

A Novel Stock Index Prediction Method Based on Financial Text Sentiment Analysis

Chengxian Li

*School of Management Science and Engineering, Central University of Finance and Economics,
Beijing, China*

Lichengxian2004@outlook.com

Abstract: With the rapid growth of artificial intelligence and machine learning, predicting stock market trends has become an increasingly important and complex task. In this paper, we propose a novel stock index prediction model, SA-BERT-LSTM, based on financial text sentiment analysis. This model integrates the BERT (Bidirectional Encoder Representations from Transformers) sentiment analysis method with the Long Short-Term Memory (LSTM) network, combining the extracted sentiment features from financial news with stock trading data. We apply this model to predict the rise and fall of the CSI 300 Index. The results demonstrate that sentiment features significantly enhance prediction accuracy. Compared to three benchmark models—BP neural network, Support Vector Machines (SVM), and XGBoost—the SA-BERT-LSTM model outperforms them in terms of prediction precision. Furthermore, the model is adaptable for predicting individual stock prices. This research contributes to the application of sentiment analysis in financial market prediction, broadening the scope of AI and machine learning in economics, and provides valuable insights for the implementation of national AI strategies.

Keywords: Sentiment analysis, Long Short-Term Memory, Pre-trained language models, Stock index prediction

1. Introduction

The development of artificial intelligence (AI) has brought significant advancements in various fields, including economics and finance. Since the State Council of China issued the "New Generation AI Development Plan" in 2017, AI technologies such as machine learning and deep learning have gained prominence as tools for solving complex problems in public decision-making, personal decision-making, urban development, and financial forecasting [1].

Stock price prediction, a critical component of the financial market, is influenced not only by quantitative data such as financial statements and historical market data but also by qualitative factors like news and media reports. These factors have traditionally been difficult to quantify, yet they can significantly impact stock market behavior. This paper explores the potential of using financial text sentiment analysis as a tool for predicting stock index movements, specifically focusing on the CSI 300 Index [2].

Despite the substantial progress in China's capital market over the past three decades, challenges remain, especially concerning the unbalanced structure of market investors. The dominance of individual investors over institutional investors, as well as their frequent trading behavior, often

results in heightened market volatility. In this context, timely and accurate predictions of stock price movements are crucial for both investors and regulatory authorities. However, financial time series data, with its nonlinearity, non-stationarity, and high noise, complicates stock market forecasting. Furthermore, significant information is often embedded in unstructured data sources, such as financial news, which traditional econometric models struggle to incorporate effectively [3].

Machine learning models, particularly those leveraging neural networks, have shown promise in overcoming these challenges. However, issues such as overfitting and local minima continue to hinder their effectiveness, particularly when dealing with large datasets and high-dimensional data. Recent advancements in deep learning, including the development of Long Short-Term Memory (LSTM) networks, have shown improved predictive capabilities, especially for time-series data. This paper proposes a hybrid model combining the BERT (Bidirectional Encoder Representations from Transformers) sentiment analysis tool with LSTM networks to predict stock index movements more effectively [4].

2. Literature review

2.1. Machine learning approaches to financial market prediction

Numerous studies have applied machine learning methods to predict stock market trends. Early works utilized Artificial Neural Networks (ANN) to model stock price movements, showing promising results. Other methods such as Support Vector Machines (SVM), Random Forests (RF), and XGBoost have been widely used in financial predictions. These models, while effective in some cases, often struggle with high-dimensional, nonlinear data, which is characteristic of financial time series [5].

In recent years, deep learning methods, particularly LSTM networks, have become more prevalent. LSTM networks have shown excellent results in modeling noisy financial time series, with studies demonstrating their ability to predict stock price volatility with high accuracy. Despite these advancements, most existing research still relies heavily on structured data, such as financial statements and historical stock prices, overlooking the potential insights offered by unstructured text data, such as news articles.

