through sentiment analysis. Both will then first undergo pre-processing, where the text data is pre-processed by natural language processing techniques to extract core features such as market sentiment and policy orientation. Both types of data enter the CNN module separately after preprocessing. After feature extraction, time series features and text features are merged in the feature fusion layer to form a multi-dimensional joint feature vector. The model combines 'hard' data such as stock price and 'soft' data such as investor sentiment through feature fusion, and the feature fused data is fed into the LSTM network, and after processing, the output of the LSTM passes through the fully-connected layer and generates a prediction of stock price. After processing, the output of LSTM goes through a fully connected layer to generate a prediction of the stock price.

4.2.2. CNN Layer Design

The design of the CNN layer includes the selection of the convolution kernel, the activation function, and the setting of the pooling layer. In this study, one-dimensional convolution is used in the CNN layer, and the step size in the convolution operation is set to 1, and for each convolution kernel, the output feature maps retain the same temporal dimension. The convolution kernel size is set to 3, then, the output of the convolution operation is nonlinearly transformed by the

ReLU activation function. The output of each layer is normalised by batch normalisation to ensure that the variations of different features during model training remain consistent. After the convolution operation, the size of the feature map is compressed by a pooling operation. The pooling layer serves to reduce the size of the feature map while retaining key local features, and two strategies, maximum pooling and average pooling, can generally be used in the pooling layer operation. However, in the study of trend forecasting for energy stocks, short-term market reactions tend to have a greater impact on the forecasts, such as the release of a relevant policy, which tends to produce an immediate reaction in the stock, which in turn has a strong short-term impact on the price. In the energy market, dramatic price fluctuations are also often driven by certain key factors, so maximum pooling is used in this study by setting both the window size and the step size to 2, which means that each pooling operation reduces the temporal dimension of the feature map by half, reducing the amount of computation while retaining significant local features. Meanwhile, the combination of convolutional and pooling layers adopts a multi-layer design, where the number and size of convolutional kernels will increase layer by layer. Since the number of convolution kernels in the initial layer is small, it mainly captures some low-level local features, and then the number of convolution kernels will be increased to deal with more complex patterns. After the convolution operation and pooling operation, the final obtained feature maps are flattened and transformed into one-dimensional vectors while retaining the local and high-level features extracted in the convolution layer. These flattened feature vectors contain both short-term fluctuations, such as stock prices and trading volumes, as well as sentiment features extracted from, e.g., the media, which will be used as inputs to the LSTM layer.

4.2.3. LSTM Layer Design

After being processed by the CNN layer, the output data serves as input for the LSTM layer, the time step is set to 8 days, slightly longer than a week, which means the model uses data from the past eight days to predict the trend on the ninth day. This study selects 128 hidden units, aiming to strike a balance between model complexity and performance. In the specific operations of the LSTM layer, the model adopts a two-layer structure, setting return sequences=True in the first layer to pass the output of each time step to the next layer. This setup allows the second LSTM layer to further extract dependencies along the time dimension and compress this part of the sequence, integrating the dependencies in the time series into the hidden state as the final feature used for prediction. Additionally, to prevent overfitting, a dropout rate of 0.2 is applied, meaning that 20% of the input data is randomly dropped during training.

5. Empirical Analysis and Model Evaluation

5.1. Design of Experiments

5.1.1. Feature Introduction and Dataset Partitioning

The purpose of constructing a stock prediction model is to use historical data to complete the prediction of future trends, and the first thing that needs to be introduced is the basic stock trading data. In this study, all the data between 2018 and 2023 are captured, excluding legal holidays, weekends, and cases where the market is closed due to other special reasons, there are a total of 1,455 sets of data, and the latter 10% or so of this data is used as the test set, and the rest of the data is used as the training set. After this, feature quantities with industry characteristics are introduced based on energy stocks, but these feature quantities do not have specific values on each trading day, so the training and test sets of these feature quantities follow the historical data feature quantities of the stocks. In addition, feature quantities generated from text such as investor sentiment are introduced to try to increase the accuracy of the prediction model, and these feature quantities again follow the corresponding trading days to divide the training and test sets.

5.1.2. Feature Combination Experiments

Gradually attempt to identify features that contribute to prediction accuracy by introducing control features. First, use only the basic trading data of stocks, such as closing prices and daily trading volume. Conduct principal component analysis on the data, select the principal components with a cumulative contribution rate of 98%, and input them into the model. The results obtained from running the model will serve as the first set of baseline data. After this, new combinations of features will be added in sequence. The second set will only add the energy price correlation index and the policy sensitivity index, while the third set will add features generated from text data, such as investor sentiment. The results obtained from each set will be compared with the baseline data following the same steps. If the results are worse than the baseline data, these features will be temporarily excluded in the next set. By comparing the results of these groups, a model with better predictive performance will ultimately be obtained.

5.2. Experimental Results and Evaluation

5.2.1. Evaluation Criteria

In order to explore the effect of the added feature quantities on the prediction model, this study will gradually add or replace the feature quantities to test the prediction situation, and ultimately to judge the prediction of the upward or downward situation, with the confusion matrix to reflect the performance of each group of feature quantities in the model. For the binary classification problem, there are the following confusion matrices:

Table 2. Confusion Matrix for Model Evaluation

	Prediction	
Actual	Positive	Negative
Positive	TP	FN
Negative	FP	TN

For the prediction of the future rise and fall of the stock, the accuracy as well as the F1-score are chosen to evaluate the model prediction results.

5.2.2. Experimental Results

Firstly, only the basic stock trading data is used as the feature quantity of the model, taking the stock trading data of seven representative companies in the energy category as samples, and counting their stock trading data on all trading days from 2018 to 2023. These data are processed and calculated, and finally more than twenty relevant feature quantities such as closing price, volume weighted average price, relative strength indicator, etc. are obtained and input into the model, and the prediction results obtained after running are as follows:

Table 3. Model Performance Metrics for Traditional Energy Companies

	PetroChina	Sinopec	Shenhua Energy
accuracy	52.41%	53.19%	52.58%
F1-score	58.34%	55.33%	56.06%

Table 4. Model Performance Metrics for Renewable Energy Companies

	Longi	Trina Solar	Goldwind	CATL
accuracy	51.53%	50.86%	51.17%	51.63%
F1-score	58.11%	53.61%	54.47%	57.87%

Using the results of the above input of conventional features as a control group, and after adding the energy price correlation indicator and the policy sensitivity indicator, the data were input into the model to obtain the new prediction results as follows:

Table 5. Updated Model Performance Metrics for Traditional Energy Companies

	PetroChina	Sinopec	Shenhua Energy
accuracy	55.39%	56.23%	55.62%
F1-score	58.77%	56.51%	56.78%

Table 6. Updated Model Performance Metrics for Renewable Energy Companies

	Longi	Trina Solar	Goldwind	CATL
accuracy	53.98%	52.48%	53.29%	54.69%
F1-score	58.32%	54.57%	55.92%	58.12%

According to the comparison of the two sets of data, it can be seen that the accuracy of the model prediction after adding the new indicators is better than that of the control group, and there is some improvement in the F1-score, so the features are retained and the text-generated indicators, such as new investor sentiment, continue to be added, and the new prediction results are obtained as follows:

Table 7. Updated Model Performance Metrics for Traditional Energy Companies (Final)

	PetroChina	Sinopec	Shenhua Energy
accuracy	60.34%	58.67%	57.66%
F1-score	61.35%	61.59%	62.27%

Table 8. Updated Model Performance Metrics for Renewable Energy Companies (Final)

	Longi	Trina Solar	Goldwind	CATL
accuracy	59.55%	56.49%	56.22%	59.48%
F1-score	62.02%	59.14%	58.88%	61.72%

With the addition of new features generated from text, the accuracy and F1-score of the prediction model were further improved.

5.2.3. Model Evaluation

It can be learnt from the analysis of the results that the prediction effect of the control group is not satisfactory, especially for renewable energy stocks, the basic stock trading data can only keep the prediction accuracy of the model only at about 51%, while after adding the feature quantity related to the industry characteristics, the accuracy is generally improved, and the F1- score is also more or less improved, and their trends are shown below:

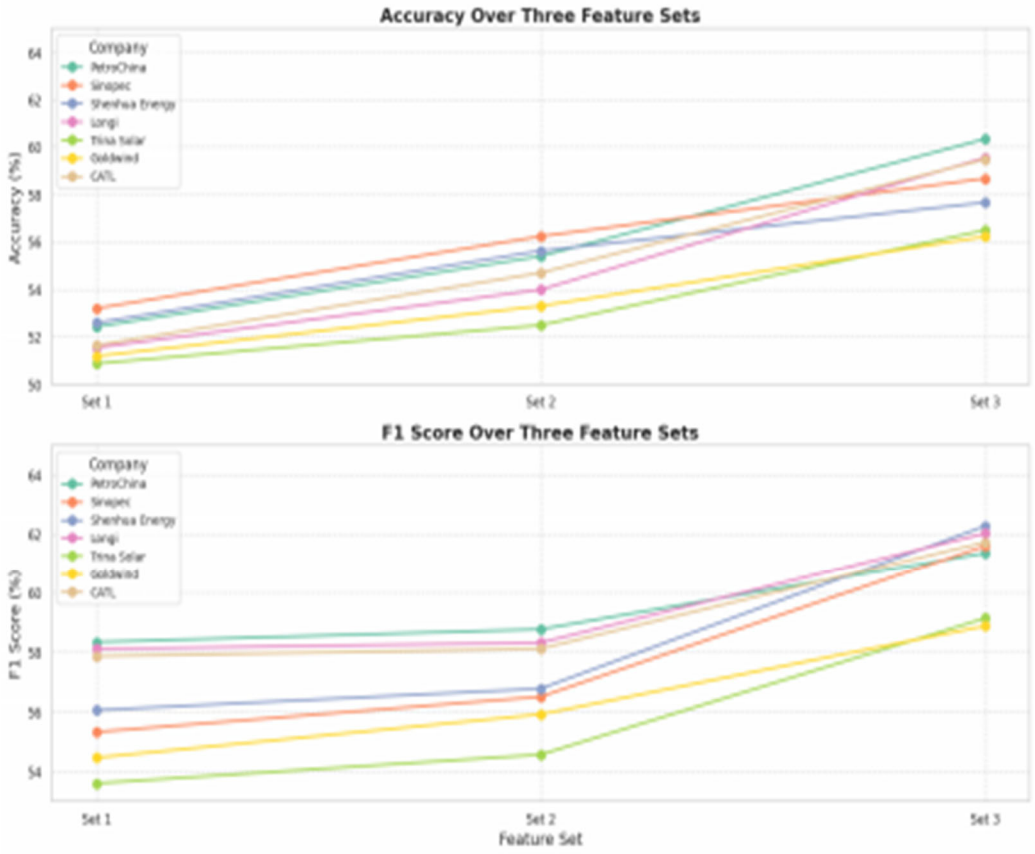

Figure 5. Accuracy and F1 Score Over Three Feature Sets for Various Companies

After the addition of text-generated features such as investor sentiment, the accuracy is again significantly improved, with an average accuracy of 51.91% in the initial control group and 58.34% in the final group. Comparing within the same group again, in the first group, the prediction accuracy of the traditional energy stocks is all above 52% and the prediction accuracy of the new energy stocks are all below 52%. In the dataset, it can be found that the data of new energy stocks is more volatile, which may cause the model to be more difficult to capture the data characteristics, while the traditional energy stock prices are relatively stable, for example, petrochina and goldwind two companies, their closing price is compared to the chart below:

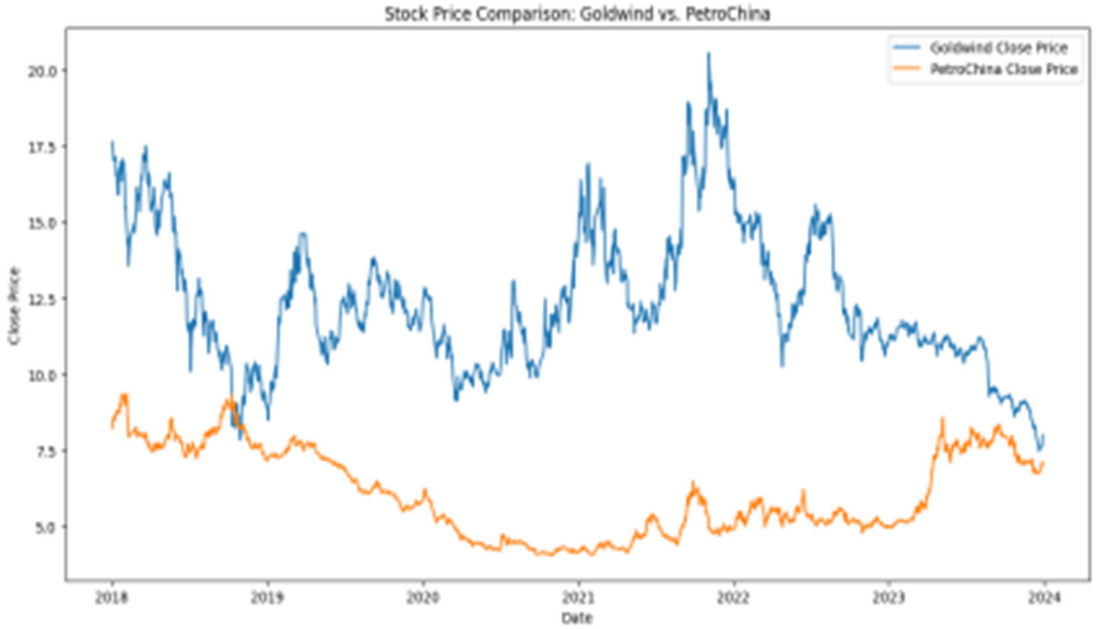

Figure 6. Stock Price Comparison: Goldwind vs. PetroChina

In the second set of data, the prediction performance for traditional energy stocks remains better. However, in the third set of data, the prediction performance for certain new energy stocks, such as CATL, is comparable to that of traditional energy stocks. As leading enterprises, they attract significant market attention, and indicators such as media and investor sentiment have a notable enhancing effect on their predictive accuracy.

6. Conclusion

Prediction for time series data such as stocks has always been a popular research topic, and in this study, we chose the specific industry of energy stocks, and utilized the characteristics of this industry to set up textual data such as investor sentiments, news media reports, and policy releases, and constructed a deep learning prediction model based on CNN-LSTM. The effectiveness of the model is verified by processing and analyzing the stock trading data of several more typical energy companies between 2018 and 2023. In the experiments, this study gradually added new feature volumes to the model, from the stock trading data at the beginning to the gradually added feature volumes generated from the text, and each new set of feature volumes gave a significant improvement in the model's prediction effect, which indicates that using industry-related text is of significant help in studying the relevant data of the industry. The experimental results of this study show that compared to traditional forecasting methods, which rely solely on technical analysis and historical data analysis, forecasting methods based on deep learning models and textual sentiment analysis are able to reflect investor behavior, and thus market changes, in a more comprehensive way, and this improves the accuracy of stock trend forecasting for specific industries.

References

[1] Chitra, R. (2011) Technical analysis on selected stocks of energy sector. International Journal of Management and Business Studies, 1(1):42–46.

[2] Li, F., Liu, C. (2009) Application study of bp neural network on stock market prediction. In Proceedings of the Ninth International Conference on Hybrid Intelligent Systems, 174–178.

[3] Gudelek, M. U., Boluk, S. A., Ozbayoglu, A. M. (2017) A deep learning based stock trading model with 2-d cnn trend detection. In Proceedings of the IEEE International Conference on Big Data (Big Data), 2565–2572.

[4] Ishwarappa, Anuradha, J. (2021) Big data based stock trend prediction using deep cnn with reinforcement-lstm model. International Journal of System Assurance Engineering and Management.

[5] Stambaugh, R. F., Yu, J., Yuan, Y. (2012) The short of it: Investor sentiment and anomalies. Journal of Financial Economics, 104(2):288–302.

[6] Awan, M. J., Rahim, M. S. M., Nobanee, H., Munawar, A., Yasin, A., Zain, A. M. (2021) Social media and stock market prediction: A big data approach. Computers, Materials and Continua, 67(2):2565–2582.

[7] Hochreiter, S., Schmidhuber, J. (1997) Long short-term memory. Neural Computation, 9(8):1735–1780.

[8] Chua, L. O., Roska, T. (1993) The cnn paradigm. IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, 40(3):147–156, 1993.

[9] Zhu, N., Zou, P., Li, W., Cheng, M. (2012) Sentiment analysis: A literature review. In Pro- ceedings of the IEEE International Symposium on Management of Technology (ISMOT), 571–576.

[10] Khoo, C. S. G., Johnkhan, S. B. (2018) Lexicon-based sentiment analysis: Comparative eval- uation of six sentiment lexicons. Journal of Information Science, 44(4):491–511.

[11] Jain, A. P., Dandannavar, P. (2016) Application of machine learning techniques to sentiment analysis. In 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 628–632.

[12] Neethu, M. S., Rajasree, R. (2013) Sentiment analysis in twitter using machine learning tech- niques. In 4th ICCCNT, 1–5.

[13] Malviya, S., Tiwari, A. K., Srivastava, R., Tiwari, V. (2020) Machine learning techniques for sentiment analysis: A review. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 12(2):72–78.