2.2. Sentiment analysis in stock market prediction

Sentiment analysis, a technique within Natural Language Processing (NLP), has gained traction as a tool for analyzing unstructured financial text data. Early research focused on analyzing media sentiment to predict stock market returns. Recent works have incorporated machine learning methods to automate sentiment extraction from financial news, social media, and other textual sources, with some studies achieving significant predictive accuracy [6].

However, traditional sentiment analysis techniques, including dictionary-based methods and simple machine learning models, often fail to capture the full depth of sentiment expressed in financial texts. In contrast, pre-trained language models like BERT, which capture contextual nuances and word relationships, have proven to be highly effective for sentiment analysis tasks in finance. Combining BERT with LSTM networks allows for better contextual understanding and improves the prediction of stock price movements.

3. Methodology

This section describes the methodology employed in this study, which combines sentiment analysis of financial news articles with stock market data to predict stock index movements. The approach integrates two powerful components: BERT for sentiment analysis and LSTM for learning temporal dependencies in stock price data. This hybrid model allows for the simultaneous processing of both

structured data and unstructured data, providing a more comprehensive and effective prediction framework.

3.1. Data collection and preprocessing

The first step in our methodology is data collection, which involves obtaining two key data types: stock trading data and financial news data. We focus on the CSI 300 Index, a representative stock index that includes the top 300 stocks listed on the Shanghai and Shenzhen stock exchanges. The stock trading data consists of daily records that include several essential features, such as the opening price (P_{open}), closing price (P_{close}), highest price (P_{high}), lowest price (P_{low}), trading volume ($V_{trading}$), and other market metrics. These features capture the daily behavior of the stock market and are critical for time-series forecasting.

In addition to the stock trading data, we collect unstructured data in the form of financial news articles. These articles are sourced from reputable financial publications such as the Shanghai Securities Journal. Key data elements from the articles include publication time, title, content, and source. The sentiment expressed in these articles is a crucial indicator of market dynamics, as positive news typically correlates with price increases, while negative news is associated with price declines. This sentiment data serves as input for the sentiment analysis process.

Once the data is collected, we move on to preprocessing. For the stock trading data, we employ Min-Max normalization to standardize the data and eliminate scale differences between features. The Min-Max normalization scales each feature to a fixed range of $[0, 1]$, using the formula:

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)},$$

Where x is the original value of the feature, $\min(x)$ and $\max(x)$ are the minimum and maximum values of the feature, respectively. This ensures that all features contribute equally to the prediction model and prevents any single feature from dominating due to differences in scale.

Next, for the financial news data, we clean the text by removing extraneous characters, such as HTML tags, punctuation, and other non-informative content. After cleaning, the text is tokenized into words or phrases, which form the basic units for sentiment analysis. Each article is then labeled as either positive or negative, based on its overall sentiment. For example, if an article reports positive earnings or optimistic market forecasts, it is labeled as positive. Conversely, articles discussing economic downturns or negative earnings reports are labeled as negative. These sentiment labels are used as features in the subsequent machine learning model.

3.2. Sentiment analysis with BERT and LSTM-based prediction

Once the data preprocessing is completed, we proceed with the sentiment analysis using BERT, a pre-trained deep learning model specifically designed for natural language understanding tasks. BERT's bidirectional architecture is crucial for capturing the contextual meaning of words, which is essential for understanding the complex and often nuanced language used in financial news. For sentiment analysis, we use the BERT-Base-Chinese model, which is pre-trained on a large Chinese text corpus and adapted for financial texts.

BERT takes the preprocessed financial news articles as input and converts each word into a contextualized word embedding. The word embeddings are generated by passing the tokenized words through BERT's transformer layers. These embeddings represent the semantic meaning of each word in its context, allowing the model to understand the sentiment of the article more effectively than traditional methods. Once the embeddings are obtained, they are passed through a classification layer

to determine the sentiment of each article. The output is a sentiment score, which is then used in conjunction with the stock trading data.

Following sentiment analysis, the next step involves combining the sentiment features with the stock trading data and feeding the combined data into an LSTM network for prediction. LSTM is a type of recurrent neural network (RNN) that is particularly well-suited for time-series prediction due to its ability to store information over long periods. This makes LSTM ideal for modeling stock market data, which exhibits temporal dependencies.

LSTM consists of a series of memory cells, each of which contains an input gate, a forget gate, and an output gate. The network learns from the historical stock trading data and sentiment features to predict future price movements. Specifically, the LSTM network processes the input sequence ($X = \{x_1, x_2, \dots, x_t\}$) where each x_t is a combination of sentiment data and stock market data at time step t . The model uses these inputs to output a sequence of hidden states h_t , which represent the model's understanding of the relationship between sentiment and stock price fluctuations over time. The LSTM's memory cells use the following update equations:

1. Forget Gate:

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

2. Input Gate:

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

3. Candidate Cell State:

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

4. Cell State Update:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t,$$

5. Output Gate:

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

6. Hidden State:

$$h_t = o_t \cdot \tanh(C_t),$$

where W_f, W_i, W_C, W_o are weight matrices. b_f, b_i, b_C, b_o are bias terms. σ is the sigmoid activation function. And $\tanh$ is the hyperbolic tangent activation function.

The final hidden state h_t is then passed through a fully connected layer followed by a softmax activation function to produce a probability distribution, indicating the likelihood that the stock index will rise or fall. The softmax function is given by:

$$P(y = 1|X) = \frac{e^{h_t}}{1 + e^{h_t}}$$

Where $P(y = 1|X)$ is the probability that the stock index will rise (upward movement), and h_t is the output of the final hidden state.

3.3. Model training and evaluation

The model is trained using historical data spanning from March 2015 to April 2021. We split the dataset into a training set (80%) and a testing set (20%). The training process involves optimizing the model's parameters using the cross-entropy loss function, which is commonly used for binary classification tasks. The Adam optimizer is employed to minimize the loss function and update the

model's weights. Hyperparameter tuning is performed using grid search and cross-validation to identify the optimal configuration of the model. The following evaluation metrics are used to assess the model's performance: Accuracy, Prediction Accuracy, Recall, and F1-score. These metrics provide insights into the model's ability to predict stock index movements and handle both positive and negative market trends.

To evaluate the effectiveness of the SA-BERT-LSTM model, we compare its performance against three baseline models: BP Neural Networks, Support Vector Machines (SVM), and XGBoost. By combining sentiment analysis and time-series prediction, the SA-BERT-LSTM model offers a powerful solution for forecasting stock index movements.

4. Results

In this section, we present the experimental results of the stock index prediction models, including both baseline models and our proposed SA-BERT-LSTM model. The performance of the models was evaluated based on four key metrics: Accuracy, Prediction Accuracy, Recall, and F1-score. These metrics provide a comprehensive view of each model's ability to predict stock index movements, with particular emphasis on how well the models balance correctly identifying both positive and negative movements.

4.1. Experimental setup

The experiments were conducted using historical data from March 2015 to April 2021, which includes daily stock trading data for the CSI 300 Index and corresponding financial news articles. The stock trading data consists of key market indicators such as opening price, closing price, trading volume, high and low prices, and price changes. The financial news data was used for sentiment analysis, with sentiment features extracted using the BERT (Bidirectional Encoder Representations from Transformers) model, as described in the methodology section. The sentiment features were then integrated with the stock trading data and used as inputs for each model.

The dataset was divided into training (80%) and testing (20%) sets, and the models were trained using standard supervised learning techniques. The performance of the models was evaluated using Accuracy, Prediction Accuracy, Recall, and F1-score. These metrics were calculated to provide insights into each model's ability to handle both the time-series nature of stock data and the unstructured financial news data.

4.2. Results comparison

The results of the experiments for each model are presented in Table 1 below:

Table 1: The experimental results of stock index prediction for each model

Model	Accuracy	Prediction	Recall	F1
BP	52.58	54.78	54.79	53.98
SVM	54.79	55.13	56.89	54.99
XGBoost	56.46	57.23	57.61	57.42
LSTM	61.82	61.94	59.92	60.99
SA-BERT-LSTM	64.23	63.66	61.75	63.12

4.3. Analysis of the results

The SA-BERT-LSTM model achieved the highest accuracy of 64.23%, surpassing all other models in this metric. This indicates that the SA-BERT-LSTM model is more successful at predicting the

correct direction of the stock index movement compared to the baseline models. The LSTM model, which was the next best performer, achieved an accuracy of 61.82%, followed by XGBoost (56.46%) and SVM (54.79%). BP, the simplest model, performed the weakest, with an accuracy of only 52.58%.

In terms of prediction accuracy, the SA-BERT-LSTM model again led with 63.66%, closely followed by the LSTM model at 61.94%. This suggests that integrating sentiment analysis into the LSTM framework helped the model not only in predicting stock price trends but also in refining its ability to predict both upward and downward movements more reliably. XGBoost and SVM achieved 57.23% and 55.13%, respectively, which are lower compared to the LSTM-based models. BP's prediction accuracy was the lowest at 54.78%.

The recall metric measures the model's ability to correctly identify positive (stock price increase) and negative (stock price decrease) movements. Here, the SVM model had the highest recall value of 56.89%, outperforming the LSTM model at 59.92%. However, despite its relatively lower recall compared to SVM, the SA-BERT-LSTM model still performed well with a recall of 61.75%. This shows that SA-BERT-LSTM was able to strike a good balance between precision and recall, correctly identifying the market movements more accurately.

The F1-score, which balances precision and recall, was also highest for the SA-BERT-LSTM model with 63.12%, indicating that the model efficiently captured both true positives and true negatives while minimizing false positives and false negatives. The LSTM model came in second with an F1-score of 60.99%, demonstrating that it was still strong at capturing the temporal patterns in stock price data. XGBoost (57.42%) and SVM (54.99%) followed in terms of F1-score, showing that the hybrid approach of SA-BERT-LSTM provided the best overall balance between precision and recall.

4.4. Discussion

The experimental results demonstrate that the SA-BERT-LSTM model outperforms all baseline models (BP, SVM, XGBoost) in terms of accuracy, prediction, recall, and F1-score. By integrating sentiment features from financial news using BERT with stock market data captured by LSTM, the model leverages both temporal dependencies and qualitative information, enabling more informed predictions of stock index movements. The LSTM model also outperformed traditional models, highlighting the advantages of time-series models in stock prediction. In conclusion, the SA-BERT-LSTM model effectively integrates sentiment analysis and time-series forecasting, improving prediction accuracy and providing a robust framework for understanding market trends.

5. Conclusion

This paper presents a novel approach to stock index prediction by integrating sentiment analysis from financial texts with time-series data. The SA-BERT-LSTM model demonstrates a significant improvement in prediction accuracy over traditional machine learning models. The combination of BERT's ability to understand context and LSTM's strength in handling time-series data makes this model highly effective for predicting stock market trends. This research provides valuable insights into the potential of using AI and machine learning in financial market forecasting and offers a promising direction for future studies.

References

- [1] Akita R, Yoshihara A, Matsubara T, et al. Deep learning for stock prediction using numerical and textual information[C]//2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS). IEEE, 2016: 1-6.
- [2] Antweiler W, Frank M Z. Is all that talk just noise? The information content of internet stock message boards[J]. *The Journal of finance*, 2004, 59(3): 1259-1294.

- [3] Atiya A, Ji C. *How initial conditions affect generalization performance in large networks*[J]. *IEEE transactions on neural networks*, 1997, 8(2): 448-451.
- [4] Bollen J M H N, Zeng X J [J]. *TwitterMoodPredictstheStockMarket*, 2011, 2(1): 1-8.
- [5] Chiu C C, Sainath T N, Wu Y, et al. *State-of-the-art speech recognition with sequence-to-sequence models*[C] 2018 *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018: 4774-4778.
- [6] Girshick R, Donahue J, Darrell T, et al. *Rich feature hierarchies for accurate object detection and semantic segmentation*[C] *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014: 580-587.